UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Applied Statistics
Friday, January 15, 2016

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. (25 PTS) Researchers are interested in comparing the effectiveness of two treatments (A and B) for severe depression. They collected the data on a random sample of n severely depressed patients and considered the following regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

where Y is a measure of the effectiveness of the treatment for a patient, X_1 is age (in years) of a patient, and X_2 is 1 if a patient received treatment A and 0 if a patient received treatment B. Note that ϵ is an error independent of X_1, with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

1) Suppose that the parameter values are $\beta_0 = 6.2$, $\beta_1 = 2.9$, $\beta_2 = 7.1$ and $\beta_3 = -1.4$. Draw the mean regression function $E(Y \mid X_1, X_2)$. Describe the type of dependence between Y and (X_1, X_2) captured by this regression model.

For parts 2-3 below, consider the following statements

(a) *For every age, there is no difference in the mean effectiveness for the two treatments.*

(b) *The effect of age on the treatment's effectiveness does not depend on treatment.*

2) Translate (a) and (b) into two hypothses on the parameters of the regression model.

3) Suppose that you have n independent observations (y_i, x_{1i}, x_{2i}), $i = 1, \ldots, n$, and you can assume the error term to be normally distributed with a mean of zero and a variance of σ^2 (i.e., $\epsilon \sim N(0, \sigma^2)$). How would you test the hypotheses for (a) and (b)? What are the test statistics and their null distributions?

2. (30 PTS) Researchers are interested in the relationship of Cognitive Level (CL) test scores to the level of psychopathology (mental or behavioral disorder). They collected the following data on a set of 20 patients in a hospital psychiatry unit:

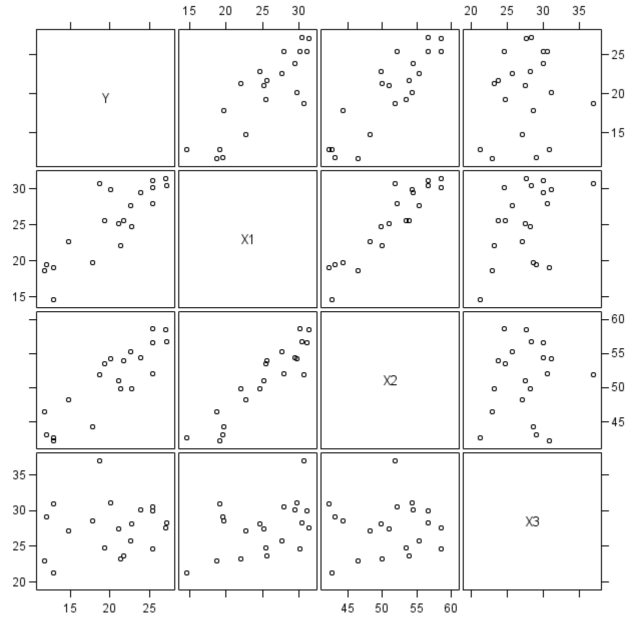
Y : CL test score

X_1 : vocabulary score

X_2 : abstraction score

X_3 : score on the symbol-digit modalities test

Figure 1 shows a scatter plot matrix of Y, X_1, X_2 and X_3.



[Figure 1] Scatter plot matrix of Y, X_1, X_2 and X_3

First, they fitted a simple regression model, denoted by $M1$, to the data:

$$M1 \; : Y = \beta_{0,M_1} + \beta_{1,M_1} X_1 + \epsilon$$

where the error term ϵ is normally distributed with a mean of zero and a variance of σ^2 (i.e., $\epsilon \sim N(0, \sigma^2)$). From the residual analysis, they found that the assumptions of the model $M1$ are met.

They next fitted another simple regression model, denoted by $M2$:

$$M2 \; : Y = \beta_{0,M_2} + \beta_{2,M_2} X_2 + \epsilon$$

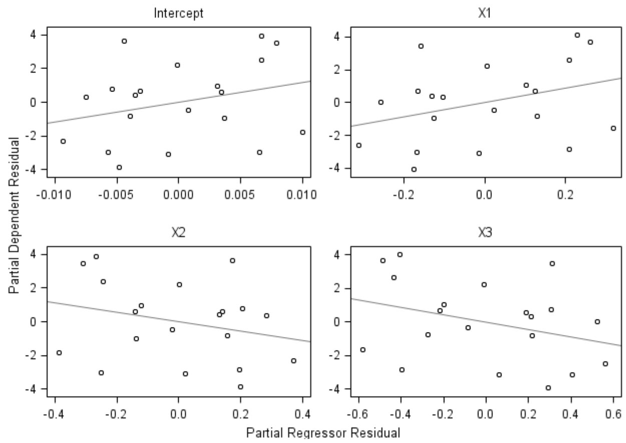
The residual analysis indicated that the assumptions of the model $M2$ are met.

Now they fitted a multiple regression model, denoted by $M3$:

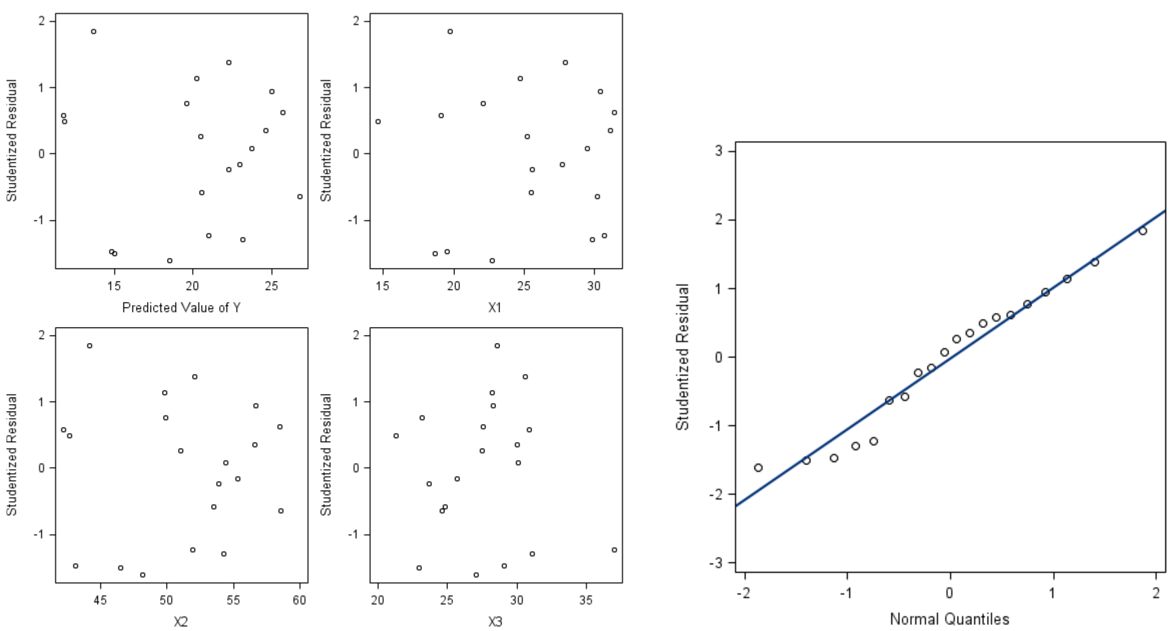
$$M3 \; : Y = \beta_{0,M_3} + \beta_{1,M_3} X_1 + \beta_{2,M_3} X_2 + \beta_{3,M_3} X_3 + \epsilon$$

1) State all the assumptions for the model $M3$ and, using Figure 1 - 3 and SAS outputs for the models $M1$, $M2$ and $M3$ provided below, comment on whether the assumptions are reasonable.

2) Is there evidence that the model $M3$ (i.e., the addition of the independent variables to the models $M1$ and $M2$) helps researchers better understand the relationship of CL test scores (Y) to the level of psychopathology (X_1, X_2, X_3)? In particular, does it make much sense to interpret each of the slope coefficients as "the change in the mean response for each additional unit increase in the predictor, when all the other predictors are held constant"? Why or Why not?



[Figure 2] Added variable plots (Partial regression plots) for X_1, X_2 and X_3



[Figure 3] Left: studentized residuals plots (residual vs. the fitted value of Y, X_1, X_2 and X_3); Right: QQ Plots for Residuals

Relevant portions of the SAS PROC REG outputs for the models [M1], [M2] and [M3] are given below:

[Model M1]

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.49610	3.31923	-0.45	0.6576
X1	1	0.85719	0.12878	6.66	<.0001

[Model M2]

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.63449	5.65741	-4.18	0.0006
X2	1	0.85655	0.11002	7.79	<.0001

[Model M3]

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
X1	1	4.33409	3.01551	1.44	0.1699
X2	1	-2.85685	2.58202	-1.11	0.2849
X3	1	-2.18606	1.59550	-1.37	0.1896

3. (20 PTS) Five groups of animals were exposed to a viral solution in varying concentration. Let n_i be the number of animals, y_i the number of animals died and $p_i = y_i/n_i$ the proportion of animals that died in the i-th group where $i = 1, \ldots, 5$.

$\log_{10}$ (Concentration)	n_i	y_i	p_i
-5	6	0	0
-4	6	1	0.167
-3	6	4	0.667
-2	6	6	1
-1	6	6	1

The goal of this study is to model the probability of death π as a function of $\log_{10}$(Concentration).

1) Do you think it is reasonable to regress the π's on $\log_{10}$(Concentration) using ordinary least squares? Discuss in detail (List each of the main assumptions of the linear regression using ordinary least squares, describe how each one may be satisfied or violated by these data, and the implications for inference about the relationship between the dose and response).

Assuming binomial response $y_i \sim Binomial(n_i, \pi_i)$ where π_i is the probability of death for the i-th group, it appears that it is more reasonable to apply a logistic regression,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \log_{10}(Concentration)$$

Relevant portions of the SAS PROC GENMOD output are given below:

```
                 Analysis Of Maximum Likelihood Parameter Estimates

                                     Standard    Likelihood Ratio 95%
Parameter    DF      Estimate         Error       Confidence Limits
Intercept     1        9.5868        3.7067       4.2753      19.4756
logconc       1        2.8792        1.1023       1.3097       5.9274
Scale         0        1.0000        0.0000       1.0000       1.0000
NOTE: The scale parameter was held fixed.

                       LR Statistics For Type 3 Analysis

                                        Chi-
             Source            DF      Square     Pr > ChiSq
             logconc            1       27.47         <.0001
```

2) Write down the estimated logistic regression model. Is there any convincing evidence supporting a dose-response relationship?

3) Interpret $\exp(\hat{\beta}_0)$ and $\exp(\hat{\beta}_1)$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1.

4) What is the estimated log-odds of death, $\log(\pi/(1-\pi))$ at $\log_{10}$(Concentration)= –3? Then find the corresponding estimate for the probability of death.

5) A parameter of interest in dose-response study is LD_{50}, the dose at which fifty percent of exposed animals would die. Estimate LD_{50}.

4. (25 PTS) Consider the R code below.

```
data <- data.frame(type=rep(c("A","B"),each=3),value=c(10,11,12,3,7,5))

result <- data.frame(type=c("A","B"),mean=0)
for (i in 1:6)
{
if (tx ype="A")
{
result$mean <- result$mean+data$value[i]
}
else
{
result$mean <- result$mean+data$value[i]
}
}
result$mean <- result$mean/3
```

(a) Describe what the code is intended to do.

(b) Find and correct syntax errors.

(c) Rewrite the code in a way that uses fewer lines.