STAT505/ST697R: Regression Analysis, Fall 2012

Contents

.

1 Introduction

1.1 Some examples

- Calibration of measurement of cholesterol.

  Have prepared samples with known concentration of cholesterol x. These are known as standards. Run a sample through the measuring device and record Y. (In some calibration problems Y might not even be in the same units as x; for example in a radioassay, the Y is a radioactive count while x is a concentration). We want to understand how Y relates to x. One reason is to use this relationship when you analyze a sample with an unknown x and want to estimate the x. This is called inverse "prediction".

- Nutritional Requirements

  An individual is measured at various levels of nitrogen intake (by controlling diet) and in each case a measure of nitrogen balance is obtained. The objective is to first build a model for how balance relates to intake and then determine an individual's nitrogen requirement; defined to be the intake at which the expected balance is 0.

- Strength of finger joints.

  One way to laminate wood together is with the use of finger joints. The process involves applying a certain amount of pressure (X_1) for a certain amount of time (X_2). The objective is to estimate how pressure and time influence the breaking strength Y of the piece of wood (or test hypotheses that a variable or variables plays no role). Once we build a model the goal is to predict breaking strength when a piece is made using a particular pressure and time and attach uncertainty to that prediction.

- LA county epidemiologic data.

  Sample of 200 employees of Los Angles County. Data obtained in 1950 and 1962 on various quantities including systolic and distolic blood pressure, height, weight, serum cholesterol, clinical status (normal, heart disease status) and year of death (if died before 1962). i) Examine relationships among various variables within a year. ii) How do certain measures in 1952 help predict

outcome variables in 1962? ii) Is presence or absence of heart disease related to blood pressure, cholesterol, etc and if so in what manner.

- Determining arterial blood pressure in the lungs.(Ex. 8.13 in book) Outcome: Y = arterial blood pressure measured by invasive procedure. X_1 = empty rate of blood pumping into heart (measured by radio-nuclide imaging). X_2 = ejection rate of blood pumped from heart into lungs (measured by radionuclide imaging). X_3 = blood gas measure. Which variables help, and how, in determining Y?
- In the contested 2004 presidential election, there were claims that their were irregularities in the balloting in Palm Beach County causing people to vote for Pat Buchanan when they meant to vote for Al Gore. Using data from other counties, a model can be built on how the proportion of votes for Buchanan relates to certain characteristics of the county. The resulting model can be used to predict what the proportion in Palm Beach County should have been (using a prediction interval), which can be used to assess how irregular the Buchanan vote in Palm Beach actually was.
- Examine the relationship of proportion of students passing MCAS to demographic and socio-economic variables associated with the school district.
- Relating years of education, age, gender and other factors to wages.

1.2 Data Structure

n "units" in the data. For the ith unit in the sample, $i = 1, \ldots, n$, we have $(Y_i, X_{i1}, \ldots, X_{ik})$, where Y = response variable and the X's are variables of potential interest in terms of their effect on Y. These will be referred to as explanatory variables, but may also be called regressors, predictors or predictor variables, independent variables, covariates, etc.

1.3 Objectives of regression analysis.

- Estimation and Model building.

  Investigate the relationship between Y and the X's; that is see how, if at all, Y changes with the X's. Which X's are most important in explaining the changes in Y?

  This could be viewed as simply a **data description/descriptive statistics problem**. In this case no attention is given to the notion of some kind of probabilistic model which generated the data. It is simply curve fitting and involves fitting a function $f(X_1, \ldots, X_k; \underline{\beta})$ to the data, where $\underline{\beta}$ contains some coefficients in the model. Or we might do the fit nonparametrically without specifying some functional form.

  Statistical Inference is a more formal venture and requires a probabilistic model for how the data was generated. This model will involve parameters which explain how Y changes as the X's change and the objective is to make inferences on those parameters. Inferences are only sensible to do if the model assumptions are reasonable.

  Within a particular model, we might want to estimate the coefficients in a model, the expected response at a particular combination of parameters or variability around the regression line.

- Prediction.

  After the data is used to fit a model relating Y to the explanatory variables, a new "unit" is to be observed with known X values and the objective is to predict the value of Y which will occur. Can do this from a simple curve fitting perspective but will not have any way to attach a measure of uncertainty to the prediction. We want to attach uncertainty to the predictions.

- Inverse "prediction" (also called calibration).

  After the data is used to fit a model relating Y to the explanatory variables, a new "unit" (or units) is observed with a response y_o, but the explanatory variables for the unit are not known. The objective is to estimate the unknown X value (or values).

1.4 Regression model for the data

Before the study is run Y_i is a random variable, while the explanatory variables may be either fixed, random or a combination of the two. Regression models for the data arise by specifying the distribution of Y_i **given** $X_{i1}, \ldots, X_{ik}$. When some of the explanatory variables can be random, this is interpreted as the conditional distribution of Y_i given that the explanatory variables take on values $X_{i1} \ldots, X_{ik}$.

Whether the model makes sense and how to interpret it depends on how the data was obtained!

The function $f(X_{i1}, \ldots, X_{ik}; \underline{\beta})$ defined by

$$E(Y_i|X_{i1}, \ldots, X_{ik}) = f(X_{i1}, \ldots, X_{ik}; \underline{\beta})$$

is called the **regression of** Y **on** $X_1, \ldots, X_k$. $E(Y_i|X_{i1}, \ldots, X_{ik})$ denotes the expected value of Y_i given $X_{i1}, \ldots, X_{ik}$. $\underline{\beta}$ is a vector containing the coefficients.

This can also be represented as:

$$Y_i = f(X_{i1}, \ldots, X_{ik}; \underline{\beta}) + \epsilon_i,$$

where ϵ_i is a random variable ("noise", "error") with mean 0; that is $E(\epsilon_i) = 0$.

Specification of the model for the data involves:

- specifying the regression function f. This models the mean behavior.

- specifying the variance of ϵ_i, denoted $\sigma^2\{\epsilon_i\}$.

- specifying any covariance/correlation structure among the ϵ_i. The covariance of ϵ_i and ϵ_j is denoted $\sigma\{\epsilon_i, \epsilon_j\}$. More generally, specifying dependence among the errors.

Note that since we are conditioning on the X's as fixed constants, $\sigma^2\{\epsilon_i\}$ is the same as the variance of Y_i given $X_{i1}, \ldots, X_{ik}$, and $\sigma\{\epsilon_i, \epsilon_j\}$ is the same as the covariance of Y_i and Y_j given $X_{i1}, \ldots, X_{ik}$.

Modeling the variance

In general the variance of ϵ_i, may change over i and, more specifically, it often changes as some function of the X values. When $\sigma^2\{\epsilon_i\}$ changes with i we refer to the resulting model as having **heteroscedasticity**. How the variance changes as the X change can sometimes be of as much interest as changes in the mean.

The **homogeneity of variance assumption** is that

$$\sigma^2\{\epsilon_i\} = \sigma^2$$

and the resulting model is referred to as a **homoscedastic model**.

We have **uncorrelated errors** if ϵ_i, and ϵ_j are uncorrelated for each pair $i \neq j$. This is implied by the assumption that the errors are independent.

Classifying regression models by the form of f.

1. **Linear regression model**.

   $$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + X_{i,p-1}\beta_{p-1} + \epsilon_i$$

   where $X_{i1}, \ldots, X_{i,p-1}$ are functions of the original explanatory variables $X_{i1}, \ldots, X_{ik}$. When $p > 2$, this is typically referred to as a **multiple linear regression model**.

   $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ is a $p \times 1$ vector of regression coefficients with β_0 being the **intercept.** *The β_0 is not essential to the definition of this as a linear regression model and in some problems is omitted.*

2. **Simple Linear Regression Model:** One explanatory variable X_1.

   $$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i.$$

3. **Polynomial model of degree q in one variable X_1:**

   $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \ldots + \beta_q X_{i1}^q + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_q X_{iq} + \epsilon_i,$$

   where $X_{ij} = X_{i1}^j$.

4. **A model with two variables and interactions:**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

Note that these models are *linear in the parameters* and not necessarily in the X values. Models which are not linear in the parameters are referred to as **nonlinear models.**

1.5 Brief comments on statistical inference

We start with a regression model as specified in section 1.4. This involves specifying the regression function f as well as the variance and covariance structure of the errors. It may also mean specifying the actual distribution of the Y's (as opposed to just specifying the mean and variance/covariance structure).

Using the model, methods of estimating the parameters in the model (the β's, σ^2, etc.) are developed. The most common approaches are least squares and maximum likelihood estimation, with maximum likelihood requiring distributional assumptions.

The estimators of the unknown parameters are random; they have uncertainty in them before the data is collected. Being random variables they have some distribution, commonly referred to as the **sampling distribution**. The sampling distribution tells us how good the estimators are . Often we will work with just the mean and the variance of the distribution.

NOTE: **Estimator** refers to the random version before the data is collected. **Estimate** is used to refer to the specific numerical value obtained after the data is collected. So, estimators are random variables, but estimates are numbers.

Using the notation of the book let b_j be the estimator of β_j. (Note that we use the b_j whether we are discussing it as an estimator or an estimate; there should really be two different notations to avoid any confusion but you should be able to tell from the context how it is being used.)

$E(b_j) - \beta_j$ equals the **bias** of b_j. If $E(b_j) = \beta_j$ then we say it is an **unbiased estimator.**

$var(b_j)$ denoted $\sigma^2\{b_j\}$ is the variance of the estimator and the square root of this $\sigma\{b_j\}$ is the standard deviation of b_j and is also called the **standard error of b_j.**

The sampling distributions are used to develop confidence intervals and tests of hypotheses about unknown parameters, develop prediction intervals, develop estimators, confidence intervals and tests for "inverse prediction".

1.6 Data Collection Schemes and More on Regression Modeling

There are many ways in which data used for regression analysis arises. The manner in which data is collected influences how we interpret the regression model and how reasonable the traditional assumptions are. Typically things will get classified as designed experiments or observational studies, but this simple classification scheme lacks enough detail to cover all the situations of interest. Here is a very broad overview.

There are n units in the study. These n units in the study are either:

1. Not sampled; that is there was no "probabilistic sampling" of units. It may be that the n units make up the whole population of interest or it may be that the n units are a convenience sample or arise in some other way.
2. Arise from a probabilistic sampling of some finite population.
   (a) Simple Random Sampling
   (b) Stratified Sampling
   (c) Other sampling schemes (can be complicated).
3. Arise from observations from some "process" (infinite population).
   For example, sampling products from a production line.

The explanatory variables arise by either

1. Being attached to the unit. In this case when the units are randomly selected the explanatory variables are random variables. When the units are fixed, the explanatory variables are fixed.

2. Being assigned to the units. In this case the X value is under the control of the experimenter and can be assigned to the units. Examples include assigning doses of fertilizers to plants, assigning doses of drugs to patients, choosing temperature and pressure settings to use in a manufacturing process, etc. The assignment of the n units to the different X values which can be assigned is often done using a *completely randomized design.*

There can be any combination of selection of units in tandem with either observing explanatory variables or experimentally assigning them to the units. This leads to a variety of different settings all leading to data used for regression analyses.

Understanding exactly what the model means and what assumptions are reasonable can be a much harder/subtler problem than you think.

Here is some brief discussion about the model in simpler cases:

1. Designed experiments.

   In a designed experiment the X values are under the control of the experimenter and can be assigned to units. This includes assigning doses to plants or people, assigning fertilizer levels to plants, setting temperature and pressure for a factory production run, etc. Actually, defining the regression model for these settings in detail is not that easy, but here is the basic idea. Consider a population of units that can be used in the experiment. Suppose that all k of the predictor variables can be manipulated and assigned to units. Suppose you pick a unit at random and assign it values $X_1, \ldots, X_k$. Once you assign these values the response Y is still random. The distribution of Y, *over the randomization to a unit and over any additional randomness after assigning the X values*, is what we mean by the distribution of $Y|X_1, \ldots, X_k$. The regression model is defined in terms of this distribution.

2. Finite Populations

   For ease of notation, consider a single explanatory variable. Consider a finite population of size N with values of the response and the explanatory variable attached to each unit in the population. What is the population regression model? Suppose that there are J distinct values $X_1^*, \ldots, X_J^*$ of the explanatory variable in the population. Further, suppose there are N_j values in the

population which have the value X_j^* for the explanatory variables and we label the values of response associated with these units as $Y_{j1}^*, \ldots, Y_{jm_j}^*$. The distribution of $Y_{j1}^*, \ldots, Y_{jm_j}^*$ is the distribution of $(Y|X_j^*)$; the mean of these values is $E(Y|X_j^*) = m(X_j^*)$ say and the variance of these values is $\sigma^2\{Y \mid X_j^*\}$, call it $v(X_j^*)$. This defines the population regression model of Y on X; the means are then typically assumed to follow a model $m(X_j^*) = f(X_j^*, \underline{\beta})$.

If the data is obtained by a simple random sample of n units out of the N units, then the explanatory variable on the ith unit is random. With X_i denoting the observed value of the explanatory variable for the ith unit selected (obviously X_i must take on one of the values from $X_1^*, \ldots, X_J^*$) it can be shown that the conditional distribution of Y_i given X_i has mean $f(X_i, \beta)$ and variance $v(X_i)$.

There are many other complex ways to sample or have regression data arise. As we encounter various examples, we will discuss the meaning of the population model and how the sampling affects the assumptions.

2 Simple Linear Regression

One explanatory variable X with:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

with $E(\epsilon_i) = 0$ (or $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$)

Model is a straight line with slope β_1 and intercept β_0.

Descriptive standpoint: fit a straight line of the form $b_0 + b_1 X$ to the data with no concern about the random nature of Y.

Statistical inference: View β_0 and β_1 and functions of them as unknown quantities (parameters) that we want to estimate from the data.

How to estimate the parameters? An estimator is a random quantity and so has some distribution with a mean (expected value) and variance. How do we judge an estimator?

- Bias

- Variance (or standard error = $\sqrt{variance}$).
- Mean squared error (bias squared + variance)
- Consistency (As sample size increases estimator "gets closer" to true value).

The behavior of an estimator (bias, standard error, etc.) is dependent on what assumptions are made about the error terms,the ϵ_i's !!! This can effect our choice of an estimator.

Constant variance assumption: $\sigma^2\{\epsilon_i\} = \sigma^2$.

Uncorrelated error assumption $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for each pair $i \neq j$.

(Note: Independence is a stronger assumption than uncorrelated. Independence implies uncorrelated but the converse is not true. The verbal description given in item 5 on page 11 of the book is for independence, not for being uncorrelated.)

2.1 Estimation and sampling properties

The **least squares estimators** are b_0 and b_1 chosen to minimize $Q(b_0, b_1) = \Sigma_i(Y_i - (b_0 + b_1 X_i))^2$.

Analytical expressions for the solution are given in equations (1.10a) and (1.10b) in the book.

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$b_0 = \bar{Y} - b_1 \bar{X}$.

Properties of the least squares estimators:

Note: For now, we are considering the sampling behavior with $X_1, \ldots, X_n$ treated as fixed; so these properties are conditional on the X values. This means we are thinking of the random behavior of the estimators as results from the random behavior of the ϵ_i with fixed X's.

- As long as $E(\epsilon_i) = 0$ **the least squares estimators are unbiased**.

$E(b_0) = \beta_0$ and $E(b_1) = \beta_1$.

This is true regardless of the variance/covariance structure of the error terms.

Define $T = \sum_{i=1}^{n}(X_i - \overline{X})^2$.

Assuming constant variance σ^2 and uncorrelated errors (Holds unless mentioned otherwise)

- Variance of b_0.

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{T}\right]$$

$\sigma\{b_0\} = \left[\sigma^2\{b_0\}\right]^{1/2}$ = **Standard error of b_0.**

- Variance of b_1.

$$\sigma^2\{b_1\} = \frac{\sigma^2}{T}$$

$\sigma\{b_1\}$ = **Standard error of b_1.**

- Covariance of b_0 and b_1.

$$cov(b_0, b_1) = \sigma\{b_0, b_1\} = \frac{-\sigma^2\overline{X}}{T}$$

- If in addition we assume that the ϵ_i are normally distributed, then b_0 and b_1 are **maximum likelihood estimators**. In addition each of b_0 and b_1 are normally distributed and jointly (b_0, b_1) follow a bivariate normal distribution.

The ith **fitted value** is $\hat{Y}_i = b_0 + b_1 X_i$ and the ith **residual** is $e_i = Y_i - \hat{Y}_i$.

Estimating σ^2:

$$\hat{\sigma}^2 = MSE = \sum_{i=1}^{n} e_i^2/(n-2)$$

is an unbiased estimator of σ^2; that is $E(MSE) = \sigma^2$. MSE stands for mean square error (which sounds a lot like but is not the same as mean squared error of an estimator).

Under normality $(n-2)MSE/\sigma^2$ follows a chi-square distribution with n-2 degrees of freedom and is independent of (b_0, b_1).

2.2 Inferences for the regression coefficients.

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{T}\right]$$

estimates the variance of b_0.

$$s\{b_0\} = \left[s^2\{b_0\}\right]^{1/2}$$

is estimated standard error of b_0 (though often referred to as just the standard error).

Similarly define $s^2\{b_1\} = \frac{MSE}{T}$ with $s\{b_1\}$ = the estimated standard error of b_0.

Under normality of the errors :

$$(b_j - \beta_j)/s\{b_j\} \sim t(n-2) \ (t \text{ with } n-2 \text{ degrees of freedom})$$

Without normality of the errors this is approximately true for n large enough in which case the confidence intervals and tests that follow will be approximately correct. How large n needs to be actually depends on the x's

- A confidence interval of level $1-\alpha$ for β_0 is given by $b_0 \pm t(1-\alpha/2; n-2)s\{b_0\}$.
- A confidence interval of level $1-\alpha$ for β_1 is given by $b_1 \pm t(1-\alpha/2; n-2)s\{b_1\}$.

Interpretation of confidence intervals

Our interpretation of confidence interval is a relative frequentist interpretation at this point. A confidence interval of level $1-\alpha$ has the property that before the experiment is run the probability is $1-\alpha$ that the random confidence interval

will contain the true parameter of interest. Hence, $1-\alpha$ can be viewed as the proportion of times the interval would be successful in the long run over repeated carrying out of the experiment. (A reminder that at this point we are conditioning on the $X_1, \ldots, X_n$ as fixed so the randomness in the experiment is just over the ϵ's.) Once we have the data and construct the interval, the interval is either successful or it isn't; you don't know. From the relative frequentist perspective it doesn't make sense to say something like "the probability is .95 that β_1 is between 4 and 10, say, since this statement is either true or false. In order to make statements like this one needs to accept the use of **subjective probability** in which probabilities can be assigned to things that are not random. In this case the probability measures belief, with 0 meaning you are sure the statement is false and 1 means you are sure the statement is true. This is implemented using **Bayesian Statistics** in which some prior information about the parameters is postulated, and after the data is collected this information is updated to create a distribution (the posterior distribution) which is used to obtain subjective probabilities about the parameters.

Testing Hypotheses

Consider $H_0 : \beta_j = c$, $H_0 : \beta_j \leq c$, or $H_0 : \beta_j \geq c$, where c is a specified constant.

Under normality a t-test of such hypotheses is based on

$$t^* = \frac{b_j - c}{s\{b_j\}}$$

If $\beta_j = c$, then t^* (viewed as a random variable before we collect the data) follows a t-distribution with n-2 degrees of freedom; that is $t^* \sim t(n-2)$. For a test of size α,

$H_0 : \beta_j = c$, $H_A : \beta_j \neq c$: reject H_0 if $|t^*| > t(1-\alpha/2, n-2)$.

$H_0 : \beta_j \leq c$, $H_A : \beta_j \geq c$: reject H_0 if $t^* > t(1-\alpha, n-2)$.

$H_0 : \beta_j \geq c$, $H_A : \beta_j \leq c$: reject H_0 if $t^* < -t(1-\alpha, n-2)$.

Special case: Testing for slope = 0 $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The test statistic is $t^* = b_1/s\{b_1\}$.

Relationship between confidence intervals and tests:

Consider the test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The test of size α can be expressed as reject H_0 if the confidence interval of level $1-\alpha$ for β_1 does not contain 0. In general, for an arbitrary parameter θ, a natural way to test $H_0 : \theta = c$ versus $H_A : \theta \neq c$ is to reject H_0 if a a confidence interval of level $1 - \alpha$ for θ does not contain the constant c. *This is completely general and is NOT limited to intervals based on t-statistics.* This can be extended to testing one-sided alternatives by using "one-sided confidence bounds" rather than confidence intervals.

The advantage of using a confidence interval rather than just testing at some desired size is that it provides more information than the test. Hypothesis testing is set up to lead to a conclusion of either reject H_0 or not reject H_0. This approach is both limited in what information it conveys and can lead to difficulties in interpretation.

P-Values.

The P-value associated with a hypothesis test can be defined in two equivalent ways.

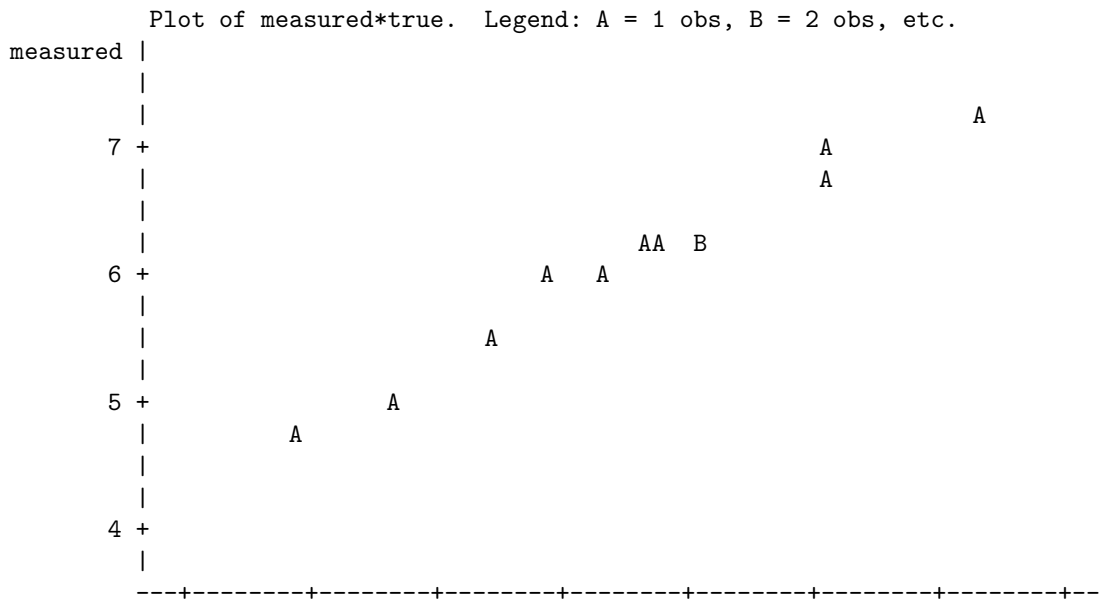
1. Suppose a test about θ rejects H_0 for t^* large where t^* is some test statistic. Let $t^*(obs)$ be the observed value of the test statistic (so this is a fixed value). Then
$$P - value = Max_{\theta \text{ satisfying } H_0} \left[Prob_\theta(t^* > t^*(obs))\right].$$
$Prob_\theta(t^* > t^*(obs))$ is the probability that t^* (considered as a random variable before the start of the experiment) is at least as big as the observed value in your data ($t^*(obs)$) when the value of the parameter is θ. A small P-value indicates the data is not consistent with H_0 while a large value indicates the data does not contradict H_0. For a test statistic which rejects H_0 for t^* small, just reverse the inequality so $P = Max_{\theta \text{ satisfying } H_0} \left[Prob_\theta(t^* < t^*(obs))\right]$.
2. Define the P-value as the smallest value of α for which a test of size α would reject H_0. That is the P-value is such that if the size α is greater than or equal to the P-value then you reject H_0 but if the α is less than the P-value you do not reject H_0.

Example: Assessment of labs in ARM (Acid Rain Monitoring) project.

Along with regular water samples, the labs are sent "blind" samples (they don't know they are special) which have know level of values of interest (pH, alkalinity) etc. Look at relationship between Y = measured = value returned by lab and X = true value. The SAS code below shows some of the basics in running SAS and performing a regression analysis. The data is for one lab using their pH measurements.

```
option ls=80 nodate;
goptions reset=all;
title 'Ph example';
data a;            /* This data set will be a temporary sas file with name
                    'a' . In this example we don't need to refer to this
                     name as if there is only one temporary sas file in use,
                     any procedure will automatically use it. */

infile 'c:\s505\ph.dat';     /* specifies file to read from */
input true measured;     /* names input variables */
proc print;
run;
proc univariate;       /* descriptive statistics on var true */
var true;
run;
proc gplot;            /* creates high resolution plot*/
plot measured*true;
run;
proc plot vpercent=50; /* low resolution plot*/
plot measured*true;
run;

/* Proc reg with some options (there are others we will explore)
 - covb produces variance-covariance matrix of coefficients.
- clb gives CI's for coefficients. The default is 95% CI's.
- p gives predicted values and residuals for each observation */

proc reg;
id true;    /* This will list the X variable also in listing
              of predicted values and residuals */
model measured=true/covb clb p;
plot measured*true;    /* this will produce scatterplot and
                          plot fitted line. Produces higher resolution
                          plot in graph window. */
run;

proc reg;
model measured=true/clb alpha=.02; /* produces 98% confidence interval*/
run;
*****OUTPUT IS EDITED. FULL ANALYSIS WILL BE SHOWN IN CLASS *****

                                Ph example
                           Obs    true    measured
                             1    7.67      7.37
```

```
                          2    6.31      6.36
                         ...
                         12    7.07      6.76

                        The UNIVARIATE Procedure
                            Variable:  true
                                Moments
      N                          12    Sum Weights               12
      Mean                6.30833333    Sum Observations        75.7
      Std Deviation       0.76386378    Variance           0.58348788
   ....
```




```
 The REG Procedure          Dependent Variable: measured
                             Analysis of Variance
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1        6.45221        6.45221     323.49    <.0001
Error                    10        0.19946        0.01995
Corrected Total          11        6.65167
            Root MSE              0.14123    R-Square     0.9700
            Dependent Mean        6.13333    Adj R-Sq     0.9670

                         Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|
     Intercept     1       -0.19161        0.35402      -0.54      0.6002
     true          1        1.00263        0.05575      17.99      <.0001
               Variable     DF       95% Confidence Limits
               Intercept     1       -0.98041        0.59719
               true          1        0.87842        1.12684
```

Covariance of Estimates

Variable	Intercept	true
Intercept	0.1253282569	-0.019603613
true	-0.019603613	0.0031075741

Output Statistics

Obs	true	Dep Var measured	Predicted Value	Residual
1	7.67	7.3700	7.4986	-0.1286
2	6.31	6.3600	6.1350	0.2250
12	7.07	6.7600	6.8970	-0.1370

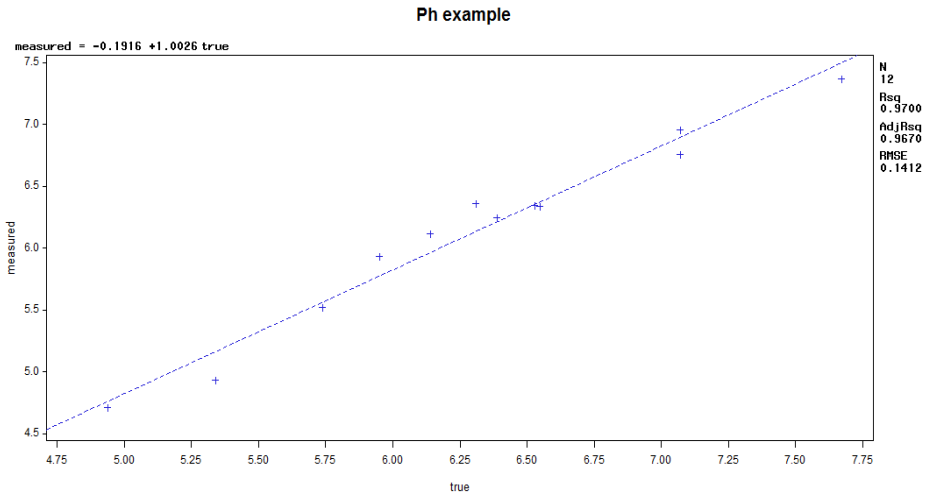


Figure 1: Plot from plot option within proc reg.

2.3 General linear combinations of the regression coefficients

$$\theta = c_0\beta_0 + c_1\beta_1,$$

for constants c_0 and c_1. Note that the individual regression coefficients are special cases; e.g., β_1 corresponds to $c_0 = 0$ and $c_1 = 1$.

$$\hat{\theta} = c_0 b_0 + c_1 b_1$$

is an unbiased estimator of θ and is the best linear unbiased estimator (BLUE) of θ. If normality is assumed for the errors, it is the maximum likelihood estimator

(and in this case is best in the sense of having the smallest possible variance among *all* possible unbiased estimators.)

$$V(\hat{\theta}) = \sigma^2\{c_0b_0 + c_1b_1\} = c_0^2\sigma^2\{b_0\} + c_1^2\sigma^2\{b_1\} + 2c_0c_1\sigma\{b_0, b_1\}.$$

$$s^2\{c_0b_0 + c_1b_1\} = MSE\left[c_0^2\left[\frac{1}{n} + \frac{\overline{X}^2}{T}\right] + c_1^2\frac{1}{T} + 2c_0c_1\frac{-\overline{X}}{T}\right]$$

is an estimate of the variance of $c_0b_0 + c_1b_1$.

Under normality, for any constants c_0 and c_1

$$\frac{c_0b_0 + c_1b_1 - (c_0\beta_0 + c_1\beta_1)}{s\{c_0b_0 + c_1b_1\}} \sim t(n-2)$$

where $\sim t(n-2)$ means follows a t-distribution with n-2 degrees of freedom.

This can be used for confidence intervals and tests as was done for the coefficients.

2.4 Estimating the regression function at a specific X_h.

At $X = X_h$ the expected value of Y is in books notation

$$E\{Y_h\} = \beta_0 + \beta_1X_h$$
$$\hat{Y}_h = b_0 + b_1X_h.$$

This notation can be confusing as it looks like we are estimating a value of Y. **THAT IS NOT THE CASE**. Rather we are estimating the expected value of Y at a value of $X = X_h$. An alternate notation we will use, which also provides some advantages later is

$$\mu(X_h) = \beta_0 + \beta_1X_h, \hat{\mu}(X_h) = b_0 + b_1X_h.$$

$E\{Y_h\} = \beta_0 + \beta_1X_h$ is of the form $c_0\beta_0 + c_1\beta_1$ with $c_0 = 1$ and $c_1 = X_h$, so results of previous section apply.

$$V(\hat{Y}_h) = \sigma^2\{\hat{Y}_h\} = \sigma^2\{b_0\} + X_h^2\sigma^2\{b_1\} + 2X_h\sigma\{b_0, b_1\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{T}\right]$$

$T = \sum_i (X_i - \bar{X})^2$.

Note that the precision is smallest when estimating the expected response at $X_h = \bar{X}$.

Estimated standard error: $s\{\hat{Y}_h\}$ (or $s\{\hat{\mu}(X_h)\}$

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{T}\right].$$

And a confidence interval for $E\{Y_h\}$ is

$$b0 + b_1 X_h \pm t(1 - \alpha/2; n - 2) s\{\hat{Y}_h\}.$$

NOTE: As n gets big the estimated standard error will get small in general and the confidence interval for will get small and therefore be tight around $E\{Y_h\}$. The uncertainty here is all due to uncertainty in the estimation of the coefficients. (To be precise we need T to get bigger as n gets bigger.)

2.5 Prediction intervals

A new Y observation will be taken on a unit with explanatory variable X_{new}. Denote this random response by Y_{new}.

Notation

Here in notes	book (section 2.5)
X_{new}	X_h
Y_{new}	$Y_{h(new)}$
$\hat{Y}_{new}$	$\hat{Y}_h$.

Predict what Y_{new} will be using $\hat{Y}_{new} = b_0 + b_1 X_{new}$.

NOTE: The point estimate is the same as if we were estimating the expected value of Y at $X = X_{new}$, but the problem (and the associated variance) is different.

Find a prediction interval of level $1 - \alpha$ for Y_{new}; that is, find a random interval (L, U) such that $P(L \leq Y_{new} \leq U) = 1 - \alpha$. Notice that unlike confidence intervals

(where the quantity in the middle is a fixed parameter) L, U and Y_{new} are all random and the probability is over all random quantities. The prediction interval is given by

$$\hat{Y}_{new} \pm t(1-\alpha/2, n-2)s\{pred\}$$

where

$$s^2\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_{new} - \overline{X})^2}{T}\right].$$

NOTES: i) $s^2\{pred\}$ is estimating the variance $\hat{Y}_{new} - Y_{new}$, which is $V(\hat{Y}_{new}) + V(Y_new)$ (why?) and equals

$$\sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{T}\right] + \sigma^2.$$

ii) Here as n gets big the later two parts of the estimated variance will generally get smaller and the major term in the estimated variance is MSE, which will converge to σ^2. The prediction interval will not keep getting smaller but always retains a component due to the noise in an individual observation (reflected in σ^2).

EXAMPLE: Returning to the pH example we will demonstrate additional features of proc reg to get confidence intervals and prediction intervals.

- clm produces a confidence interval for the expected response at the X in that observation

- cli produces a prediction interval for a future response taken at the X in that observation.

- The p option prints out predicted values even if there were no confidence intervals specified.

- Note that in the data listing that std error mean predict is the standard error associated with the estimated mean, what we denoted $s\{\hat{Y}_h\}$ (or $s\{\hat{\mu}(X_h)\}$). It has nothing to do with prediction. The standard error for prediction cannot be printed directly in proc reg. It can however be saved to an output file using stdi= as illustrated below.

- If you want to get confidence or prediction intervals associated with an X_h not in the data, you can add a case to the original data file that has a missing value (denoted with a .) for the Y value and the value of X_h for X.

- The output command within proc reg will create a temporary SAS file which has the original data plus what is specified to be saved. In this example we used p = , r = , and stdi = for predicted values, residuals and standard errors for prediction. The name on the right side is of your choosing. There are many other quantities that can be saved in the output statement (see the online documentation for the names used and other details).

```
-  data step not listed again -

proc reg;
id true;
model measure=true/ cli clm p;
   /* The following creates a temporary sas file result will have
    the original data plus variables called yhat, resid  and stdi */
output out=result p = yhat r=resid stdi=sepred;
run;
title 'listing of SAS file result';
proc print data =result nooobs;
run;

*** these come from the clm, cli and p options (not from the output command).
```

Output Statistics

Obs	true	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	7.67	7.3700	7.4986	0.0862	7.3066	7.6906
2	6.31	6.3600	6.1350	0.0408	6.0442	6.2258
3	6.14	6.1200	5.9646	0.0418	5.8713	6.0578
4	7.07	6.9600	6.8970	0.0589	6.7658	7.0282
5	6.39	6.2500	6.2152	0.0410	6.1238	6.3066
6	5.95	5.9300	5.7741	0.0454	5.6729	5.8752
7	6.53	6.3500	6.3556	0.0426	6.2607	6.4505
8	6.55	6.3400	6.3756	0.0429	6.2800	6.4713
9	5.34	4.9300	5.1625	0.0676	5.0117	5.3132
10	5.74	5.5200	5.5635	0.0516	5.4485	5.6785
11	4.94	4.7100	4.7614	0.0865	4.5687	4.9541
12	7.07	6.7600	6.8970	0.0589	6.7658	7.0282

Obs	true	95% CL Predict		Residual
1	7.67	7.1300	7.8672	-0.1286
2	6.31	5.8075	6.4625	0.2250
3	6.14	5.6364	6.2927	0.1554
4	7.07	6.5561	7.2379	0.0630
5	6.39	5.8875	6.5429	0.0348
6	5.95	5.4435	6.1046	0.1559
7	6.53	6.0269	6.6843	-0.005584
8	6.55	6.0467	6.7045	-0.0356
9	5.34	4.8135	5.5114	-0.2325
10	5.74	5.2285	5.8986	-0.0435
11	4.94	4.3924	5.1304	-0.0514
12	7.07	6.5561	7.2379	-0.1370
Sum of Residuals				0

listing of SAS file result

Obs	true	measure	yhat	resid	sepred
1	7.67	7.37	7.49859	-0.12859	0.16544
2	6.31	6.36	6.13500	0.22500	0.14700
3	6.14	6.12	5.96456	0.15544	0.14729
4	7.07	6.96	6.89701	0.06299	0.15300
5	6.39	6.25	6.21522	0.03478	0.14707
6	5.95	5.93	5.77406	0.15594	0.14835
7	6.53	6.35	6.35558	-0.00558	0.14751
8	6.55	6.34	6.37564	-0.03564	0.14761
9	5.34	4.93	5.16245	-0.23245	0.15659
10	5.74	5.52	5.56350	-0.04350	0.15037
11	4.94	4.71	4.76140	-0.05140	0.16561
12	7.07	6.76	6.89701	-0.13701	0.15300

Plotting options in SAS:

You can do some plots within proc reg. The following illustrate some plots that would come after the model statement in proc reg.

```
 plot measure*true/conf;  /* smoothed conf. intervals for mean */
plot measure*true/pred;  /* smoothed prediction intervals */
plot measure*true/conf pred; /* smooth conf and prediction intervals both*/
plot measure*true p.*true lclm.*true uclm.*true/overlay;
plot (measure p. lcl. ucl.)*true/overlay; /*(equivalent to line above)*/
```

You can also save certain quantities to a SAS file via the output command and then use either proc plot (for lineprinter plots/low resolution) or proc gplot (for high resolution plots). You can gain full control over the plot using gplot using various commands including symbol statements, axis statements and goptions. The next illustration does a plot using gplot from the output data set. Note that within the plot in gplot you can use y*x=n where n indicates explicitly which symbol to use. You can't do this directly with the plot within reg.

```
/* shows how to save other things to a sas file (here
```

```
result2) and then control plotting */
proc reg data=a;
model measure=true;
output out=result2 p = yhat r=resid lclm=lowerm uclm=upperm
lcl=lowerp ucl=upperp;
run;
proc sort data=result2;
by true;
run;
goptions reset=all;
symbol1 v=star  color=black;
symbol2 v=p i=spline color= red;
symbol3 v=plus i=spline color=blue;
symbol4 v=plus i=spline color=blue;
proc gplot data=result2;
 plot measure*true=1 yhat*true=2 lowerm*true=3 upperm*true=4/overlay;
run;
```

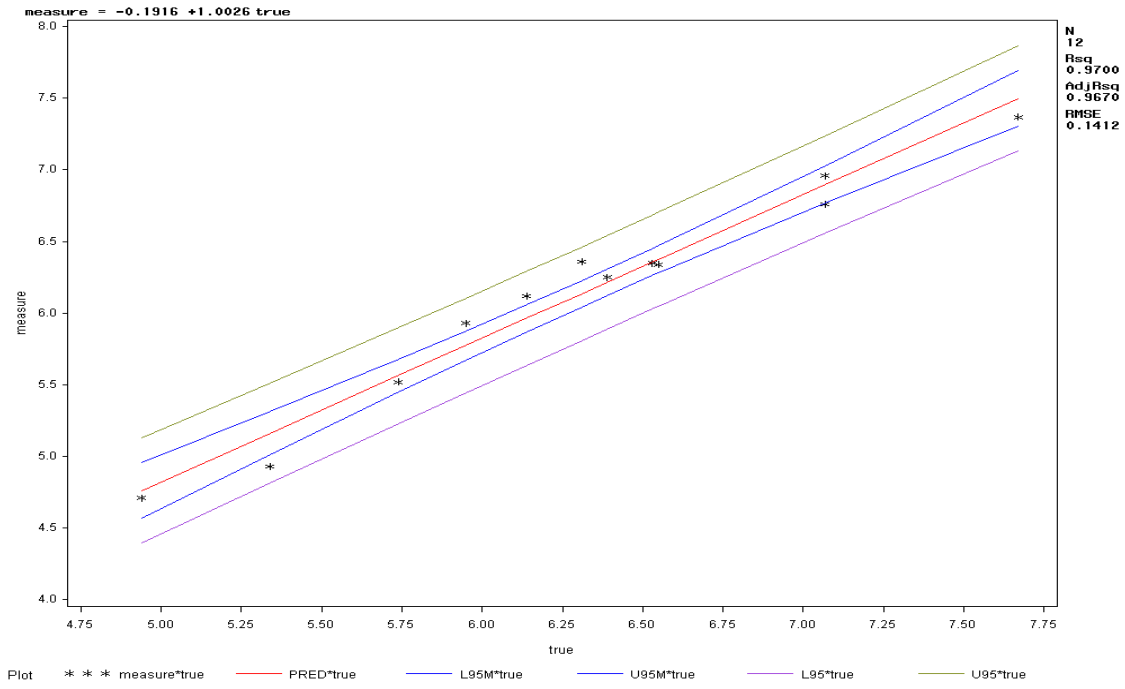


Figure 2: Plot from plot within proc reg using pred and conf options.

2.6 More on hypothesis testing:

- Does a large P-value mean that there is strong evidence that H_0 is true? NO.

  Do not interpret not rejecting H_0 as proof that H_0 is true, even with a large P-value associated with the test. When we do not reject H_0, we might have made a Type II error (not rejecting H_0 when in fact H_0 is false). We have not controlled the probabilities of a Type II error. Confidence intervals will assist us in our interpretation as will knowing something about the power of the test; more about power later.

- Does a small P-value (highly significant test) mean that an important relationships exists? NO.

  While you may have concluded that $\beta_1 \neq 0$, the magnitude of β_1 may be small enough that it is not scientifically significant. Once again confidence intervals on β_1 will assist in the scientific interpretation.

  The two points above, point out that whether we reject H_0 or do not reject H_0 there can often be difficulty with the interpretation of the conclusion of the test. **A simple reporting of the P-value associated with a test with no further information is often useless**. One could also argue that H_0 will rarely be exactly true when it contains a single point (such as $H_0 : \beta_1 = 0$) and with enough data we will almost always reject H_0. This suggests we should not approach the problem via testing but utilize confidence intervals or if testing is to be done it be done via a confidence interval.

- Does a small P-value (highly significant test) mean that X is good at predicting Y? NO.

  The key component in the ability of X to predict Y is σ^2, the variance of Y given X, the variability around the regression line. A highly significant test of $\beta_1 = 0$ simply means that we have concluded the slope is not 0, it says nothing about the size of σ^2. The ability of X to predict Y should be assessed via examination of MSE (which estimates σ^2) and prediction intervals. Note though that part of the width of the prediction interval comes from uncertainty in knowing the coefficients and you should examine how this piece relates to the piece due to MSE.

2.7 Simultaneous inferences

Suppose we have g items of interest (either parameters or future values) and the interval for the jth item will be estimated using an interval C_j, then the **simultaneous confidence level** is the probability that all g intervals are successful, where being successful means the confidence interval contains the parameter it is estimating or the prediction interval contains the random future value it is trying to predict; that is the simultaneous confidence level is

$$P(C_1 \text{ successful} \cap C_2 \text{ successful} \cap ...C_g \text{ successful}).$$

If each interval has level $1-\alpha^*$ when considered by itself, the simultaneous confidence level *is not* $1-\alpha^*$. A lower bound is $1-g\alpha*$

A quick and easy way to get the intervals to have simultaneous confidence level at least $1-\alpha$ can be developed using **Bonferroni's method.**. For g intervals get confidence level $1-\alpha/g$ for each separate interval.

Simultaneous confidence intervals for β_0 and β_1. $b_0 \pm t(1-\alpha/4; n-2)s\{b_0\}$, $b_1 \pm t(1-\alpha/4; n-2)s\{b_1\}$.

Simultaneous Bonferroni intervals for g mean values $\beta_0+\beta_1 X_j^* = \mu\{X_j^*\}$, j = 1 to g.

$b_0 + b_1 X_j^* \pm t(1-\alpha/(2g); n-2)s\{\hat{\mu}\{X_j^*\}\}$.

Simultaneous Bonferroni prediction intervals for g future values, Y_{new_j} corresponding to X_{new_j}, j = 1 to g,

$b_0 + b_1 X_j^* \pm t(1-\alpha/(2g); n-2)s\{pred_j\}$.

REMARKS:

1. The Bonferroni method is completely general and can be applied in any situation, not just those utilizing t-type intervals as above.
2. This method can also be used to test simultaneously a collection of g null hypotheses, say $H_{01}, H_{02}, \ldots, H_{0g}$. Consider

   $H_0 : H_{01}$ true $\cap H_{02}$ true $\cap \ldots \cap H_{0g}$ true .

We want to test H_0 the null hypothesis that all g null hypotheses are true. If we have individual test for each H_{0j}, each of size α^*, then a test of H_0 which rejects H_0 if we reject any H_{0j} has a simultaneous size less than or equal to $g\alpha^*$. So, if we want a test of this form of H_0 of size less than or equal to α, then the individual tests should be carried out at size $\alpha^* = \alpha/g$.

3. All the confidence intervals and/or tests do not have to use a common confidence level or size. If the confidence interval C_j has confidence level $1 - \alpha_j$, then the simultaneous coverage rate is greater than or equal to $1 - \sum_j \alpha_j$. If the test of H_{0j} has size α_j then the simultaneous size is $\leq \sum_j \alpha_j$. So choosing the α_j in any way so $\sum_j \alpha_j = \alpha$ will yield a simultaneous confidence level $\geq 1 - \alpha$ or a simultaneous test of size $\leq \alpha$.

2.7.1 Confidence Band for the regression line

It is possible in some problems to construct simultaneous confidence intervals for infinitely many values. In the simple linear regression problem, suppose we want intervals for $\{\mu\{X\} = \beta_0 + \beta_1 X, \text{ for all } X\}$. If $C(X)$ is the interval for $\mu\{X\}$ then we want $P(\mu\{X\} \in C(X), \text{ for all } X) = 1 - \alpha$. When we consider the intervals $C(X)$ over all X, this creates a two-dimensional region. In two-dimensions we can interpret the simultaneous confidence intervals as a **confidence band for the regression line** since the probability the line $\{\beta_0 + \beta_X, \text{ for all } X\}$ is contained by the region is $1 - \alpha$. This is done using what is generally known as **Scheffe's method** and in the case of simple linear regression yields what are known as **Working-Hotelling Bands**; these are given by

$$b_0 + b_1 X \pm (2F(1 - \alpha, 2, n - 2))^{1/2} s\{\hat{\mu}\{X\}\}.$$

2.7.2 Scheffe type intervals for finite number of means or predictions

When g gets very big the Bonferroni intervals can perform badly in that the intervals are bigger than is needed (although we want simultaneous level $1 - \alpha$, we get something quite a bit bigger.)

Simultaneous Scheffe intervals for g mean values $\beta_0 + \beta_1 X_j^* = \mu\{X_j^*\}$, j = 1 to g.

$$b_0 + b_1 X_j^* \pm (2F(1-\alpha, 2, n-2))^{1/2} s\{\hat{\mu}\{X_j^*\}\}.$$

Simultaneous Scheffe prediction intervals for Y_{new_j} corresponding to X_{new_j}, j = 1 to g,

$$b_0 + b_1 X_{new_j} \pm (gF(1-\alpha, g, n-2))^{1/2} s\{pred_j\}.$$

EXAMPLE:In this example we model nitrogen balance versus nitrogen intake for the purpose of determining nitrogen requirement. The data is from Kishi et al. (J. of Nutrition, 1978; "Requirement and utilization of egg protein by Japanese young men with marginal intakes of energy"). Individuals were randomized to one of three nitrogen intake levels (using a controlled diet). Nitrogen balance was determined. There is one observation per individual so the error term in the regression model consist of both among and within individual variability. The data consists of kcal (caloric input), ni (nitrogen intake), niq (an adjusted intake measure based on the protein source used) and balance (nitrogen balance).

In this example we will use the built in options in proc reg to get various confidence and prediction limits. We will then demonstrate how you can use SAS as a "calculator" and use built in functions to compute things. We will then use these features to get simultaneous confidence and prediction intervals and show how to plot a simultaneous confidence band (which is not an automatic option in SAS).

```
options linesize=80;
title 'Nitrogen intake-balance study';
data a;
infile 'kishi.dat';
input kcal ni niq nbal;
proc reg;
id ni;
model nbal=ni/covb clb clm cli;
plot nbal*ni/conf;
plot nbal*ni/pred;
run;
```

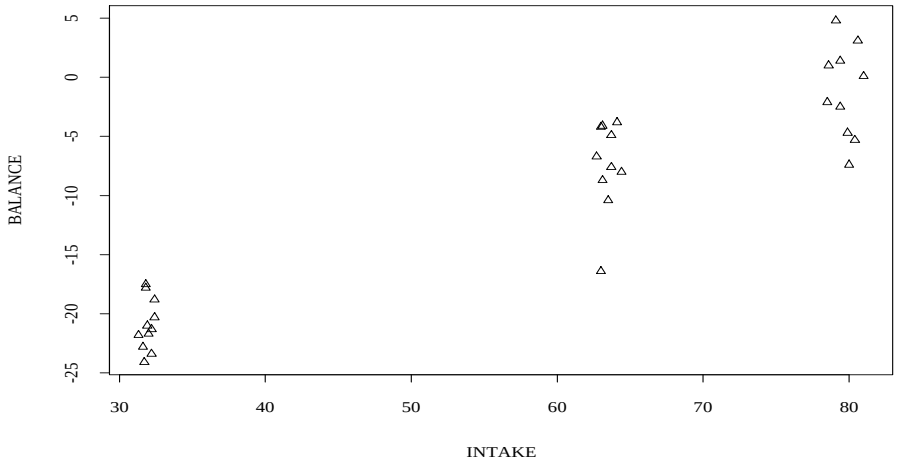


Figure 3: Nitrogen balance versus nitrogen intake from Kishi et al. (1978.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2160.44968	2160.44968	195.03	<.0001
Error	29	321.24452	11.07740		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-34.05202	1.81476	-18.76	<.0001
ni	1	0.41616	0.02980	13.97	<.0001

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-37.76362	-30.34042
ni	1	0.35522	0.47711

Covariance of Estimates

Variable	Intercept	ni
Intercept	3.2933527705	-0.051061172
ni	-0.051061172	0.0008880204

Output Statistics

Obs	ni	Dep Var nbal	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	31.6	-22.7000	-20.9012	0.9762	-22.8979	-18.9046

Obs	ni	95% CL Predict		Residual
1	31.6	-27.9951	-13.8074	-1.7988

... other cases omitted

```
/* COMPUTING PREDICTION INTERVALS AND CONFIDENCE INTERVALS FOR THE
INTAKE-BALANCE DATA DIRECTLY. THIS SHOWS HOW TO CALCULATE WITHIN THE
```

```
DATA STEP IN SAS AND TAKE ADVANTAGE OF SOME BUILT IN FEATURES OF IT.
CALCULATING THIS WAY THE SAS DATA SET a HAS ONE RECORD IN IT WITH ANYTHING
WHICH IS CALCULATED BEING A VARIABLE IN THAT RECORD */

options ls=80  nodate;
data a;
/* enter various values from proc reg output*/
b0=  -34.05202;
b1= .41616;
s2b0=3.2933527705;
s2b1=0.0008880204;
sb0b1 = -0.051061172;
mse =  11.07740;
n = 31;
seb0=sqrt(s2b0);
seb1=sqrt(s2b1);
/* GET 95% CONFIDENCE INTERVAL AND PREDICTION INTERVAL AT INTAKE = 31.6
WHICH IS THE INTAKE FOR FIRST OBSERVATION IN THE SAMPLE. */
tval = tinv(.975,n-2);   /* gets value of t distribution with
                            n-2 degrees of freedom with area
                            .975 to left of it */
yhat = b0 + b1*31.6;
seyhat = sqrt(s2b0 + (31.6**2)*s2b1 + 2*31.6*sb0b1);
l95m= yhat - tval*seyhat;
u95m = yhat + tval*seyhat;
sepred = sqrt(seyhat**2 + mse);
l95i = yhat - tval*sepred;
u95i = yhat + tval*sepred;

/* Simultaneous 95% ci's on beta0 and beta 1*/
alpha = .05;
tval2 = tinv(1-alpha/4,n-2);
lbeta0s=b0-(tval2*seb0);   ubeta0s=b0+(tval2*seb0);
lbeta1s=b1-(tval2*seb1);   ubeta1s=b1+(tval2*seb1);

/* SIMULTANEOUS CONFIDENCE INTERVALS ON THE MEAN/EXPECTED VALUE
FOR BALANCE  AT intake = 30,60 and 80*/

/* USING BONFERONNI */

tbon = tinv(1-(alpha/(2*3)),n-2);
yhat30 = b0 + b1*30;
se30 = sqrt(s2b0 + (30**2)*s2b1 + 2*30*sb0b1);
l30m = yhat30 - tbon*se30;   u30m = yhat30 + tbon*se30;
yhat60 = b0 + b1*60;
se60 = sqrt(s2b0 + (60**2)*s2b1 + 2*60*sb0b1);
l60m = yhat60 - tbon*se60;  u60m = yhat60 + tbon*se60;
yhat80 = b0 + b1*80;
se80 = sqrt(s2b0 + (80**2)*s2b1 + 2*80*sb0b1);
l80m = yhat80 - tbon*se80;  u80m = yhat80 + tbon*se80;
        /* USING SCHEFFE */
```

```
f= finv(1-alpha,2,n-2);
      /* gets value of F distribution with 2 and
       n-2 degrees of freedom with area
       1 - alpha to left of it */
mult = sqrt(2*f);
l30ms = yhat30 - mult*se30;   u30ms = yhat30 + mult*se30;
l60ms = yhat60 - mult*se60;  u60ms = yhat60 + mult*se60;
l80ms = yhat80 - mult*se80;  u80ms = yhat80 + mult*se80;
run;
title 'confidence intervals and pred. intervals at intake=31.6';
proc print;     var yhat  l95m u95m l95i u95i;
run;
title 'simultaneous CIs on coefficients';
proc print; var b0 lbeta0s ubeta0s lbeta1s ubeta1s;
run;
title 'simultaneous CIs for means at 30,60,80 Bonferonni';
proc print;   var  l30m u30m l60m u60m l80m u80m;
run;
title 'simultaneous CIs for means at 30,60,80 Scheffe';
proc print;  var  l30ms u30ms l60ms u60ms l80ms u80ms;
run;
```

```
            confidence intervals and pred. intervals at intake=31.6
        Obs      yhat        l95m        u95m        l95i        u95i
         1     -20.9014    -22.8980    -18.9047    -27.9952    -13.8075

                        simultaneous CIs on coefficients
         Obs       b0        lbeta0s     ubeta0s    lbeta1s    ubeta1s
          1     -34.0520    -38.3418    -29.7622    0.34572    0.48660

               simultaneous CIs for means at 30,60,80 Bonferonni
   Obs      l30m        u30m        l60m        u60m        l80m        u80m
   1     -24.1446    -18.9899    -10.6131    -7.55178    -3.04165    1.52321

                  simultaneous CIs for means at 30,60,80 Scheffe
   Obs      l30ms       u30ms       l60ms       u60ms       l80ms      u80ms
    1     -24.1840    -18.9504    -10.6365    -7.52836    -3.07658    1.55814
```

Now we show how you can get values to plot a confidence band for the regression line. These are the Working-Hotelling bands

```
options ls=80  nodate;
title 'Getting and plotting simultaneous confidence band';
data a;
b0=  -34.05202;   b1= .41616;
s2b0=3.2933527705;      s2b1=0.0008880204; sb0b1 = -0.051061172;
mse =  11.07740;        n = 31;
seb0=sqrt(s2b0);        seb1=sqrt(s2b1);
f = finv(.95,2,n-2);    mult = sqrt(2*f);
do x=30 to 80 by .5;
yhat=b0+b1*x;
```

```
sey = sqrt(s2b0 + (x**2)*s2b1 + 2*x*sb0b1);
l95c = yhat - mult*sey;
u95c = yhat + mult*sey;
output;
end;
proc print;    var x yhat l95c u95c;
run;
goptions reset=all hsize=3 vsize=3;
title1 c=black f=swiss 'Simultaneous Confidence band';
symbol1 l=1 i=spline;
symbol2 l=2 i=spline;
proc gplot data=a;
plot yhat*x =1 l95c*x=2 u95c*x=2/overlay;
run;
```

Getting and plotting simultaneous confidence band

Obs	x	yhat	l95c	u95c
1	30.0	-21.5672	-24.1840	-18.9504
2	30.5	-21.3591	-23.9450	-18.7733
47	53.0	-11.9955	-13.5760	-10.4151
101	80.0	-0.75922	-3.0766	1.55814

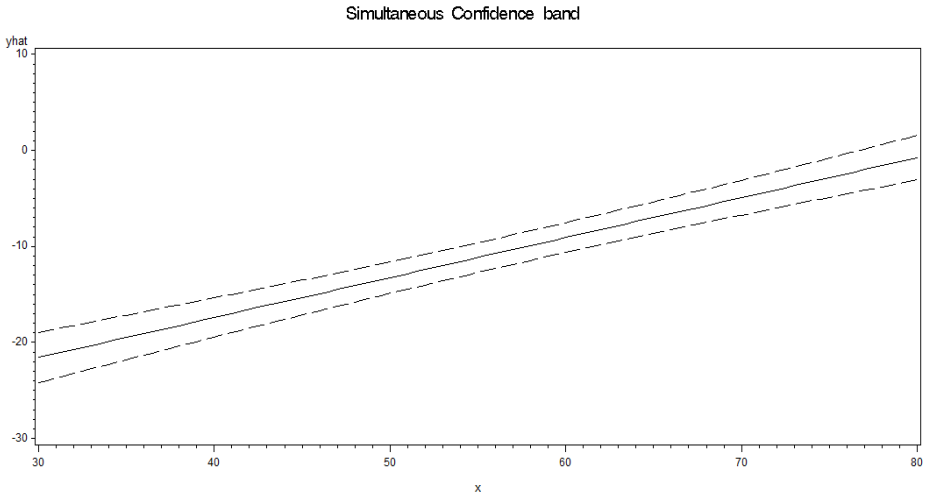


Figure 4: Kishi et al. example, confidence band

2.8 Inverse Prediction/estimation

New unit has unknown value X_{new} which we want to estimate. Observe Y_{new} from:

$$Y_{new} = \beta_0 + \beta_1 X_{new} + \epsilon_{new}$$

$\hat{X}_{new} = (Y_{new} - b_0)/b_1$.

- Approximate variance to attach to this estimate?

$$s^2\{predX\} = \frac{1}{b_1^2}(MSE + s^2\{b_0\} + \hat{X}_{new}^2 s^2\{b_1\} + 2\hat{X}_{new} s\{b_0, b_1\})$$

$$= \frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{X}_{new} - \overline{X})^2}{T}\right]$$

  (This is based on a first order Taylor series expression for $\hat{X}_{new}$ as an approximate linear function of b_0, b_1 and Y_{new}. Since $\hat{X}_{new}$ is a non-linear function of these variables we can't get the variance exactly).

- Approximate confidence interval for X_{new}: $\hat{X}_{new} \pm t(1-\alpha/2, n-2)s\{predX\}$. This is often called a "Wald" interval. If there is much uncertainty in b_1 (as often happens with small samples) this method is unreliable.

- A better method using "Fieller's method.

  This derives from the fact that

$$h(X_{new}) = \frac{Y_{new} - (b_0 + b_1 X_{new})}{[\sigma^2 + \sigma^2\{b_0\} + X_{new}^2\sigma^2\{b_1\} + 2X_{new}\sigma\{b_0, b_1\}]^{1/2}}$$

  is distributed t with $n-2$ degrees of freedom. The set of X_{new} where $h(X_{new})^2 \leq t^2$, where $t = t(1-\alpha/2, n-2)$, turns out to be a confidence set for X_{new}. The result is:

  If a test of size α of $H_0 : \beta_1 = 0$ rejects H_0 (equivalent to $c_1 > 0$ with c_1 as below) the Fieller interval is

$$\frac{c_{01}}{c_1} \pm \frac{\left[c_{01}^2 - c_0 c_1\right]^{1/2}}{c_1},$$

$c_0 = (Y_{new} - b_0)^2 - t^2(MSE + s^2\{b_0\})$

$c_1 = (b_1)^2 - t^2 s^2\{b_1\}$

$c_{01} = (Y_{new} - b_0)b_1 + t^2 s\{b_0, b_1\}$

When the P-value for the test of $H_0 : \beta_1 = 0$ is rather small the approximate interval will be similar to Fieller's method. In general it is recommended that the Fieller method be used. If β_1 is not signficantly different than 0 ($c_1 \leq 0$) then you are in trouble and should really not be doing inverse prediction.

Intuitively a reasonable way to get a confidence "set" for X_{new} is to take the set

$$C = \{X \text{ such that the prediction interval at } X \text{ contains the observed } Y_{new}\}$$

This gives a nice graphical interpretation and in fact when $c_1 > 0$ it gives exacly Fieller's interval above.

Example: Return to the pH example. Suppose that lab analyzes a sample and returns a response $Y_{new} = 6$. X_{new} is the true pH of the sample. $\hat{X}_{new} = (6 - (-.10161))/1.002633 = 6.1735$. The approximate standard error for this is 0.146796, The approximate CI is [5.8482687, 6.5024316] and the Fieller interval is [5.8436504, 6.5029045]. The two intervals are very close since the test for $\beta_1 = 0$ is highly signficant.

2.9 Regulation/inverse estimation

Estimate X such that $E(Y|X) = m$ (fixed), or estimate

$$X(m) = (m - \beta_0)/\beta_1$$

Estimate is $\hat{X}(m) = (m - b_0)/b_1$ (but biased).

Approximate confidence interval:

$$\hat{X}(m) \pm t(1 - \alpha/2)s\{\hat{X}(m)\},$$

$$s^2\{\hat{X}(m)\} = \frac{1}{b_1^2}[s^2\{b_0\} + \hat{X}(m)^2 s^2\{b_1\} + 2\hat{X}(m)s\{b_0, b_1\}$$

$$= \frac{MSE}{b_1^2}\left[\frac{1}{n} + \frac{(\hat{X}(m) - \overline{X})^2}{T}\right].$$

Fieller's method: Get an interval for $X(m)$ if the slope is significantly different than 0. Use the earlier formula for the Fieller interval in inverse prediction but now with $c_0 = (m - b_0)^2 - t^2 s^2\{b_0\}$ and $c_{01} = (m - b_0)b_1 + t^2 s\{b_0, b_1\}$ and c_1 as before.

The Fieller confidence set is the same as getting set of X values where confidence interval for $E(Y|X)$ contain m.

Example: In nitrogen intake-balance example, the objective was to determine the intake X at which the expected value balance equals 0 ($m = 0$).

$\hat{X}(0) = -(-37.76362)/.41516 = 81.8235$, $s\{\hat{X}(0)\} = 2.257597$

Approximate CI is (77.206, 86.441). Fieller interval: (77.659, 87.055).

For the use of Fieller's method in both the inverse prediction and regulation problems, if we reject $H_0 : \beta_1 = 0$ with a test of size α then the confidence set C will be a finite interval. If not (i.e., we we fail to reject H_0) then the confidence set will be either the whole real line $(-\infty, \infty)$ or the complement of a finite interval. We show in class why such regions would occur when we view the confidence set as arising from a graphical interpretation utilizing the prediction intervals or confidence intervals for the means.

3 Decomposing variability and the Analysis of Variance

(Will be looked at more generally later for multiple linear regression.)

Total uncorrected sum of squares: $SSTOU = \sum_i Y_i^2$.

Total corrected sum of squares: $SSTO = \sum_i (Y_i - \overline{Y})^2$
($SSTO/(n-1)$ = sample variance of Y_i's)

$SST0 = SSTOU - SS(mean)$, where $SS(mean) = n\overline{Y}^2$.

Error sum of squares: $SSE = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i e_i^2$.

Sum of squares due to regression: $SSR = SSTO - SSE$

$$SSTO = SSR + SSE \text{ and } SSTOU = SS(mean) + SSTO + SSR + SSE.$$

Define $MSR = SSR/(2-1)$, $MSE = SSE/(n-2)$, $F^* = MSR/MSE$.

$H_0 : \beta_1 = 0$: Can $F^* = t^{*2}$ and a test based on F^* is equivalent to the t-test that we already have.

General result: If $t^* \sim t(d)$ (read $\sim$ as "is distributed as"), then $t^{*2} \sim F(1, d)$ (is distributed F with 1 and d degrees of freedom) and $t(1-\alpha/2, d)^2 = F(1-\alpha, d)$.

3.1 R^2 and correlation

Coefficient of multiple determination: $R^2 = SSR/SSTO = 1 - SSE/SSTO$ = proportion of "total variablility" in the Y's explained by X.

Sample correlation between X and Y is

$$r = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2\right]^{1/2}}.$$

It can be shown that $r = sign(b_1)R$ and so $R = |r|$. r is a measure of the linear relationship between X and Y.

R^2 is by far the most popular measure that users like to use to assess whether they have a good model or not.

- It is difficult to interpret the implications of an R^2 of a certain size except for the extreme case of $R = 1$ (perfect linear relationship between Y and X so $SSE = 0$). - A large value of R^2 does not mean you can predict well nor does a small value of R^2 mean you cannot predict very well?

- When you can control the X values, the expected value of R^2 depends on which X_i's are used.

R^2 of limited value in assessing a model by itself.

The below demonstrates how the positioning of the X values can alter R^2. In each case the regression fit (coefficients and MSE) are identical, but the R^2 changes dramatically with higher R^2 for the more spread out X's and smaller R^2 when the X's are compressed.

OBS	TRUE	TRUE2	TRUE3	MEASURE1	MEASURE2	MEASURE3
1	7.67	10.3933	6.98917	7.11280	9.84337	6.43016
2	6.31	6.3133	6.30917	6.81000	6.81335	6.80916
3	6.14	5.8033	6.22417	6.43080	6.09325	6.51519
4	7.07	8.5933	6.68917	7.08600	8.61339	6.70415
5	6.39	6.5533	6.34917	6.31960	6.48337	6.27866
6	5.95	5.2333	6.12917	6.24180	5.52324	6.42144
7	6.53	6.9733	6.41917	6.33884	6.78336	6.22771
8	6.55	7.0333	6.42917	6.26880	6.75342	6.14764
9	5.34	3.4033	5.82417	4.46500	2.52320	4.95045
10	5.74	4.6033	6.02417	5.43300	4.29332	5.71792
11	4.94	2.2033	5.62417	4.60720	1.86327	5.29318
12	7.07	8.5933	6.68917	6.48600	8.01339	6.10415

Dependent Variable: MEASURE1

Root MSE	0.42369	R-square	0.7823

Dependent Variable: MEASURE2

Root MSE	0.42369	R-square	0.9700

Dependent Variable: MEASURE3

Root MSE	0.42369	R-square	0.4733

BELOW IS SAME FOR ALL THREE RUNS

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Error	10	1.79510	0.17951		

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-0.191766	1.06205091	-0.181	0.8603
TRUE	1	1.002656	0.16723670	5.995	0.0001

4 Random Regressors and Correlation models

Suppose now that we have a a pair of random variables (Y, X^*) and X is the realized/observed value of the predictor. X^* is the random variable and X is the value it takes on.

The book begins with using Y_1 and Y_2 rather than Y and X and then makes the connection to regression using $Y_1 = Y$ and $Y_2 = X$. They use X for both the random variable and the observed value. This is notationally convenient but it can lead to confusion in understanding what is going on. After awhile I will do the same thing, but be aware there is a difference.

Define: $E(Y) = \mu_Y$, $V(Y) = \sigma^2_Y$, $E(X^*) = \mu_X$, $V(X^*) = \sigma^2_X$ $Cov(Y, X^*) = \sigma_{XY}$,

$\rho = \sigma_{XY}/\sigma_X\sigma_Y$. (Population correlation)

The book makes it sound a bit like you need to choose between a regression model, with the X's treated as fixed, or a correlation model, with random X's. That is not the case. You can consider a regression model in either case. In the case where the predictor was random then when we talk about the behavior of $Y|X$ we are if we are more careful about writing it referring to the conditional distribution of Y given $X^* = X$.

If we assume the conditional regression model

$$Y_i|X_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$ holds, **then we can proceed with inferences for the β's and σ^2 as before.**

What relationships are there between the parameters in the conditional regression model and the parameters in the joint distribution of (Y, X^*)?

If the conditional regression model holds then (*regardless of the actual distributions involved*):

- $\beta_1 = \sigma_{XY}/\sigma^2_X$, implying $\rho = \beta_1(\sigma_x/\sigma_Y)$.
- $\beta_0 = \mu_y - \beta_1\mu_X$.

- $\sigma^2 = \sigma_Y^2 - \beta_1^2\sigma_X^2$.

Going in the other direction. If we assume that (Y, X^*) have a bivariate normal distribution (this is the model focused on in section 2.11 of the book) then this implies the condition regression model above with the additional assumption that ϵ_i has a normal distribution. The relationship between the two sets of parameters are as above (expressed in different notation in equations (2.80abc) in the book.

The parameters in the bivariate/joint model are estimated unbiasedly by

$\hat{\mu}_Y = \bar{Y}, \quad \hat{\mu}_X = \bar{X}, \quad \hat{\sigma}_Y^2 = s_Y^2, \quad \hat{\sigma}_X^2 = s_X^2, \quad \hat{\sigma}_{XY} = S_{XY}$

$S_{XY} = \Sigma_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)$, $S_X^2 = \Sigma_{i=1}^n (X_i - \bar{X})^2/(n-1)$ and S_Y^2 similar.

Notice that from our least squares approach (and a little algebra) $b_1 = S_{XY}/S_X^2$, $b_0 = \bar{Y} - b_1\bar{X}$

$$MSE = \hat{\sigma}^2 = \frac{n-1}{n-2}\left[S_Y^2 - b_1^2 S_X^2\right]$$

$\hat{\rho} = r = S_{XY}/S_X S_Y$ and $\hat{\rho}^2 = R^2$.

So estimating the parameters from the perspective of the bivariate model for (Y, X) agrees (with a little modification in $\hat{\sigma}^2$) with the approach using the conditional regression model.

- Under the bivariate normal model, independence of Y and X is equivalent to $\rho = 0$ (but in general $\rho = 0$ does not imply independence). Since $\rho = \beta_1(\sigma_X/\sigma_Y)$, where β_1 is the slope in the conditional regression model for $Y|X$. Hence our t-test of $\beta_1 = 0$ that we developed under fixed predictors is also testing $\rho = 0$ (and independence) under the normal model.
- Without joint normality, the t-test for β_1 is not necessarily testing independence. A distribution free test of independence is based on the use of the Spearman rank correlation. See example below.

Example: Brain Size and Intelligence. This is taken from DASL (Data and Story Library) web site at Carnegie Mellon.

Abstract: Are the size and weight of your brain indicators of your mental capacity? In this study by Willerman et al. (1991) the researchers use Magnetic Resonance Imaging (MRI) to determine the brain size of the subjects. The researchers take into account gender and body size to draw conclusions about the connection between brain size and intelligence.

Willerman et al. (1991) conducted their study at a large southwestern university. They selected a sample of 40 right-handed Anglo introductory psychology students who had indicated no history of alcoholism, unconsciousness, brain damage, epilepsy, or heart disease. These subjects were drawn from a larger pool of introductory psychology students with total Scholastic Aptitude Test Scores higher than 1350 or lower than 940 who had agreed to satisfy a course requirement by allowing the administration of four subtests (Vocabulary, Similarities, Block Design, and Picture Completion) of the Wechsler (1981) Adult Intelligence Scale-Revised. With prior approval of the University's research review board, students selected for MRI were required to obtain prorated full-scale IQs of greater than 130 or less than 103, and were equally divided by sex and IQ classification. The MRI Scans were performed at the same facility for all 40 subjects. The cans consisted of 18 horizontal MR images. The computer counted all pixels with non-zero gray scale in each of the 18 images and the total count served as an index for brain size.

Since the sampling was done in a way that sampled from four different groups; two IQ levels crossed by gender, we need to account for that. Can't just view as a sample of 40 from the total population since controlled the sample sizes so there were 10 from each category. We will carry out regression analysis separately for each group. The 10 in the group are a random sample from the associated population (e.g., females with FSIQ scores less than 103).

```
data a;
infile 'c:\s597\data\Brain.dat';
input Gender $ FSIQ VIQ PIQ Weight Height mriCount;
if fsiq > 129then iqgroup=1;
if fsiq < 104 then iqgroup=2;  run;
proc print;
var gender fsiq mricount iqgroup;  run;
proc sort;
by gender iqgroup;      run;
```

```
proc reg;
model fsiq=mricount;
plot fsiq*mricount/conf;
plot fsiq*mricount/pred;
by gender iqgroup;       run;
proc reg;
model fsiq=mricount;
plot fsiq*mricount/conf;
plot fsiq*mricount/pred;     run;
proc corr;
var fsiq mricount;
by gender iqgroup;    run;
--------------------------- Gender=Female iqgroup=1 -----------------------
                                    Sum of           Mean
 Source                   DF        Squares         Square
 Error                     8       76.73939        9.59242
              Root MSE              3.09716    R-Square     0.1290
                       Parameter       Standard
 Variable     DF       Estimate          Error    t Value      Pr> |t|
 Intercept     1      119.16618       13.94076       8.55      <.0001
 mriCount      1     0.00001727     0.00001586       1.09      0.3082
--------------------------- Gender=Female iqgroup=2 -----------------------
                                    Sum of           Mean
 Source                   DF        Squares         Square
 Error                     8      437.82194       54.72774
              Root MSE              7.39782    R-Square     0.1839
                       Parameter       Standard
 Variable     DF       Estimate          Error    t Value      Pr> |t|
 Intercept     1       25.55433       47.67909       0.54      0.6066
 mriCount      1     0.00007534     0.00005611       1.34      0.2162
--------------------------- Gender=Male iqgroup=1 -------------------------
                                    Sum of           Mean
 Source                   DF        Squares         Square
 Error                     8      113.78297       14.22287
              Root MSE              3.77132    R-Square     0.0557
                       Parameter       Standard
 Variable     DF       Estimate          Error    t Value      Pr> |t|
 Intercept     1      121.44475       24.84643       4.89      0.0012
 mriCount      1     0.00001749     0.00002545       0.69      0.5114
--------------------------- Gender=Male iqgroup=2 -------------------------
                                    Sum of           Mean
 Source                   DF        Squares         Square
 Error                     8      340.80641       42.60080
              Root MSE              6.52693    R-Square     0.5107
                       Parameter       Standard
 Variable     DF       Estimate          Error    t Value     Pr> |t|
 Intercept     1      -11.92568       35.85256      -0.33      0.7480
 mriCount      1     0.00011069     0.00003831       2.89      0.0202
```

```
------------------------------  EVERYBODY  ------------------------------
                                  Sum of           Mean
Source                    DF     Squares         Square
Error                     38       19725      519.07660
             Root MSE            22.78325    R-Square     0.1279
                    Parameter       Standard
Variable     DF       Estimate          Error    t Value      Pr> |t|
Intercept     1        5.16770       46.00819       0.11      0.9112
mriCount      1     0.00011915     0.00005047       2.36      0.0235

--------------------------- Gender=Female iqgroup=1 ----------------------
                           The CORR Procedure
                    Pearson Correlation Coefficients, N = 10
                           Prob > |r| under H0: Rho=0
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.35910
                                                   0.3082
                      mriCount       0.35910       1.00000
                                     0.3082
                    Spearman Correlation Coefficients, N = 10
                           Prob > |r| under H0: Rho=0
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.25849
                                                   0.4708
--------------------------- Gender=Female iqgroup=2 ----------------------
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.42887
                                                   0.2162

                    Spearman Correlation Coefficients, N = 10
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.38298
                                                   0.2747
--------------------------- Gender=Male iqgroup=1 ------------------------
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.23610
                                                   0.5114
                    Spearman Correlation Coefficients, N = 10
                                      FSIQ      mriCount
                      FSIQ           1.00000      -0.04295
                                                   0.9062
--------------------------- Gender=Male iqgroup=2 ------------------------
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.71462
                                                   0.0202
                    Spearman Correlation Coefficients, N = 10
                                      FSIQ      mriCount
                      FSIQ           1.00000       0.67684
                                                   0.0316
```

Recall that the Root MSE is $\hat{\sigma}$, the estimate of the standard deviation in FSIQ scores at a particular MRI count. We do not do very well in prediction as indicated by the Root MSE values and the prediction plots (the accompanying plots will be shown in class.) Notice that the significance of the slope is not indicative of how "good" the relationship is. For example, group 2 males have a p-value of .0202 for the slope, but an R-square of .5107 and $\hat{\sigma} = 6.53$. This also seen in the run with everybody, but that analysis would not be appropriate since we sampled by groups. We will look later at how to compare regressions across groups. The output from proc corr has been edited. It gives the sample correlation r as well as the P-value associated with a test for $\rho = 0$ under the assumption of bivariate normality. This test is equivalent to the t-test for $\beta_1 = 0$.

5 Simple residual analysis and other diagnostic measures for assessing model assumptions in simple linear regression

ith residual: $e_i = Y_i - \hat{Y}_i$.

The residuals are random variables and hence have some expected value, variance and covariance/correlation structure. Here we ignore the variance and covariance structure but these will be used with more advanced diagnostic measures used later.

As random variables, each residual has expected value 0, that is $E(e_i) = 0$ (**if the model is correct**).

Also true that $\sum_{i=1}^{n} e_i = 0$ (but this will not be true for a model fit with no intercept.)

The theoretical variance of e_i is not σ^2 (which is the variance of ϵ_i) but is approximately σ^2 for reasonable size n. (More on more sophisticated procedures not utilizing this approximation later in the course.)

ith semi-studentized residual: $e_i/MSE^{1/2} = e_i/\hat{\sigma}$.

In any of discussion below the plots can be with either residuals or semi-studentized residuals. (There will be more sophisticated studentized residuals used later.)

- Assessing Linearity.

  - Plot of Y versus X

  - Plot of residual e_i versus X_i or $\hat{Y}_i$.

  - Test for lack of fit when there are multiple values (details later).

  A systematic trend in the plot of the residuals versus an explanatory variable or the fitted values indicates that there is a problem with the assumed regression model. If the model is correct the residuals should, in general, center around 0 across the explanatory variables and the fitted values.

  Without linearity we have the wrong model for how $E(Y|X)$ depends on X and obviously inferences do not make sense.

- Assessing constant variance.

  - Plot of residual e_i versus X_i or $\hat{Y}_i$.

  - Plot of $|e_i|$ or e_i^2 versus X_i or $\hat{Y}_i$.

  - Tests for equal variance (later).

  If the regression model looks okay, a violation of the constant variance assumption will be indicated by a change in the spread of the residuals as X or $\hat{Y}$ changes.

  This can also be seen (sometimes more easily) by plotting the absolute or squared residual rather than the residual itself. *In this case look to see constant variance is indicated by the average absolute or squared residual not changing over X or $\hat{Y}$.*

  If variances are not constant, then standard errors for coefficients are incorrect as are standard errors for confidence intervals on means, prediction, etc. All the inferences are off, but by how much depends on how severe the assumption is violated. The same is true if the errors are in fact correlated and we assume they are not.

- Assessing independence or uncorrelatedness of error terms

  How to do this often depends on knowing what might be causing the correlation. One place this is a concern is when the observations are collected over different times. *Assuming data is ordered by time.*

  - Plot e_i versus i.

  - Plot e_{i+1} versus e_i for $i = 1, \ldots, n-1$

  An ad-hoc approach to testing is to analyze the correlation between e_{i+1} and e_i. We'll return to handling data collected over time in more detail later.

- Assessing normality.

  If the errors are not normal, tests, confidence intervals and prediction intervals will not behave as advertised for small samples. As the sample size gets bigger (and with some restrictions on how the X's are chosen) inferences for the coefficients and linear combinations of them (including the mean of Y at a given X) are approximately correct even without normality of the errors. For these purposes we need normality most for small n, but this is where it is difficult to evaluate. Prediction intervals as we have given them depend heavily on the normality assumption regardless of n. This is because the distribution of the single Y_{new} is critical. Also inferences for σ^2 based on a chi-square depend on normality regardless of sample size.

  Get descriptive statistics on the residuals:

  - stem-and-leaf plot, box plot, histogram and smoothed histogram.

  - normal probability plots and associated tests for normality (but tests for normality can be of limited value since often have little power (ability to detect non-normality) with small or moderate sample size or can be overpowered (pick up small but unimportant deviations from normality) with large sample sizes.

Note: If there are unequal variances or correlation in the errors these need to be accounted for (done later) before the normality assessment.

Cholesterol example: Calibration of serum cholesterol. Notice there are replicates (multiple samples with the same true value). Residual plots indicates the linear fit has a problem, but need to judge just how bad the consequences are in this case since the residuals are small relative to what we are estimating.

Working in SAS. Note: SAS 9.3 (newest version) automatically produces a number of residual plots when you run the regression.

```
option ls=80 nodate;
title 'calibration of serum cholesterol';
data a;
infile 'e:\s505\chol.dat';
input true measured;
proc print;
run;
proc reg;
model measured=true/covb clb cli clm p;
plot measured*true;
plot r.*(true p.);
run;
```

```
                          Obs    true    measured
                           1       50        55
                           2       50        54
                           3       50        53
                           4      200       204
                           5      200       203
                           6      200       207
                           7      400       385
                           8      400       382
                           9      400       382
                             The REG Procedure
                            Analysis of Variance
                                   Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      1        162559         162559    5625.07    <.0001
Error                      7     202.29279       28.89897
Corrected Total            8        162761
             Root MSE              5.37578    R-Square     0.9988

                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|
     Intercept     1       10.78829        3.24719       3.32      0.0127
     true          1        0.93739        0.01250      75.00      <.0001
```

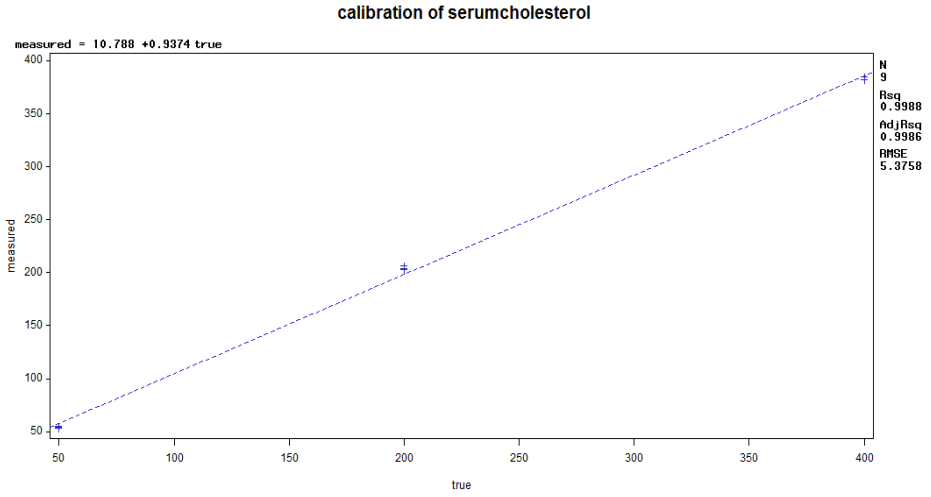


Figure 5: Cholesterol fit

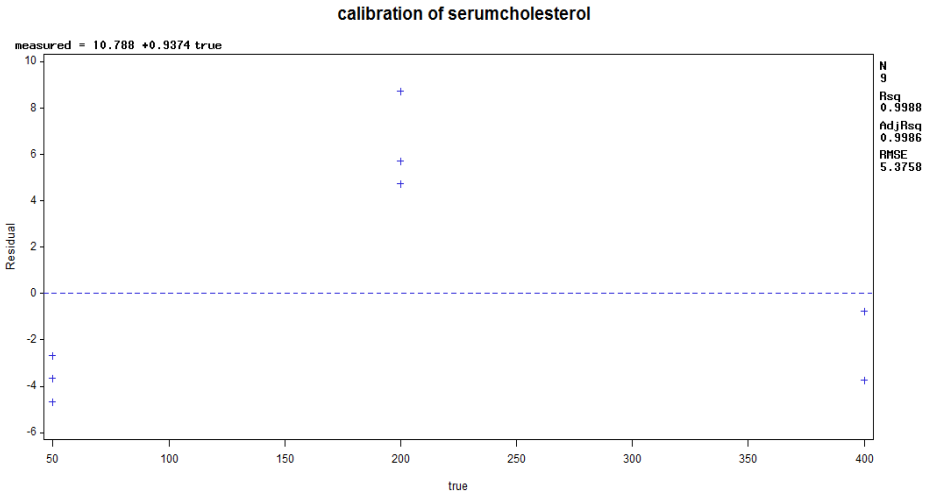


Figure 6: cholesterol residual versus x

Example: Puffins: This example uses data from an article entitled "Breeding success of puffins on different habitats on Great Island, Newfoundland" from Ecological Monographs. The variables are Y = nesting frequency of puffins, X_1 = grass cover, X_2 = soil depth, X_3 = angle of slope and X_4 = distance from cliff edge.

Here we examine just a regression of Y on X = *slope*;

SAS code:

```
title 'puffin example ';
options pagesize=60 linesize=80;
data a;
infile 'e:\s505\data\puffins.dat';
input  success grass soil slope distance;
run;
proc reg;
model success=slope/covb clb cli clm p;
plot success*slope/conf pred;
plot r.*(slope p.);
output out =result p=yhat r = resid;
   /* saves predicted (now called yhat) and residual
     (now called resid) to sass file result. This will
     also have the original data. you could use these
     to customize your own plots. For example
     proc gplot data=result;
     plot resid*yhat;
     will plot the residual versus predicted value and you
    can use other gplot options to make the graph fancier.
   See below for plot using absolute residual */
run;
ods graphics off;   /* in 9.3 this will send graphs to graph window
                     rather than to results window which is in html form.
                    Don't use this if just want to save out of results window*/
proc univariate data=result plot normal;  /* normal will do tests for
                                      normality */
var resid;
hist resid/kernel(color = black); /* gives a histogram with smoothing*/
run;
  /* get an isolated probability plot of the residuals*/
proc capability data=result noprint;
probplot resid;
run;
/* plot of absolute residual versus x */
data b;
set result;
absresid = abs(resid);
run;
proc gplot data=b;
plot absresid*slope;
run;
```

```
                                puffin example                                1
```

```
                          Analysis of Variance
                                 Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F
Model                     1    1336.49664    1336.49664      83.28    <.0001
Error                    36     577.71389      16.04761
Corrected Total          37    1914.21053

            Root MSE              4.00595    R-Square     0.6982

                        Parameter       Standard
     Variable     DF      Estimate          Error    t Value    Pr > |t|
     Intercept     1      -0.02635        1.06591      -0.02      0.9804
     slope         1       0.51404        0.05633       9.13      <.0001

                       The UNIVARIATE Procedure
                      Variable:  resid  (Residual)
                               Moments
       N                          38    Sum Weights                38
       Mean                        0    Sum Observations            0
       Std Deviation      3.95144136    Variance           15.6138888
       Skewness           0.43683797    Kurtosis             -0.63306
                       Basic Statistical Measures
             Location                      Variability
          Mean      0.00000     Std Deviation            3.95144
          Median   -0.52545     Variance                15.61389
          Mode     -4.08595     Range                   14.59826
                                Interquartile Range      5.87900

                          Tests for Normality
          Test                 --Statistic---    -----p Value------
          Shapiro-Wilk         W      0.95571    Pr < W      0.1374
          Kolmogorov-Smirnov   D     0.104559    Pr > D     >0.1500
          Cramer-von Mises     W-Sq  0.071923    Pr > W-Sq  >0.2500
          Anderson-Darling     A-Sq  0.510691    Pr > A-Sq   0.1933
```

Other SAS plots and output in class.

R code and output.

```
# THIS FITS A SLR USING THE PUFFIN DATA AND DOES DIAGNOSTICS
# WITH THE RESIDUALS. IT ALSO HAS A NUMBER OF THE PREVIOUS
# THINGS WE HAVE DONE, PROGRAMMED IN TERMS OF JUST y and x
# SO IT CAN BE REUSED.

data<-read.table("e:/s505/data/puffins.dat") # no names
attach(data)
success<-V1; grass<-V2;  soil<-V3; slope<-V4; distance<-V5  # rename the variables

# for convenience just relabel to y and x. Then you can
# reuse code without making many changes in variable names.
```

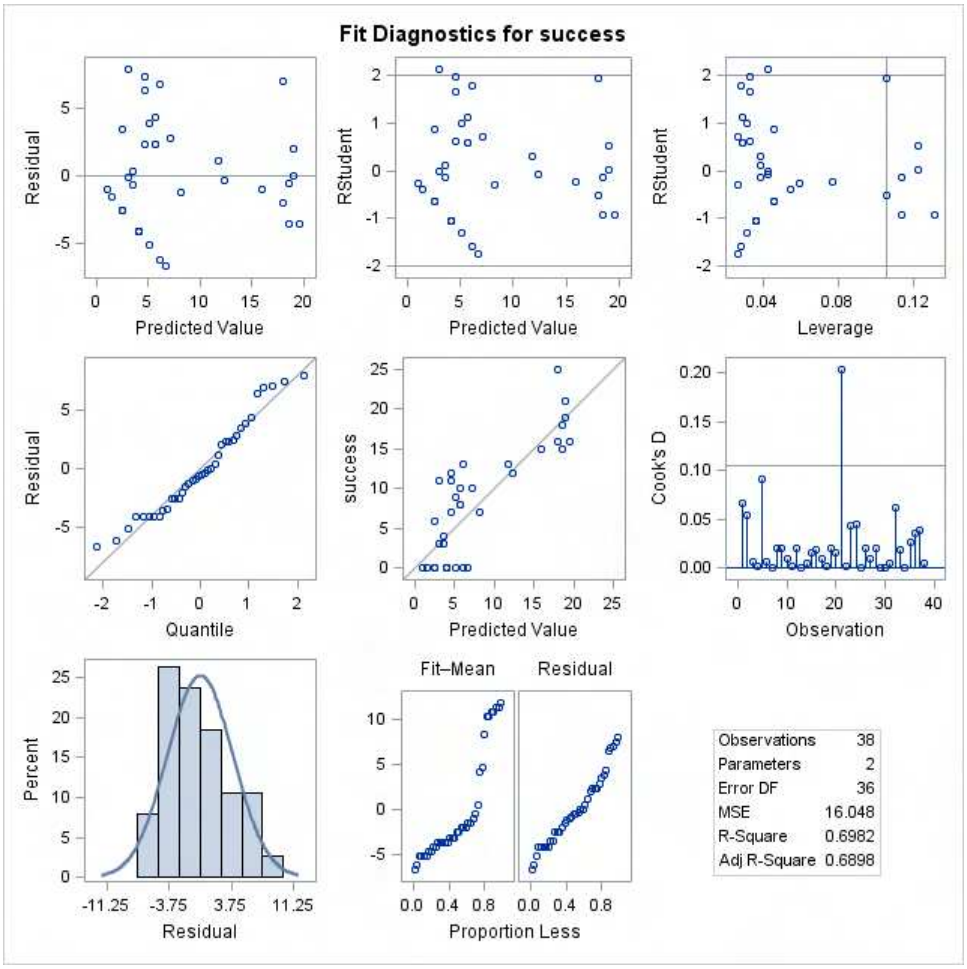


Figure 7: Diagnostic plots from SAS 9.3

```
y<-success
x<-slope

regout<-lm(y ~x)
summary(regout)
anova(regout)
confint(regout)
vcov(regout)
fits <-fitted(regout)
residual<-residuals(regout)
  par(mfrow=c(1,1))  #reset graph to single plot per page
plot(x,y,xlab="slope", ylab="success",main = "Puffin example")
lines(x,fits)

 # confidence and prediction internvals
```

```
xsort<-sort(x)                    # use xsort so don't change
                                  # order of original x's
xvalues<-data.frame(xsort)
 cat("confidence intervals for E(Y|X)")
cintervals<- predict(regout,xvalues,interval = "confidence","\n")
 # cintervals      #uncomment this if want a listing
 cat("prediction intervals for E(Y|X)")
pintervals<- predict(regout,xvalues,interval = "predict","\n")
 # pintervals      #uncomment this if want a listing
lines(x, cintervals[,2],type="l",lty=2)
lines(x, cintervals[,3],type="l",lty=2)
lines(x, pintervals[,2],type="l",lty=3)
lines(x, pintervals[,3],type="l",lty=3)

# if want to plot and save the confidence and prediction
# intervals stop here or they will get overwritten. Then
# run block below to get residual plots

########Residual analysis: plots: one way ##########
par(mfrow=c(2,2))  # this makes the graph window have 4 panels
                   # with 2 rows and 2 columns

plot(x,residual)
plot(fits,residual)
hist(residual)
lines(density(residual)) # gets smooth histogram via
                         # density(residual), then overlays it.
qqnorm(residual)

## residual analysis, summary statistics and tests for normality
summary(residual)
shapiro.test(residual) # this just shows the Shapiro-Wilks test.
                       # other tests requiring installing the
                       # package nortest

 # plot of absolute residual
 par(mfrow=c(1,1))
 absresid<-abs(residual)
 plot(x,absresid)
```

Here is some of the output.

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.6561 -3.2452 -0.5255  2.3930  7.9421

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02635    1.06591  -0.025     0.98
x            0.51404    0.05633   9.126 6.75e-11 ***
```

```
---

Residual standard error: 4.006 on 36 degrees of freedom
Multiple R-squared: 0.6982,     Adjusted R-squared: 0.6898
F-statistic: 83.28 on 1 and 36 DF,  p-value: 6.745e-11

Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1336.50 1336.50  83.283 6.745e-11 ***
Residuals 36  577.71   16.05
---

> ## residual analysis, summary statistics and tests for normality

 summary(residual)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.6560 -3.2450 -0.5255  0.0000  2.3930  7.9420

 shapiro.test(residual) # this just shows the Shapiro-Wilks test.

        Shapiro-Wilk normality test
data:  residual
W = 0.9557, p-value = 0.1374
```

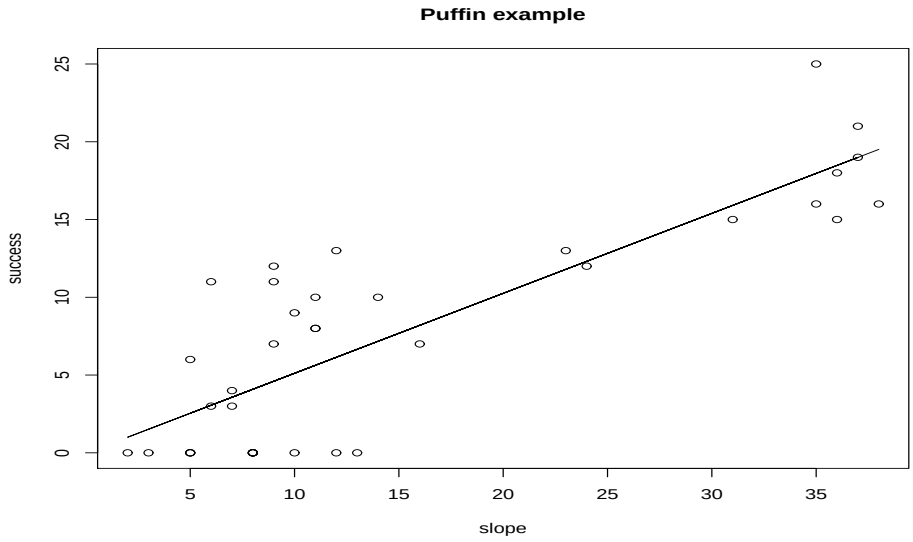


Figure 8: Data and fit for Puffin example; from R.

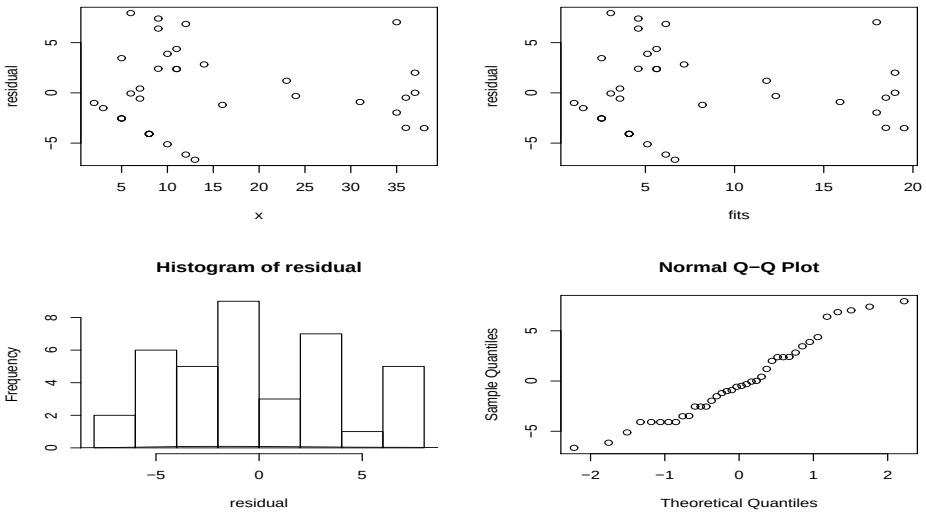


Figure 9: Residual plots for puffin data; from R.

Assessing Normality.

The normal probability plot plots the residual versus its approximate expected value under normality (given by equation (3.6)). The book describes a correlation test for normality (section 3.5) where they tests for zero correlation between the residual and its expected value under normality. It is better to use the tests

for normality (Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling) given in SAS (and many other programs). Anderson-Darling is generally thought to be the best of these. These tests are primarily based on the difference between the cumulative distribution function of the observed data and the one under normality. As always we have to be careful with testing. With small to moderate sample sizes any of these tests can have low power (probability of correctly rejecting H_0 : when the errors are not normally distributed). Unfortunately it is with smallish sample sizes that we most need normality for our inferences to be valid, but it is here that it is most difficult to assess. With large sample sizes the tests can detect small, but possibly unimportant, deviations from normality.

More on assessing and testing for constant variance.

There are a variety of tests for equality of variance, all approximate. They essentially break up depending on either

- There is a natural grouping of the data by the X values where the variance could be considered essentially constant within each group and the test is designed to test equal variances across the groups. Levene's test and the Brown-Forsythe test fall in this category. On pages 116-117 the book uses this for two groups. Two notes though:

  - This can be easily generalized (details later) to multiple groups.

  - This should not be used unless the X's within a group are very similar. It is not recommended that you arbitrarily break the data into groups and use this procedure.

- The variance is modeled as a function of either X or the mean of $Y|X$ and one tests for constant variance within this model. Having a model for the variance will also be helpful when we account for unequal variances, if it exists.

Modelling the variance and using regression to assess constant variance.

Assume a model for $\sigma_i^2 = V(\epsilon_i)$ or σ_i as a function of X_i or $\mu_i (= \beta_0 + \beta_1 X_i)$. Examine the model using e_i^2 in place of σ_i^2, $|e_i|$ in place of σ_i and/or $\hat{Y}_i$ in place of μ_i. Can get a rough test for constant variance within this framework.

The model used in section 3.6 is $log(\sigma_i^2) = \gamma_0 + \gamma_1 X_i$. One approach is to regress $log(e_i^2)$ versus X_i and use the t-test for $\gamma_1 = 0$. This is easy to implement. The **Breusch-Pagan** test in section 3.6 uses this model but gives a specialized test of constant variance which is *not!* just testing γ_1 in the regression of $log(e_i^2)$ on X_i.

A very popular model for the variance (the so-called power model) is

$$\sigma_i = \theta_1 \mu_i^{\theta_2},$$

or $log(\sigma_i) = log(\theta_1) + \theta_2 log(\mu_i)$, where $\mu_i = E(Y_i|X_i)$. The variance is constant if $\theta_2 = 0$.

Examine the model through a plot of $log|e_i|$ versus $\log(\hat{Y}_i)$, which should be roughly linear if the model is correct. A test for slope = 0 in the regression of $log|e_i|$ on $\log(\hat{Y}_i)$ is approximate test of the null hypothesis of constant variance (assuming the model is okay.)

Esterase Assay example. The data is from Carroll and Ruppert's *Transformation and Weighting in Regression*. Ester is the amount of esterase in a sample and count is the number of bindings observed in a radioiumnoassay. The objective is to model the radioactive binding counts as a function of the level of esterase. In this example the variance is clearly increasing as the level of esterase increases. This is seen from the original scatterplot, the residual plot and the plot of absolute residual.

The plot of $log(e_i^2)$ versus X_i does not look linear. A plot of $log|e_i|$ on $\log(\hat{Y}_i)$ looks pretty linear so we'll entertain the power model. A linear regression of $log|e_i|$ on $\log(\hat{Y}_i)$ yields

Dependent Variable: labsr

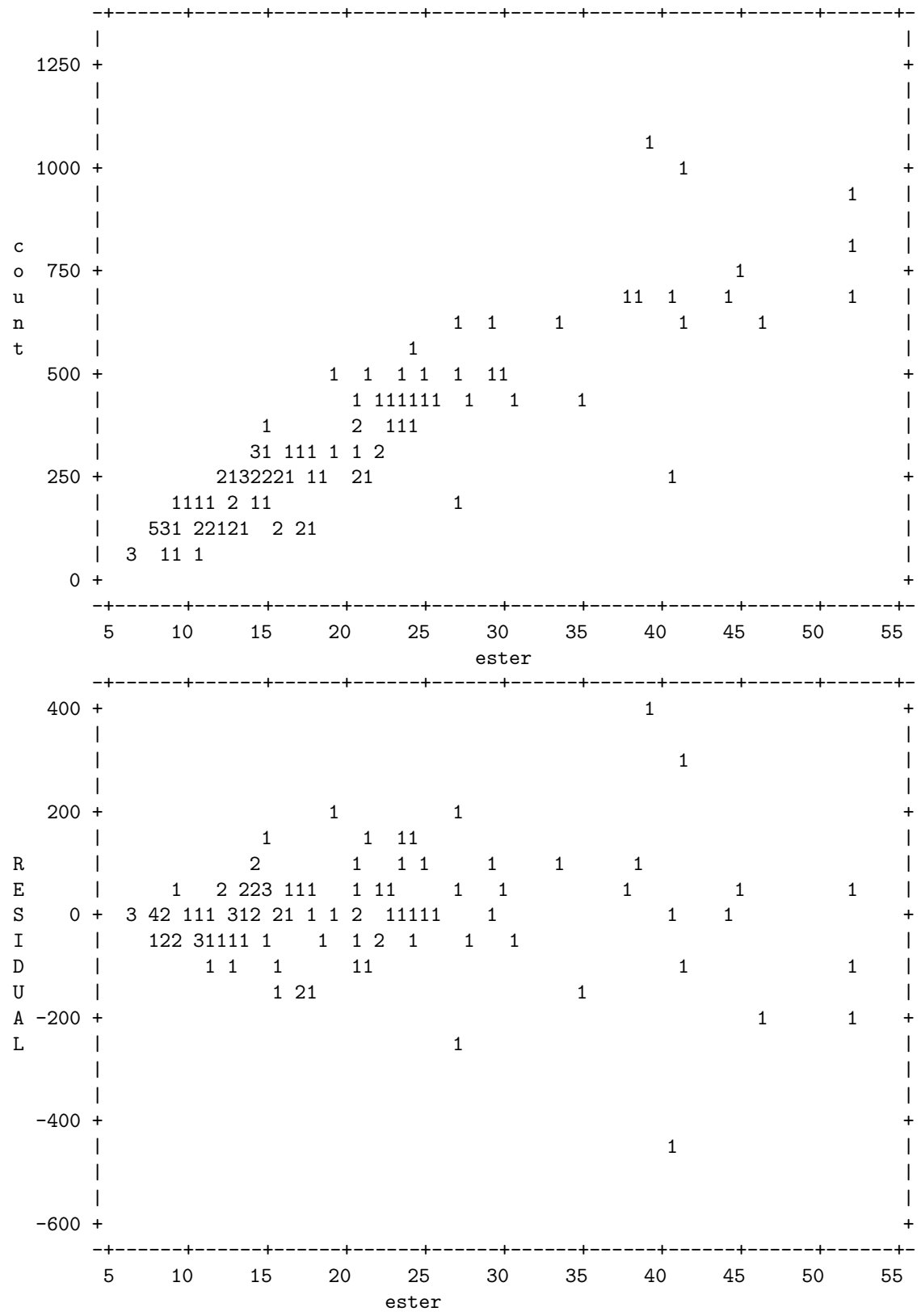
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.44942	1.22119	-1.19	0.2380
lpred	1	0.89486	0.21516	4.16	<.0001

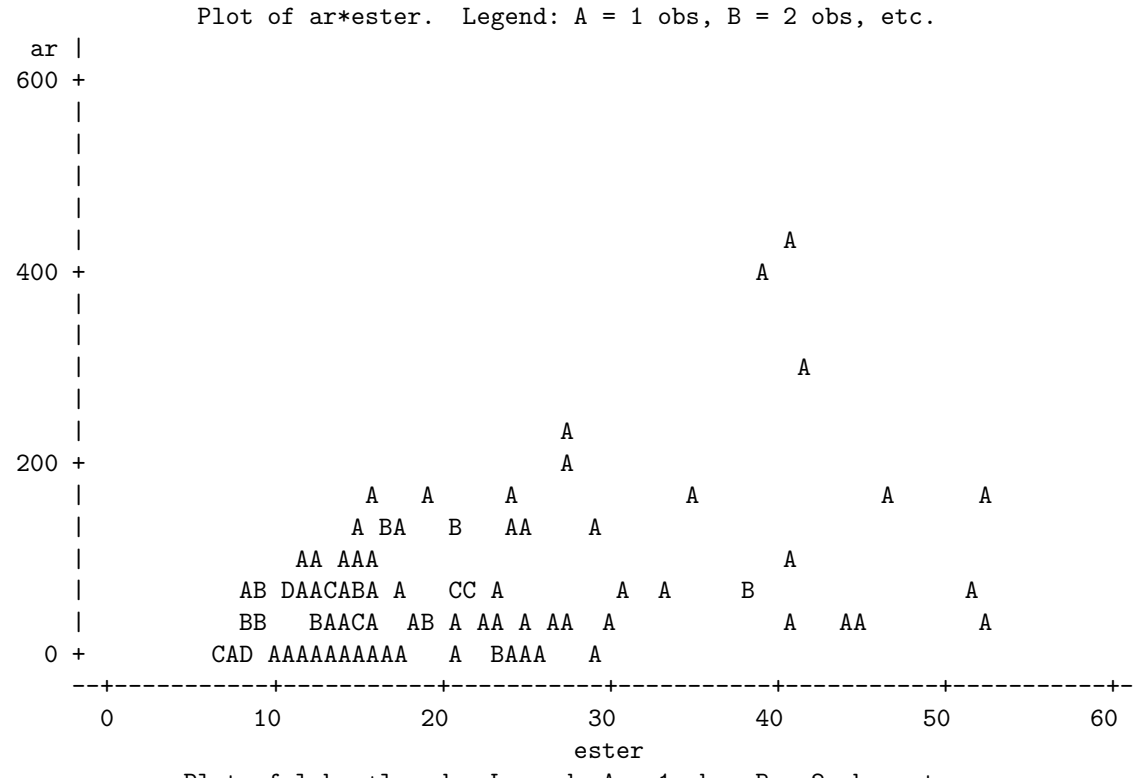
The slope .89486 is a rough estimate of θ_2, while -1.44942 estimates $log(\theta_1)$. The t-test is testing constant variance via an approximate test of $H_0 : \theta_2 = 0$.

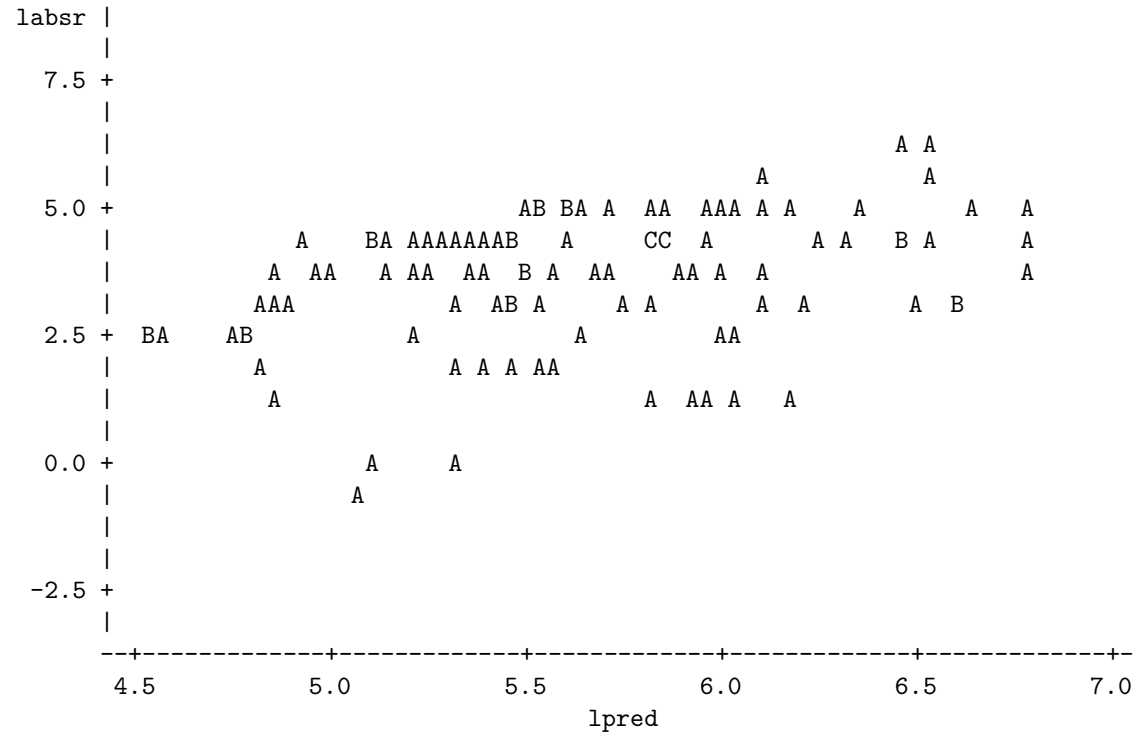
```
title 'Esterase Hormone data ';
data a;
infile 'c:\s505\ester.dat';
```

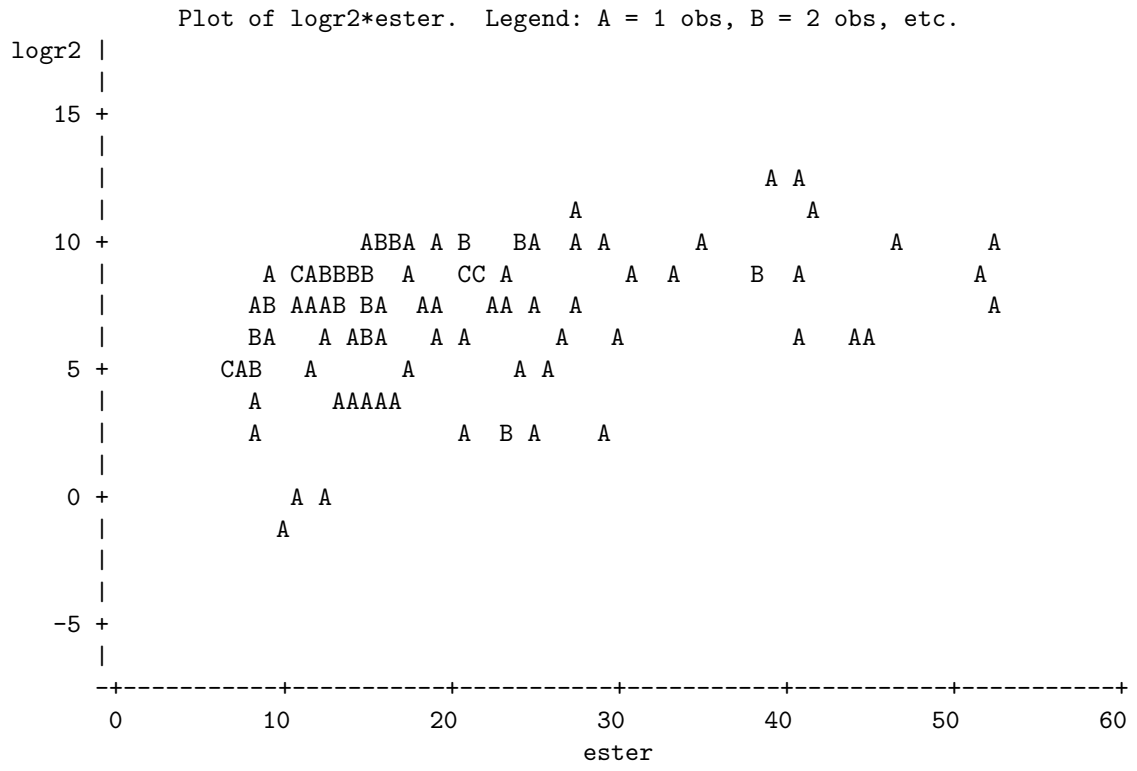
```
input ester count;
run;
proc reg;
model count=ester/covb;
plot  r.*(ester p.);
output out=result p=yhat r=resid;     run;
data c;
set result;
r2=resid**2;
ar=abs(resid);
labsr=log(abs(resid));
lpred=log(yhat);     run;
proc gplot data=c;
title 'plots for esterase assay example';
plot (count r2 ar)*ester labsr*lpred;    run;
proc reg data=c;
model labsr=lpred;
run;
```

NOTE: Here and some later examples, I am leaving in low resolution line printer plots (this is for convenience in terms of formatting). We'll look at the good plots in class and I'll also have the R code and plots for these examples.


count
1250
1000
750
500
250
0
5 10 15 20 25 30 35 40 45 50 55
ester
RESIDUAL
400
200
0
-200
-400
-600
5 10 15 20 25 30 35 40 45 50 55
ester


Plot of ar*ester. Legend: A = 1 obs, B = 2 obs, etc.
ar
600
400
200
0
0
10
20
30
40
50
60
ester


Plot of labsr*lpred. Legend: A = 1 obs, B = 2 obs, etc.
labsr
7.5
5.0
2.5
0.0
-2.5
4.5
5.0
5.5
6.0
6.5
7.0
lpred



5.1 Replicate data, tests for lack of fit and tests for constant variance.

Y_{ij} = response for ith observation in group j, $i = 1$ to n_j, $j = 1$ to c groups.

One-way model: $Y_{ij} = \mu_j + \epsilon_{ij}$

$\hat{\mu}_j = \bar{Y}_j = \sum_i Y_{ij}/n_j.$

$SSPE = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \qquad MSPE = SSPE/(n-c).$

Testing equal means in the one-way model/The one-way analysis of variance.

Assuming uncorrelated errors with constant variance and normality (or large sample sizes). Consider the null model that all μ_j are equal. This is tested with $F_{anova} = MS_G/MSPE$, where $MS_G = SSG/(c-1)$ with $SSG = \sum_j n_j(\bar{Y}_j - \bar{Y})^2$. Compare using an F with $c-1$ and $n-c$ degrees of freedom. (This can be motivated using the General linear test approach which we will do later). For two groups this

is equivalent to the two sample t-test. In SAS, this test can be run using proc anova or proc glm.

Testing for linearity in regression framework.

$H_0 : \theta_j = \beta_0 + \beta_1 X_j$, $H_A : \theta_j$ not specified.

(I've used θ_j rather than μ_j in book, since we have used μ_i for the expected value of Y_i where i indexes the overall order in the sample.)

$SSLF = SSE - SSPE$, $MSLF = SSLF/(c-2)$,$F_{lof} = MSLF/MSPE$

Compare using an F with c-2 and n-c degrees of freedom.

Levene and Brown-Forsythe test for equal variances.

Working with the residuals and assuming have the model correct. Divide observations into c groups, where the X's are similar in each group. (Later with more than one predictor we could group on the fitted values.) Form residuals with e_{ij} being the ith residual in group j. The various Levene type tests run a one way anova on values d_{ij} chosen such that under equal variance we would expect the d's to have common expected value. This means using F_{anova} above but with d's in place of Y's.

Levene's (not modified) test uses either $d_{ij} = |e_{ij} - \bar{e}_j|$ or $d_{ij} = (e_{ij} - \bar{e}_j)^2$, where $\bar{e}_j = \sum_i e_{ij}/n_j$. The modified Levene's test, also called Brown-Forsythe test uses $d_{ij} = |e_{ij} - \tilde{e}_j|$ where $\tilde{e}_j$ is the median of the e_{ij} in group j. These test can be carried out by constructing the appropriate d value and running a one-way analysis of variance.

NOTE: If we run Levene's test in Anova directly with the grouping (no regression model and residuals used) then it corresponds to running a one-way anova F-test on $d_{ij} = |Y_{ij} - \bar{Y}_j|$ or $d_{ij} = (Y_{ij} - \bar{Y}_j)^2$.

Example: Using the cholesterol data to test for lack of fit. Also runs Levene's test for equal variance across the three groups (within a model that allows a separate mean for each group, not within the linear regression framework.)

```
option ls=80 nodate;
data a;
infile 'a:\chol.dat';
input true measured;  run;
proc means;
class true;
var measured;  run;
proc reg;  model measured=true;   run;
  /* The proc anova fits a model with a different mean allowed for each
    value of true (need replications at some of the true values for
     this to work). The SSE from the anova here is what is called
     SSPE in the test for lack of fit.*/
proc anova;
class true;
model measured=true;
means true/ hovtest=levene;    /* This is one of many tests available in the
                                  hovtest to test for equal variances. */
run;
    data b;
/* carry out test for lack of fit using values from the regression
fit and from the one-way anova. */
sspe= 16.6667;   sse=202.29279;   n=9;   c=3;
 mse=sse/(n-2);   mspe=sspe/(n-c);   sslf = sse-sspe;
 mslf=sslf/(c-2);    f=mslf/mspe;   fpvalue= 1 - probf(f,c-2,n-c);
 proc print;    run;
```

```
                              The MEANS Procedure
                  N
          true  Obs  N          Mean       Std Dev       Minimum       Maximum
            50    3  3    54.0000000     1.0000000    53.0000000    55.0000000
           200    3  3   204.6666667     2.0816660   203.0000000   207.0000000
           400    3  3   383.0000000     1.7320508   382.0000000   385.0000000
FROM PROC REG
                                    Sum of           Mean
 Source                   DF        Squares         Square    F Value    Pr > F
 Model                     1         162559         162559    5625.07    <.0001
 Error                     7      202.29279       28.89897
 Corrected Total           8         162761
                              The ANOVA Procedure
                      Class         Levels    Values
                      true               3    50 200 400
                          Number of observations    9
Dependent Variable: measured
                                      Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F
 Model                       2    162744.2222     81372.1111   29294.0   <.0001
 Error                       6        16.6667         2.7778
```

```
Corrected Total             8    162760.8889
Source                     DF       Anova SS    Mean Square   F Value   Pr > F
true                        2    162744.2222     81372.1111   29294.0   <.0001

            Levene's Test for Homogeneity of measured Variance
               ANOVA of Squared Deviations from Group Means
                              Sum of        Mean
        Source        DF     Squares      Square    F Value    Pr > F
        true           2      7.5062      3.7531       1.17    0.3714
        Error          6     19.1852      3.1975

   Obs   sspe    sse   n c   mse     mspe    sslf    mslf     f      fpvalue
    1  16.6667 202.293 9 3 28.8990 2.77778 185.626 185.626 66.8253 .000180449
```

Example: Using the Kishi data on nutritional requirements. This groups the data by similar ni values. Then constructs Levene's modified test.

```
option ls=80 nodate;
data a;
infile 'kishi.dat';
input kcal ni niq nbal;
if 30 < ni < 35 then group=1;
if 60 < ni <65 then group=2;
if 75 < ni < 85 then group=3;
proc print;  run;
proc reg;
model nbal=ni;
output out=result r=resid;   run;
proc means mean median;
class group;  var resid;  run;
data b;
set result;
 if group=1 then median = -0.5484660;
 if group =2 then median = 0.7004449;
 if group=3 then median =  -0.0370720;
 d = abs(resid-median);     run;
proc anova;
class group;
model d =group;    run;
proc anova data=result;
class group;
model resid=group;
means group/hovtest=levene;    run;
 proc anova data=result;
class group;
model nbal=group;
means group/hovtest=levene;    run;
                Obs    kcal     ni      niq     nbal    group
                  1    49.5    31.6    30.7    -22.7      1
                 31    48.0    81.0    79.4      0.2      3
                      Analysis Variable : resid Residual
```

```
                group    NObs            Mean          Median
          ------------------------------------------------------
                    1      11      -0.0932954      -0.5484660
                    2      10       0.2747271       0.7004449
                    3      10      -0.1721022      -0.0370720
          ------------------------------------------------------

                              The ANOVA Procedure
Dependent Variable: d
                                        Sum of
 Source                     DF          Squares    Mean Square   F Value   Pr > F
 Model                       2       14.5820939      7.2910469      1.78   0.1865
 Error                      28      114.4087282      4.0860260
 Corrected Total            30      128.9908221

  THE F-TEST ABOVE IS BROWN-FORSYTHE TEST FOR EQUAL VARIANCE WITHIN THE LINEAR
REGRESSION MODEL.

                              The ANOVA Procedure
Dependent Variable: resid   Residual
 ....
               Levene's Test for Homogeneity of resid Variance
                ANOVA of Squared Deviations from Group Means
                              Sum of        Mean
        Source        DF     Squares      Square    F Value    Pr > F
        group          2       641.1       320.5       1.31    0.2865
        Error         28      6865.1       245.2

The above test is another test for equal variance within the regression model
 using squared deviations around the mean residual in the group. This is
the unmodified Levene's test.

  The ANOVA Procedure  Dependent Variable: nbal
               Levene's Test for Homogeneity of nbal Variance
                ANOVA of Squared Deviations from Group Means
                              Sum of        Mean
        Source        DF     Squares      Square    F Value    Pr > F
        group          2       602.0       301.0       1.20    0.3175
        Error         28      7049.5       251.8
 THE ABOVE TEST FOR EQUAL VARIANCES AMONG GROUPS WITHIN THE ONE-WAY ANOVA MODEL.
```

5.2 Plotting residuals versus other variables not in the model.

Can serve two purposes. Assessing correlation of error terms and assessing whether the linear regression model is correct (or can be improved on).

Does seeing a pattern indicate that the linear model we have is wrong? Not necessarily.

- If both the X_i and the other variable are fixed values associated with some unit then seeing a pattern indicates the model is incorrect and needs that other variable. Examples: other variables set to some value in a production process where the values of this other variable sometime change. Or units in the experiment are fixed with certain

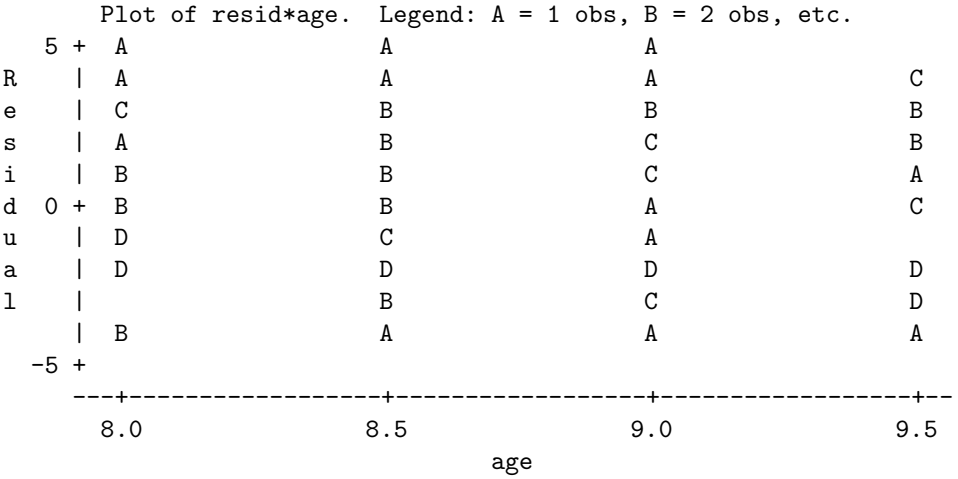
characteristics (other variables) and we do NOT randomly assign X to them.

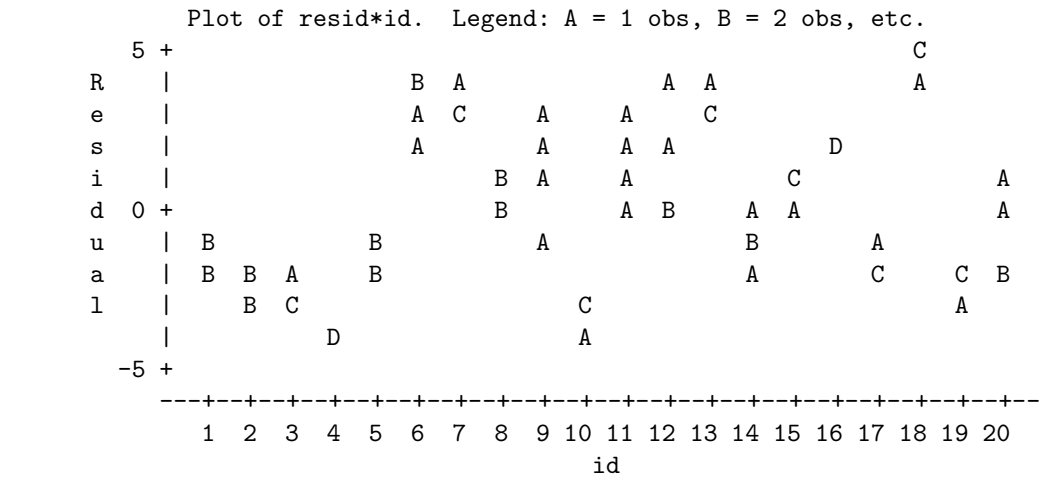
- If the other variable is random (which can happen in various ways), then seeing a pattern does not indicate that the model is wrong. The model is for $Y|X$ and the influence of the other random quantity will be part of the error term. It would mean that we might do better at predicting Y to include that other variables, but it doesn't mean the model is wrong.

- The other variable might be a blocking variable (such as a day, machine, person) where there are multiple observations associated with that blocking variable. If the omitted variable is random, seeing the residuals related to it would be a sign of correlated errors (e.g., the repeated observations on a particular block are correlated through a common block effect). If the other variable(s), are fixed (as in the brain size-intelligence example) then seeing a pattern indicates the model is wrong.

Ramus example: Repeated measures data. Sample of boys, length of ramus bone measured at 4 ages on each boy. With the boys random we are fitting a model for $Y|X$ where the distribution of Y given X (age) and its expected value includes variation among boys. The plot of residuals versus id (which identifies the individual boys) shows a clear pattern. This doesn't mean the model which says $E(Y|X)$ is linear in X is wrong (in fact it looks fine), but that the errors are correlated and standard regression analysis should not be used. With the boys fixed, and not sampled, the pattern in the residuals indicate the model is wrong and that different boys might have different coefficients for how the expected response changes with time.

```
title 'ramus data';
options ps=60 ls=80 nodate;
data ram;
infile 'a:ram2.dat';
input id length age; /* age in years, length=lenght of ramus bone*/
run;
proc reg;
model length=age;
var id;
plot  length*age/conf;
plot r.*(age id);
run;
```





GPA example: Using data from Appendix C4 in the text. 705 students. GPA in freshman year, hsrank = hsrank (percentile; lower is better); act = ACT entrance exam score, ayear = year of entry. Look at regression of GRP on ACT, examine residuals, assess the possible effects of high school rank and/or year of entry.

- Have a significant liner relationship between ACT and GPA, but note that we can't predict GPA very well (we'll plot prediction intervals in class), because of the amount of noise.
- Looks like high school rank might possible play some role. Appear to be more positive residuals at higher end of rank. Does these mean linear model for GPA on ACT is wrong? No. Just that we might be able to do a better job of predicting by adding high school rank.
- The academic year of entry is a blocking variable that is fixed. Whatever year effect is there appears to be minor. In assessing this though we also need to be aware that the hsrank may change over years, so we can pick up that effect.
- We will look soon at how to add the other variables to the model through multiple regression.

```
title 'GPA data from Appendix C4 ';
options ls=80 nodate;
data a;
infile 'APPENC04.txt';
input id gpa hsrank act ayear;
proc print;
run;
proc reg lineprinter;
var ayear hsrank;
model gpa=act;
plot gpa*act r.*(act hsrank ayear)/vplots=3;
run;
proc means;
class ayear;
var resid;
run;
```

GPA data from Appendix C4

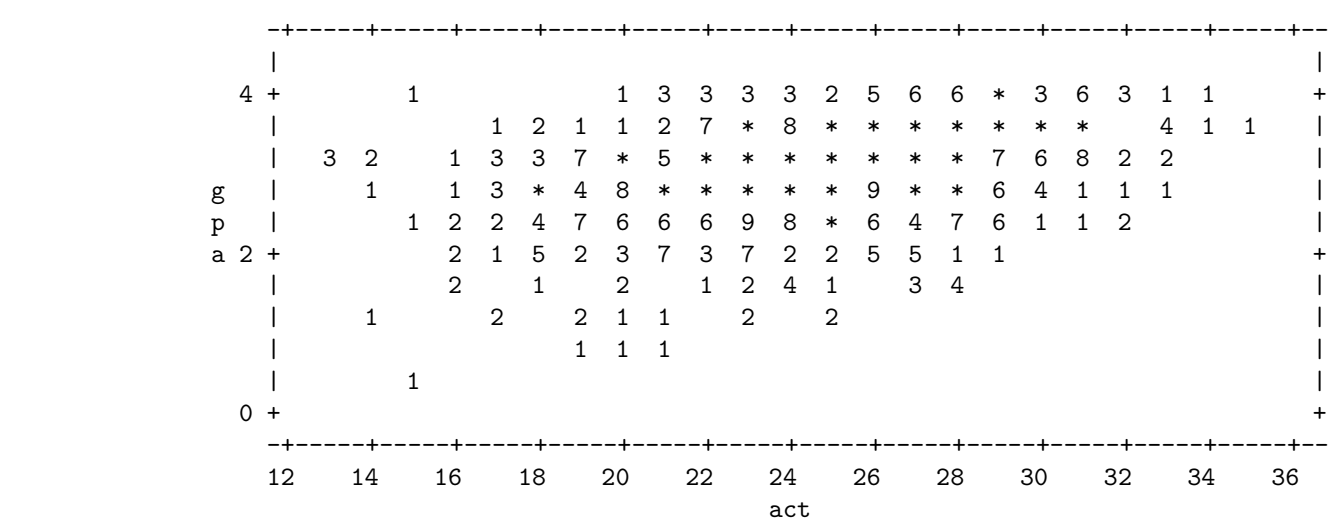
Obs	id	gpa	hsrank	act	ayear
1	1	0.98	61	20	1996
2	2	1.13	84	20	1996
705	705	4.000	99	32	2000

The REG Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	37.88888	37.88888	108.47	<.0001
Error	703	245.55952	0.34930		
Corrected Total	704	283.44840			

Root MSE 0.59102 R-Square 0.1337

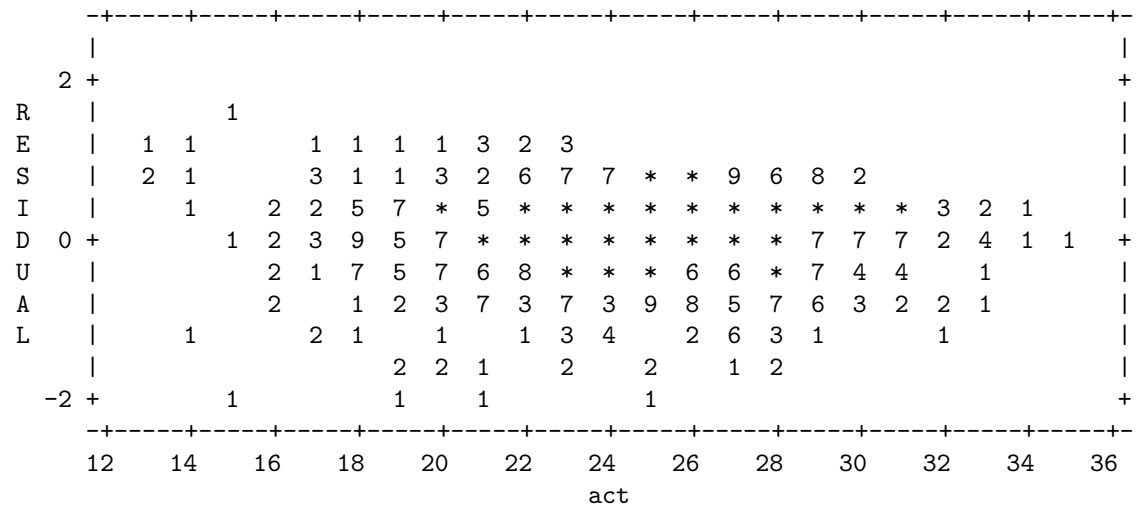
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.55870	0.13802	11.29	<.0001
act	1	0.05780	0.00555	10.41	<.0001

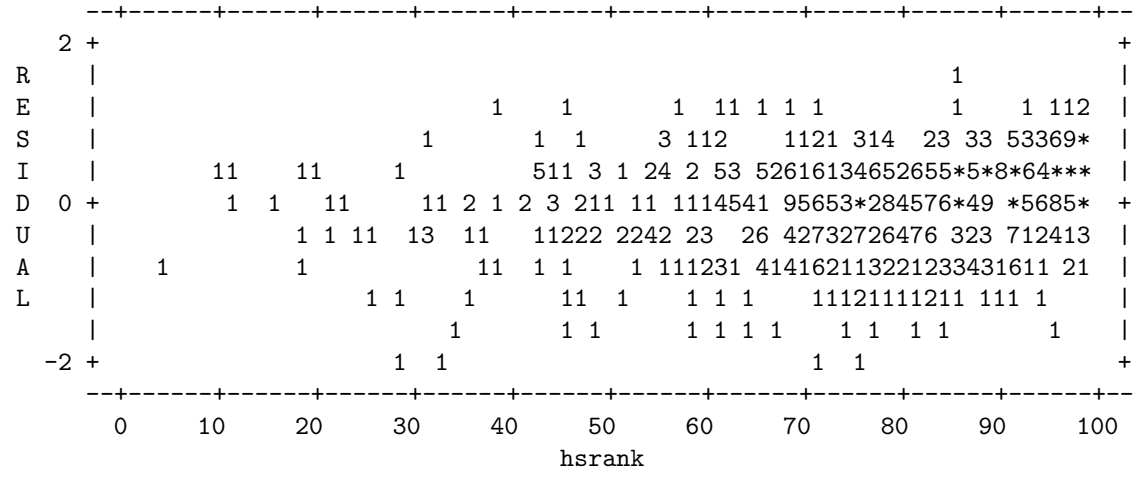


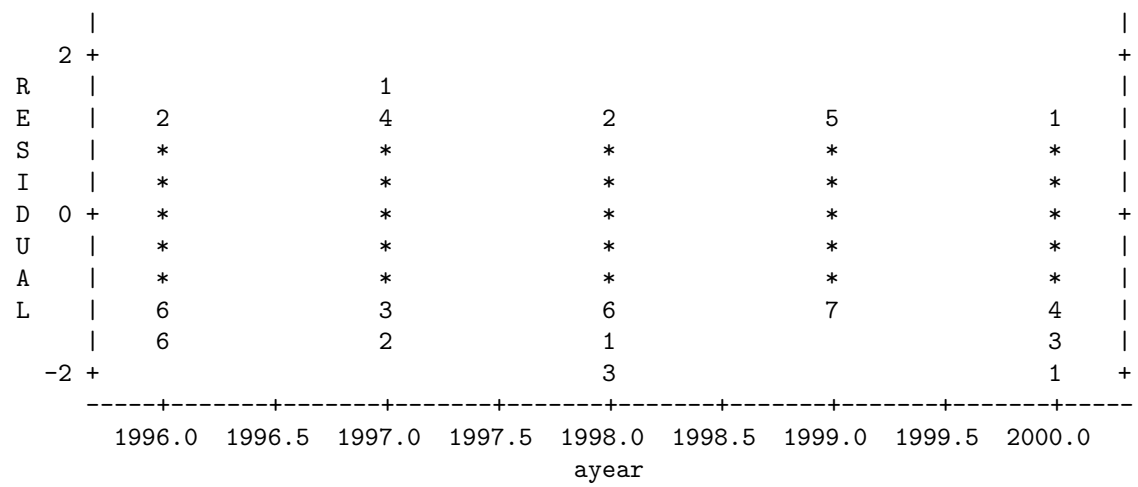
The MEANS Procedure

Analysis Variable : resid Residual

ayear	NObs	N	Mean	Std Dev	Minimum	Maximum
1996	142	142	-0.0570721	0.6206586	-1.7347121	1.1908889
1997	131	131	0.0220141	0.6019719	-1.7325126	1.4942904
1998	154	154	0.0251350	0.6020389	-1.9157096	1.1874874
1999	141	141	0.0316997	0.5433192	-1.3991160	1.0566869
2000	137	137	-0.0227742	0.5848413	-1.8395126	1.0374874


RESIDUAL
2
0
-2
12 14 16 18 20 22 24 26 28 30 32 34 36
act


RESIDUAL
2
0
-2
0 10 20 30 40 50 60 70 80 90 100
hsrank


RESIDUAL
2
0
-2
1996.0 1996.5 1997.0 1997.5 1998.0 1998.5 1999.0 1999.5 2000.0
ayear

5.3 General Linear Tests

A unified way to think about the F-tests we've been using.

Full Model. A regression model with p coefficients in it. SSE(F) = Error sum of squares under full model. $df_F = n - p$. (For simple linear regression $p = 2$.

Reduced/null model: A special case of the full model. A regression model with $p_0 < p$ coefficients. SSE(R) or sometimes denoted $SSE(H_0)$ is error sum of squares under the reduced/null model. $df_R = n - p_0$.

H_0: Null model is true

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} = \frac{(SSE(R) - SSE(F))/(p - p_0)}{SSE(F)/(n-p)}.$$

has an F distribution with $d_1 = df_R - df_F = p - p_0$ and $d_2 = df_F = n - p$ degrees of freedom under the null hypothesis. Reject H_0 if $F > F(\alpha, d_1, d_2)$ or get P-value by area to right of observed F under the F with d_1 and d_2 degrees of freedom.

Model	Error sum of squares	df
SLR model	SSE	n-2
Group means model $Y_{ij} = \theta_j + \epsilon_{ij}$	$\sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2$ (SSPE)	n-c
Single mean $Y_{ij} = \theta + \epsilon_{ij}$ (replicates) or	$\sum_{ij}(Y_{ij} - \bar{Y})^2 = SSTO$	n-1
Single mean $Y_i = \beta_0 + \epsilon_i$ (regression)	$\sum_i (Y_i - \bar{Y})^2 = SSTO$	n-1

Note that SSTO = error sum of squares under model with a single mean can be expressed as $\sum_i Y_i^2 - n\bar{Y}^2$.

- Test for zero slope in regression framework:
  Full Model: Regression model. Null Model: Single mean.
- Test for equal means in group means model:
  Full model: Group means model Null Model: Single mean.
- Test for lack of fit with replicates:
  Full Model: Group means model. Null Model: SLR model.

6 What to do about model violations?

Most of these we will handle in detail after we cover multiple regression.

- Nonlinearity.

  - Tranform X to $X' = g(X)$ such that model for Y is linear in X'. (If doing inverse prediction or regulation then convert back to X scale).

  - Consider polynomial model in X (requires multiple linear regression) or use a non-linear model (non-linear in parameters).

  - Investigating shape of regression function using nonparametric regression. Fits a smooth curve to the family without making assumptions about the parametric form of the regression function. In SAS can use the LOESS procedure.

- Nonconstant variance.

  - Use robust estimate of variances of b_0 , b_1 and $Cov(b_0, b_1)$.

  - Model the variance and use weighted least squares.

- Non-normality.

  * The non-normality of the errors is not much of an issue for estimating coefficients, functions of them and regulation unless the sample size is rather small. This is because the distribution of the estimated coefficients becomes normal as the sample size increases regardless of the distribution of the errors and the various inferences based on the t and F are approximately correct.

  * It is an important issue for prediction and inverse prediction since these depend on the normality of an individual observation.

  * Can transform Y to Y' to achieve normality and this is fine for prediction and inverse prediction. If you have predicted Y' then can back transform to predict Y.

  NOTE: In terms of the regression function if you use Y' as the response and fit a linear regression for $E(Y'|X)$ you CANNOT just backtransform to get an estimate of $E(Y|X)$. So for example if you fit the model $E(Y'|X) = log(X) = \beta_0 + \beta_1 X$, with estimated coefficients b_0 and b_1, then $e^{b_0 + b_1 X}$ is a biased estimator of $E(Y|X)$. *The models* $E(Y'|X) = log(X) = \beta_0 + \beta_1 X$ *and* $E(Y|X) = e^{\beta_0 + \beta_1 X}$ *cannot both hold at the same time.*

- Correlated errors. Often need to model and remove the correlation in some manner. Sometimes there are techniques available that are robust to the correlation.

Accounting for unequal variances.

- Use least squares estimator but get an estimate of variance-covariance that does not depend on the constant variance assumption. *This is fine for making inferences about the coefficients and functions of them, but does not handle prediction. For prediction we need to know the form of* $V(Y_{new})$.
- Model the variance and use weighted least squares.

White's robust estimator of the covariance:

In SAS, White's estimator is obtained with the ACOV option. This gives quantities of $s^2\{b_0\}$, $s^2\{b_1\}$ and $s\{b_0, b_1\}$ that are robust to the constant variance assumption) Use these in previous formulas for inferences on β_0, β_1 and $\beta_0 + \beta_1 X$. *This will not work very well with small sample size.*

The spec option in the model statement will also provide another approximate test of the hypothesis that the errors have constant variance (assuming all of the other assumptions hold). This test is related to White's procedure.

Weighted Least squares

Suppose $\sigma_i^2 = \sigma^2 a_i^2$ where the a_i are known. Consider $Y_{*i} = (1/a_i)Y_i$,

$$Y_{*i} = (1/a_i)\beta_0 + (1/a_i)X_i\beta_1 + \epsilon_{*i}$$

where ϵ_{*i} has mean 0 and variance σ^2 (since $var(\epsilon_{*i}) = (1/a_i)^2 var(\epsilon_i) = (1/a_i)^2\sigma^2 a_i^2 = \sigma^2$.)

Multiple linear regression on the transformed model is equivalent to using weighted least squares in which we minimize $\sum_i w_i(Y_i - (\beta_0 + \beta_1 X_i)))^2$, where $w_i = 1/a_i^2$.

Weighted least squares yields estimates that have smaller variance than the simple least squares estimates (assuming we have the variance model right).

Most regression routines will allow you to specify a weighting variable and run weighted least squares so you do not need to actually create transformed variables. In SAS proc reg this is done using the weight option.

Estimated variances and weights.

Often need to get estimated variances $\hat{\sigma}_i^2$ and then treat these as if they were known variances and run weighted least squares with $w_i = 1/\hat{\sigma}_i^2$. (This process could be iterated between updated regression coefficients and updated estimates of the variances, and hence the weights, leading to what is known as iteratively reweighted least squares.)

A common approach (especially when there are multiple predictors) is to try and model the variance as a function of the mean, rather than directly as a function of the predictors.

Example: Consider $\sigma_i = \theta_1|\mu_i|^{\theta_2}$, where $\mu_i = E(Y_i|X_i)$ or $log(\sigma_i) = \eta_0 + \eta_1 log(|\mu_i|)$, $\eta_0 = log(\theta_1)$ or $\theta_1 = e^{\eta_0}$ and $\eta_1 = \theta_2$. Regress $log|e_i|$ on $\log(|\hat{Y}_i|)$ to get estimate $\hat{\eta}_0$ and $\hat{\eta}_1$ and take $\hat{\sigma}_i = e^{\hat{\eta}_0}|\hat{Y}_i|^{\hat{\eta}_1}$. Would then run weighted least squares with weights $w_i = 1/\hat{\sigma}_i^2$.

If we assume that

$$\sigma_i^2 = \sigma^2 g(\mu_i)$$

where the function $g(\mu_i)$ is fully specified; that is has no unknown parameters then iteratively reweighted least squares can be carried out in SAS proc nlin.

Weighted least squares with replicates

If there are replicates then an observation in group k can be given weight $1/s_k^2$, where s_k^2 = sample variance of the Y values in group k.

The esterase assay example.

1. Rerun the regression of count on ester, but now include the acov option as well as the clb and clm option. (Even though the acov option is there, note that all of the output is under the assumption of constant variance except the estimated covariance matrix associated with the acov option and results of using test option). Using the results from acov, get confidence interval on the coefficients and on the expected count at ester = 6.4 (which corresponds to the first case). Compare these to what are obtained under unequal variances.

2. Earlier there is a fit of $log|e_i|$ on $log(\hat{Y}_i)$. Use this to create weights and run a weight least

squares analysis under the assumption that $\sigma_i = \theta_1 \mu_i^{\theta_2}$. Again include the clb and clm option and compare the intervals for the coefficients and the expected count at ester=6.4 to those in the previous part.

```
title 'Esterase Hormone data ';
options pagesize=60 linesize=80;
data a;
infile 'a:ester.dat';
input ester count;  run;
proc reg;
model count=ester/covb acov p clm cli spec;
output out=result p=yhat r=resid;
run;
data c;
set result;
wt2 = 1/((exp(-1.44942)*(abs(yhat)**.89486))**2);
run;
proc reg data=c;
model count=ester/clm cli;
weight wt2;
plot count*ester/pred;
plot count*ester/conf;
run;
```

From spec option

```
                    Test of First and Second
                          Moment Specification
                        DF    Chi-Square    Pr > ChiSq
                        2          6.74        0.0345
```

Using least squares under constant variance assumption leads to estimated coefficients and 95% confidence intervals

```
 Variable     DF       Estimate       St. Error    t Value    Pr > |t|
Intercept     1      -15.99447       20.73967      -0.77      0.4423
ester         1       17.04141        0.89739      18.99      <.0001
           Variable     DF       95% Confidence Limits
           Intercept     1      -57.12202       25.13307
           ester         1       15.26186       18.82096
```

and 95% CI's and prediction intervals at ester=6.4 and ester=17.1 as below.

```
                        Output Statistics

                    Dep Var  Predicted     Std Error
 Obs  ester          count      Value  Mean Predict       95% CL Mean
   1       6.4    84.0000    93.0706       15.9330    61.4749   124.6662
  48      17.1   340.0000   275.4137       10.3198   254.9492   295.8781
  72      46.6   599.0000   778.1353       25.5974   727.3748   828.8959
           Obs  ester         95% CL Predict         Residual
```

1	6.4	-111.7490	297.8901	-9.0706
48	17.1	72.0137	478.8137	64.5863
72	46.6	569.4983	986.7723	-179.1353

Allowing unequal variances but still using least squares, White's robust estimator of the variance-covariance of the estimated coefficients is:

Consistent Covariance of Estimates		
Variable	Intercept	ester
Intercept	427.31188264	-24.9806879
ester	-24.9806879	1.6580846777

This yields approximate 95% confidence intervals:

For β_0, $-15.99447 \pm 1.98304(427.31188264)^{1/2}$ = [-56.9869 24.9979]

For β_1, $17.04141 \pm 1.98304(1.6580847)^{1/2} = [14.4879, 19.5949]$.

For $E(Y|6.4) : 93.0706pm1.98304(13.2467) = [66.8019, 119.339]$

where $13.2467 = (s2b0 + (6.4 * *2) * s2b1 + 2 * 6.4 * sb0b1)^{1/2}$, with s2b0= 427.31188264, sb0b1= -24.9806879, s2b1= 1.6580846777.

Despite what looks like a serious issue with unequal variances the confidence intervals have not changed too much when done using either covb or acov.

b) Using weighted least squares with weight $w_i = 1/([exp(-1.44942) * \hat{Y}_i^{.89846}]^2$ leads to

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-38.91551	13.22844	-2.94	0.0040
ester	1	18.27004	0.92985	19.65	<.0001

Variable	DF	95% Confidence Limits	
Intercept	1	-65.14800	-12.68302
ester	1	16.42610	20.11397

Obs	ester	Weight Variable	Dep Var count	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	6.4	0.005437	84.0000	78.0335	8.3989	61.3781	94.6888
48	17.1	0.000780	340.0000	273.4930	7.2935	259.0297	287.9563
72	46.6	0.000122	599.0000	812.3768	32.0906	748.7400	876.0136

Obs	ester	95% CL Predict		Residual
1	6.4	19.0995	136.9675	5.9665
48	17.1	123.5402	423.4458	66.5070
72	46.6	428.9912	1196	-213.3768

The inferences for the intercept have changed quite a bit from using unweighted least squares. The interval for the slope has shifter a bit, but is a bit smaller than the one from least squares with acov. The interval for $E(Y|6.4)$ is shifted down a bit and quite a bit tighter than the $[66.8019, 119.339]$ from least squares with acov.

The prediction intervals are dramatically effected in some cases, as they should be because of the changing variance. The prediction intervals under constant variance are generally too large at low ester and too small at high ester.

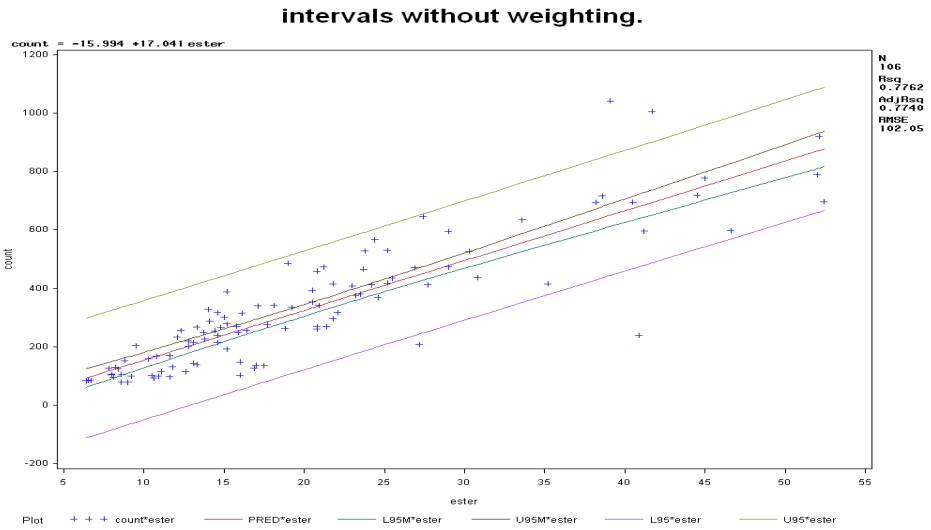


Figure 10: Ester example, no weighting.

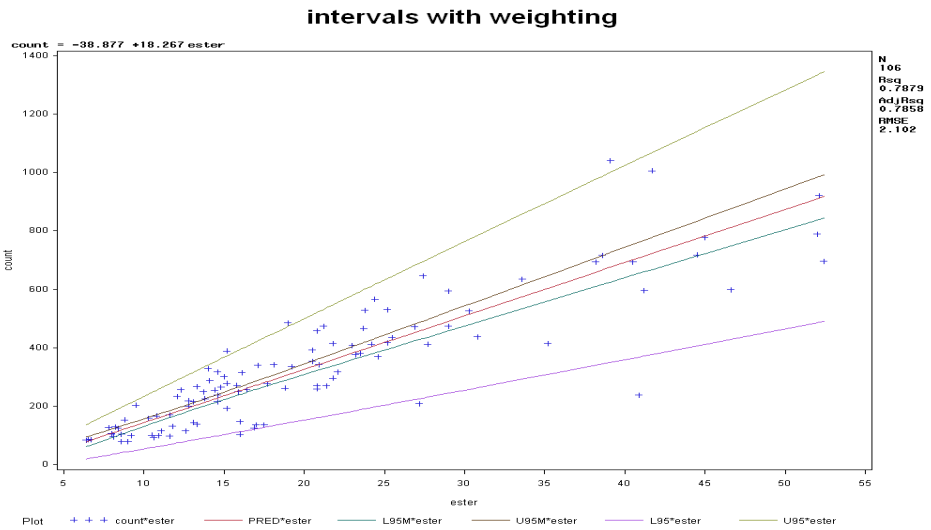


Figure 11: Ester example, with weighting.

7 The simple linear regression in matrix form and a few asides about matrices.

The SLR model in matrix form, used to introduce basic idea matrices and vectors, addition, multiplication and the mean and covariance matrix of a random vector.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ 1 & X_{n-1} \\ 1 & X_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}.$$

$\mathbf{Y}$ is a $n \times 1$ matrix (or vector; with no modifier vector means column vector)

$\boldsymbol{\epsilon}$ is a $n \times 1$.

$\mathbf{X}$ is a $n \times 2$ matrix

$\boldsymbol{\beta}$ is a 2×1 vector.

Each of $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are random vectors. Each has a mean vector and a variance-covariance (or dispersion) matrix. Illustrate with $\mathbf{Y}$

$$\boldsymbol{\mu}_Y = E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ . \\ . \\ . \\ E(Y_n) \end{bmatrix}$$

Variance-covariance matrix of Y: $\sigma^2\mathbf{Y}$ (also often denotes $D(\mathbf{Y})$, $Cov(\mathbf{Y})$ or $\boldsymbol{\Sigma}_y$) is $n \times n$ matrix with $(i,i)th$ element = variance of Y_i and $(i,j)th$ element equal to $\sigma\{Y_i, Y_j\}$ (covariance of Y_i and Y_j); see (5.42).

$E(\boldsymbol{\epsilon}) = \mathbf{0}$ (a vector of 0's), $\quad \sigma^2\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}_n$

$\mathbf{I}_n = n \times n$ identity matrix; 1's on the diagonal and 0's on the off diagonals.

The transpose of a column vector is a row vector and is denoted with a $'$; so $\mathbf{Y}' = (Y_1, \ldots, Y_n)$ is a $1 \times n$ vector.

Defining $\boldsymbol{\beta}' = (\beta_0, \beta_1)$ and $\mathbf{b}' = (b_0, b_1)$ then:

$$E(\mathbf{b}) = \boldsymbol{\beta}, \quad \sigma^2\{\mathbf{b}\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} \end{bmatrix}.$$

8 Multiple Regression models.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + X_{i,p-1}\beta_{p-1} + \epsilon_i$$

where $X_{i1}, \ldots, X_{i,p-1}$ are known values. These may come from $p-1$ different explanatory variables (in which case the model is linear in the parameters and the original X's both) or they may contain functions of an original set of explanatory variables as seen below with squares, products, etc.

- **Polynomial model of degree q in one variable**:

  $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_q X_i^q + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_q X_{iq} + \epsilon_i,$

  where $X_{ij} = X_i^j$.

- **A first order (linear) model with two variables**:

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i.$$

  - β_1 measures the change in $E\{Y_i\}$ per unit increase in X_1 with X_2 held fixed. Same for all levels of X_2.

  - β_2 measures the change in $E\{Y_i\}$ per unit increase in X_2 with X_1 held fixed. Same for all X_1.

- **A model with two variables and interaction:**

  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i,$

  where $X_{i3} = X_{i1} X_{i2}$.

Change in $E(Y)$ for unit change in X_1 depends on the value of X_2.

$E(Y_i) = \beta_0 + [\beta_1 + \beta_3 X_{i2}]X_{i1} + \beta_2 X_{i2}$

At $X_2 = x_2$, the change in $E\{Y_i\}$ per unit increase in X_1 is $\beta_1 + \beta_3 x_2$.

Interaction means that the effect of one variable is dependent on what level of the other variable is present.

Similarly if $X_1 = x_1$, the change in $E\{Y_i\}$ per unit increase in X_2 is $\beta_2 + \beta_3 x_1$.

- **A model with two variables with quadratic and interaction effects:** Sometimes called second order response surface model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 X_{i1}^2 + \beta_5 X_{i2}^2 + \epsilon_i,$$

$X_{i3} = X_{i1}X_{i2}$, $X_{i4} = X_{i1}^2$, $X_{i5} = X_{i2}^2$.

Writing the regression model in matrix form.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & ... & X_{1,p-1} \\ . & . & ... & . \\ 1 & X_{i1} & ... & X_{i,p-1} \\ . & . & ... & . \\ 1 & X_{n1} & ... & X_{n,p-1} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}.$$

With no intercept the first column of 1's in $\mathbf{X}$ would be eliminated. The linear regression model in matrix form is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

$E(\boldsymbol{\epsilon}) = \mathbf{0}$ (an $n \times 1$ vector of 0's.).

Estimating the parameters

The least squares estimates $b_0, b_1, ...b_{p-1}$ minimize $\sum_i (Y_i - (b_0 + b_1 X_{i1} + ... b_{p-1} X_{i,p-1}))^2$.

The least squares estimates are uniquely determined if the matrix $\mathbf{X'X}$ is non-singular (or equivalently $\mathbf{X}$ which is $n \times p$ is of rank p) in which case the least squares estimates are obtained via $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$.

When $\mathbf{X}'\mathbf{X}$ is singular (equivalently the rank of $\mathbf{X}$ is less than p) there are infinitely many sets of coefficients which minimize the sum of squared deviations. The β's are *not identifiable or are said to be not estimable.*

- For a polynomial model of degree q, we need at least q+1 distinct values of the explanatory variable.

- In general it is necessary that $n \geq p$.

- If there are linear restrictions among the X variables (e.g., one of the X's can be written as a linear combination of the others) then there will be a problem.

The **ith residual:** $e_i = Y_i - (b_0 + b_1 X_{i1} + \ldots + b_{p-1} X_{i,p-1}) = Y_i - \hat{Y}_i$.

Vector of residual: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - X\mathbf{b}$

$\hat{\sigma}^2 = MSE = \sum_i e_i^2/(n-p)$ is unbiased for σ^2

$E(b_j) = \beta_j$ for each j, or in matrix form $E(\mathbf{b}) = \boldsymbol{\beta}$. (This comes from the fact that if we fix $\mathbf{X}$ then
$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a constant matrix and so

$$E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

The Variance-covariance matrix of the estimated coefficients (this is $p \times p$ matrix):

$$\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

.

This is estimated by $s^2\{\mathbf{b}\} = MSE(\mathbf{X}'\mathbf{X})^{-1}$.

The square root of the jth diagonal element of this matrix is $s\{b_j\}$, the estimated standard error of b_j.

Inferences for the individuals coefficients

Same form as for simple linear regression.

Confidence interval for β_j: $b_j \pm t(1-\alpha/2, n-p)s\{b_j\}$.

Hypothesis testing: $H_0 : \beta_j = 0$ versus $H_A : \beta_1 \neq 0$.

Reject H_0 if $|t^*| > t(1-\alpha/2; n-p)$, where $t^* = b_j/s\{b_j\}$.

Can easily generalize to one-sided tests or to null values other than 0.

This tests the significance of a coefficient for a variable *given that the other variables are in the model.*

Confidence intervals for the mean value.

$$E(Y|\mathbf{X}_h) = \mu\{\mathbf{X}_h\} = \beta_0 + \beta_1 X_{h1} + \ldots + X_{hp-1}\beta_{p-1} = \mathbf{X}_h'\boldsymbol{\beta}$$

where $\mathbf{X}_h' = (1, X_{h1}, \ldots, X_{hp-1})$.

$$\hat{\mu}\{\mathbf{X}_h\} = b_0 + b_1 X_{h1} + \ldots + b_{p-1} X_{hp-1} = \mathbf{X}_h'\mathbf{b}$$

= estimate of $E(Y)$ at $\mathbf{X}_h$.

$$V(\hat{\mu}\{\mathbf{X}_h\}) = \sigma^2\{\hat{\mu}\{\mathbf{X}_h\}\} = \mathbf{X}_h'\sigma^2\{\mathbf{b}\}\mathbf{X}_h$$

$s^2\{\hat{\mu}\{\mathbf{X}_h\}\} = \mathbf{X}_\mathbf{h}'s^2\{\mathbf{b}\}\mathbf{X}_h$.

$s\{\hat{\mu}\{\mathbf{X}_h\}\}$ = estimated standard error of $\hat{\mu}\{\mathbf{X}_h\}$.

Confidence interval: $\hat{\mu}\{\mathbf{X}_h\} \pm t(1-\alpha/2; n-p)s\{\hat{\mu}\{\mathbf{X}_h\}\}$

NOTE: The book uses $E\{Y_h\}$ for what we denote here by $\mu\{\mathbf{X}_h\}$ and uses $\hat{Y}_h$ for what we denote here by $\hat{\mu}\{\mathbf{X}_h\}$. The μ notation is just a reminder that it is the "mean" of Y at $\mathbf{X}_h$.

Simultaneous confidence intervals for g different mean values, say $\mu\{\mathbf{X}_{hk}\}$ $k = 1$ to g.

Bonferroni's method: $\hat{\mu}\{\mathbf{X}_{hk}\} \pm t(1-\alpha/2g; n-p)s\{\hat{\mu}\{\mathbf{X}_{hk}\}\}$

Scheffe's method: $\hat{\mu}\{\mathbf{X}_{hk}\} \pm (pF(1-\alpha, p, n-p))^{1/2}s\{\hat{\mu}\{\mathbf{X}_{hk}\}\}$.

Scheffe method work for as many intervals as you want, including infinitely many which gives a confidence "band" for the regression surface.

Prediction intervals.

Predict outcome Y_{new} at a fixed set of X values contained in $\mathbf{X}_{new}' = (1, X_1, \ldots, X_{p-1})$.

$\hat{Y}_{new} = b_0 + b_1 X_1 + \ldots + b_{p-1} X_{p-1} = \mathbf{X}'_{new}\mathbf{b}$.

The variance of $\hat{Y}_{new} - Y_{new}$ is $\sigma^2 + \mathbf{X}'_{\mathbf{new}}\sigma^2\{\mathbf{b}\}\mathbf{X}_{new}$, estimated by

$s^2\{pred\} = MSE + \mathbf{X}'_{\mathbf{new}}s^2\{\mathbf{b}\}\mathbf{X}_{new}$.

prediction interval: $\hat{Y}_{new} \pm t(1-\alpha/2; n-p)s\{pred\}$.

Multiple prediction of g values; for kth prediction use

Bonferroni's method: $\hat{Y}_{newk} \pm t(1-\alpha/2g; n-p)s\{pred_k\}$.

Scheffe's method: $\hat{Y}_{newk} \pm (gF(1-\alpha, g, n-p)^{1/2}s\{pred_k\}$,

The Analysis of Variance

Total uncorrected sum of squares is $SSTOU = \sum_i Y_i^2$.

Total corrected sum of squares is $SSTO = \sum_i (Y_i - \overline{Y})^2$.

Error sum of Squares $SSE = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i e_i^2$

Sum of squares due to regression (on $X_1, \ldots, X_{p-1}$**)** is $SSR = SSTO - SSE$:

There are a variety of computational formulas for these sums of squares.

Define $MSR = SSR/(p-1)$.

Under $H_0 : \beta_1 = \beta_2 = \ldots = B_{p-1} = 0$. $F^* = MSR/MSE$ follows an F-distribution with $p-1$ and $n-p$ degrees of freedom respectively. A test of size α rejects H_0 if $F^* > F(1-\alpha, p-1, n-p)$. Or equivalent if the P-value (area to the right of the observed F^* under the density for the F with p-1 and n-p degrees of freedom) is less than α. Rejecting H_0 indicates that at least one of the coefficients other than the intercept is 0.

$E(MSR) = \sigma^2 + k$ where $k \geq 0$ and equals 0 if and only if H_0 is true.

Assessing assumptions

All the inferences above depend on the assumptions on the ϵ_i's be being correct. Same exact issues here as in SLR and we evaluate the assumptions in a similar

manner. There are more advanced diagnostics we'll return to later using modified residuals.

- Assess whether model is correct (the errors have mean 0) via plot of residual versus each of the predictor and versus the fitted values. In addition we might want to plot the residuals versus products of variables for products not in the model which would pick up interactions. This is not often done with more than a few predictors but typically if an important interaction was missed it will show up in the plot versus fitted values.

  With a few X's and a designed experiment where replication is used at combinations of X's, a test for lack of fit can be used.

- Can examine constant variance by plotting the residual versus X's or the fitted values, but better to plot absolute or squared residuals. The latter will also help suggest a model for the variance, within which we could test for constant variance (and that model could be used for weighted least squares). See examples with SLR including homework problems. Typically with multiple predictors we are not in a situation where we can exploit grouping to use Levene's test. (There are other test including something called the spec test associated with White's covariance estimator. This is discussed later).

- Assess normality of errors using residuals as before (and as before, we should be sure we've dealt with any violations of the model and of the constant variance assumption).

- If the data is collected over time or space or over some clustering of the data (e.g., on trees, a person, etc.) and these things are not in the model then the residuals should be plotted versus them. Patterns can be indicative of correlations among errors or the need for other terms in the model (e.g., tree effects) depending on whether these other factors are viewed as fixed or random.

Allowing unequal variances. The two options here are to continue to use least squares but fix up the estimate of the covariance matrix of $\mathbf{b}$. This uses White's estimator. The other approach is to create weights and use the weights to do weighted least squares. See pages 71-74 of the notes and page 427 of the text.

White's estimator. We can now give a general expression for White's estimator of the variance-covariance matrix of $\mathbf{b}$, which is produced by the ACOV option in SAS.

Using results on linear transformations

$$\sigma^2\{\mathbf{b}\} = (\mathbf{X}'\mathbf{X})^{-1}X'\sigma^2\{\mathbf{Y}\}X(\mathbf{X}'\mathbf{X})^{-1}.$$

where $\sigma^2\{\mathbf{Y}\}$ is a diagonal matrix (assuming the ϵ_i are uncorrelated) with σ_i^2 used as the ith diagonal element. White's estimator produced by ACOV uses

$$s^2_{white}\{\mathbf{b}\} = (\mathbf{X}'\mathbf{X})^{-1}X'\mathbf{D}X(\mathbf{X}'\mathbf{X})^{-1}$$

$\mathbf{D}$ is a diagonal matrix with $r_1^2, r_2^2, \ldots, r_n^2$ on the diagonal. This is what is labeled "consistent covariance of estimates".

For confidence intervals on the coefficients and linear combinations of them (including for $\mu\{\mathbf{X}_h\}$) we use the earlier results but with $s^2_{white}\{\mathbf{b}\}$ rather than $s^2\{\mathbf{b}\}$. This is true for either one-at-a-time or simultaneous intervals. The computing will often have to be customized to do this. As with a single variable earlier, you cannot get prediction intervals under unequal variance unless you model the variance.

Example: using House pricing data.

```
Reference: Albuquerque Board of Realtors
Description: The data are a random sample of records of resales
of homes from Feb 15 to Apr 30, 1993 from the files maintained by
the Albuquerque Board of Realtors. This type of data is collected
by multiple listing agencies in many
cities and is used by realtors as an information base.

Number of cases: 117
   1.PRICE = Selling price ($hundreds)
   2.SQFT = Square feet of living space
   3.AGE = Age of home (years)
   4.FEATS = Number out of 11 features (dishwasher, refrigerator,
microwave, disposer, washer, intercom, skylight(s),
     compactor, dryer, handicap fit, cable TV access
   5.NE = Located in northeast sector of city (1) or not (0)
   6.COR = Corner location (1) or not (0)
   7.TAX = Annual taxes ($)
```

The population of interest is the collection of all samples from which these were sampled. Note that here Y and all of the X's are random together. We'll first fit a model regressing Y =price on X_1 = sqft and X_2 = tax. The model is $Y_i|X_{i1}, X_{i2} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$. This is a model for the price conditional on X_1 and X_2. It doesn't assume that the other variables might not influence price, but is looking at a model conditioning on just these two variables. The effects of any other variables are part of the error term. There can be many different regression models, one for each set of X variables that we could consider as predictors. These can all be correct models. An important point though is that the true coefficients associated with a variable will depend on what other variables are in the model. For example, there is no unique coefficient attached to sqft. It will be on thing in a model with sqft by itself, something different with sqft and tax both in the model, etc., changing depending on what other predictors are in the model.

There is evidence of changing variance via the test from the regression of $log(|e_i|)$ on $log(|\hat{Y}_i|)$. (Note: Testing for equal variance in the original regression by assessing the slope in this latter model can be rather approximate since there is often unequal variance in the regression of $log(|e_i|)$ on $log(|\hat{Y}_i|)$, but it will give us a rough test.)

```
title 'Illustrating multiple regression with house price data';
options linesize=70 pagesize=60 nodate;
data values;
infile 'house.dat';
input PRICE SQFT AGE FEATS NE CUST COR TAX;
prod=sqft*tax;   /* used to assess interaction later*/
run;
run;
proc g3d;   scatter sqft*tax=price;   run;
proc gplot;  plot price*sqft;  plot price*tax;  run;
proc reg;
model price =sqft tax/clb covb cli clm p acov;
/* The acov will be accomodate unequal variances */
output out=result p=yhat r=resid;   run;
proc gplot data=result;
plot resid*yhat;    plot resid*prod;   run;
/* could also do plots within proc reg */
proc univariate  plot normal data=result;
var resid;
hist resid/kernel(color = black); /* gives a histogram with smoothing*/
run;
data b;  set result;
absr=abs(resid);   logar=log(absr);    logp=log(yhat);  run;
proc reg;     model absr=yhat;    model logar=logp;   run;

proc reg data=values;
model price=sqft tax prod/acov;   /* include an interaction */
run
                      Dependent Variable: PRICE
                                 Sum of         Mean
 Source                 DF      Squares       Square  F Value  Pr > F
 Model                   2     12476918      6238459   205.79  <.0001
 Error                 104      3152660        30314
 Corrected Total       106     15629578
         Root MSE            174.10927    R-Square     0.7983
         Dependent Mean     1077.34579    Adj R-Sq     0.7944
         Coeff Var            16.16095
                      Parameter       Standard
 Variable     DF       Estimate          Error    t Value    Pr > |t|
 Intercept     1       88.33552       55.80885       1.58      0.1165
 SQFT          1        0.25019        0.06208       4.03      0.0001
 TAX           1        0.72072        0.10703       6.73      <.0001
                        Parameter Estimates
            Variable     DF       95% Confidence Limits
```

Intercept	1	-22.33552	199.00657
SQFT	1	0.12708	0.37329
TAX	1	0.50847	0.93298

Covariance of Estimates

Variable	Intercept	SQFT	TAX
Intercept	3114.627641	-1.898751665	0.4214277592
SQFT	-1.898751665	0.0038540083	-0.005705064
TAX	0.4214277592	-0.005705064	0.0114562926

Heteroscedasticity Consistent Covariance of Estimates (THIS IS WHITE'S ESTIMATOR)

Variable	Intercept	SQFT	TAX
Intercept	7691.7364195	-10.12658657	10.587005822
SQFT	-10.12658657	0.0172232383	-0.022020091
TAX	10.587005822	-0.022020091	0.0316636978

***SAS 9.3 WILL GIVE CONFIDENCE INTERVALS AND TEST FOR ZERO COEFFICIENTS THAT MAKE USE OF THE STANDARD ERRORS THAT ACCOMODATE UNEQUAL VARIANCE IN THE ERRORS. THESE SHOW UP IN THE RESULTS WINDOW (A FEATURE OF 9.3) RATHER THAN IN THE USUAL OUTPUT WINDOW.

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
5	10.88	0.0538

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	2050	1933	52.1014	1829	2036
2	2080	1523	38.6728	1446	1600
3	2150	1615	37.3783	1541	1689
4	2150	1998	49.1515	1900	2095
5	1999	1982	61.7311	1860	2105

etc.

Obs	95% CL Predict		Residual
1	1572	2293	117.4048
2	1169	1877	557.0325
3	1262	1968	535.3446
4	1639	2356	152.4871

etc.

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.905047	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.122071	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.480528	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.914371	Pr > A-Sq	<0.0050

The REG Procedure

```
                    Dependent Variable: absr
                                                Parameter      Standard
Variable    Label                        DF      Estimate         Error
Intercept   Intercept                     1     -52.16382      38.10094
yhat        Predicted Value of PRICE      1       0.15345       0.03371
                       Parameter Estimates
   Variable    Label                        DF   t Value    Pr > |t|
   Intercept   Intercept                     1     -1.37      0.1739
   yhat        Predicted Value of PRICE      1      4.55      <.0001

                          Model: MODEL2
                    Dependent Variable: logar
                                                Parameter      Standard
Variable    Label                        DF      Estimate         Error
Intercept   Intercept                     1      -4.20288       2.75780
logp                                      1       1.19376       0.39724
   Variable    Label                        DF   t Value    Pr > |t|
   Intercept   Intercept                     1     -1.52      0.1305
   logp                                      1      3.01      0.0033
                       The REG Procedure
                    Dependent Variable: PRICE
                    Parameter       Standard
 Variable     DF       Estimate          Error    t Value    Pr > |t|
 Intercept     1      194.93093      159.89757       1.22      0.2256
 SQFT          1        0.18923        0.10588       1.79      0.0768
 TAX           1        0.59103        0.21149       2.79      0.0062
 prod          1     0.00006698     0.00009413       0.71      0.4783
               Consistent Covariance of Estimates
Variable       Intercept           SQFT            TAX           prod
Intercept   19818.356545   -11.47268544   -24.43223646   0.0123630282
SQFT        -11.47268544   0.0152922789    -0.01028262   -3.295305E-6
TAX         -24.43223646    -0.01028262   0.1032641467    -0.00002851
prod        0.0123630282   -3.295305E-6    -0.00002851    1.086783E-8
            Test 1 Results for Dependent Variable PRICE
                                     Mean
      Source             DF         Square    F Value    Pr > F
      Numerator           1          15424       0.51    0.4783
      Denominator       103          30459

                      Test 1 Results using
                         ACOV estimates
                    DF    Chi-Square    Pr > ChiSq
                     1          0.41        0.5205
```

Anaysis using R. Here we compute White's estimate of $\sigma^2\{\underline{\}}$ (which is the covariance of the estimated coefficients) directly using matrix calculations.

```
detach(data)
rm(list=ls())   # clear workspace
data<-read.table("f:/s505/data/house.dat",na.strings=".") #no names
```

```
     #na.strings = indicates that a . is a missing value
attach(data)
price <- V1; sqft<-V2; age<-V3; feats<-V4
ne<- V5; cust<-V6; cor<-V7; tax<- V8
prod=sqft*tax
par(mfrow=c(2,1))
plot(sqft, price)
plot(tax,price)
regout<-lm(price ~ sqft+ tax, na.action=na.exclude)
    # if we don't use na.action = na.exclude then
    # the residual vector will have only cases with
    # non-missing and be of a different length than
    # sqft and tax. But the regression only fits
    # using the 107 cases with price, sqft and tax all not missing
summary(regout)
anova(regout)
confint(regout)
fits <-fitted(regout)
resids<-residuals(regout)      #saves residuals
par(mfrow=c(2,2))
plot(sqft,resids)
plot(tax,resids)
plot(fits,resids)
plot(prod,resids)
xvalues<-data.frame(sqft,tax)
cintervals<- predict(regout,xvalues,interval = "confidence")
cintervalsum<-cbind(sqft,tax,price,cintervals)
cintervalsum
pintervals<- predict(regout,xvalues,interval = "predict")
pintervals
absresid<-abs(resids)
logabresid<-log(absresid)
logfit<-log(abs(fits))
plot(fits,absresid)
par(mfrow=c(2,2))
plot(logfit,logabresid)
summary(lm(logabresid~logfit))   # regress the log of the absolute residual
summary(lm(price~sqft + tax + prod))

# Get the X matrix for fit with sqft and tax and construct White's estimator.

xmat<-model.matrix(regout)   #xmat is the x matrix
xmat
regout2<-lm(price ~ sqft+ tax) #need residuals for just cases with no missing
r2<-residuals(regout2)^2   # this has squared residuals
D<-diag(r2)    # creates a diagonal matrix with squared residuals on diagonal
xpxi<-solve(t(xmat) %*% xmat)   #  %*% is the matrix multiplication operator
                                # solve finds the inverse, in this case of X'X
acov<-xpxi %*% t(xmat) %*% D %*% xmat %*% xpxi
cat("White's robust estimate of Cov(b)", "\n")
acov
```

```
Showing commands that produce output and output (some edited).

> summary(regout)

 Call:
 lm(formula = price ~ sqft + tax, na.action = na.exclude)

 Residuals:
    Min      1Q  Median      3Q     Max
  -596.40  -82.46   -6.50   59.73  610.92

  Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.33552   55.80885   1.583 0.116498
sqft         0.25019    0.06208   4.030 0.000106 ***
tax          0.72072    0.10703   6.734 9.35e-10 ***

Residual standard error: 174.1 on 104 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.7983,     Adjusted R-squared: 0.7944
F-statistic: 205.8 on 2 and 104 DF,  p-value: < 2.2e-16

> anova(regout)

Analysis of Variance Table
Response: price
           Df   Sum Sq  Mean Sq F value    Pr(>F)
sqft        1 11102445 11102445 366.248 < 2.2e-16 ***
tax         1  1374474  1374474  45.341 9.347e-10 ***
Residuals 104  3152660    30314
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> confint(regout)
                  2.5 %      97.5 %
(Intercept) -22.3355189 199.0065654
sqft          0.1270784   0.3732950
tax           0.5084704   0.9329755


> cintervalsum

    sqft  tax price       fit       lwr       upr
1   2650 1639  2050 1932.5952 1829.2762 2035.9142
2   2600 1088  2080 1522.9675 1446.2779 1599.6572
3   2664 1193  2150 1614.6554 1540.5329 1688.7779
 etc.

> pintervals
          fit       lwr       upr
1   1932.5952 1572.2025 2292.9879
```

```
2   1522.9675 1169.2878 1876.6473
3   1614.6554 1261.5234 1967.7874
4   1997.5129 1638.7535 2356.2724


> summary(lm(logabresid~logfit))   # regress the log of the absolute residual

Call:
lm(formula = logabresid ~ logfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.2029     2.7578  -1.524  0.13052
logfit        1.1938     0.3972   3.005  0.00332 **
Residual standard error: 1.237 on 105 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.0792,     Adjusted R-squared: 0.07043
F-statistic: 9.031 on 1 and 105 DF,  p-value: 0.00332

> summary(lm(price~sqft + tax + prod))
lm(formula = price ~ sqft + tax + prod)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.949e+02  1.599e+02   1.219   0.2256
sqft        1.892e-01  1.059e-01   1.787   0.0768 .
tax         5.910e-01  2.115e-01   2.795   0.0062 **
prod        6.698e-05  9.413e-05   0.712   0.4783

Residual standard error: 174.5 on 103 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.7993,     Adjusted R-squared: 0.7934
F-statistic: 136.7 on 3 and 103 DF,  p-value: < 2.2e-16

> xmat
    (Intercept) sqft  tax
1             1 2650 1639
2             1 2600 1088
...
117           1  970  541

> cat("White's robust estimate of Cov(b)", "\n")
White's robust estimate of Cov(b)
> acov
            (Intercept)         sqft         tax
(Intercept)  7691.73642 -10.12658657 10.58700582
sqft          -10.12659   0.01722324 -0.02202009
tax            10.58701  -0.02202009  0.03166370
```

Continuing the example: CI's for $E(Y)$ and prediction intervals.

One-at-a-time **confidence intervals for $E(Y)$ and prediction intervals for a new Y** associated with the **X** vectors in the original data can be obtained automatically in either SAS or R, as shown above. These are, however computed under the assumption of constant variance. For example, the CI for $E(Y)$ associated with the first observation (with $X_1 =$ sqft $= 2650$ and $X_2 =$ tax $= 1639$) is (1829,2036) and the prediction interval associated with those X values is (1572,2293).

If you want a CI or PI for a new set of X values (say sqft= 2500 and tax = 1000) not in the data, in SAS you can do it by entering a new line in the day containing the values for sqft and tax but having the other values missing. In R, you can use the technique that was in the homework 2 solutions. Here's code you would add with partial output. Note that when you use interval="predict" it gives you the correct prediction interval but what it calls se.fit is just the standard error associated with estimating $E(Y)$ (i.e., the same as when you use interval = "confidence".)

```
#how to get CI for E(Y) and PI for Y at sqft=2500 and tax = 1000
newx<-data.frame(sqft= c(2500),tax = c(1000))

> predict(regout,newx,interval="confidence",se.fit=TRUE)

       fit      lwr      upr
1 1434.525 1358.178 1510.872
$se.fit
[1] 38.49996

> predict(regout,newx,interval="predict",se.fit=TRUE)

       fit     lwr      upr
1 1434.525 1080.92 1788.131
```

For simultaneous comparisons once you have the estimated mean and predicted value and associated standard error you can compute the Bonferroni or Scheffe intervals similar to how we did for simple linear regression. It is easier in this case to automate the whole process computing using the matrix-vector expressions (see pages 80-81 of the notes). Also, to carry out computations to get various confidence intervals and tests of hypotheses based on the use of $s^2_{white}\{\mathbf{b}\}$ then we need to do some additional programming using matrix calculations. **We're putting off for now doing these matrix caclulations.**

8.1 Weighted least squares

Weighted least squares proceeds in the same manner as done for simple linear regression. For example, in the house price example, if we assume that $log(\sigma_i) = \gamma_0 + \gamma_1 log(\mu_i)$ then using the fit of $log(e_i)$ on $log(\widehat{Y}_i)$ yields $\widehat{\gamma}_0 = -4.2029$ and $\widehat{\gamma}_1 = 1.19376$, leading to using an estimate of σ_i of $\widehat{\sigma}_i = e^{-4.2029+1.19376log(\widehat{Y}_i)}$ and a weight of $w_i = 1/\widehat{\sigma}_i^2$.

NOTE: If you want to assess whether the weights corrected the problem of unequal variance you have to be careful not to use the regular residuals even when those residuals come from using the weighted least squares coefficients; i.e., $Y_i - \widehat{Y}_{i,wls}$. This residual will still demonstrate the original heteroscedasticity (i.e, changing variance). Instead the residuals have to be weighted $e_{Wi} = e_i\sqrt{w_i}$. These are what need to be used in residual plots.

8.2 Testing general linear hypotheses

We can test general linear hypotheses using the general linear/full-reduced model approach described earlier (see page 69 of the notes and Section 2.8 of the text). In general this involves fitting the full and reduced model (with the reduced model being a subset of the full model). We saw in the house example (see class handout) how we can do this in R via the anova command. In SAS, there is a test command that can used within proc reg. Often the test of interest is that some subset of the β's equals 0.

There are also ways to construct these test using general matrix expressions, which can be applied when we allow for unequal variances. Will return to later.

Example. Consider the house price example with model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$ where $X_1 =$ sqft and $X_2 =$ tax. Consider testing $H_0 : \beta_2 = 0$ and $\beta_3 = 0$ (so null model is a simple linear regression model in sqft).

```
proc reg;
model price=sqft;
run;
proc reg;
model price=sqft tax prod/acov;
test tax = 0, prod=0;    /* Test that coefficients for sqft and prod = 0
                            this will automatically carry out the test
                            under constant variance and under unequal variance */

** MODEL WITH SQFT ONLY. SO SSE(R) = 4527133 with 105 dof.
                                 Sum of          Mean
 Source                   DF     Squares        Square  F Value  Pr > F
 Model                     1    11102445      11102445   257.50  <.0001
 Error                   105     4527133         43116
                        Parameter       Standard
 Variable     DF        Estimate          Error    t Value    Pr > |t|
 SQFT          1         0.60910        0.03796      16.05      <.0001

** MODEL WITH SQFT ONLY. SO SSE(F) = 3137236 with 103 dof. MSE = 30459.
                                 Sum of          Mean
 Source                   DF     Squares        Square  F Value  Pr > F
 Model                     3    12492343       4164114   136.71  <.0001
 Error                   103     3137236         30459
                        Parameter       Standard
 Variable     DF        Estimate          Error    t Value    Pr > |t|
 Intercept     1       194.93093      159.89757       1.22      0.2256
 SQFT          1         0.18923        0.10588       1.79      0.0768
 TAX           1         0.59103        0.21149       2.79      0.0062
 prod          1      0.00006698     0.00009413       0.71      0.4783

                Test 1 Results for Dependent Variable PRICE
                                          Mean
           Source             DF         Square    F Value    Pr > F
           Numerator           2         694949      22.82    <.0001
           Denominator       103          30459
```

```
        Test 1 Results using
           ACOV estimates
DF    Chi-Square    Pr > ChiSq
 2         21.06        <.0001

* THIS CHI-SQUARE TEST IS DOING THE TEST ALLOWING UNEQUAL VARIANCES.
```

The test option gives the F-test (under constant variance) and the chi-square test (allowing unequal variances) that the coefficients for both tax and prod are 0. The F statistic can also be calculated as
$F = ((SSE(R) - SSE)/(105 - 103))/MSE = ((4527133 - 3137236)/2)/30459. = 22.82$

To compute in R we do the following:

```
full<-lm(price~sqft+ tax + prod)
reduced<-lm(price~sqft)
anova(reduced,full)

Analysis of Variance Table
Model 1: price ~ sqft
Model 2: price ~ sqft + tax + prod
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1    105 4527133
2    103 3137236  2   1389898 22.816 6.27e-09 ***
```

8.3 Additional Sums of Squares and R^2

$SSR(X_k, \ldots, X_{p-1}|X_1, \ldots, X_{k-1})$ is the **additional sums of squares** due to the inclusion of $X_k, \ldots, X_{p-1}$ given $X_1, \ldots, X_{k-1}$ in the model.

$$\begin{aligned} SSR(X_k, \ldots, X_{p-1}|X_1, \ldots, X_{k-1}) &= SSR(X_1, \ldots, X_{p-1}) - SSR(X_1, \ldots, X_{k-1}) \\ &= SSE(X_1, \ldots, X_{k-1}) - SSE(X_1, \ldots, X_{p-1}) \geq 0. \end{aligned}$$

If we consider sets of variables (either individually or in blocks) entering the model sequentially we can decompose SSR into additive pieces. This is useful, later for model building. In particular

$$SSR(X_1, \ldots, X_{p-1}) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + ...SSR(X_{p-1}|X_1, \ldots, X_{p-2}).$$

In house example above with $X_1 = sqft$, $X_2 = tax$ and $X_3 = sqft * tax$
$SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1) = 12492343 - 11102445 = 1389898 = SSE(X_1) - SSE(X_1, X_2, X_3) = 4527133 - 3137236$ (differ by 1 because of rounding).

Note that when we add another variable then $R^2 = SSR/SSTO$ is increasing. R^2 is always biggest with all variables in the model.

Note also that the additional sum of squares can be used to test about subsets of coefficients being 0, as seen with the general linear test above. If we want to test $H_0 : \beta_k = \beta_{k+1} = \ldots = \beta_{p-1} = 0$ then the F-statistic can also be written as $F^* = SSR(X_k, \ldots, X_{p-1}|X_1, \ldots, X_{k-1})/(p-k)MSE$.

8.4 Multicollinearity

Patient satisfaction example.

Modeling patient satisfaction as a function of severity of illness and patient's anxiety. The two predictors have a relatively high correlation. This is exercise 6.15 in text, but the analysis here uses just the 23 cases that were in the fourth edition.

- The t-test for each coefficient individually are non-significant but the overall F-test is significant.
- The high correlation between the two predictors translates into high correlation in the two estimated coefficients. (The correlation matrix of the estimated coefficients is obtained using the corrb option. An element of the corrb matrix is obtained by taking an element of covb and dividing by the product of the estimated standard errors of the two coefficients involved.)

  The high correlation in the predictors leads to high standard errors on the fitted coefficients.
- Using either severity or anxiety, there is not much of a drop in R-squared, and the fitted values are not all that different from what is obtained using both.

```
title 'patient example, prob 6.15 in NWNK ';
options pagesize=20 linesize=80;
data a;
infile 'patient.dat';
input satis age severity anxiety;  run;
proc corr;   run;
proc g3d;
scatter severity*anxiety=satis;  run;
proc reg;
model satis = severity anxiety/covb corrb;
output out=result1 p=yhat1 r=resid;   run;
proc plot data=result;
plot resid*(yhat severity anxiety);   run;
proc reg data=a;
model satis = severity;
output out=result2 p=yhat2;  run;
proc reg;
model satis = anxiety;
output out=result3 p=yhat3; run;
data all;
merge result1 result2 result3; run;
proc print data=all;
var satis yhat1 yhat2 yhat3;
run;
```

```
                          The CORR Procedure
                    Pearson Correlation Coefficients
                       Prob > |r| under H0: Rho=0
                          Number of Observations
                        satis           age      severity       anxiety
      satis          1.00000      -0.77368      -0.58744      -0.60231
      severity      -0.58744       0.46661       1.00000       0.79453
      anxiety       -0.60231       0.49769       0.79453       1.00000
```

The REG Procedure
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2426.96719	1213.48359	6.53	0.0066
Error	20	3718.25021	185.91251		
Corrected Total	22	6145.21739			

Root MSE 13.63497 R-Square 0.3949

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	164.23055	34.15344	4.81	0.0001
severity	1	-1.11137	1.07796	-1.03	0.3148
anxiety	1	-20.23141	15.76162	-1.28	0.2140

Correlation of Estimates

Variable	Intercept	severity	anxiety
Intercept	1.0000	-0.7611	0.2140
severity	-0.7611	1.0000	-0.7945
anxiety	0.2140	-0.7945	1.0000

** Using severity only

Root MSE 13.84361 R-Square 0.3451

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	173.61403	33.87240	5.13	<.0001
severity	1	-2.21072	0.66458	-3.33	0.0032

** Using anxiety only

Root MSE 13.65540 R-Square 0.3628

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	137.43188	22.18783	6.19	<.0001
anxiety	1	-33.14267	9.58524	-3.46	0.0024

Obs	satis	Both yhat1	severity only yhat2	anxiety only yhat3
1	48	61.0183	60.8672	61.2037
2	57	66.5751	71.9208	61.2037
3	66	66.3755	67.4994	64.5180
4	70	78.9136	76.3423	77.7751
5	89	80.0250	78.5530	77.7751
23	60	62.1296	63.0780	61.2037

The patient example demonstrates what is known as multicollinearity. Multicollinearity refers to the case where there are "high" correlations (in some sense) among the predictor variables, which means there are strong linear relationships among some of the variables.

- In the case where there is at least one perfect linear relationship among the variables, then $X'X$ is singular and there are infinitely many solutions to the least squares problem (coefficients are not identifiable).
- If you do not have enough distinct combinations of the X's relative to the type of model you are fitting then you will get perfect linear relationships among the columns of the X matrix and have identifiability issues. For example, with two predictors X_1 and X_2 and a model with linear and quadratic terms and a product (so six coefficients in all), you need at least 6 distinct (X_1, X_2) combinations (but not just any six distinct combinations will do).
- With severe multicollinearity there may be numerical issues in getting the inverse of $X'X$ since the $X'X$ matrix is close to singular.
- With high multicollinearity there will often be large standard errors in some of the individual estimated coefficients.
- Multicollinearity can translate into high correlations in the estimated coefficients. This means that although the individual coefficients may have large standard errors (indicating large uncertainty regarding the individual coefficients), it may be still be possible to determine linear combinations with less uncertainty. There are ways to determine joint two dimensional confidence ellipses for the pairs of coefficients.
- When we go to do model building we may want to screen out some predictors that are highly correlated with other predictors.
- Section 7.6 has some additional discussion.

8.5 Polynomial and interaction models.

Models with quadratic and higher order terms and with products (leading to "interactions") fall into our previous developments with the right definitions of additional

variables. The only issues here are i) interpreting the parameters and ii) potential numerical issues when the variables (including original and higher order or product terms) may be highly correlated. This is a case of multicollinearity With high correlation among the predictors computation of least squares estimators is potentially unstable because $(\mathbf{X}'\mathbf{X})$ can be close to singular.

With polynomial models numerical issues can be avoided to some extent with the use of centered variables or the use of orthogonal polynomials. This is not always necessary, but it is often a good precaution to center the variables .If the centered polynomial model is used we can convert back to get coefficients in the original unscaled version. (The book uses centered values with coefficients β_0, etc. but below I use the β's for the model with uncentered values.)

For example if $E(Y_i|X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$ and

$E(Y_i|X_i) = \beta_0^* + \beta_1^*(X_i - \bar{X}) + \beta_2^*(X_i - \bar{X})^2$,

then $\beta_0 = \beta_0^* - \beta_1^*\bar{X} + \beta_2^*\bar{X}^2$, $\beta_1 = \beta_1^* - 2\beta_2^*\bar{X}$ and $\beta_2 = \beta_2^*$

Example: Using yield data from problem 36 in chapter 7 of fourth edition of our text. Response is yield and predictors are moisture (in inches) and temp (in degrees centigrade). We'll being by looking at yield as a function of moisture only first for illustration. *Think about the possible ways in which this data might have arose and what the models means. Is data from different fields in a natural setting? If so, what about natural fertility of the soil? Is it from a greenhouse experiment with moisture and temp controlled. If so, were they randomized? etc.*

```
title 'yield example ';
data a;
infile 'f:/s505/data/yield.dat';
input yield moist temp;
moistc=(moist-10);  /* centered moisture value */
m2c = moistc**2;
m2=moist*moist;  t2=temp*temp; mt=moist*temp;
run;
proc g3d;
scatter moist*temp=yield;
run;
proc sort;
by moist;
run;
```

```
proc gplot;
plot yield*temp;
by moist;
run;
proc corr;     /* shows correlation of the variables*/
var yield moist m2 moistc m2c;  run;
proc reg;
model yield=moist m2/covb;
plot r.*(moist temp);
model yield=moistc m2c/covb;
run;
```

```
            Plot of yield*moist.  Legend: A = 1 obs, B = 2 obs, etc.
      55 +
         |
         |
         |
         |                 B             A
         |                 A             A
      50 +  A              A             A
         |  A                            A
         |  B              A             A              B
   yield |
         |  A                                           B
         |                                              A
      45 +
         |                                                            A
         |                                                            A
         |                                                            B
         |
         |                                                            A
      40 +
         ---+--------------+--------------+--------------+--------------+--
            6              8             10             12             14
                                        moist
```

```
                          The CORR Procedure
                          Simple Statistics
Variable         N        Mean     Std Dev         Sum     Minimum     Maximum
moist           25    10.00000     2.88675   250.00000     6.00000    14.00000
 ...
```

Pearson Correlation Coefficients, N = 25
Prob > |r| under H0: Rho=0

	yield	moist	m2	moistc	m2c
yield	1.00000	-0.69359	-0.76291	-0.69359	-0.63081
moist	-0.69359	1.00000	0.99307	1.00000	0.00000
m2	-0.76291	0.99307	1.00000	0.99307	0.11750
moistc	-0.69359	1.00000	0.99307	1.00000	0.00000
m2c	-0.63081	0.00000	0.11750	0.00000	1.00000

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	212.18594	106.09297	79.90	<.0001
Error	22	29.21166	1.32780		
Corrected Total	24	241.39760			

Root MSE 1.15230 R-Square 0.8790

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	28.44114	3.27893	8.67	<.0001
moist	1	5.09514	0.69344	7.35	<.0001
m2	1	-0.29286	0.03443	-8.51	<.0001

Covariance of Estimates

Variable	Intercept	moist	m2
Intercept	10.751407317	-2.247780111	0.1090694991
moist	-2.247780111	0.4808542263	-0.023710761
m2	0.1090694991	-0.023710761	0.001185538

Model: MODEL2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	212.18594	106.09297	79.90	<.0001
Error	22	29.21166	1.32780		
Corrected Total	24	241.39760			

Root MSE 1.15230 R-Square 0.8790

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50.10686	0.35915	139.52	<.0001
moistc	1	-0.76200	0.08148	-9.35	<.0001
m2c	1	-0.29286	0.03443	-8.51	<.0001

Covariance of Estimates

Variable	Intercept	moistc	m2c
Intercept	0.128986538	0	-0.009484304
moistc	0	0.006639013	0
m2c	-0.009484304	0	0.001185538

Here the centering is not necessary to deal with any numerical instability.

Fitting a model with moisture and temperature with quadratic terms and product.

```
proc reg;
model  yield = moist temp;
plot (r.)*(moist temp);
run;
proc reg;
model  yield = moist temp m2 t2 mt/acov;
output out=result2 r=resid p=yhat;
plot r.*(moist temp p.);
run;
tvars:    test temp=0, t2=0, mt=0;   run;
proc g3d data=result2;
plot moist*temp=yhat;  /* plot fitted surface */
run;
```

```
FULL FIT
                              Sum of          Mean
        Source          DF      Squares       Square      F Value       Prob>F
        Model            5    227.58719     45.51744       62.622       0.0001
        Error           19     13.81041      0.72686
        C Total         24    241.39760
            Root MSE       0.85256     R-square       0.9428
                         Parameter      Standard    T for H0:
      Variable  DF       Estimate         Error   Parameter=0    Prob > |T|
      INTERCEP   1    -27.900286   50.13253545        -0.557        0.5843
      MOIST      1      5.216143    1.06898733         4.880        0.0001
      TEMP       1      5.622143    4.50546693         1.248        0.2272
      M2         1     -0.292857    0.02547520       -11.496        0.0001
      T2         1     -0.138571    0.10190079        -1.360        0.1898
      MT         1     -0.005500    0.04262816        -0.129        0.8987
                Test tvars Results for Dependent Variable yield
                                        Mean
           Source             DF         Square    F Value    Pr > F
           Numerator           3        5.13375       7.06    0.0022
           Denominator        19        0.72686

           Test tvars Results using Heteroscedasticity
                 Consistent Covariance Estimates
                   DF    Chi-Square    Pr > ChiSq
                    3         53.62        <.0001

ABOVE TESTING USING WHITE'S ESTIMATOR.
```

Using THE full/reduced model approach the F-test of the null hypothesis that the coefficient for temp, temp-squared and temp*moist are all 0 can also be constructed via $F^* = (29.2116 - 13.81041)/3 * .72686 = 7.06$ and compared to an F with 3 and 19 degrees of freedom. 29.2116 is SSE(R) from model with just m and m2, while SSE(F) = 13.8104 is from the full fit.

Summary of the key points in the yield, moisture, temperature example.

- When fitting a quadratic model in moisture alone.

  - The center X and X^2 had zero correlation

  - The fitted values (which can be used as estimated means or predictions at the corresponding X) are the same whether the fit is done in terms of the centered or uncentered values.

  - If $\mathbf{b}$ are the fitted coefficients with original X's and $\mathbf{b}^*$ the fitted coefficients using centered X's, then using the fact that $\beta_0 = \beta_0^* - \beta_1^*\bar{X} + \beta_2^*\bar{X}^2$, $\beta_1 = \beta_1^* - 2\beta_2^*\bar{X}$ and $\beta_2 = \beta_2^*$,
  $\mathbf{b} = A\mathbf{b}^*$, where

$$A = \begin{bmatrix} 1 & -10 & 100 \\ 0 & 1 & -20 \\ 0 & 0 & 1 \end{bmatrix}.$$

  With $s^2\{\mathbf{b}\}$ and $s^2\{\mathbf{b}^*\}$ denoting the estimated variance-covariance matrix for $\mathbf{b}$ and $\mathbf{b}^*$ respectively, $s^2\{\mathbf{b}\} = As^2\{\mathbf{b}^*\}A'$.

  These results let you get the results for coefficients in the uncentered X's from those of the centered X's which are generally more accurate (numerically).

- Test about various subsets of coefficients equaling 0 can be obtained either directly using the full-reduced model approach or using the test option in SAS.

- One should not look through the collection of individual t-tests to decide as a group whether some variables should be eliminated. In the full fit (with linear, quadratic and product) the three t-test associated with coefficients for variables involving temp are all non-significant, but the F-test that all three of these coefficients are 0 simultaneously is significant (with a p-value of .0022).

- The residual plots indicate that there is still a little problem with the model even using quadratic models plus interaction. If we drop the observations at the highest temp the fit is improved.

Illustration of issue with identifiability/estimability of regression coefficients. Trying to fit second order model with interaction using different combinations of moisture and temperature. There are 6 parameters in this model. There is a unique solution/parameters are identifiable if X matrix has rank 6 or equivalently $X'X$ nonsingular (i.e. has rank 6 which is true if and only if determinant is non-zero). We know we need at least six distinct X_1, X_2 combinations, but not just and combination will do.

THESE TWO LEAD TO BEING ABLE TO ESTIMATE THE PARAMETERS

3 x 3	
X1	X2
6.0	20.0
6.0	22.0
6.0	24.0
10.0	20.0
10.0	22.0
10.0	24.0
14.0	20.0
14.0	22.0
14.0	24.0

6 distinct combos with center point	
6.0	20.0
6.0	22.0
6.0	24.0
10.0	22.0
14.0	20.0
14.0	24.0

CANNOT ESTIMATE WITH THESE TWO DESIGNS

3 x 2	
6.0	20.0
6.0	22.0
10.0	20.0
10.0	22.0
14.0	20.0
14.0	22.0

4 x 2	
6.0	20.0
6.0	22.0
8.0	20.0
8.0	22.0
10.0	20.0
10.0	22.0
14.0	20.0
14.0	22.0

Creating scatterplot matrices: Using the house data to illustrate

In R

```
 pairs(~price+sqft+tax+age)
```

Using SAS

```
proc corr plots=matrix(histogram);
var price sqft tax age;
```

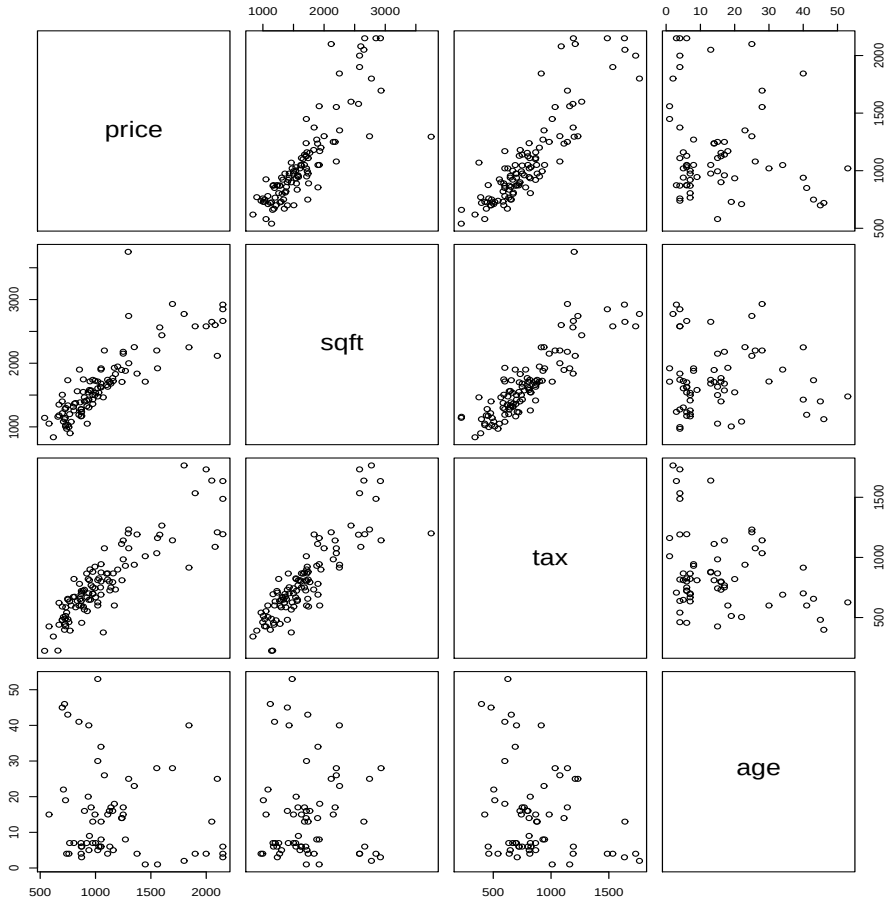


Figure 12: Scatterplot matrix for patient data from R

8.6 Qualitative Predictors and regression for different groups

Suppose we have g groups based on a categorical/qualitative predictor (race, gender, state) and a single quantitative predictor X_1 (for simplicity to begin with). If observation i is from group j, consider the model

$$Y_i = \beta_{j0} + \beta_{j1} X_{i1} + \epsilon_i,$$

which allows different coefficients for each group.

Define: $Z_{ij} = 1$ if observation i is in group j and 0 otherwise. The model can be expressed as:

$$Y_i = \beta_{10} Z_{i1} + \beta_{20} Z_{i2} + \ldots \beta_{g0} Z_{ig} + \beta_{11} Z_{i1} X_{i1} + \beta_{21} Z_{i2} X_{i1} + \ldots + \beta_{g1} Z_{ig} X_{i1} + \epsilon_i.$$

If we fit this model directly note that it will have no overall intercept term.

Alternatively,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_{i1} + \beta_{g+1} Z_{ig} + \beta_{g+2} Z_{i1} X_{i1} + \beta_{g+3} Z_{i2} X_{i1} + \ldots + \beta_{2g+1} Z_{ig} X_{i1} + \epsilon_i.$$

where $\beta_0 + \beta_{j+1} = \beta_{j0}$, $\beta_1 + \beta_{g+1+j} = \beta_{j1}$.

Will have some trouble fitting the last model as is. This version has a total of 1 + 1 + g + g = 2g +2 terms while the first has 2g terms. Since $\sum_{j=1}^{g} Z_{ij} X_{i1} = X_{i1}$ and $\sum_{j=1}^{g} Z_{ij} = 1$ (corresponding to the first two terms in the model, 1 for the intercept and the X_{i1}) there are exact linear restrictions among the columns of the $\mathbf{X}$ matrix. The $\mathbf{X'X}$ will be singular. The problem is in retaining X and all XZ products in the model and the intercept and all of the Z terms. This can be alleviated by dropping one of the XZ terms and dropping one of the Z terms. If you drop $Z_{ig} X_{i1}$ and Z_{ig}, then $\beta_{g0} = \beta_0$ and $\beta_{g1} = \beta_1$ (so β_0 and β_1 are the intercept and slope for the last group). The other coefficients represent differences (effects) compared to this group; e.g., $\beta_0 + \beta_2 = \beta_1$ so the coefficient β_2 for Z_{i1} is $\beta_1 - \beta_0 = \beta_1 - \beta_g$ and the coefficient for $X_{i1} Z_{ij}$ is $\beta_{j1} - \beta_{jg}$.

All of the above is easily extended to handle additional quantitative X variables or more than one categorical predictor.

The objective is to estimate the regression coefficients for each group and compare them. In the general model,

- The slope is constant if and only if all of the coefficients for the ZX products equal 0.
- The intercept is constant if and only if all of the coefficients of the Z terms equal 0.

We can use proc reg or lm in R with suitable dummy variables and fit either model. We can also use glm in SAS or lm in R and do fits without explicitly creating the dummy variables.

Example: Using the brain data. Relate fsiq (full-scale IQ) to mricount (indicator of brain size) for four groups (gender crossed with FSIQ grouping, since students were stratified based on FSIQ). Below shows fits and inferences for a variety of models, first for SAS and then with R.

```
option ls=80 nodate;
title 'Brain Data';
data a;
infile 'Brain.dat';
input Gender $ FSIQ VIQ PIQ Weight Height mriCount;
z1=0;
z2=0;
z3=0;
z4=0;
if fsiq > 129 and Gender='Male' then group='M1';
if fsiq > 129 and Gender='Female' then group='F1';
if fsiq < 104 and Gender='Male' then group='M2';
if fsiq < 104 and Gender='Female' then group='F2';
if group='M1' then z1=1;
if group='F1' then z2=1;
if group='M2' then z3=1;
if group='F2' then z4=1;
mz1=mricount*z1;
mz2=mricount*z2;
mz3=mricount*z3;
mz4=mricount*z4;
run;
proc print;
var gender fsiq  group z1 z2 z3 z4;
run;
proc sort; by group; run;
```

```
proc gplot; plot fsiq*mricount; by group; run;

proc reg;  model fsiq=mricount;      by group;    run;
proc reg;
model fsiq=mricount z1 z2 z3 z4 mz1 mz2 mz3 mz4; /* will have trouble*/
model fsiq=mricount z1 z2 z3 mz1 mz2 mz3;
model fsiq=z1 z2 z3 z4 mz1 mz2 mz3 mz4/noint;
model fsiq = mz1 mz2 mz3 mz4;  /* model with common intercept*/
model fsiq= z1 z2 z3 z4 mricount/noint; /* model with common slope */
model fsiq= mricount; /* model with common intercept and slope */
run;
/* The test statement in proc reg will test linear hypotheses.
If the acov option is expressed earlier, these tests are based
on the robust estimate of the variance-covariance of the coeffiicents*/
proc reg;
model fsiq=mricount z1 z2 z3 mz1 mz2 mz3;
commoni: test z1=0,z2=0,z3=0;  /* test for equal intercepts*/
commons: test mz1=0,mz2=0,mz3=0; /* test for equal slopes*/
run;
```

Obs	Gender	FSIQ	group	z1	z2	z3	z4
1	Female	133	F1	0	1	0	0
2	Male	140	M1	1	0	0	0
5	Female	137	F1	0	1	0	0
6	Female	99	F2	0	0	0	1

group=F1

Root MSE	3.09716	R-Square	0.1290

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	119.16618	13.94076	8.55	<.0001
mriCount	1	0.00001727	0.00001586	1.09	0.3082

group=F2

Root MSE	7.39782	R-Square	0.1839

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25.55433	47.67909	0.54	0.6066
mriCount	1	0.00007534	0.00005611	1.34	0.2162

group=M1

Root MSE	3.77132	R-Square	0.0557

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	121.44475	24.84643	4.89	0.0012
mriCount	1	0.00001749	0.00002545	0.69	0.5114

group=M2

Root MSE	6.52693	R-Square	0.5107

Variable	DF	Parameter	Standard

Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	-11.92568	35.85256	-0.33	0.7480
mriCount	1	0.00011069	0.00003831	2.89	0.0202

The REG Procedure
Model: MODEL1
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	21649	3092.67847	102.12	<.0001
Error	32	969.15071	30.28596		
Corrected Total	39	22618			

Root MSE	5.50327	R-Square	0.9572

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

z4 = Intercept - z1 - z2 - z3
mz4 = mriCount - mz1 - mz2 - mz3

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	B	25.55433	35.46867	0.72	0.4765
mriCount	B	0.00007534	0.00004174	1.81	0.0805
z1	B	95.89042	50.72072	1.89	0.0678
z2	B	93.61185	43.26231	2.16	0.0381
z3	B	-37.48001	46.60315	-0.80	0.4272
z4	0	0	.	.	.
mz1	B	-0.00005785	0.00005586	-1.04	0.3081
mz2	B	-0.00005807	0.00005037	-1.15	0.2574
mz3	B	0.00003535	0.00005278	0.67	0.5078
mz4	0	0	.	.	.

Model: MODEL2
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	21649	3092.67847	102.12	<.0001
Error	32	969.15071	30.28596		
Corrected Total	39	22618			

Root MSE	5.50327	R-Square	0.9572

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25.55433	35.46867	0.72	0.4765
mriCount	1	0.00007534	0.00004174	1.81	0.0805
z1	1	95.89042	50.72072	1.89	0.0678
z2	1	93.61185	43.26231	2.16	0.0381
z3	1	-37.48001	46.60315	-0.80	0.4272
mz1	1	-0.00005785	0.00005586	-1.04	0.3081

```
      mz2           1    -0.00005807     0.00005037      -1.15      0.2574
      mz3           1     0.00003535     0.00005278       0.67      0.5078
-------
                                Model: MODEL3
             NOTE: No intercept in model. R-Square is redefined.
                                   Sum of           Mean
 Source                   DF        Squares         Square    F Value    Pr > F
 Model                     8         536485          67061    2214.25    <.0001
 Error                    32      969.15071       30.28596
             Root MSE              5.50327    R-Square     0.9982
                           Parameter       Standard
      Variable     DF        Estimate          Error    t Value    Pr > |t|
      z1            1       121.44475       36.25693       3.35      0.0021
      z2            1       119.16618       24.77096       4.81      <.0001
      z3            1       -11.92568       30.22956      -0.39      0.6958
      z4            1        25.55433       35.46867       0.72      0.4765
      mz1           1     0.00001749     0.00003713       0.47      0.6409
      mz2           1     0.00001727     0.00002819       0.61      0.5446
      mz3           1     0.00011069     0.00003230       3.43      0.0017
      mz4           1     0.00007534     0.00004174       1.81      0.0805
----
                                Model: MODEL4
                                   Sum of           Mean
 Source                   DF        Squares         Square    F Value    Pr > F
 Model                     4          21199     5299.67376     130.70    <.0001
 Error                    35     1419.20496       40.54871
             Root MSE              6.36779    R-Square     0.9373
                           Parameter       Standard
      Variable     DF        Estimate          Error    t Value    Pr > |t|
      Intercept     1        68.66363       17.68744       3.88      0.0004
      mz1           1     0.00007148     0.00001821       3.93      0.0004
      mz2           1     0.00007460     0.00002021       3.69      0.0008
      mz3           1     0.00002473     0.00001899       1.30      0.2014
      mz4           1     0.00002467     0.00002092       1.18      0.2463
-------
                                Model: MODEL5
             NOTE: No intercept in model. R-Square is redefined.
                                   Sum of           Mean
 Source                   DF        Squares         Square    F Value    Pr > F
 Model                     5         536303         107261    3262.29    <.0001
 Error                    35     1150.76201       32.87891
             Root MSE              5.73401    R-Square     0.9979
                           Parameter       Standard
      Variable     DF        Estimate          Error    t Value    Pr > |t|
      z1            1        87.52915       17.23346       5.08      <.0001
      z2            1        88.49221       15.50821       5.71      <.0001
      z3            1        42.67086       16.51753       2.58      0.0141
      z4            1        45.14449       15.02337       3.00      0.0049
      mriCount      1     0.00005226     0.00001757       2.97      0.0053
----
                                Model: MODEL6
```

```
                                  Sum of           Mean
 Source                   DF       Squares         Square    F Value    Pr > F
 Model                     1     2892.98916     2892.98916       5.57    0.0235
 Error                    38          19725      519.07660
             Root MSE              22.78325    R-Square     0.1279
                           Parameter       Standard
      Variable     DF       Estimate          Error    t Value    Pr > |t|
      Intercept     1        5.16770       46.00819       0.11      0.9112
      mriCount      1     0.00011915     0.00005047       2.36      0.0235
------
FROM LAST REG WITH TEST STATEMENTS
               Test commoni Results for Dependent Variable FSIQ
                                       Mean
           Source             DF         Square    F Value    Pr > F
           Numerator           3      150.01808       4.95    0.0062
           Denominator        32       30.28596

               Test commons Results for Dependent Variable FSIQ
                                       Mean
           Source             DF         Square    F Value    Pr > F
           Numerator           3       60.53710       2.00    0.1339
           Denominator        32       30.28596
```

You can use proc GLM to do some of the analysis without having to construct the indicator/dummy variables yourself. proc glm does not have all of the same features of proc reg, but it easily handles qualitative variables through use of the class statement. It also has an estimate command that lets us get inferences for any linear combination of the coefficients.

```
proc glm;
class group;
model fsiq = group group*mricount/solution noint;
run;
```

PARTIAL OUTPUT

```
 Source                    DF    Type III SS    Mean Square   F Value   Pr > F
 group                      4    1061.138247     265.284562      8.76   <.0001
 mriCount*group             4     472.449289     118.112322      3.90   0.0109
                                          Standard
   Parameter                Estimate          Error    t Value    Pr > |t|
   group          F1      119.1661770     24.77096356       4.81      <.0001
   group          F2       25.5543305     35.46867007       0.72      0.4765
   group          M1      121.4447461     36.25692733       3.35      0.0021
   group          M2      -11.9256789     30.22956491      -0.39      0.6958
   mriCount*group F1        0.0000173      0.00002819       0.61      0.5446
   mriCount*group F2        0.0000753      0.00004174       1.81      0.0805
   mriCount*group M1        0.0000175      0.00003713       0.47      0.6409
   mriCount*group M2        0.0001107      0.00003230       3.43      0.0017
```

```
proc glm;
model fsiq=z1 z2 z3 z4 mz1 mz2 mz3 mz4/noint;
estimate 'slope1 versus slope 2' mz1 1 mz2 -1;
run;
                                The GLM Procedure
A LOT OF OTHER OUTPUT ELIMINATED
                                                        Standard
  Parameter                     Estimate                  Error    t Value    Pr > |t|
  slope1 versus slope 2    2.2112845E-7          0.00004662          0.00      0.9962
```

Working with qualitative variables in R. Doing the Brain example. Exploiting the use of model matrices to computed dummy variables.

```
> brain<-read.table("f:/s505/Brain_name.dat",header=T)
> brain  #list the data
   Gender FSIQ VIQ PIQ Weight Height mriCount
1       F  133 132 124    118   64.5   816932
2       M  140 150 124      .   72.5  1001121
3       M  139 123 150    143   73.3  1038437

38      F   88  86  94    139   64.5   893983
39      M   81  90  74    148   74.0   930016
40      M   89  91  89    179   75.5   935863
> attach(brain)

> # defines iqgroups
> iqgroup <-rep(2,length(FSIQ))  # initializes vector to 2
> for (i in 1:length(FSIQ))
  + {if (FSIQ[i] > 129) {iqgroup[i]=1}} # changes iqgroup to 1
> groupi<-factor(iqgroup)  # turn iqgroup into a factor variable
> # create dummy variables by using the model matrices that
> # come if fit a model with a qualitative grouping variable only
> # as a predictor. The -1 in the model indicates a grouping variable

> x1<-model.matrix(lm(FSIQ~-1 + groupi)) # no intercept is specified by -1
> x2<-model.matrix(lm(FSIQ~-1 + Gender))
> x1  # the first col of x1 is the indicator for group1, the second for group2

   group1 group2
1       1      0
2       1      0
3       1      0
4       1      0


37      1      0
38      0      1
39      0      1
40      0      1
```

```
> x2  # the first col of x2 is the indicator for female the second for male

   GenderF GenderM
1        1       0
2        0       1
3        0       1
4        0       1


37       0       1
38       1       0
39       0       1
40       0       1

> z1<-x1[,1]*x2[,2]
> z2<-x1[,1]*x2[,1]
> z3<-x1[,2]*x2[,2]
> z4<-x1[,2]*x2[,1]
> mz1<-mriCount*z1
> mz2<-mriCount*z2
> mz3<-mriCount*z3
> mz4<-mriCount*z4

> # Fit a model with separate coefficients for each group. Assumes common
> # variance
> full<-lm(FSIQ~-1+z1+z2+z3+z4 + mz1+mz2+ mz3 + mz4)
> # could also have used
> # full<-lm(-1+z1+z2+z3+z4 + I( mriCount*z1)+ I(mriCount*z2) + I(mriCount *z3) +I(mriCount*z4))
> summary(full)

Call:
lm(formula = FSIQ ~ -1 + z1 + z2 + z3 + z4 + mz1 + mz2 + mz3 +
    mz4)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0187  -2.9737   0.2036   2.8841  14.5693

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
z1   1.214e+02  3.626e+01   3.350  0.00209 **
z2   1.192e+02  2.477e+01   4.811 3.44e-05 ***
z3  -1.193e+01  3.023e+01  -0.395  0.69583
z4   2.555e+01  3.547e+01   0.720  0.47646
mz1  1.749e-05  3.713e-05   0.471  0.64088
mz2  1.727e-05  2.819e-05   0.612  0.54456
mz3  1.107e-04  3.230e-05   3.427  0.00170 **
mz4  7.534e-05  4.174e-05   1.805  0.08048 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 5.503 on 32 degrees of freedom
Multiple R-squared: 0.9982,     Adjusted R-squared: 0.9977
F-statistic:  2214 on 8 and 32 DF,  p-value: < 2.2e-16

> # try to fit with intercept, mriCount and  all z's and mz's
> lm(FSIQ~ z1+z2+z3+z4 + mriCount+ mz1+mz2+ mz3 + mz4)

Call:
lm(formula = FSIQ ~ z1 + z2 + z3 + z4 + mriCount + mz1 + mz2 +
    mz3 + mz4)

Coefficients:
(Intercept)           z1           z2           z3           z4     mriCount
  2.555e+01    9.589e+01    9.361e+01   -3.748e+01           NA    7.534e-05
        mz1          mz2          mz3          mz4
 -5.785e-05   -5.807e-05    3.535e-05           NA

> # fit model with intercept, mriCount and eliminate last z and mz
> lm(FSIQ~ z1+z2+z3 + mriCount+ mz1+mz2+ mz3)

Call:
lm(formula = FSIQ ~ z1 + z2 + z3 + mriCount + mz1 + mz2 + mz3)

Coefficients:
(Intercept)           z1           z2           z3     mriCount          mz1
  2.555e+01    9.589e+01    9.361e+01   -3.748e+01    7.534e-05   -5.785e-05
        mz2          mz3
 -5.807e-05    3.535e-05

> # model with constant intercept
> commoni<-lm(FSIQ~ mz1+mz2+ mz3 + mz4)
> # model with constant slope
> commons<-lm(FSIQ~ -1+ z1+ z2+ z3 +z4 + mriCount)
> # model with common intercept and slope
> commonboth<-lm(FSIQ ~ mriCount)
> # test for equal intercepts

> anova(commoni,full)

Analysis of Variance Table
Model 1: FSIQ ~ mz1 + mz2 + mz3 + mz4
Model 2: FSIQ ~ -1 + z1 + z2 + z3 + z4 + mz1 + mz2 + mz3 + mz4
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     35 1419.20
2     32  969.15  3    450.05 4.9534 0.006165 **

> # test for equal slopes
> anova(commons,full)

Analysis of Variance Table
Model 1: FSIQ ~ -1 + z1 + z2 + z3 + z4 + mriCount
```

```
Model 2: FSIQ ~ -1 + z1 + z2 + z3 + z4 + mz1 + mz2 + mz3 + mz4
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     35 1150.76
2     32  969.15  3    181.61 1.9989 0.1339

>  # working with group which has four categories
> group<-paste(Gender, groupi)
> group
 [1] "F 1" "M 1" "M 1" "M 1" "F 1" "F 2" "F 1"   etc.
[23] "F 1" "M 1" "F 2" "M 1" "F 2" "M 2" "F 2"   etc.

> lm(FSIQ ~ group*mriCount)  # This will use an intercept and mriCoung plus interactions.

Coefficients:
      (Intercept)           groupF 2           groupM 1           groupM 2           mriCount
        1.192e+02         -9.361e+01          2.279e+00         -1.311e+02          1.727e-05

groupF 2:mriCount  groupM 1:mriCount   groupM 2:mriCount
  5.807e-05          2.211e-07          9.343e-05

> lm(FSIQ ~-1 +  group + mriCount*group)

Coefficients:
         groupF 1           groupF 2           groupM 1           groupM 2
        1.192e+02          2.555e+01          1.214e+02         -1.193e+01
mriCount  groupF 2:mriCount  groupM 1:mriCount  groupM 2:mriCount
 1.727e-05      5.807e-05          2.211e-07          9.343e-05
```

Above assumes constant variance throughout. *Under this assumption*

- The various ways of fitting the full model (different intercepts and slopes) lead to the same estimated coefficients.

- When SAS or R encounters a model "not of full rank" (there are linear restrictions among the predictors) it sets certain parameters to 0. Here that corresponds to dropping the last Z and the last ZX terms.

- The test for equal slopes and coefficients are obtained directly using the test option in proc reg in SAS or the anova command in R. They can also be obtained using the full-reduced/general linear test approach. For example, for equal slopes, $F = [(1150.75201 - 969.15071)/3]/30.28596 = 2.00$.

- Note that estimated coefficients from different groups are uncorrelated so for comparisons between groups you can also construct tests and confidence intervals

directly from the estimates and standard errors. For example to compare the slope in group 1 to the slope in group 2 the standard error for the estimated difference in slopes, $b_{11} - b_{21}$ is $SE = \sqrt{(s\{b_{11}\}^2 + s\{b_{21}\}^2}$. From this you can form a confidence interval or a test. The degrees of freedom involved is degrees of freedom associated with MSE (recall assuming equal variance throughout at this point). Earlier, this was done using the estimate option in GLM in SAS. In R there are various ways to extract the information you need to compute this.

The Brain example. Allowing unequal variances.

There is some evidence of different variances among the groups. To accout for this

1. Fit one of big modesl above allowing different variances everywhere and use ACOV. If you run each group with acov or one overall regression with acov, you get the same variance-covariance matrix associated with each pair of coefficients for a group. Estimates from different groups are uncorrelated. The test statements will be carried out under acov (using a chi-square test) if that option is chosen.

2. Assume equal variance within a group but different variance among groups. This won't change the estimated coefficients but will change variances-covariances of estimated coefficients. Can run proc reg by group and get variance-covariance for each. An easier, but equivalent approach, is to give weight $1/MSE_j$ to an observation from group j and run one big model with weighting. Can now use one big model in reg or glm and carry out various test, estimates, etc.

9 Variable Selection/model building

Have a response Y and a collection of $P-1$ predictors $X_1, \ldots, X_{P-1}$ (which may be made up of some original variables and functions of them). Note the use of P here. The number of coefficients in a particular subset of interest will be denoted by p. The goal is to determine which of the predictors to retain in a final model. There are two perspectives here:

- Trying to determine the "correct" model.
  $E(Y|X_1, \ldots, X_{P-1}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots \beta_{P-1} X_{P-1}$. The true model has some β_j's equal to 0 and we are trying to decide which those are. Note that this is the "correct" model in the sense of the regression function given *all of the* $P-1$ *predictors*. There can still be other correct models based on conditioning on a different set of predictors. Just because the coefficient of X_j is 0 in the conditional model with all of the predictors does not mean it has to be 0 in a conditional model on a different subset of predictors.

- Trying to come up with a good model for prediction.

  We are not concerned with estimating coefficients associated with particular variables, but simply in coming up with a parsimonious model that does a good job of prediction. Here we want to penalize models that have too many parameters.

$p-1$ will indicate the number of variables in a particular subset under consideration (so p parameters in total).

1. ALL POSSIBLE SUBSETS. Examine fits over all subsets over various sizes ($p = 1$ to P). There are many proposed measures to base decision on: .

   (a) $R^2 = 1 - (SSE/SSTO)$. (Always largest with all variables.)

   (b) Adjusted R-square: $R^2_a = 1 - \frac{MSE}{(SSTO/(n-1))} = 1 - a + aR^2$,
   where $a = (n-1)/(n-p)$. Larger R^2_a corresponds to smaller MSE.

   (c) AIC (Akaike Information criterion) $= nlog(SSE_p) - nlog(n) + 2p$. Has become very popular. Smaller is better

   (d) Mallows C. For a subset with p variables

   $$C = \frac{SSE_{subset}}{MSE_P} - (n - 2p),$$

   where MSE_P is MSE from fit with all variables.
   The motivation for Mallow's C is to minimize to total expected squared error

   $$\sum_i E[(\hat{Y}_i - \mu_i)^2] = \sum_i (E(\hat{Y}_i - \mu_i)^2 + \sigma^2\{\hat{Y}_i)\},$$

   where $\hat{Y}_i$ is an estimator of $\mu_i = E(Y_i)$ based on the subset under consideration.
   This measure trades off between bias and variance. Good subsets are ones with small C *and* with C less than or close to p. A C less than or close to p indicates little bias. When you use all variables then $C_p = p$.
   Note that SAS gives you NUMBER IN MODEL which is $p-1$ so you need to compare SAS's C(p) to NUMBER IN MODEL + 1.

   (e) Others: Press, SBC (Bayesian criteriaon; smaller is better)

2. Forward selection

3. Backward selection

4. Stepwise selection

Using example from the book. Will work with logsurv = log(survival) as they did there since linearity is more reasonable on this scale. For prediction can convert back to survival by exponentiating. There are issues though in interpreting the back transformation as an estimator of expected survival because the transformation is non-linear. After doing model building here with log-survival there is more to be done if objective is to estimate expected survival and variability in survival. Only four variables here so good for illustration. Can list all subsets easily.

```
title 'surgical unit example in ch 8.';
data a;
infile 'surg.dat';
input blood prog enz liver surv logsurv;
run;
proc gplot;
plot (logsurv surv)*(blood prog enz liver);
run;
proc corr; run;
```

The CORR Procedure

6 Variables: blood prog enz liver surv logsurv

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
blood	54	5.78333	1.60303	312.30000	2.60000	11.20000
prog	54	63.24074	16.90253	3415	8.00000	96.00000
enz	54	77.11111	21.25378	4164	23.00000	119.00000
liver	54	2.74426	1.07036	148.19000	0.74000	6.40000
surv	54	197.16667	145.29940	10647	34.00000	830.00000
logsurv	54	2.20614	0.27378	119.13180	1.53150	2.91910

Pearson Correlation Coefficients, N = 54
Prob > |r| under H0: Rho=0

	blood	prog	enz	liver	surv	logsurv
blood	1.00000	0.09012	-0.14963	0.50242	0.37252	0.34640
		0.5169	0.2802	0.0001	0.0055	0.0103
prog	0.09012	1.00000	-0.02361	0.36903	0.55398	0.59289
	0.5169		0.8655	0.0060	<.0001	<.0001
enz	-0.14963	-0.02361	1.00000	0.41642	0.58024	0.66512
	0.2802	0.8655		0.0017	<.0001	<.0001
liver	0.50242	0.36903	0.41642	1.00000	0.72233	0.72621
	0.0001	0.0060	0.0017		<.0001	<.0001

```
surv        0.37252     0.55398     0.58024     0.72233     1.00000     0.91310
             0.0055      <.0001      <.0001      <.0001                  <.0001

logsurv     0.34640     0.59289     0.66512     0.72621     0.91310     1.00000
             0.0103      <.0001      <.0001      <.0001      <.0001
proc reg;
model logsurv=blood prog enz liver/selection =rsquare;  run;
```

The REG Procedure
Dependent Variable: logsurv
R-Square Selection Method

Number in Model	R-Square	Variables in Model
1	0.5274	liver
1	0.4424	enz
1	0.3515	prog
1	0.1200	blood
2	0.8130	prog enz
2	0.6865	enz liver
2	0.6496	prog liver
2	0.6458	blood enz
2	0.5278	blood liver
2	0.4381	blood prog
3	0.9723	blood prog enz
3	0.8829	prog enz liver
3	0.7192	blood enz liver
3	0.6500	blood prog liver
4	0.9724	blood prog enz liver

```
proc reg data=a outest=selectarsq;
model logsurv=blood prog enz liver/aic sbc  selection =adjrsq;
run;
```

Adjusted R-Square Selection Method

Number in Model	Adjusted R-Square	R-Square	AIC	SBC	Variables in Model
3	0.9707	0.9723	-326.6674	-318.71149	blood prog enz
4	0.9701	0.9724	-324.7107	-314.76583	blood prog enz liver
3	0.8759	0.8829	-248.7297	-240.77375	prog enz liver
2	0.8056	0.8130	-225.4455	-219.47859	prog enz
3	0.7023	0.7192	-201.4980	-193.54211	blood enz liver
2	0.6742	0.6865	-197.5576	-191.59067	enz liver
2	0.6358	0.6496	-191.5392	-185.57221	prog liver
2	0.6319	0.6458	-190.9594	-184.99243	blood enz
3	0.6290	0.6500	-189.6024	-181.64646	blood prog liver
1	0.5183	0.5274	-177.3845	-173.40655	liver

```
        2     0.5093    0.5278   -175.4366   -169.46964  blood liver
        1     0.4317    0.4424   -168.4549   -164.47695  enz
        2     0.4160    0.4381   -166.0369   -160.06993  blood prog
        1     0.3390    0.3515   -160.3025   -156.32457  prog
        1     0.1031    0.1200   -143.8168   -139.83885  blood
```

The SAS file selectarsq contains the different measures and other things.
Can be sorted by differnt measures

```
proc print data=selectarsq;
```

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	blood	prog	enz
1	MODEL1	PARMS	logsurv	0.04687	0.48362	0.069225	.009294538	.009523639
2	MODEL1	PARMS	logsurv	0.04733	0.48876	0.068520	.009254111	.009474546
3	MODEL1	PARMS	logsurv	0.09646	0.94226	.	.007898656	.006999735
4	MODEL1	PARMS	logsurv	0.12070	0.90742	.	.009863328	.008753048
5	MODEL1	PARMS	logsurv	0.14937	1.16811	0.040120	.	.006966215
6	MODEL1	PARMS	logsurv	0.15626	1.38878	.	.	.005652541
7	MODEL1	PARMS	logsurv	0.16522	1.40853	.	.006092336	.
8	MODEL1	PARMS	logsurv	0.16611	1.02711	0.077905	.	.009447119
9	MODEL1	PARMS	logsurv	0.16676	1.39156	0.004029	.006134480	.
10	MODEL1	PARMS	logsurv	0.19002	1.69638	.	.	.
11	MODEL1	PARMS	logsurv	0.19178	1.71206	-0.004216	.	.
12	MODEL1	PARMS	logsurv	0.20640	1.54547	.	.	.008567887
13	MODEL1	PARMS	logsurv	0.20922	1.33433	0.050447	.009172350	.
14	MODEL1	PARMS	logsurv	0.22258	1.59881	.	.009603518	.
15	MODEL1	PARMS	logsurv	0.25929	1.86399	0.059163	.	.

Obs	liver	logsurv	_IN_	_P_	_EDF_	_RSQ_	_AIC_	_SBC_
1	.	-1	3	4	50	0.97235	-326.667	-318.711
2	0.00193	-1	4	5	49	0.97237	-324.711	-314.766
3	0.08185	-1	3	4	50	0.88290	-248.730	-240.774
4	.	-1	2	3	51	0.81297	-225.446	-219.479
5	0.09796	-1	3	4	50	0.71919	-201.498	-193.542
6	0.13901	-1	2	3	51	0.68653	-197.558	-191.591
7	0.15025	-1	2	3	51	0.64958	-191.539	-185.572
8	.	-1	2	3	51	0.64579	-190.959	-184.992
9	0.14698	-1	3	4	50	0.64999	-189.602	-181.646
10	0.18575	-1	1	2	52	0.52737	-177.385	-173.407
11	0.18893	-1	2	3	51	0.52783	-175.437	-169.470
12	.	-1	1	2	52	0.44239	-168.455	-164.477
13	.	-1	2	3	51	0.43805	-166.037	-160.070
14	.	-1	1	2	52	0.35152	-160.303	-156.325
15	.	-1	1	2	52	0.12000	-143.817	-139.839

```
proc reg;
model logsurv=blood prog enz liver/selection =cp;   run;
```

C(p) Selection Method

Number in Model	C(p)	R-Square	Variables in Model
3	3.0393	0.9723	blood prog enz
4	5.0000	0.9724	blood prog enz liver
3	161.6625	0.8829	prog enz liver
2	283.6695	0.8130	prog enz
3	451.9870	0.7192	blood enz liver
2	507.8964	0.6865	enz liver
2	573.4372	0.6496	prog liver
3	574.7100	0.6500	blood prog liver
2	580.1453	0.6458	blood enz
1	788.1481	0.5274	liver
2	789.3404	0.5278	blood liver
1	938.8651	0.4424	enz
2	948.5500	0.4381	blood prog
1	1100.012	0.3515	prog
1	1510.590	0.1200	blood

```
proc reg;
model logsurv=blood prog enz liver/selection= forward; run;
```

Forward Selection: Step 1

Variable liver Entered: R-Square = 0.5274 and C(p) = 788.1481

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.09514	2.09514	58.02	<.0001
Error	52	1.87763	0.03611		
Corrected Total	53	3.97277			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.69638	0.07174	20.18803	559.10	<.0001
liver	0.18575	0.02439	2.09514	58.02	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable enz Entered: R-Square = 0.6865 and C(p) = 507.8964

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.72744	1.36372	55.85	<.0001
Error	51	1.24533	0.02442		
Corrected Total	53	3.97277			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.38878	0.08447	6.60079	270.32	<.0001

```
      enz            0.00565      0.00111      0.63230    25.89  <.0001
      liver          0.13901      0.02206      0.96994    39.72  <.0001
                  Bounds on condition number: 1.2098, 4.8392
--------------------------------------------------------------------------------
                          Forward Selection: Step 3
         Variable prog Entered: R-Square = 0.8829 and C(p) = 161.6625
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     3        3.50756        1.16919     125.66    <.0001
Error                    50        0.46521        0.00930
Corrected Total          53        3.97277
                    Parameter     Standard
      Variable      Estimate        Error   Type II SS  F Value  Pr > F
      Intercept      0.94226      0.07139      1.62089   174.21  <.0001
      prog           0.00790   0.00086260      0.78012    83.85  <.0001
      enz            0.00700   0.00070128      0.92694    99.63  <.0001
      liver          0.08185      0.01498      0.27780    29.86  <.0001
                  Bounds on condition number: 1.4642, 11.822
--------------------------------------------------------------------------------
                          Forward Selection: Step 4
         Variable blood Entered: R-Square = 0.9724 and C(p) = 5.0000
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     4        3.86300        0.96575     431.10    <.0001
Error                    49        0.10977        0.00224
Corrected Total          53        3.97277
                    Parameter     Standard
      Variable      Estimate        Error   Type II SS  F Value  Pr > F
      Intercept      0.48876      0.05023      0.21208    94.67  <.0001
      blood          0.06852      0.00544      0.35544   158.66  <.0001
      prog           0.00925   0.00043673      1.00583   448.99  <.0001
      enz            0.00947   0.00039625      1.28075   571.71  <.0001
      liver          0.00193      0.00971   0.00008809     0.04  0.8436
                  Bounds on condition number: 2.5553, 29.286
                All variables have been entered into the model.
                        Summary of Forward Selection
       Variable    Number    Partial     Model
Step   Entered     Vars In   R-Square   R-Square    C(p)     F Value   Pr > F
  1    liver          1      0.5274     0.5274    788.148     58.02   <.0001
  2    enz            2      0.1592     0.6865    507.896     25.89   <.0001
  3    prog           3      0.1964     0.8829    161.662     83.85   <.0001
  4    blood          4      0.0895     0.9724     5.0000    158.66   <.0001
--------------------------------------------------------------------------------
```

```
proc reg;
model logsurv=blood prog enz liver/selection = backward;  run;

                        Backward Elimination: Step 0
          All Variables Entered: R-Square = 0.9724 and C(p) = 5.0000

                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F
       Intercept     0.48876      0.05023      0.21208    94.67  <.0001
       blood         0.06852      0.00544      0.35544   158.66  <.0001
       prog          0.00925   0.00043673      1.00583   448.99  <.0001
       enz           0.00947   0.00039625      1.28075   571.71  <.0001
       liver         0.00193      0.00971   0.00008809     0.04  0.8436
                  Bounds on condition number: 2.5553, 29.286
--------------------------------------------------------------------------------
                        Backward Elimination: Step 1
         Variable liver Removed: R-Square = 0.9723 and C(p) = 3.0393
                            Analysis of Variance
                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F
       Intercept     0.48362      0.04263      0.28279   128.71  <.0001
       blood         0.06923      0.00408      0.63315   288.17  <.0001
       prog          0.00929   0.00038250      1.29732   590.45  <.0001
       enz           0.00952   0.00030641      2.12263   966.07  <.0001
                  Bounds on condition number: 1.0308, 9.1864
--------------------------------------------------------------------------------
      All variables left in the model are significant at the 0.1000 level.
                        Summary of Backward Elimination
        Variable    Number    Partial     Model
 Step   Removed     Vars In   R-Square   R-Square    C(p)     F Value   Pr > F
   1    liver          3      0.0000     0.9723     3.0393      0.04   0.8436
--------------------------------------------------------------------------------

proc reg;
model logsurv=blood prog enz liver/selection = stepwise; run;

                         Stepwise Selection: Step 1
          Variable liver Entered: R-Square = 0.5274 and C(p) = 788.1481
                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F
       Intercept     1.69638      0.07174     20.18803   559.10  <.0001
       liver         0.18575      0.02439      2.09514    58.02  <.0001
                        Bounds on condition number: 1, 1
--------------------------------------------------------------------------------
                         Stepwise Selection: Step 2
          Variable enz Entered: R-Square = 0.6865 and C(p) = 507.8964
                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F
       Intercept     1.38878      0.08447      6.60079   270.32  <.0001
       enz           0.00565      0.00111      0.63230    25.89  <.0001
       liver         0.13901      0.02206      0.96994    39.72  <.0001
```

```
                Bounds on condition number: 1.2098, 4.8392
--------------------------------------------------------------------------------
                        Stepwise Selection: Step 3
        Variable prog Entered: R-Square = 0.8829 and C(p) = 161.6625
                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F
      Intercept     0.94226      0.07139      1.62089   174.21  <.0001
      prog          0.00790   0.00086260      0.78012    83.85  <.0001
      enz           0.00700   0.00070128      0.92694    99.63  <.0001
      liver         0.08185      0.01498      0.27780    29.86  <.0001
                Bounds on condition number: 1.4642, 11.822
--------------------------------------------------------------------------------
                        Stepwise Selection: Step 4
        Variable blood Entered: R-Square = 0.9724 and C(p) = 5.0000
                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F
      Intercept     0.48876      0.05023      0.21208    94.67  <.0001
      blood         0.06852      0.00544      0.35544   158.66  <.0001
      prog          0.00925   0.00043673      1.00583   448.99  <.0001
      enz           0.00947   0.00039625      1.28075   571.71  <.0001
      liver         0.00193      0.00971   0.00008809     0.04  0.8436
--------------------------------------------------------------------------------
                        Stepwise Selection: Step 5
        Variable liver Removed: R-Square = 0.9723 and C(p) = 3.0393
                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F
      Intercept     0.48362      0.04263      0.28279   128.71  <.0001
      blood         0.06923      0.00408      0.63315   288.17  <.0001
      prog          0.00929   0.00038250      1.29732   590.45  <.0001
      enz           0.00952   0.00030641      2.12263   966.07  <.0001
                Bounds on condition number: 1.0308, 9.1864

     All variables left in the model are significant at the 0.1500 level.
 No other variable met the 0.1500 significance level for entry into the model.
                      Summary of Stepwise Selection
       Variable  Variable  Number  Partial   Model
  Step Entered   Removed   Vars In R-Square R-Square  C(p)   F Value Pr > F
    1  liver                   1    0.5274   0.5274  788.148   58.02 <.0001
    2  enz                     2    0.1592   0.6865  507.896   25.89 <.0001
    3  prog                    3    0.1964   0.8829  161.662   83.85 <.0001
    4  blood                   4    0.0895   0.9724   5.0000  158.66 <.0001
    5            liver         3    0.0000   0.9723   3.0393    0.04 0.8436

**Regression run with no selection
                                    Sum of           Mean
 Source                   DF       Squares         Square    F Value    Pr > F
 Model                     4       3.86300        0.96575     431.10    <.0001
 Error                    49       0.10977        0.00224
 Corrected Total          53       3.97277
             Root MSE              0.04733    R-Square     0.9724
             Dependent Mean        2.20614    Adj R-Sq     0.9701
```

```
                          Parameter       Standard
      Variable     DF       Estimate          Error    t Value    Pr > |t|
      Intercept     1        0.48876        0.05023       9.73      <.0001
      blood         1        0.06852        0.00544      12.60      <.0001
      prog          1        0.00925     0.00043673      21.19      <.0001
      enz           1        0.00947     0.00039625      23.91      <.0001
      liver         1        0.00193        0.00971       0.20      0.8436
SAS's defaults for entering and leaving the models.
   SLE=value      specifies the significance level for entry into the model
   used in the FORWARD and STEPWISE methods. The defaults are 0.50 for
   FORWARD and 0.15 for STEPWISE.
SLS=value    specifies the significance level for staying in the model for
the BACKWARD and STEPWISE methods. The defaults are 0.10 for BACKWARD and
0.15 for STEPWISE.
```

Following is from the SAS documentation:

Criteria Used in Model-Selection Methods: When many significance tests are performed, each at a level of, for example, 5 percent, the overall probability of rejecting at least one true null hypothesis is much larger than 5 percent. If you want to guard against including any variables that do not contribute to the predictive power of the model in the population, you should specify a very small SLE= significance level for the FORWARD and STEPWISE methods and a very small SLS= significance level for the BACKWARD and STEPWISE methods.

In most applications, many of the variables considered have some predictive power, however small. If you want to choose the model that provides the best prediction using the sample estimates, you need only to guard against estimating more parameters than can be reliably estimated with the given sample size, so you should use a moderate significance level, perhaps in the range of 10 percent to 25 percent.

In addition to R2, the Cp statistic is displayed for each model generated in the model-selection methods. The Cp statistic is proposed by Mallows (1973) as a criterion for selecting a model. It is a measure of total squared error defined as $Cp = [(SSEp)/(s2)] - (N - 2p)$*, where s2 is the MSE for the full model, and SSEp is the sum-of-squares error for a model with p parameters including the intercept, if any. If Cp is plotted against p, Mallows recommends the model where Cp first approaches p. When the right model is chosen, the parameter estimates are unbiased, and this is reflected in Cp near p. For further discussion, refer to Daniel and Wood (1980).*

The Adjusted R2 statistic is an alternative to R2 that is adjusted for the number of parameters in the model. The adjusted R2 statistic is calculated as $ADJRSQ = 1 - [((n - i)(1 - R2))/(n - p)]$*, where n is the number of observations used in fitting the model, and i is an indicator variable that is 1 if the model includes an intercept, and 0 otherwise.*

Variable Selection in R. It takes a little digging to figure out how to do this. We'll use the leaps package first to do all possible subsets. With the output from that we can save and order the models by whichever criterion we want.

```
# EXAMPLE OF VARIABLE SELECTION USING R.
# FIRST PART DOES ALL POSSIBLE SUBSETS
# THIS REQUIRES THE LEAPS PACKAGE WHICH MUST
# BE INSTALLED FIRST.
library(leaps)
surgical<-read.table("f:/s505/data/surg.dat",na.strings=".")
names(surgical) <- c("blood", "prog", "enz", "liver", "surv", "logsurv")
attach(surgical)
subsets<-regsubsets(logsurv~blood+prog+enz+liver,nbest=10,data=surgical)
       #the nbest = 10 says to get up to 10 best models for each
       # possible number of variables. Since 10 > 4 this will do
       # all possible subsets
summary(subsets)
info<-summary(subsets)
str(info)     #shows you what is in info
info$which    # shows what is in each model
whichvm<-1*info$which   # converts info on which variables to a 0/1
                        # matrix form (in general multiplying a logical
                        # true false/matrix times a number will convert it

p<-rowSums(whichvm)     #p = number of variables in  model
n<-subsets$nn           #total samples size. This was part of what is stored in subsets
                        #these next set extract measures from info
rsquared<-info$rsq
 sse<-info$rss
 adjR2<-info$adjr2
 Cp<-info$cp
  AIC<-n*log(sse/n) +2*p
subsetinfo<-cbind(info$which,rsquared,sse,adjR2,Cp,AIC)
subsetinfo
        # you can sort this dataframe by the various measures and list
        # Here we sort by adjusted R2
sortadjr2<-subsetinfo[order(adjR2),]
sortadjr2

> # EXAMPLE OF VARIABLE SELECTION USING R.
> # FIRST PART DOES ALL POSSIBLE SUBSETS
> # THIS REQUIRES THE LEAPS PACKAGE WHICH MUST
> # BE INSTALLED FIRST.
> library(leaps)
> surgical<-read.table("f:/s505/data/surg.dat",na.strings=".")
> names(surgical) <- c("blood", "prog", "enz", "liver", "surv", "logsurv")
> attach(surgical)

> subsets<-regsubsets(logsurv~blood+prog+enz+liver,nbest=10,data=surgical)
>        #the nbest = 10 says to get up to 10 best models for each
>        # possible number of variables. Since 10 > 4 this will do
>        # all possible subsets
> summary(subsets)
```

```
Subset selection object
Call: regsubsets.formula(logsurv ~ blood + prog + enz + liver, nbest = 10,
    data = surgical)
4 Variables  (and intercept)
      Forced in Forced out
blood     FALSE      FALSE
prog      FALSE      FALSE
enz       FALSE      FALSE
liver     FALSE      FALSE
10 subsets of each size up to 4
Selection Algorithm: exhaustive
         blood prog enz liver
1  ( 1 ) " "   " "  " " "*"
1  ( 2 ) " "   " "  "*" " "
1  ( 3 ) " "   "*"  " " " "
1  ( 4 ) "*"   " "  " " " "
2  ( 1 ) " "   "*"  "*" " "
2  ( 2 ) " "   " "  "*" "*"
2  ( 3 ) " "   "*"  " " "*"
2  ( 4 ) "*"   " "  "*" " "
2  ( 5 ) "*"   " "  " " "*"
2  ( 6 ) "*"   "*"  " " " "
3  ( 1 ) "*"   "*"  "*" " "
3  ( 2 ) " "   "*"  "*" "*"
3  ( 3 ) "*"   " "  "*" "*"
3  ( 4 ) "*"   "*"  " " "*"
4  ( 1 ) "*"   "*"  "*" "*"
> info<-summary(subsets)
> str(info)     #shows you what is in info
List of 8
 $ which : logi [1:15, 1:5] TRUE TRUE TRUE TRUE TRUE TRUE ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:15] "1" "1" "1" "1" ...
  .. ..$ : chr [1:5] "(Intercept)" "blood" "prog" "enz" ...
 $ rsq   : num [1:15] 0.527 0.442 0.352 0.12 0.813 ...
 $ rss   : num [1:15] 1.878 2.215 2.576 3.496 0.743 ...
 $ adjr2 : num [1:15] 0.518 0.432 0.339 0.103 0.806 ...
 $ cp    : num [1:15] 788 939 1100 1511 284 ...
 $ bic   : num [1:15] -32.49 -23.56 -15.41 1.08 -78.56 ...
 $ outmat: chr [1:15, 1:4] " " " " " " " " "*" ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:15] "1  ( 1 )" "1  ( 2 )" "1  ( 3 )" "1  ( 4 )" ...
  .. ..$ : chr [1:4] "blood" "prog" "enz" "liver"
 $ obj   :List of 28
  ..$ np       : int 5
  ..$ nrbar    : int 10
  ..$ d        : num [1:5] 54 60.7 13079.9 18918.7 75.7
  ..$ rbar     : num [1:10] 2.74 63.24 77.11 5.78 5.83 ...
  ..$ thetab   : num [1:5] 2.20614 0.18575 0.00609 0.007 0.06852
  ..$ first    : int 2
  ..$ last     : int 5
  ..$ vorder   : int [1:5] 1 5 3 4 2
  ..$ tol      : num [1:5] 3.67e-09 1.77e-08 2.99e-07 3.72e-07 2.71e-08
  ..$ rss      : num [1:5] 3.973 1.878 1.392 0.465 0.11
  ..$ bound    : num [1:5] 1e+35 1e+35 1e+35 1e+35 1e+35
```

```
  ..$ nvmax    : int 5
  ..$ ress     : num [1:5, 1:10] 3.973 1.878 0.743 0.11 0.11 ...
  ..$ ir       : int 5
  ..$ nbest    : int 10
  ..$ lopt     : int [1:15, 1:10] 1 1 5 1 4 3 1 2 3 4 ...
  ..$ il       : int 15
  ..$ ier      : int 0
  ..$ xnames   : chr [1:5] "(Intercept)" "blood" "prog" "enz" ...
  ..$ method   : chr "exhaustive"
  ..$ force.in : Named logi [1:5] TRUE FALSE FALSE FALSE FALSE
  .. ..- attr(*, "names")= chr [1:5] "" "blood" "prog" "enz" ...
  ..$ force.out: Named logi [1:5] FALSE FALSE FALSE FALSE FALSE
  .. ..- attr(*, "names")= chr [1:5] "" "blood" "prog" "enz" ...
  ..$ sserr    : num 0.11
  ..$ intercept: logi TRUE
  ..$ lindep   : logi [1:5] FALSE FALSE FALSE FALSE FALSE
  ..$ nullrss  : num 3.97
  ..$ nn       : int 54
  ..$ call     : language regsubsets.formula(logsurv ~ blood + prog + enz + liver, nbest = 10, data = surgical)
  ..- attr(*, "class")= chr "regsubsets"
 - attr(*, "class")= chr "summary.regsubsets"
> info$which    # shows what is in each model
  (Intercept) blood  prog   enz liver
1        TRUE FALSE FALSE FALSE  TRUE
1        TRUE FALSE FALSE  TRUE FALSE
1        TRUE FALSE  TRUE FALSE FALSE
1        TRUE  TRUE FALSE FALSE FALSE
2        TRUE FALSE  TRUE  TRUE FALSE
2        TRUE FALSE FALSE  TRUE  TRUE
2        TRUE FALSE  TRUE FALSE  TRUE
2        TRUE  TRUE FALSE  TRUE FALSE
2        TRUE  TRUE FALSE FALSE  TRUE
2        TRUE  TRUE  TRUE FALSE FALSE
3        TRUE  TRUE  TRUE  TRUE FALSE
3        TRUE FALSE  TRUE  TRUE  TRUE
3        TRUE  TRUE FALSE  TRUE  TRUE
3        TRUE  TRUE  TRUE FALSE  TRUE
4        TRUE  TRUE  TRUE  TRUE  TRUE
> whichvm<-1*info$which   # converts info on which variables to a 0/1
>                         # matrix form (in general multiplying a logical
>                         # true false/matrix times a number will convert it
>
> p<-rowSums(whichvm)     #p = number of variables in  model
> n<-subsets$nn           #total samples size. This was part of what is stored in subsets
>                         #these next set extract measures from info
> rsquared<-info$rsq
>  sse<-info$rss
>  adjR2<-info$adjr2
>  Cp<-info$cp
>   AIC<-n*log(sse/n) +2*p
> subsetinfo<-cbind(info$which,rsquared,sse,adjR2,Cp,AIC)
> subsetinfo
  (Intercept) blood prog enz liver  rsquared       sse     adjR2          Cp       AIC
1           1     0    0   0     1 0.5273749 1.8776320 0.5182859  788.148136 -177.3845
1           1     0    0   1     0 0.4423867 2.2152705 0.4316634  938.865126 -168.4549
```

```
1           1     0    1   0     0 0.3515172 2.5762746 0.3390464 1100.012198 -160.3025
1           1     1    0   0     0 0.1199959 3.4960560 0.1030727 1510.589506 -143.8168
2           1     0    1   1     0 0.8129742 0.7430109 0.8056399  283.669453 -225.4455
2           1     0    0   1     1 0.6865344 1.2453274 0.6742416  507.896397 -197.5576
2           1     0    1   0     1 0.6495765 1.3921529 0.6358344  573.437188 -191.5392
2           1     1    0   1     0 0.6457938 1.4071804 0.6319034  580.145259 -190.9594
2           1     1    0   0     1 0.5278304 1.8758225 0.5093139  789.340380 -175.4366
2           1     1    1   0     0 0.4380533 2.2324863 0.4160162  948.550025 -166.0369
3           1     1    1   1     0 0.9723471 0.1098586 0.9706879    3.039323 -326.6674
3           1     0    1   1     1 0.8829008 0.4652086 0.8758748  161.662489 -248.7297
3           1     1    0   1     1 0.7191890 1.1155980 0.7023404  451.987024 -201.4980
3           1     1    1   0     1 0.6499865 1.3905238 0.6289857  574.709962 -189.6024
4           1     1    1   1     1 0.9723693 0.1097705 0.9701137    5.000000 -324.7107
>         # you can sort this dataframe by the various measures and list
>         # Here we sort by adjusted R2
> sortadjr2<-subsetinfo[order(adjR2),]
> sortadjr2
  (Intercept) blood prog enz liver  rsquared       sse     adjR2          Cp       AIC
1           1     1    0   0     0 0.1199959 3.4960560 0.1030727 1510.589506 -143.8168
1           1     0    1   0     0 0.3515172 2.5762746 0.3390464 1100.012198 -160.3025
2           1     1    1   0     0 0.4380533 2.2324863 0.4160162  948.550025 -166.0369
1           1     0    0   1     0 0.4423867 2.2152705 0.4316634  938.865126 -168.4549
2           1     1    0   0     1 0.5278304 1.8758225 0.5093139  789.340380 -175.4366
1           1     0    0   0     1 0.5273749 1.8776320 0.5182859  788.148136 -177.3845
3           1     1    1   0     1 0.6499865 1.3905238 0.6289857  574.709962 -189.6024
2           1     1    0   1     0 0.6457938 1.4071804 0.6319034  580.145259 -190.9594
2           1     0    1   0     1 0.6495765 1.3921529 0.6358344  573.437188 -191.5392
2           1     0    0   1     1 0.6865344 1.2453274 0.6742416  507.896397 -197.5576
3           1     1    0   1     1 0.7191890 1.1155980 0.7023404  451.987024 -201.4980
2           1     0    1   1     0 0.8129742 0.7430109 0.8056399  283.669453 -225.4455
3           1     0    1   1     1 0.8829008 0.4652086 0.8758748  161.662489 -248.7297
4           1     1    1   1     1 0.9723693 0.1097705 0.9701137    5.000000 -324.7107
3           1     1    1   1     0 0.9723471 0.1098586 0.9706879    3.039323 -326.6674
```

Doing Forward, backward and stepwise selection in R. There is a step function but the values of AIC are odd and it doesn't seem to do what it should. We'll use stepAIC which is part of the MASS package.

```
library(MASS)
null<-lm(logsurv~1,data=surgical)
full<-lm(logsurv~blood+prog+enz+liver,data=surgical)
 stepAIC(null,scope=list(lower=null,upper=full) ,direction="forward")  #forward selection
 stepAIC(full,direction="backward")   #backward selection
 stepAIC(full)                       # stepwise selection

> null<-lm(logsurv~1,data=surgical)
> full<-lm(logsurv~blood+prog+enz+liver,data=surgical)
>  stepAIC(null,scope=list(lower=null,upper=full) ,direction="forward")  #forward selection
Start:  AIC=-138.91
logsurv ~ 1

        Df Sum of Sq    RSS     AIC
+ liver  1   2.09514 1.8776 -177.38
+ enz    1   1.75750 2.2153 -168.46
```

```
+ prog   1   1.39650 2.5763 -160.30
+ blood  1   0.47672 3.4961 -143.82
<none>               3.9728 -138.91

Step:  AIC=-177.38
logsurv ~ liver

        Df Sum of Sq    RSS     AIC
+ enz    1   0.63230 1.2453 -197.56
+ prog   1   0.48548 1.3921 -191.54
<none>               1.8776 -177.38
+ blood  1   0.00181 1.8758 -175.44

Step:  AIC=-197.56
logsurv ~ liver + enz

        Df Sum of Sq     RSS     AIC
+ prog   1   0.78012 0.46521 -248.73
+ blood  1   0.12973 1.11560 -201.50
<none>               1.24533 -197.56

Step:  AIC=-248.73
logsurv ~ liver + enz + prog

        Df Sum of Sq     RSS     AIC
+ blood  1   0.35544 0.10977 -324.71
<none>               0.46521 -248.73

Step:  AIC=-324.71
logsurv ~ liver + enz + prog + blood


Call:
lm(formula = logsurv ~ liver + enz + prog + blood, data = surgical)

Coefficients:
(Intercept)        liver          enz         prog        blood
   0.488756     0.001925     0.009475     0.009254     0.068520

>  stepAIC(full,direction="backward")   #backward selection
Start:  AIC=-324.71
logsurv ~ blood + prog + enz + liver

        Df Sum of Sq     RSS     AIC
- liver  1   0.00009 0.10986 -326.67
<none>               0.10977 -324.71
- blood  1   0.35544 0.46521 -248.73
- prog   1   1.00583 1.11560 -201.50
- enz    1   1.28075 1.39052 -189.60

Step:  AIC=-326.67
logsurv ~ blood + prog + enz

        Df Sum of Sq     RSS     AIC
<none>               0.10986 -326.67
```

```
- blood  1   0.63315 0.74301 -225.45
- prog   1   1.29732 1.40718 -190.96
- enz    1   2.12263 2.23249 -166.04

Call:
lm(formula = logsurv ~ blood + prog + enz, data = surgical)

Coefficients:
(Intercept)        blood         prog          enz
   0.483621     0.069225     0.009295     0.009524

>  stepAIC(full)                      # stepwise selection
Start:  AIC=-324.71
logsurv ~ blood + prog + enz + liver

        Df Sum of Sq     RSS     AIC
- liver  1   0.00009 0.10986 -326.67
<none>               0.10977 -324.71
- blood  1   0.35544 0.46521 -248.73
- prog   1   1.00583 1.11560 -201.50
- enz    1   1.28075 1.39052 -189.60

Step:  AIC=-326.67
logsurv ~ blood + prog + enz

        Df Sum of Sq     RSS     AIC
<none>               0.10986 -326.67
- blood  1   0.63315 0.74301 -225.45
- prog   1   1.29732 1.40718 -190.96
- enz    1   2.12263 2.23249 -166.04

Call:
lm(formula = logsurv ~ blood + prog + enz, data = surgical)

Coefficients:
(Intercept)        blood         prog          enz
   0.483621     0.069225     0.009295     0.009524
```

Forcing variables into the model. In either SAS or R you can force certain variables into the model and then do variable selection (forward, backward, stepwise) with the remaining variables. In SAS as an option in the model statement you use *include* $= j$ where j is an integer. This will include the first j variables in the list (so you need to order things in the right way). In R using step AIC you can set the inital object and/or lower model (used in scope) to be one containing variables you want to force in.

10 Additional topics on residual analysis and diagnostics.

Even if model assumptions are met exactly, the residuals have unequal variance and are correlated (this is because the residual e_i is not that same as ϵ_i. In fact The variance-covariance matrix of the residuals

$\sigma^2(\underline{e}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, is called the **hat matrix.**

Recall that $\sigma^2(\boldsymbol{\epsilon}) = \sigma^2 I_n$.

If the elements of H are small then $Cov(\boldsymbol{\epsilon}) \approx \sigma^2\mathbf{I}_n$, so residuals behave similar to the ϵ's. This usually works fine for moderate to large sample sizes (although this depends on how many predictors there are).

Leverage value: $h_{ii} = ith$ diagonal element of $\mathbf{H}$. Large values of h_{ii} indicate an observation is an outlier in the X space.

Can be shown that $\bar{h}$ = average of the h_{ii} equals p/n. One criteria takes values with high leverage have $h_{ii} > 2p/n$ as potential outliers.

Studentized residual:

$$r_i = \frac{e_i}{[MSE(1 - h_{ii})]^{1/2}}.$$

This tries to account for the unequal variance of the residuals and is generally preferred over use of unstudentized residuals for residual plots.

Studentized deleted residuals: (Delete ith case, get residual and then "stu-

dentize"). This simplifies to:

$$t_i = \frac{Y_i - \hat{Y}_{i(i)}}{s\{d_i\}}$$

where $s^2\{d_i\} = MSE_{(i)}(1 + \mathbf{X}_i'(X_{(i)}'X_{(i)})^{-1}\mathbf{X}_i)$ estimates the variance of $\hat{Y}_{i(i)}$ = fitted value for case i based on regression model without using case i.

$_{(i)}$ indicates the ith case has been deleted. $\mathbf{X}_i'$ is the row of X corresponding to the ith observation.

One rule: absolute(studentized deleted residual) $> t(1-\alpha/2n, n-p-1)$ indicates outlier. Sometimes same rule used for studentized residuals.

$$\text{COOK'S DISTANCE: } D_i = \frac{\sum_j(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p * MSE}$$

Measure of influence of the ith case on the overall fit.

One rule: If Cook's distance $> F(.5, p, n-p)$, consider an outlier. With Cook's distance between $F(.2, p, n-p)$ and $F(.5, p, n-p)$ look a little closer.

- There are various computational shortcuts for these measures, as described in text.
- There are many other diagnostic measures/tools including DFFITS, DFBETAS, Variance inflation factors, added-variable plots,
- SAS 9.3 will automatically do some plots involving these diagnostic measures when you run proc reg.

Example

```
title 'Illustrating multiple with house price data';
options linesize=70 pagesize=60 nodate;
data values;
infile 'house.dat';
input PRICE SQFT AGE FEATS NE CUST COR TAX;
original=_n_;  /* gives orginal case number */
run;
proc reg;
id sqft tax;
```

```
model price =sqft tax/ p r influence;
output out=result p=yhat r=resid h=hvalue student=stres
rstudent=stdelres cookd=cookd;     run;
/* first name is sas convention, after = is name of your
choosing */

proc print data=result; run;
proc gplot data=result;
plot resid*yhat;
plot stres*yhat;
run;
/* Below is a way to sort and get ordered values of the statistic*/
%macro sortp(vars);
proc sort data=result;
by &vars; proc print;
var original &vars;
run;
%mend;
%sortp(stres); /* sort on studentized residual*/
%sortp(stdelres);
%sortp(hvalue);
%sortp(cookd);
```

```
                        Output Statistics
                           Dep Var Predicted    Std Error
     Obs SQFT     TAX          PRICE     Value Mean Predict  Residual
       1     2650     1639       2050      1933      52.1014  117.4048
       2     2600     1088       2080      1523      38.6728  557.0325

                         Std Error  Student                    Cook's
     Obs SQFT     TAX       Residual Residual   -2-1 0 1 2           D
       1     2650     1639     166.1    0.707 |      |*     |    0.016
       2     2600     1088     169.8    3.281 |      |******|    0.186

                                       Hat Diag       Cov
       Obs SQFT     TAX       RStudent         H     Ratio    DFFITS
         1     2650     1639    0.7050    0.0895    1.1144    0.2211
         2     2600     1088    3.4489    0.0493    0.7801    0.7857

 Obs  PRICE  SQFT  AGE  FEATS  NE  CUST  COR   TAX    yhat      resid
   1   2050  2650   13    7     1    1    0   1639  1932.60   117.405
   2   2080  2600    .    4     1    1    0   1088  1522.97   557.032

Sorted on studentized residual
                     Obs    original      stres
                      10        97         .
                      11        79      -4.10774
                      12        50      -2.14396

                     116         2       3.28129
                     117        89       3.56268
```

```
sorted on
                      Obs    original    stdelres
                       10        97        .
                       11        79      -4.46629
                       12        50      -2.18240

                      116         2       3.44889
                      117        89       3.78392

sorted on leverage value
                       Obs    original     hvalue
                       115        7       0.11706
                       116        5       0.12571
                       117       79       0.30461

sorted on cooks distance

                       Obs    original     cookd
                       115        3       0.15962
                       116        2       0.18626
                       117       79       2.46375

** Values to compare things to ***
options ls=80 nodate;
data a; n= 107; p=3;
t=tinv(1-(.05/214),n-p-1);
cleverage = 2*p/n;   f20 = finv(.20,p,n-p);  f50 = finv(.50,p,n-p);
proc print;  run;
  Obs  n  p    t    cleverage    f20       f50
   1  107 3 3.61462  0.056075  0.33507    0.79386
```

Working the example in R. Note that here I used data without missing values in price, sqft or tax; so the case numbers differ. In particular case 79 in the SAS analysis is case 72 here.

```
data<-read.table("f:/s505/data/housenm.dat",sep=",",na.strings=".")
  #housenm.dat got rid of cases with price, sqft or tax missing. data is comma separated.
attach(data)
price <- V1; sqft<-V2; age<-V3; feats<-V4
ne<- V5; cust<-V6; cor<-V7; tax<- V8
regout<-lm(price ~ sqft+ tax, na.action=na.exclude)
 summary(regout)
xmat<-model.matrix(regout)   $ gets the X matrix for the model
resid<-residuals(regout)
stresid<-rstudent(regout)     #gets studentized residual
cooks<-cooks.distance(regout) #gets cook's distances
leverage<-hat(xmat)           #gets leverage values
n<-length(price)
```

```
obs<-seq(1,n)
sumall<-cbind(obs,price,sqft,tax,resid,stresid,cooks,leverage)
sumall
plot(obs,cooks)
cooksort<-sumall[order(cooks),]   # sort on Cook's distance
cooksort
```

```
> sumall<-cbind(obs,price,sqft,tax,resid,stresid,cooks,leverage)
> sumall
    obs price sqft  tax       resid      stresid        cooks    leverage
1     1  2050 2650 1639  117.404785  0.704989331 1.637370e-02 0.089547729
2     2  2080 2600 1088  557.032462  3.448885055 1.862559e-01 0.049336442
3     3  2150 2664 1193  535.344603  3.293871072 1.596174e-01 0.046088693
4     4  2150 2921 1635  152.487078  0.912209130 2.405843e-02 0.079694812
5     5  1999 2580 1732   16.890621  0.103257123 5.159151e-04 0.125708475
6     6  1900 2580 1534   60.593764  0.358999024 3.165551e-03 0.068093156
7     7  1800 2774 1765 -254.429458 -1.565998645 1.068848e-01 0.117060056
8     8  1560 1920 1161  154.546652  0.902108539 9.433703e-03 0.033549589

71   71   619  837  342   74.770952  0.434992617 2.164387e-03 0.032927037
72   72  1295 3750 1200 -596.403228 -4.466288194 2.463747e+00 0.304607706
73   73   975 1500  700    6.878348  0.039520283 5.481375e-06 0.010319911
74   74   939 1428  701  -11.828932 -0.068006905 1.815214e-05 0.011527391
75   75   820 1375  585  -33.965175 -0.195519189 1.786157e-04 0.013697375
76   76   780 1080  600  -10.970939 -0.063494274 3.417898e-05 0.024571271


105 105   869 1165  694  -10.984766 -0.063632369 3.688098e-05 0.026350631
106 106   766 1200  634  -79.497924 -0.459168008 1.347048e-03 0.018666779
107 107   739  970  541   18.072253  0.104834032 1.103983e-04 0.028983815

> cooksort
    obs price sqft  tax       resid      stresid        cooks    leverage
41   41   725 1140  490   -1.702618 -0.009827031 6.385757e-07 0.019268303
89   89  1109 1740  816   -2.770325 -0.015911260 8.253758e-07 0.009593614
104 104   870 1273  638    3.355554  0.019327006 1.936631e-06 0.015170696
20   20   995 1500  743   -4.112739 -0.023634455 2.030970e-06 0.010687430
73   73   975 1500  700    6.878348  0.039520283 5.481375e-06 0.010319911
92   92  1045 1630  750    8.317928  0.047774992 7.467273e-06 0.009627074


81   81  2100 2116 1209  610.915355  3.783920790 1.309123e-01 0.030013384
3     3  2150 2664 1193  535.344603  3.293871072 1.596174e-01 0.046088693
2     2  2080 2600 1088  557.032462  3.448885055 1.862559e-01 0.049336442
72   72  1295 3750 1200 -596.403228 -4.466288194 2.463747e+00 0.304607706
```

11 Non-parameteric regression

The goal of nonparametric regression is to fit a model $E(Y|\mathbf{x})$ without forcing a particular parametric functional form for the regression. For a single variable this is also useful as a diagnostic way to assess linearity and to suggest a parametric model if so desired. There are many approaches to doing this. The Lowess (or Loess) method is discussed in sections 3.10 and 11.4 of the book. Here's a quick look at what it does.

Local Linear smoothing/Loess.

There are three ingredients that go into this:

- A distance function:

  Consider a point x (can be multivariate) and define a distance function where $d(x, x_i) \geq 0$ (and $= 0$ when $x = x_i$) is the distance between x and x_i.

- Smoothing parameter. This a number s, $0 < s \leq 1$ which determines the neighborhood around the point x in that the $100s$ percent of the x's that are closest to x are included in the local neighborhood around x. If $s = 1$ all point are included in the neighborhood.

- Weighting function. Suppose there are q points in the neighborhood around x with a maximum distance $D(x)$. A weight is assigned to point i, with predictors x_i in a manner that the weight decreases as the distance $d_i = d(x, x_i)$ increases. For $s < 1$ the weight is $w_i = [1 - (d_i/D(x))^3]^3$ if $d_i < D(x)$ and $= 0$ otherwise while for $s = 1$ (or > 1 $D(x)$ is replaced by $D(x)s^{1/p}$, where p is the number of predictors in the model. (SAS uses a rescaled version of these weights which does not effect the fitted values).

For a range of x's, fit a weighted least squares of Y_i on x_i with weights as above. Typically a linear model or second order model (includes quadratic and, with multiple predictors, products) is fit. Note that the weights and points involved change with x. At each x obtain $\hat{m}(x)$ = fitted value at x from the weighted least squares fit associated with point x. This is the nonparametric estimate of $E(Y|x)$.

Other nonparametric methods

- Splines. An approach which is not fully nonparametric is to model the regression function using piecewise polynomials (splines). See proc tpspline and transreg.

- Neural Networks. See Section 13.6 of Kutner et al.

  Here

  $$E(Y_i) = g_Y(\beta_0 + \beta_1 H_{i1} + \beta_{m-1} H_{i,m-1})$$

  where $H_{ij} = g_j(\underline{x}_i' \underline{alpha}_j)$. Often the various g functions are all taken as the logistic model $g(Z) = 1/(1+e^{-Z})$. This just leads to a non-linear models with parameters α's and β's, but special fitting techniques are used due to overparameterization.

- Computer Assisted Regression Trees (CART). See page 453 of Kutner et al. for a basic introduction.

Example 1: Star Data. Y = measure of color and X = log(P) where P is a pulsation period. Each observation is a star (Kanbur and Ngeow Period-color and amplitude-color relations in classical Cepheid variables, Mon. Not. R. Astonomical Society, 2003). A plot of the data indicates two things of interest. One is possible heteroscedasticity, the other is a possible bend in the regression line. In fact, theory suggest a bend at x = 1

Nonparametric regression via Loess using SAS uusing the default smoothing parameter ($s = .5$).

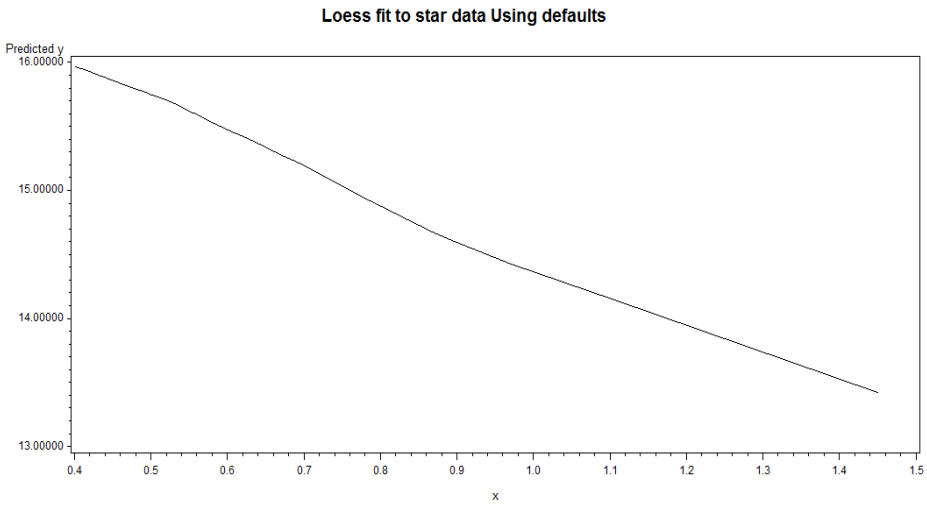


Figure 13: Star data. Nonparametric fit.

Example: Moth capture data

Here we use nonparametric regression to model the expected number of captures of moths (collected by Ring Carde and colleagues in Russia) as a function of time, where we run time sequentially over 1, 2, ... for data collected over a number of days. The first observation is at 11 a.m. on the first day of collection. One of their main interests was whether there were two humps in the model within a day. There are also various non-linear models (including trigonometric models with unknown periods) that could be tried here. The SAS code is similar to the star example and is posted. It shows how to get confidence limits for the estimated mean at each x via the clm option.

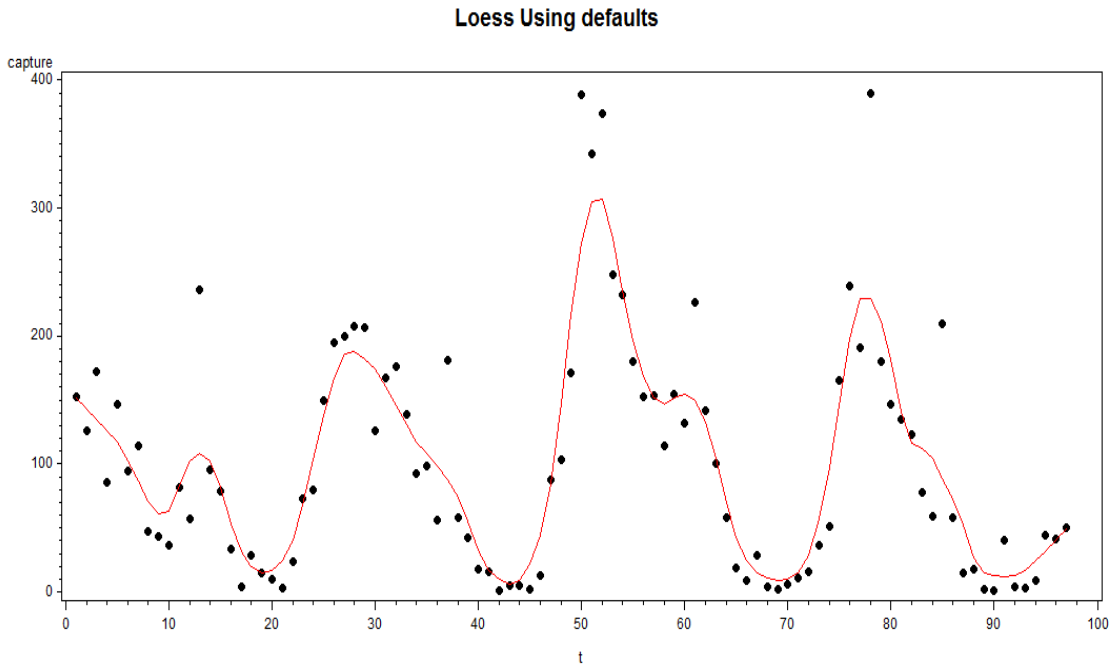


Figure 14: Nonparametric fit to Moth captures

12 Autocorrelation in the error terms

t indexes time order and we have a model $Y_t|\mathbf{x}_t = \mathbf{x}_t'\boldsymbol{\beta} + \epsilon_t$.

(We will not consider dynamic models where x can contain previous values of y which comes under the heading of time series.)

Illustrate the problem of autocorrelation of errors over time and remedies in the context of **a first order autoregressive, AR(1)** model for the errors. Assume observations are equally spaced over time.

$$\epsilon_t = \rho\epsilon_{t-1} + u_t,$$

where the u_t are assumed to be independent and identically distributed with mean 0 and variance σ_u^2. This is an autoregressive 1 model for serial/auto correlation.

If you consider stretching back infinitely in time, then it can be shown that

$\epsilon_t = \sum_{s=0}^{\infty} \rho^s u_{t-s}$ so $E(\epsilon_t) = 0$,

$V(\epsilon_t) = \sigma_\epsilon^2 = \sum_{s=0}^{\infty} \rho^{2s} V(u_{t-s}) = \sigma_u^2/(1-\rho^2), cov(\epsilon_t, \epsilon_{t+k}) = \rho^{|k|}\sigma_\epsilon^2$ and $corr(\epsilon_t, \epsilon_{t+k}) = \rho^{|k|}$.

Variance-covariance of $\mathbf{Y}$ is

$$\Sigma = \sigma^2\{\mathbf{Y}\} = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & . & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & . & \rho^{n-2} \\ . & . & . & . & . & . \\ \rho^{n-1} & \rho^{n-2} & . & . & . & 1 \end{bmatrix}.$$

Least squares estimators of $\boldsymbol{\beta}$'s are still unbiased. A plot of the residuals versus time is helpful in detecting serial correlation over time .

Durbin-Watson test for serial correlation: $H_0 : \rho = 0$.

Test statistic is

$$D = \sum_{t=1}^{n} (r_t - r_{t-1})^2 / \sum_{t=1}^{n} r_t^2,$$

critical values and p-values based on the distribution of D under the null hypothesis. The test is based on normality.

A commonly used estimate of ρ is

$$\hat{\rho} = \sum_{t=2}^{n} r_t r_{t-1} / \sum_{t=2}^{n} r_{t-1}^2,$$

which is based on a linear regression with no intercept treating the r_t as if they are the ϵ_t. There are alternate estimators.

A simple estimator that corrects for autocorrelation.

For $t = 2, \ldots, n$,

$Y_t^* = Y_t - \hat{\rho} Y_{t-1} = (\mathbf{x}_t'\boldsymbol{\beta} + \epsilon_t) - \hat{\rho}(\mathbf{x}_{t-1}'\boldsymbol{\beta} + \epsilon_{t-1}) = (\mathbf{x}_t' - \hat{\rho}\mathbf{x}_{t-1}')\boldsymbol{\beta} + \epsilon_t - \hat{\rho}\epsilon_{t-1} = \mathbf{x}_t^{*'}\boldsymbol{\beta} + \hat{u}_t$ where $\mathbf{x}_t^* = \mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1}$ and $\hat{u}_t = \epsilon_t - \hat{\rho}\epsilon_{t-1}$, which if $\hat{\rho}$ were ρ would be exactly u_t.

Cochran-Orcutt estimator. Run a multiple linear regression with Y_t^* as the response and $\mathbf{x}_t^*$ as vector of predictors but note *there is no overall intercept in there.* This is for $t = 2$ to T.

A modification of that estimator is to argue for a way to get values to use for $t = 1$. This is based on the idea of doing generalized weighted least squares based on the variance-covariance of $\mathbf{Y}$. Skipping the details this leads to $Y_1^* = (1 - \hat{\rho}^2)^{1/2} Y_1$ and $\mathbf{x}_1^* = (1 - \hat{\rho}^2)^{1/2} \mathbf{x}_1$. This leads to what are called **Yule-Walker estimates.**

- If autocorrelation is still present, the process can be repeated.

- As with our earlier weighted least squares, the standard errors that are given are generally underestimates since they do not account for uncertainty arising from estimation of ρ. This can be important and the standard errors/covariance matrix for the coefficients can be evaulated via the bootstrap.

There are other estimation techniques that can be used including maximum likelihood estimators under normality.

Serial Correlation Example: industry versus company sales.

Example from book (page 488). 20 quarters of data (5 years).

x = obseved industry sales and y = observed company sales.

X and Y are random over time.

There may be some random quantities which contribute to the error which may have influence over more than one year. This would lead to correlation in the errors. The objective is to look at the conditional behavior of Y given X, not accounting for past values of either series, rather than to model dynamic behavior. The latter concern would take us into multivariate time series. We assume that given X_i, $Y_i = \beta_0 + \beta X_i + \epsilon_i$.

Note: I used low-level plots via proc plot in SAS for easy inclusion here.

- First straight least squares is run in proc reg. The residual plot shows serial correlation. At estimate of the autocorrelation is given as is the DW test statistic.

- Then GLS/Yule-Walker estimates are obtained directly via proc reg under the AR(1) model using transformed values.

- Proc autoreg in SAS and similar procedures in other statistics packages will automatically fit these models and allow tests for higher order serial correlations and make corrections for higher order serial correlations (e.g., lag 2, lag 3, ...)

In this particular example with quarterly data a lag 4 correlation may be more suitable. Here I've shown the output using autoreg allowing a lag 4 model.

```
1 20.96 127.3
2 21.40 130.0
3 21.96 132.7
4 21.52 129.4
5 22.39 135.0
6 22.76 137.1
7 23.48 141.2
8 23.66 142.8
9 24.10 145.5
10 24.01 145.3
11 24.54 148.3
12 24.30 146.4
13 25.00 150.2
14 25.64 153.1
15 26.36 157.3
16 26.98 160.7
17 27.52 164.2
18 27.78 165.6
19 28.24 168.7
20 28.78 171.7
```

```
  A=1, B=2, etc.  Plot of COMP*IND.

COMP |
30.0 +
     |
     |                          A
     |                        A
     |                      AA
     |                    A
     |                   A
     |                 A
25.0 +               A
     |             BA
     |           B A
     |
     |        AA
     |      A
     |    AB
     |
20.0 +
     |
     -+---------+---------+---------+-
     120       140       160       180
                    IND
```

```
title 'autocorrelation example';
options pagesize=60 linesize=80;
data a;
infile 'blais.dat';
input time comp ind;
proc plot hpercent=50 vpercent=50;
plot comp*ind;
run;
title 'LEAST SQUARES WITH RESIDUALS AND DW VIA PROC REG';
proc reg;
model comp = ind/ dw;
var time;
plot residual.*time/vplots=3; /* the vplots=3 means to use size of 3 per page*/
run;
data b;
set a;
ind1 = lag(ind);  /* uses lag 1 of ind */
comp1= lag(comp);
newx1=1-.626005;
newx2=ind-(.626005*ind1);
newy=comp - (.626005*comp1);
if ind1=. then newy = sqrt(1-(.626005**2))*comp; /* deals with  observation 1*/
if ind1=. then newx1= sqrt(1-(.626005**2));
```

```
if ind1=. then newx2= sqrt(1-(.626005**2))*ind;
proc print;
run;
title 'YULE WALKER/GLS VIA REG DIRECTLY';
proc reg;
model newy=newx1 newx2/ noint;
var time;
plot residual.*time/vplots=3;
run;

title 'YULE-WALKER WITH AR4 ERRORS';
proc autoreg;
model comp = ind/nlag=4 method=yw;
run;
```

LEAST SQUARES WITH RESIDUALS AND DW VIA PROC REG

Dependent Variable: COMP

Analysis of Variance

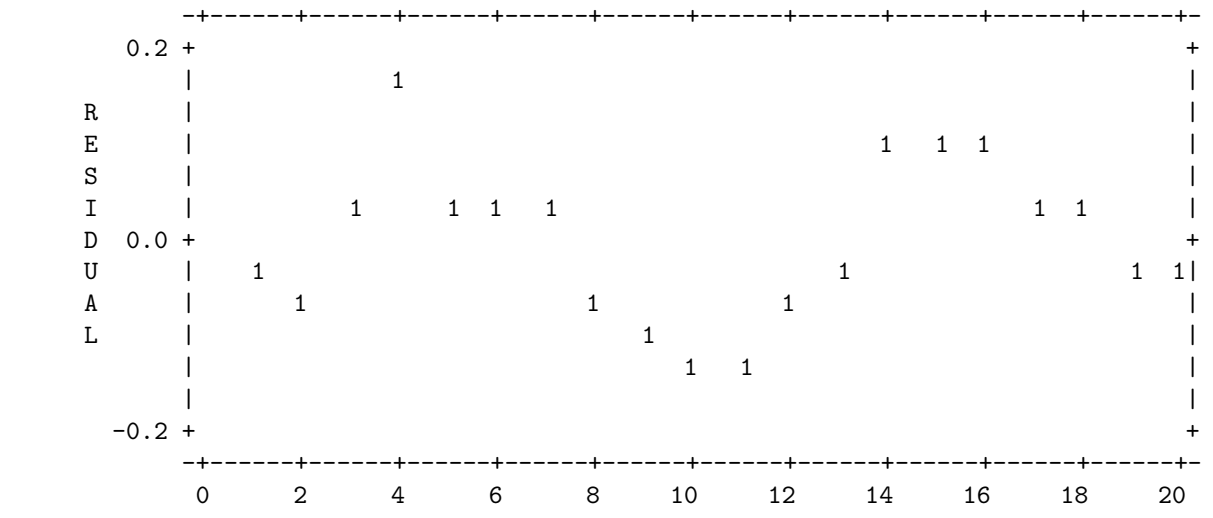
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	110.25688	110.25688	14888.144	0.0001
Error	18	0.13330	0.00741		
C Total	19	110.39018			

Root MSE	0.08606	R-square	0.9988
Dep Mean	24.56900	Adj R-sq	0.9987
C.V.	0.35026		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-1.454750	0.21414605	-6.793	0.0001
IND	1	0.176283	0.00144474	122.017	0.0001

Durbin-Watson D	0.735
(For Number of Obs.)	20
1st Order Autocorrelation	0.626



TIME

OBS	TIME	COMP	IND	IND1	COMP1	NEWX1	NEWX2	NEWY
1	1	20.96	127.3	.	.	0.77982	99.2710	16.3450
2	2	21.40	130.0	127.3	20.96	0.37399	50.3096	8.2789

etc.

YULE WALKER/GLS VIA REG DIRECTLY

NOTE: No intercept in model. R-square is redefined.

Dependent Variable: NEWY

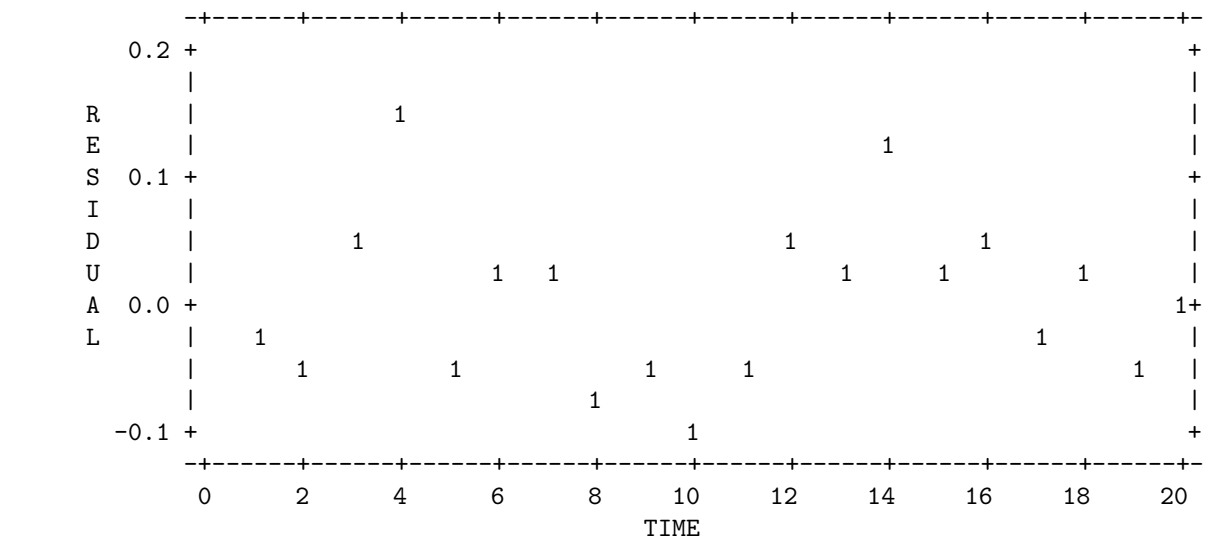
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	2004.12498	1002.06249	227702.453	0.0001
Error	18	0.07921	0.00440		
U Total	20	2004.20419			

Root MSE	0.06634	R-square	1.0000
Dep Mean	9.85875	Adj R-sq	1.0000
C.V.	0.67289		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
NEWX1	1	-1.290313	0.33959411	-3.800	0.0013
NEWX2	1	0.175142	0.00228275	76.724	0.0001



YULE-WALKER WITH AR4 ERRORS

Estimates of the Autoregressive Parameters

```
                Lag    Coefficient      Std Error        t Ratio
                  1    -0.56816055       0.249753         -2.275
                  2    -0.05758150       0.291282         -0.198
                  3     0.09080571       0.291282          0.312
                  4     0.35599058       0.249753          1.425
        Variable     DF        B Value   Std Error   t Ratio Approx Prob
        Intercept     1      -1.473603      0.2227    -6.618      0.0001
        IND           1       0.176379     0.00150   117.219      0.0001
```

MORE IN SAS

```
title 'LEAST SQUARES WITH DW TESTS VIA AUTOREG';
proc autoreg;
model comp = ind/dw=4 dwprob;
run;
```

```
  LEAST SQUARES WITH DW TESTS VIA AUTOREG                    9
                           The AUTOREG Procedure
                         Dependent Variable    comp
                         Durbin-Watson Statistics
                   Order            DW    Pr < DW    Pr > DW
                     1          0.7347     0.0002     0.9998
 NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is
               the p-value for testing negative autocorrelation.
                                       Standard                 Approx
      Variable        DF     Estimate        Error    t Value    Pr > |t|
      Intercept        1      -1.4548       0.2141      -6.79      <.0001
      ind              1       0.1763     0.001445     122.02      <.0001
```

```
title 'YULE-WALKER VIA PROC AUTOREG';
proc autoreg;
model comp = ind/nlag=1 method=yw;
run;
```

```
                           The AUTOREG Procedure
                                       Standard                 Approx
      Variable        DF     Estimate        Error    t Value    Pr > |t|
      Intercept        1      -1.2903       0.3494      -3.69      0.0018
      ind              1       0.1751     0.002349      74.56      <.0001
```

STANDARD ERRORS COMPUTED SLIGHTLY DIFFERENTLY THAN WHEN RUN THROUGH PROC REG DIRECTY

Return to the house price example. Run without case 79 which was influential.

Notice the significant change in the coefficients without case 79. MSE has also dropped from 30314 to 25642. You should repeat the diagnostics that we did earlier, but now without case 79.

Notice also that with case 79 dropped, the changes from the analysis assuming constant variance and not assuming constant variance are not as dramatic as before.

House price data with no case 79 1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12940603	6470302	252.33	<.0001
Error	103	2641155	25642		

Root MSE	160.13207	R-Square	0.8305

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.68646	55.91224	-0.19	0.8488
SQFT	1	0.41167	0.06758	6.09	<.0001
TAX	1	0.51631	0.10856	4.76	<.0001

Covariance of Estimates

Variable	Intercept	SQFT	TAX
Intercept	3126.1782108	-2.407745932	1.3711932883
SQFT	-2.407745932	0.0045673184	-0.006480621
TAX	1.3711932883	-0.006480621	0.0117854133

Consistent Covariance of Estimates

Variable	Intercept	SQFT	TAX
Intercept	4503.7683118	-4.876928617	3.8393201953
SQFT	-4.876928617	0.0082492374	-0.010200804
TAX	3.8393201953	-0.010200804	0.015857791

```
                  J     COEFF        SE     LOWER     UPPER
                  0 -10.68646 55.912237 -121.5752 100.20228
                  J     COEFF  SEROBUST    LOWERR    UPPERR
                  0 -10.68646 67.110121 -143.7836 122.41064

                  J     COEFF        SE     LOWER     UPPER
                  1   0.41167 0.0675819 0.2776372 0.5457028
                  J     COEFF  SEROBUST    LOWERR    UPPERR
                  1   0.41167 0.0908253 0.2315394 0.5918006

                  J     COEFF        SE     LOWER     UPPER
                  2 0.5163116 0.1085606 0.3010072 0.7316161
                  J     COEFF  SEROBUST    LOWERR    UPPERR
                  2 0.5163116 0.1259277 0.2665637 0.7660596

  get confidence interval on expected  price at sqft=2650,tax=1639
                               YHAT    SEYHAT       CML       CMU
   under constant variance 1926.4738 47.938359 1831.3994 2021.5483
                                 YHAT   SEYHATR      CMLR      CMUR
 allowing unequal variances 1926.4738 56.208582 1814.9974 2037.9503
```

Summary: Emphasis has been on linear regression models with focus on:

1. Components of the model and interpretation of parameters.
2. Matrix formulation of the regression model and expression of methods in matrix form. (Assists in reading documentation and is the language in which many statistical problems are expressed.)
3. Methods for estimation, confidence intervals (one-at-a-time and simultaneous) and test of hypotheses about parameters and linear combinations of them (and ratios in the case of inverse prediction and regulation).
4. Prediction.
5. Model assessment and model building.
6. Some additional Diagonostics (outliers, influential observations, ...)
7. Introduction to dealing with serially correlated errors.

Important additional Topics.

- More additional diagonostics (diagnostics for multicollinearity, added variable plots, etc.)
- Small sample techniques without normality (e.g. bootstrap).
- Assesssments for correlated errors.
- Regression methods allowing correlated errors (data over time, time series/cross sectional regression, repeated measures, longitudinal data).
- Regression with other distributions or models specific to counts, survival data, etc. (Generalized linear models)
- Measurement error in the predictor variables and/or the response.
- Design of experiments. Sample size, choice of X's in designed experiments.
- Transformations.
- Non-linear regression (including binary regression).
- Neural Networks and non-parametric regression.