SOLUTION

ST505/697R: Fall 2012 EXAM 2/Final: PART I, CLOSED BOOK

Be sure to read carefully and follow the instructions given in a problem!

1. This problem concerns relating body fat Y (as measured by an expensive, but essentially exact, measure) to a skin measure (X_1) and thigh measure (X_2) using linear regression. The Body Fat Example output fits the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ with output from both SAS and R.

   (a) What assumptions are being made on the error terms for this analysis to be valid?

   $E(\epsilon_i) = 0$

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

   ϵ_i and ϵ_j uncorrelated/independent for $i \neq j$

   ϵ_i distributed NORMAL

   (b) First, state precisely what null hypothesis is being tested by the F statistic of 29.797 in the Analysis of Variance table (SAS), or equivalently the F-statistic at the end of the summary in the R output?

   $H_0: \beta_1 = \beta_2 = 0$

   -What is your conclusion? What does does it say about the coefficients?

   REJECT Ho. Conclude at least one of β_1 or β_2 is NOT 0

   (c) What hypothesis is being tested by the P-value of .0369 associated with thigh?

   $H_0: \beta_2 = 0$

   (d) Notice that the standard errors and the t-statistics for the coefficients are missing. Show *with a number involved* how with one calculation you can get the standard error for skin from other information on the output. Then *set-up* (with numbers) how you would then get the t-statistic for testing that the coefficient for skin is 0. *no need to carry out the calculations.*

   $S.E = \sqrt{.092075}$ $\qquad t = \dfrac{.222}{\sqrt{.092075}}$

   (e) What is the -0.081628463 in the covariance matrix an estimate of?

   ESTIMATES $Cov[b_1, b_2]$

   (f) Suppose we fit the model assuming $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. If $\beta_3 \neq 0$ we say there is interaction. State **IN WORDS** what the meaning of this is.

   Effect (slope) of X_2 depends on what the level of X_1 is

   or " " " X_1 " " " " X_2 is.

   (g) Suppose we have a third variable $X_3 = 1$ if the subject is white 0 if the subject is non-white. Suppose we fit a model of the form $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_1 * X_3$ (note that X_2 is not involved here). Give **verbal** interpretations to each of the parameters β_0, β_1, β_2 and β_3, referring to race as you do so.

   β_0 = intercept when RACE = White

   β_1 = Slope/Coefficient for X_3 if RACE = White

   β_2 = Intercept for RACE = NONWhite − Intercept for RACE = White

   β_3 = Slope for X_3 when RACE = White − Slope for X_3 when RACE = NONWhite

SEE BACK

2. A friend of yours took a course which covered an introduction to statistics with just a little regression. Looking over what you have been doing she sees a number of new things that she doesn't know about. (Let's call her Florence after Florence Nightingale, who provided the motivation for the creation of the Red Cross, and also happened to be an early member of the American Statistical Association.) Provide a brief answer to Florence's questions below. Your answers should be: **ONE SENTENCE FOR EACH QUESTION AND SHOULD BE JUST IN WORDS. NO FORMULAS!**

   (a) I saw this thing labeled consistent covariance of estimates in your SAS output, or what was also referred to as White's robust estimate of the covariance of the estimates in R. Why would I want to use that?

   I would use that since it estimates THE COVARIANCE-VARIANCE of b allowing for the fact that The VARIANCES MAY be UNEQUAL

   (b) I see that you are doing model building. I remember that R^2 was used as a measure of how good the model was. Why don't you just compute R^2 for each combination of variables and chose the combination of variables with the highest R^2?

   Because the HIGHEST R^2 ALWAYS OCCURS WITH ALL VARIABLES IN THE MODEL.

   (c) I happened to see that example you had modelling yield as a function of temperature and moisture. Since it involved temperature- squared and moisture-squared, why do you refer to is as a multiple linear model when the model for expected yield is a non-linear function of yield and moisture, or did you make a mistake?

   IT MAY BE NONLINEAR IN THE X'S BUT IT IS LINEAR IN THE PARAMETERS (WHICH IS WHAT LINEAR IS REFERRING TO)

   (d) What is this C_p statistic used for?

   IT IS used for VARIABLE SELECTION

   REMAINING PARTS OF THIS PROBLEM FOR 697R STUDENTS ONLY. You can use two sentences here if needed

   (e) I know what residuals are. Why do you want to use studentized residuals rather than the regular residuals?

   Even if The ε_i have CONSTANT VARIANCE The VARIANCES Of The e_i (residuals) have NON CONSTANT VARIANCE. By "STUDENTIZING" The VARIANCE of e_i's will Be CONSTANT

   (f) What is the leverage value used for? Is the leverage value related to the outcome Y?

   The leverage determines OUTLIERS IN THE X SPACE. IT IS ONLY A FUNCTION of The X'S, NOT The Y'S

   (g) What is Cook's distance used for?

   TO DETERMINE INFLUENTIAL OBSERVATIONS.

Set-up means **WITH ALL NUMBERS! BUT NO NEED TO CALCULATE OUT. THE ONLY SELECTION IS FOR TABLE VALUES WHICH YOU CAN LEAVE SYMBOLIC BUT THERE MUST BE NUMBERS FOR THE DEGREES OF FREEDOM INVOLVED.**

1. This problem returns to the Fat example introduced in the closed book portion. See the output. Assume the ϵ_i are uncorrelated and normally distributed with mean 0 and variance σ^2.

I DIDN'T GRADE THIS SINCE QUESTION SAID Simply adding +4 TO everyone

a) Set-up how *by simply adding two numbers in the output* you can find the standard error associated with getting prediction interval at skin = 19.5 and thigh = 43.1 (which are the values for the first case in the data).

$SEPRE = \sqrt{MSE + s^2\{\hat{Y}_h\}} = \sqrt{6.46 + 1.14}$

b) Suppose we wanted simultaneous 90% prediction intervals for body fat at 5 different skin/thigh combinations. The predictions intervals will be of the form $\widehat{Y} \pm mult * SE_{pred}$. Describe what the multiplier is for the Bonferroni and Scheffe methods and state how you would decide which is better.

BONFERRONI : $t(1 - \frac{.10}{10}, 17)$

Scheffe : $\sqrt{5F(.90, 5, 17)}$

BETTER HAS SMALLER MULTIPLIER

c) Consider estimating $\mu(X_1, X_2) = E(Y|X_1, X_2)$ for a value X_1 for skin and X_2 for thigh.

- Set-up how you would get an estimate of $\mu(X_1, X_2)$, call this $\widehat{\mu}(X_1, X_2)$. Your answer should involve X_1 and X_2 and numbers,

$-19.174 + .222 \cdot X_1 + .6594 \cdot X_2 = \hat{\mu}(X_1, X_2)$

- Set-up how you would get an estimate of the standard error of $\widehat{\mu}(X_1, X_2)$. Your answer can be in a form using a matrix with numbers and a vector, which could involve numbers and possibly X_1 and X_2.

$\sqrt{X_h' s^2\{b\} X_h}$

$$\sqrt{\begin{pmatrix} 1 & X_1 & X_2 \end{pmatrix} \begin{pmatrix} 69.9 & 1.84 & -2.27 \\ 1.84 & .092 & -.0182 \\ -2.27 & -.0182 & .0848 \end{pmatrix} \begin{pmatrix} 1 \\ X_1 \\ X_2 \end{pmatrix}} = SE(\hat{\mu})$$

- Using $\widehat{\mu}$ from above and it's estimated standard error, set-up how you would calculate simultaneous 95% confidence intervals for $\mu(X_1, X_2)$ over all values of X_1 and X_2.

$\hat{\mu}(X_1, X_2) \pm \sqrt{3F(.95, 3, 17)} \cdot SE(\hat{\mu})$

d) There was a third variable, X_3, which was a midarm measurement. All you are told is that the SSE for the fit with all three variables is 98.4. Set-up how to carry out a test of size .10 of $H_0 : \beta_3 = 0$, where β_3 is the coefficient for X_3 in the model with all three variables. Set-up the test statistic (all numbers) and state your decision rule in terms of comparing your test statistic to a table value (specified in as much detail as possible).

$SSE(F) = 98.4$ $\quad SSE(R) = 109.95 = 17(2.53)^2$ From R

$F = \dfrac{SSE(R) - SSE(F)}{MSE(F)}$ $\quad MSE(F) = \dfrac{98.4}{16} = \dfrac{SSE(F)}{n-p}$

$17 = n - 3$
$n = 20$
MODEL WITH ADDITIONAL X_3 has $p = 4$

Reject Ho if $F > F(.90, 1, 16)$

NOTE: IN PROBLEM 2 ONLY give Code necessary TO PROVIDE ~~WHAT~~ WHAT IS ASKED FOR.

With X=DOSE, X2=DOSE^2 and Z1=VARIETY-1
MOST of THIS PROBLEM IS MUCH LIKE HW 8

2. There is a data file called yield.dat from a study on tomato yields for two varieties over different doses of phosphorous (other nutrients held fixed). The file has 30 observations including variety (given values of 1 and 2), dose and yield, in that order, separated by spaces with no missing values. Consider a model where for variety j ($=1$ or 2) with $X =$ dose, $E(Y|X) = \beta_{j0} + \beta_{j1}X + \beta_{j2}X^2$. This allows a separate quadratic regression for each variety. Assume the variance is constant throughout.

(a) Provide code needed to read this data, define any additional variables (if necessary) and run a regression so that the output contains six estimated coefficients which are the estimates of $\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}$ and β_{22}.

R

Z1=0 if VARIETY 1
=1 if Variety 2

```
data<-read.table("yield.dat")
attach(data)
VARIETY<-V1; dose<-V2; YIELD<-V3
Z1<-VARIETY-1; Z2<-1-Z1
REGOUT<-lm(YIELD~-1+Z1+Z2+I(DOSE*Z1)
+I(DOSE*Z2)+I(Z1*(DOSE^2))+I(Z2*(DOSE^2))
OR DEFINE DZ1=DOSE*Z1; DZ2=DOSE*Z2
DZZ1=Z1*(DOSE^2); DZZ2<-Z2*(DOSE^2)
lm(y~-1+Z1+Z2+DZ1+DZ2+DZZ1+DZZ2)
```

SAS

```
data a;
infile
input VARIETY,DOSE,YIELD;
Z1=VARIETY-1;
Z2=1-Z1;
D2=DOSE**2;
DZ1=DOSE*Z1;
DZ2=DOSE*Z2;
DZZ1=D2*Z1;
DZZ2=D2*Z2;
proc reg;
model YIELD = Z1 Z2 DZ1
DZ2 DZZ1 DZZ2/noint;
```

(b) Provide any additional code needed to run a regression that also has six coefficients but where three of them are estimates of $\beta_{10} - \beta_{20}$, $\beta_{11} - \beta_{21}$, and $\beta_{12} - \beta_{22}$. Indicate which coefficients in the output would be estimating these differences. (If you need any additions to the data step beyond what you gave in i) then give them also.)

```
D2<-DOSE^2
lm(yield~Z1+Dose+D2+I(Z1*DOSE)+
      I(Z1*D2))
```

$\beta_{10}-\beta_{20}$ = Coefficient of Z1

$\beta_{11}-\beta_{21}$ = " " Z1*DOSE

$\beta_{12}-\beta_{22}$ = " " Z1*D2

(b)

```
Proc reg;
model YIELD = Z1 DOSE
D2 DZ1 DZZ1;
```

(c) Provide any additional code to test the hypothesis that the linear and quadratic terms are the same for each variety; i.e. the expected value for an observation from variety j is $\beta_{j0} + \beta_1X + \beta_2X^2$.

```
regnull<-lm(YIELD~Z1+DOSE+D2)
anova(regnull, regout)
```

regout from lm in either a) or b)

SAS using proc reg in b)

```
test DZ1=0, DZ2=0;
```

What degrees of freedom will be associated with the resulting test? Give numbers.

n=30 d.f. of full model SSE = 30-6=24

d.f. of null model SSE is 30-4=26

Under Ho F statistic ~F(2,24)

2 and 24 degrees of freedom

3. The second part of the output gives results from working with the Puffin data, used earlier in class. This has an outcome Y and four predictors. The output shows fits from all possible regressions (one variable, two variables, etc.) Use this to build a model using stepwise selection. **Explain each step clearly** and what the final model is . Use a p-value of .15 for entering or removing variables from the model.

STEP 1: START WITH NOTHING. Consider p-values for each variable by itself. Both X3 and X4 have p-values < .0001. X4 has smaller p-value since it has larger |t| value. ENTER X4.

STEP 2: Consider each of X1, X2 or X3 with X4 with p-values of .4675, .0018 and .1139 respectively. Enter X2 since it has smallest p-value and is < .15.

Now need to consider possible removal of X4 once X2 is in model (since procedure is stepwise not forward). P-value is now still < .0001 so keep X4.

STEP 3 Consider entering either X1 or X3 with X2 and X4. P-values are .8691 and .7965 respectively. Do not enter either. Stop

Final model has just X2 and X4 in it.