Midterm I Statistics 500 (Winter 2014)

Name: Key

Instructions: Answer the following questions carefully. You must **show all work** where needed to receive credit. Partial credit will be given where work is shown. Space is provided for your answers. You may use scrap paper to arrive at solutions, but **show all work** on the test.
Total points: 65

1) Suppose $\boldsymbol{Y}$ is a bivariate normal random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. That is, $\boldsymbol{Y} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & -1 \\ -1 & 4 \end{bmatrix}$$

Let $\boldsymbol{A}$ be a 1×2 vector with fixed elements

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \end{bmatrix}$$

a) Calculate $E[\boldsymbol{AY}]$. (3 points)

$$E[AY] = A\,E[Y] = A\mu$$

$$= \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 6 + 1 = \boxed{7}$$

b) Calculate $Var(\boldsymbol{AY})$. (4 points)

$$Var(AY) = A\,Var(Y)\,A'$$

$$= \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 6 + 2 = \boxed{8}$$

c) What is the distribution of $(\boldsymbol{Y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})$? (2 points)

$$\chi^2_2$$

2) Data were collected on housing values in suburbs of Boston. Definitions of six variables are provided:

medv	median value of owner-occupied homes in \$1000s
zn	proportion of residential land zoned for lots greater than 25,000 sq. ft.
ptratio	pupil-teacher ratio by town
dis	weighted mean of distances to five Boston employment centers
rm	average number of rooms per dwelling
lstat	lower status of the population (percent)

A normal linear regression model was fit with `medv` as the response variable and all other variables as predictors. Results are shown below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.60981    4.06012   6.061 2.66e-09 ***
zn           0.03294    0.01400   2.352    0.019 *
ptratio     -0.89528    0.12022  -7.447 4.21e-13 ***
dis         -0.77659    0.15841  -4.902 1.28e-06 ***
rm           4.06211    0.42747   9.503  < 2e-16 ***
lstat       -0.67258    0.04664 -14.420  < 2e-16 ***
---
Residual standard error: 5.11549 on 500 degrees of freedom
F-statistic: 226.5 on 5 and 500 DF,  p-value: < 2.2e-16
```

a) Explain how you can determine the number of observations in the data set. How many observations are there? (3 points)

$n - p = 500$

$p = 6$

$\boxed{n = 506}$

b) What is $\hat{\sigma}$ (the estimate for σ)? (2 points)

$\boxed{\hat{\sigma} = 5.115}$ from output

c) Calculate an approximate 95% confidence interval for β_{rm}. Assume a critical t-value of 2.00. (4 points)

$\hat{\beta}_{rm} \pm 2.00\, se(\hat{\beta}_{rm})$

$4.0621 \pm 2.00\,(0.4275)$

$\boxed{(3.207,\ 4.917)}$

d) State the null hypothesis that the F-statistic is testing. (2 points)

$H_0: \beta_{zn} = \beta_{ptratio} = \beta_{dis} = \beta_{rm} = \beta_{lstat} = 0$

e) An ANOVA table is shown below. SST is given. Fill in the rest of the table. Show work.(4 points)

Source	DF	Sum Sq	Mean Sq	F
Regression	5	29632.2	5926.4	226.5
Error	500	13084.1	26.17	
Total	505	42716.3		

$MSE = \hat{\sigma}^2 = \boxed{26.17}$

$\hat{\sigma}^2 = \frac{SSE}{n-p}$ $\quad SSE = (n-p)\hat{\sigma}^2 = 500(5.11549)^2 = \boxed{13084.1}$ SSE

$SSR = SST - SSE = 42716.3 - 13084.1 = \boxed{29632.2}$ SSR

$MSR = \frac{29632.2}{5} = \boxed{5926.4}$

f) Calculate R^2. (3 points)

$R^2 = \frac{SSR}{SST} = \frac{29632.2}{42716.3} = \boxed{0.69}$

A reduced model was fit to test a certain hypothesis. Answer parts g) and h) using results shown below.

```
Analysis of Variance Table

Model 1: medv ~ zn + I(ptratio + dis) + rm + lstat
Model 2: medv ~ zn + ptratio + dis     + rm + lstat
  Res.Df     SSE      Df    Sum of Sq     F
1    501   13092.59
2    500
```

g) State the null hypothesis for the test being conducted. (3 points)

$H_0: \beta_{ptratio} = \beta_{dis}$

h) Calculate the value of the F-statistic for the test using a formula that you know. What are the degrees of freedom in the numerator and in the denominator? Show how you calculated the F-statistic using the formula. (4 points)

$F = \frac{(SSE_\omega - SSE_\Omega)/(p-q)}{SSE_\Omega/(n-p)} = \frac{(13092.59 - 13084.1)/(6-5)}{13084.1/500} = \boxed{0.324}$

$F \sim F_{\boxed{1, 500}}$ when Ho true

Consider the hypothesis

$$H_0 : \beta_{ptratio} = \beta_{dis} = \beta_{lstat} = 0$$

and answer parts i), j), and k) below.

i) Write down the A matrix and the c vector for this hypothesis using the notation

$$H_0 : \boldsymbol{A\beta} = \boldsymbol{c}\,.$$

Use the order of the variables given in the model fit above. That is, (3 points)

$$\boldsymbol{\beta} = (\text{Intercept}, \beta_{zn}, \beta_{ptratio}, \beta_{dis}, \beta_{rm}, \beta_{lstat})'$$

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad c = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

j) Continuing with part i) what is the distribution of

$$\frac{(\boldsymbol{A\hat\beta} - \mathbf{c})'[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A\hat\beta} - \mathbf{c})}{\sigma^2}$$

when H_0 is true? Explain carefully, starting with the distribution of $\hat{\boldsymbol{\beta}}$. (4 points)

$$\hat\beta \sim N_p(\beta, \sigma^2 (X'X)^{-1})$$

$$A\hat\beta - c \sim N_3(A\beta - c, \sigma^2 A(X'X)^{-1}A')$$

$$(A\hat\beta - c)' \frac{[A(X'X)^{-1}A']^{-1}}{\sigma^2}(A\hat\beta - c) \sim \boxed{\chi^2_3}$$

k) Using the fact that $\dfrac{(n-p)\,\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-p}$ what is the distribution of

$$\frac{(\boldsymbol{A\hat\beta} - \mathbf{c})'[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A\hat\beta} - \mathbf{c})}{3\,\sigma^2}\,?$$

Be sure to give supporting evidence. (3 points)

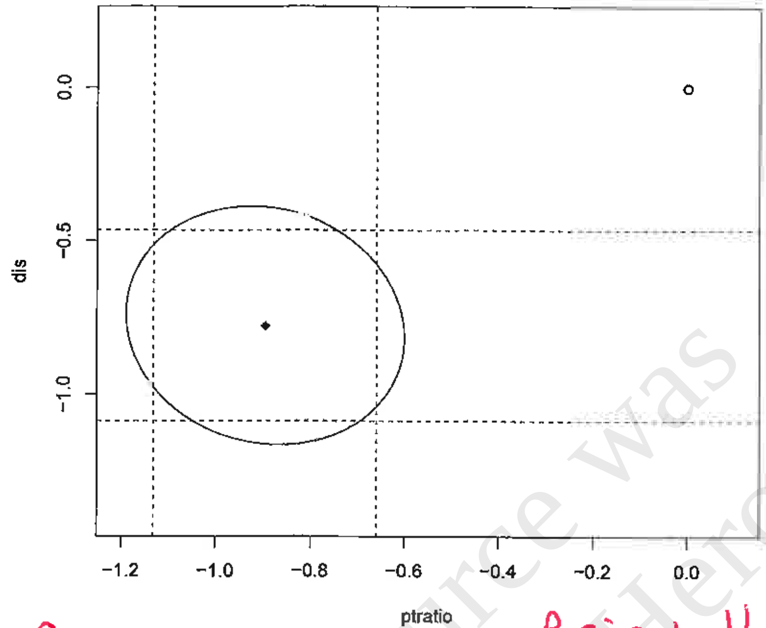
$$\left[\begin{array}{l} (A\hat\beta - c)'[A(X'X)^{-1}A']^{-1}(A\hat\beta - c) \sim \chi^2_3 \quad H_0 \text{ true} \\ (n-p)\dfrac{\hat\sigma^2}{\sigma^2} \sim \chi^2_{500} \end{array}\right.$$

independent

$$F = \frac{\dfrac{(A\hat\beta - c)'[A(X'X)^{-1}A']^{-1}(A\hat\beta - c)}{3\cancel{\sigma^2}}}{\dfrac{\cancel{(n-p)}\,\hat\sigma^2}{\cancel{(n-p)}\,\cancel{\sigma^2}}} = \frac{(A\hat\beta - c)'[A(X'X)^{-1}A']^{-1}(A\hat\beta - c)}{3\,\hat\sigma^2}$$

$$\sim \boxed{F_{3,500} \text{ when } H_0 \text{ true}}$$

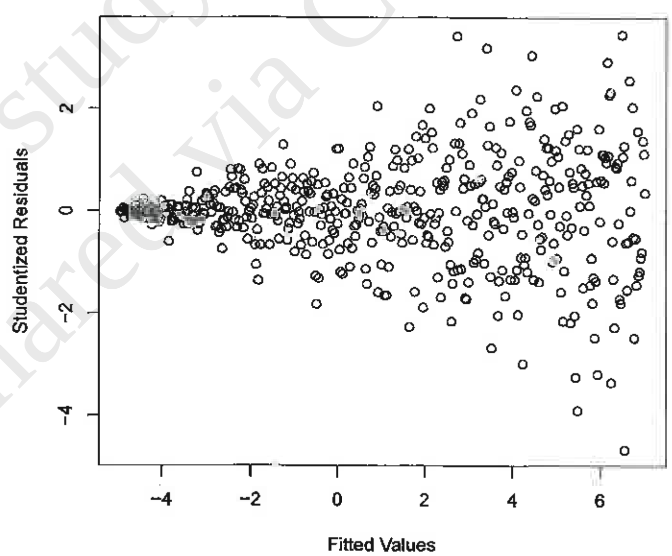
1) Below is a plot of the 95% joint confidence region for the parameters associated with ptratio and dis. The location of the origin on the display tells us the outcome of a certain hypothesis test. State the null hypothesis of that test and your conclusion. (3 points).



$H_0: \beta_{ptratio} = \beta_{dis} = 0$

Reject H_0 at $\alpha = .05$
the origin (0,0) is outside of the ellipse

3) A normal theory linear model was fit and the figure below shows a plot of the residuals against the fitted values. Do any assumptions of the model appear to be violated? Explain. (3 points)



The assumption of constant variance appears violated

4) Consider the linear model

$$y_i = \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \qquad i = 1, \ldots, n$$

where β is the unknown regression parameter, x is the predictor variable, and the ϵ_i are independent. Let

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_{i=1}^{n}(y_i - \beta x_i)^2$$

be the least squares criterion and find the least squares estimator for β. Be sure to show all work. (5 points)

$$\frac{\partial \epsilon' \epsilon}{\partial \beta} = -2\sum (y_i - \beta x_i) x_i = 0$$

$$\sum x_i y_i - \beta \sum x_i^2 = 0$$

$$\sum x_i y_i = \beta \sum x_i^2$$

$$\boxed{\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}}$$

5) Suppose $y_1, y_2, \ldots, y_n$ are independent random variables with $E[y_i] = \mu_i$ and $Var(y_i) = \mu_i^4$ where $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ is the mean in a regression model with predictor vector $\boldsymbol{x}_i$ and unknown parameter vector $\boldsymbol{\beta}$. Since the variance depends on the mean, the variance is not constant. Find a variance-stabilizing transformation $f(y)$ for the response y. (4 points)

$$f(y) = \int \frac{dy}{\sqrt{Var(y)}} \simeq \int \frac{dy}{\sqrt{y^4}} = \int y^{-2}\, dy$$

$$= -y^{-1} + c \quad \therefore \boxed{f(y) \propto \frac{1}{y}}$$

is the variance-stabilizing transformation

6) Assuming the normal theory linear model, calculate the following: (3 points each)

a) $(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{e}$

$$(I_n - P)e = (I_n - P)(I_n - P)y$$

$$= (I_n - P)y = \boxed{e}$$

b) $Var(\boldsymbol{e})$

$$Var(e) = Var((I_n - P)y) = (I_n - P)Var(y)(I_n - P)'$$

$$= \sigma^2(I_n - P)(I_n - P) = \boxed{\sigma^2(I_n - P)}$$