SOLUTION

ST505/697R: Fall 2012 . MIDTERM EXAM: PART I, CLOSED BOOK (52 pts)

READ THE QUESTIONS CAREFULLY!!

1. A shipment of biological material from a company is sent in a carton containing 1000 ampules. Data was collected on a number of shipments where X = the number of times the carton was transferred between aircraft and Y = the number of broken ampules. The output page contains some results from a fitting a simple linear regression of Y on X. This was used in one of the homeworks.

   (a) (10 pts) Write down the simple linear regression model and state clearly what assumptions are made about the error terms in order for the statistical analyses given in the output to be correct. Include all assumptions!

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

   (1) $E(\epsilon_i) = 0$

   (2) $V(\epsilon_i) = \sigma^2$

   (3) ϵ_i distributed NORMAL (needed for everything since n is SMALL)

   (4) for $i \neq j$, ϵ_i and ϵ_j are uncorrelated (more strongly independent)

   (b) (20 pts). Give a NUMERICAL VALUE, for each of the following. Except for a simple calculation in viii) and ~~ix~~) vii the remaining answers should be found directly from the output, with no additional calculations.

   i. The estimated slope. 4

   ii. The estimated intercept. 10.2

   iii. A 95% confidence interval for the slope. [2.918, 5.082]

   iv. An estimate of the change in the expected number of broken ampules for each additional transfer. SAME AS SLOPE = 4

   v. An estimate of the variance in the number of broken ampules at a fixed number of transfers.
   MSE = 22 ESTIMATES σ^2 = VARIANCE of Y at a given X

   vi. An estimate of the variance of the estimated intercept.
   .44

   vii. An estimate of the expected number of broken ampules when 2 transfers are made.
   $10.2 + 4(2) = 18.2$

   viii. The residual for the first observation which had an outcome of $Y = 16$ and for which $X = 1$.
   $16 - (10.2 + 4(1)) = 1.8$

   ix. An estimate of the covariance between the estimated intercept and the estimated slope.
   −.22

x. A measure of the proportion of variability in the number of broken ampules explained by the number of transfers made.

$R^2 = .9009$ (90.1%)

(c) (4 pts) The P-value in the analysis of variance is listed as less than .0001. i) What hypothesis is being tested by this p-value? What would be your conclusion.

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

WOULD REJECT H_0 FOR ANY REASONABLE α (CERTAINLY ANY $\alpha \geq .0001$)

(d) (8 pts) For the simple linear regression model, define the problem of inverse prediction and then of regulation; so two different definitions. For each problem, explicitly use the setting about (with X = number of transfers and Y = number of broken ampules) to illustrate each of these problems. (Note that I am NOT asking you how to carry out the analysis but just describe what the problem is in each case.) Each definition and illustration should be just one sentence each.

Inverse prediction: ESTIMATE THE X VALUE ASSOCIATED WITH A NEW UNIT FOR WHICH WE OBSERVED AN OUTCOME Y_{new}.

EX: A SHIPMENT IS RECEIVED WITH 10 BROKEN AMPULES. ESTIMATE THE NUMBER OF TRANSFERS.

Regulation: ESTIMATE THE X AT WHICH $E(Y) = \beta_0 + \beta_1 X$ EQUALS A SPECIFIED CONSTANT m.

EX: ESTIMATE THE NUMBER OF TRANSFERS AT WHICH THE EXPECTED NUMBER OF BROKEN AMPULES IS 20.

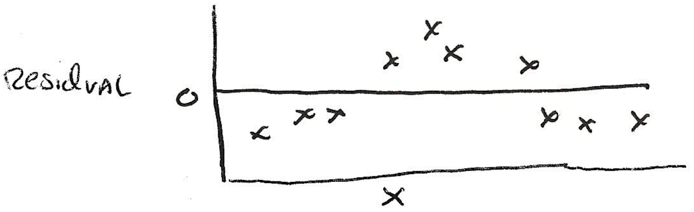
The remaining questions in this part of the exam are NOT tied to problem 1 or the data there.

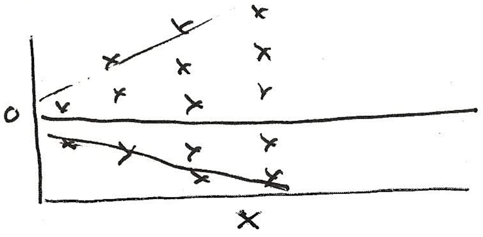
2. (4 pts) The statistic R^2 is often reported as a measure of how good the fit to the data is. Even supposing the linear regression model is a good model for the data, give two reasons R^2 not always useful?

① WITH FIXED X'S THE MAGNITUDE OF R^2 CAN BE ALTERED BY CHANGING THE SPREAD IN X'S

② BIG R^2 DOES NOT NECESSARILY CORRESPOND TO GOOD PREDICTION OF Y (i.e. σ^2 OR MSE CAN BE BIG)

3. (4 pts) On the left below, provide a sketch illustrating what a plot of the residuals versus X could look like if assumption that $E(Y)$ is linear in X was not true. On the right provide a sketch below illustrating what a plot of the residuals versus X could look like if the assumption that the variance of the error is constant was not true.





4. (2 pts) What do we use Levene's test for? (A one sentence anwser in words is sufficient).

IT TESTS FOR CONSTANT VARIANCE OF THE ERRORS (USING GROUPED DATA)

SOLUTION

ST505/697R: Fall 2005. MIDTERM EXAM: PART II, OPEN BOOK AND NOTES (72 points).

IMPORTANT: **SET-UP means to set-up the computation with all numerical values in place but you don't need to calculate the answer. In fact, don't try to calculate things out or you may run out of time.**

Where a table value is needed, you can leave it symbolically, for example as $t_{.95,3}$, but make sure there are numerical values involved in the arguments. Other than table values all things in a set-up should involve numerical values, except where indicated otherwise.

IF YOU DON'T KNOW HOW TO EVEN START ON SOMETHING, MOVE ON!

1. (60 pts) This problem uses the transfer/breakage output that was used in the close book part. We will make the usual assumptions about the error terms.

   Note that in parts a) - d) **I do NOT want computer code.**

   (a) Set-up how the confidence interval (8.670, 11.7296) for the intercept is calculated.

   $10.2 \pm t(.975, 8)(.66332)$

   (b) Set-up simultaneous 95% confidence intervals for β_0 and β_1. $\alpha = .05 \quad 1 - \frac{\alpha}{4}$

   $10.2 \pm t(1 - \frac{.05}{4}, 8)(.66332) \qquad 4 \pm t(1 - \frac{.05}{4}, 8)(.4690)$

   (c) Three shipments will be made in the next week that involve 2, 3 and 5 transfers.

   - First show how you would get a predicted value and an associated standard error to use for prediction; your answers should be in terms of numbers and a general X (where X could be set to 2, 3 or 5)

   $\hat{Y} = 10.2 + 4 \cdot X \qquad s_{pred}\{X\} = \sqrt{2.2 + .44 + X^2(.22) - 2 \cdot X(.22)}$

   - Set-up how you would you get a prediction interval at a single X. (In doing so you can refer to the predicted value and standard error from above).

   $\hat{Y} \pm t(.975, 8)\, s_{pred}\{X\}$

   - Now set up simultaneous prediction intervals for all three predictions using two techniques. How would you decide which of the two is better? (You don't need to calculate anything and do the comparison but indicate how you would decide).

   Bonferroni $\hat{Y} \pm t(1 - \frac{.05}{6}, 8)\, s_{pred}\{X\}$

   Scheffe $\hat{Y} \pm \sqrt{3F(.95, 3, 8)}\, s_{pred}\{X\}$

   Use whichever one has smaller multiplier. So, Bonferroni if $t(1 - \frac{.05}{6}, 8) < \sqrt{3F(.95, 3, 8)}$

   (d) Set-up simultaneous confidence intervals for $\mu(X)$ = the expected/mean number of broken ampules obtained with X transfers for many (possibly infinitely many) values of X. Your set-up should involve numbers except for X and a table value (see earlier note about table values).

   Use Working-Hotelling to handle estimating $\beta_0 + \beta_1 X$ for possibly infinitely many X's

   $10.2 + 4 \cdot X \pm \sqrt{2F(.95, 2, 8)}\sqrt{.44 + X^2(.22) - 2 \cdot X(.22)}$

(e) Provide either an R or SAS program that does the following (indicating clearly which code is responsible for which part.

- produces the output given in the attachment.
- Plots the original data and the fitted line on the same graph.
- Plots the residuals and the absolute residuals versus the fitted values.
- Gets a histogram (with a smoothed fit overlayed) for the residuals and gets a test of the hypothesis that the errors are normally distributed.

R CODE

Gives info on output sheet

```
Regout <- lm(number ~ transfers)
Summary(regout)
Anova(regout)
Confint(regout)
vcov ( regout)
```

Plot original data and fitted line

```
attach(data)
plot(transfers, number)
fits <- fitted(regout)
lines (transfers, fits)
```

get residuals and absolute residuals. plot versus fitted values

```
resids <- residuals(regout)
abresid <- abs(resids)
Plot (fits, resids)
plot ( fits, absresid )
```

Histogram with smooth curve

```
hist(resids, freq = FALSE)
lines(density(resids))
```

Test of normality

```
Shapiro.test(resids)
```

SAS CODE

C.I.'s for β's

Gets $s^2\{\underline{b}\}$

Output on sheet

```
proc reg;
model number = transfers/clb covb;
run;
plot number*transfers;
```

Inside reg. plots data and fit

In REG use

```
output out=result r=resid p=yhat;
```

(*) SEE BELOW

(f) There were replicate values for some of the transfer values. The following one-way analysis of variance grouping on transfers was run. Use this to carry out a test for lack of fit. . State what the null and alternative hypotheses are, Set-up the test statistic (all numbers) specify the degrees of freedom involved (again with numbers) and describe how you would carry out the test for a specified α.

```
This is SAS output;  in R, Model will be labeled group and Error as Residual.

Dependent Variable: number
                                   Sum of
 Source                 DF        Squares     Mean Square   F Value   Pr > F
 Model                   3    160.9333333      53.6444444     19.31   0.0017
 Error                   6     16.6666667       2.7777778
```

SSPE = 16.667 d.f. = 6

SSE = 17.6 df = 8

c must = 4 since c-1 = 3

2 = 8-6 also = c-2

$$F = \frac{(17.6 - 16.667)/2}{16.667/6}$$

F statistic to test lack of fit

Reject H_0 if $F > F(1-\alpha, 2, 6)$

e) Rest of SAS Code

```
data b;
set result;
absresid =abs(resid);
run;
proc univariate normal;
var resid;
hist resid/KERNEL;
run;
```

does normal test

gets histogram with smoothing.

Plot e_i and $|e_i|$ versus $\hat{Y}_i$

```
Proc gplot data=b;
Plot resid*yhat;
Plot absresid*yhat;
run;
```

2. (12 pts) A study is to be carried out to estimate the relationship between Y = breaking strength of a cable on X = diameter of the cable in inches. A simple linear regression model is assumed to hold over the range of X from 1 to 4, with the usual assumptions including normality of the ϵ_i's. Suppose that n = 20 observations will be made using cables with diameters $X_1, \ldots X_{20}$, where $\bar{X}$ = 2. and $\sum_i (X_i - \bar{X})^2 = 10$ and we know that σ^2= the variance of the error term is 25.

$\sigma^2\{b_1\} = \frac{\sigma^2}{10}$

$\sigma^2\{b_1\} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{10}\right)$

$\sigma\{b_0, b_1\} = -\frac{\sigma^2 2}{10}$

(a) What is the variance of b_1? (give a number for your final answer)

$$\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} = \frac{25}{10} = 2.5$$

(b) What is the variance of $b_0 + 3b_1$? (which estimates the mean breaking strength for 3 inch cables.) Set-up with numerical values.

$$V(b_0 + 3b_1) = \sigma^2\{b_0\} + 3^2\sigma^2\{b_1\} + 2\cdot3\cdot\sigma\{b_0, b_1\}$$

$$= 25\left(\frac{1}{20} + \frac{(3-2)^2}{10}\right) \quad \underline{\text{OR}} \quad 25\left[\frac{1}{20} + \frac{2^2}{10}\right] + 3^2 \cdot \frac{25}{10} + 2\cdot3\left(-\frac{2}{10}\right)$$

(c) Consider a single cable with a diameter X.

- What is $E(Y)$? (this will involve parameters and X)
- What is the probability that the breaking strength Y is within 5 of its expected value? You can leave your answer in a form involving probabilities about a random variable with a standard normal distribution. Show your work (and in doing so you will also show that the answer doesn't depend on what X is or on the coefficients.)

$$E(Y) = \beta_0 + \beta_1 X$$

$$P[E(Y) - 5 < Y < E(Y) + 5]$$

$$P\left[\frac{E(Y) - 5 - E(Y)}{5} \leq \frac{Y - E(Y)}{5} \leq \frac{E(Y) + 5 - E(Y)}{5}\right]$$

$$= P[-1 < Z < 1]$$ where $Z \sim$ STANDARD NORMAL