COVER PAGE

STATISTICS 608 - EXAM 2

July 10/11 2017

Student's Name (PRINT) RAJAN KAPOOR

Student's Email Address (PRINT) r.kapoor@tamu.edu

INSTRUCTIONS TO STUDENTS:

(1) Write your answers in the spaces provided on the examination paper. The last two (blank) pages can be used for rough work or as additional space for answers. You have exactly 90 minutes to complete the exam within the time frame 12:01 PM, CDT, 7/10/2017 to 12:01 PM, CDT 7/11/2017. **The exam starts AFTER you have downloaded and printed it.** If you have been granted extra time by Disability Services, your proctor will have been informed accordingly.

(2) Upon completing the exam, you have a 30-minute buffer in which to scan and upload it to Webassign. You may not work on the exam during this time.

(3) You may use your own computer together with a pocket calculator and/or any software package already loaded on your computer to do calculations.

(4) The exam is OPEN BOOK. You may make use of the textbook and any other material that you saw fit to prepare beforehand, either as hard copy or on your PC/Laptop.

(5) **You may not access the internet other than to download and to upload the exam.**

I attest that I spent no more than 90 minutes to complete the exam. I did not access the internet during the exam nor did I receive assistance from anyone during the exam. I promise not to discuss or provide any information to anyone concerning any aspect of this exam until after 3:31 PM on 7/11/2017.

Student's Signature Rajan

INSTRUCTIONS TO PROCTOR:

The exam starts only **after** the student has downloaded and printed it. **Immediately** after the exam ends, have the student scan the exam **with this cover sheet on top** to a PDF file and upload it to Webassign.

(1) I certify that the student's exam start time was 4:39 AM, and the exam completion time was 6:08 AM

(2) I certify that the student has followed all the **INSTRUCTIONS TO STUDENTS** listed above.

(3) I certify that the exam was scanned into a PDF and uploaded to Webassign **in my presence**.

Proctor's Printed Name REMOTE PROCTOR

Proctor's Signature Rajan Date 07/11/2017

STATISTICS 608
Examination 2 Summer 2017

Duration: **90 MINUTES** Total points available: **35** (**32** points = 100%)

SHOW ALL CALCULATIONS AND EXPLANATIONS. PARTIAL CREDIT WILL ACCRUE FOR ALL *RELEVANT* WORK SHOWN.

24/35

This paper consists of seven (7) pages.

2

Question 1 [3] Look at the following output from fitting to observed responses a three - covariate linear model with intercept:

covariate	est. coef.	st.err.	t	sig
X_1	0.5	0.5	1.0	0.33
X_2	1.0	0.25	4.0	0.00
X_3	0.4	0.4	1.0	0.33

The statement:"Both X_1 and X_3 should be dropped from the model because neither has a significant t-statistic" is (choose one) (a) justified (b) not justified. Explain *briefly* your reasoning.

X_1 and X_3 donot have significant t-statistic but they (both of them) cannot be dropped simultaneously on the basis of t-statistic because the significance might change when one of them is dropped. Partial-f test should be used to make this decision.

(b) not justified.

12

Question 2 **[12]** The *Question 2 Exam 2* data show the numerical values of a predictor y and three covariates x_1, x_2 and x_3. Assume that a valid linear regression

$$E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is in force.

Test the null hypothesis $H_0: \beta_2 = 2\beta_1$ at the **10%** level of significance.
For partial credit you need to show your calculations and briefly outline your method: numerator and denominator of your test statistic **[4+2=6]**; *applicable degrees of freedom* **[2+1=2]**, *numerical value of your test statistic* **[1]**, *critical value* **[1]** *and decision* **[1]**.

$\beta_0 + \beta_1(x_1 + 2x_2) + \beta_3 x_3 \quad \alpha = 0.1$

$df_{full} = 20 - 4 = 16$ ✓

$RSS_{full} = 12.7244$ ✓

$df_{red} = 20 - 3 = 17$ ✓

$RSS_{red} = 16.9028$ ✓

$$F = \left(\frac{RSS_{red} - RSS_{full}}{df_{red} - df_{full}} \right) \div \frac{RSS_{full}}{df_{full}}$$

(Nr. / Den.)

$= (4.1784) \div (0.7953)$

$= 5.2539$ ✓

$f_c = F_{16,17,0.1} = 1.8997 \approx 1.9$ (critical value)

$F > f_c \Rightarrow$ reject null hypothesis ✓

1

Question 3 [4+4=8] Consider the *Question 3 Exam 2* data. These data come from Table 8.1 in Section 8.1 of the textbook. *The questions to follow have* **nothing** *to do with logistic regression.*

Set $z_i = y_i/m_i$ and

$$V_i = arcsin(\sqrt{z_i}). \tag{1}$$

It is known that $var(V_i)$ is, to good approximation, of the form C/m_i, for some constant C which does not depend upon x and that, also to good approximation,

$$E[V_i|x] = arcsin(\sqrt{\theta_i(x)})$$

where $\theta_i(x) = E[z_i|x]$. The following linear model has been proposed:

$$arcsin(\sqrt{\theta_i(x)}) = \gamma_0 + \gamma_1 x$$

where x denotes the Michelin Guide food rating.

3.1 Find the least squares estimates of γ_0 and γ_1 and of the constant C. Explain briefly how you go about this.
Note: the arcsine function is denoted by asin in ***R***.

$\gamma_0 = -2.94$ ×
$\gamma_1 = 0.244$ ×
$C = 0.0834$ ×

1

$l \leftarrow asin\left(\sqrt{\left(\frac{y}{m}\right)}\right) * \sqrt{m}$

$\ln(1 \sim x)$

What about x - is it also multiplied by $\sqrt{m}$?

3.2 Notwithstanding your answers in **3.1**, **assume for this part of the question** *that*

$$\hat{\gamma}_0 = 1.75,\ \hat{\gamma}_1 = 0.117,\ \hat{C} = 0.254.$$

Find a **90%** *prediction* interval for z^*, the predicted z-value for a new restaurant with food rating $x^* = 21$.

You need to show some numbers / formulas used to get partial credit.

prediction interval for V_i

$(-0.1396 \times\ ,\ 4.5076 \times)$

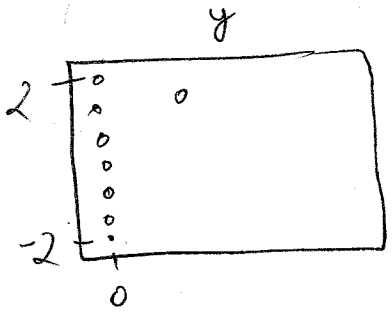
?

$z_i = y_i/m_i \rightarrow \left(\left[\arcsin^{-1}(-0.1396)\right]^2, \left[\arcsin^{-1}(4.5076)\right]^2\right)$

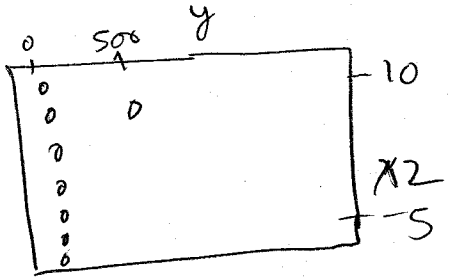
5

Question 4 **[2+3=5]** The *Question 4 Exam 2* data show the numerical values of a response variable y and two covariates x_1 and x_2. Show a plot which indicates that the response variable y should be transformed if a linear regression model is to be fit (a rough, hand drawn version, is acceptable). Then find an appropriate transformation. *You need not show any R or SAS code.*

Scatter plot of y v/s x_1 and y v/s x_2 should be linear. but it is concentrated around 0.

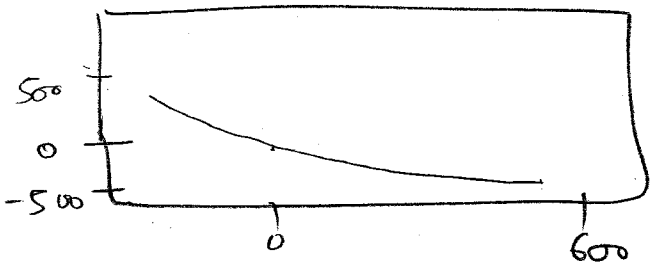


These two plots are not appropriate.



Also,

Residuals v/s fitted value



Wide variation
Non constvariance

2

$y^{0.0726}$ ✓ for response variable

i.e. $\log y$ Using inverse repone plot

3

4

Question 5 [4+3=7]

In the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ ($\mathbf{X}$ is $n \times q$ and $\boldsymbol{\beta}$ is $q \times 1$) denote the true, unknown to you, value of the parameter vector $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^0$. The residual vector is defined by $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ denotes the least squares estimator of $\boldsymbol{\beta}^0$. *Assume that* $\mathbf{X}'\mathbf{X} = \mathbf{I}$, *the identity matrix.*

5.1 Show that

$$\mathbf{e}'\mathbf{e} = \hat{\mathbf{e}}'\hat{\mathbf{e}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0).$$

5.2 Suppose $q = 3$. Then there are *three* possible fitted models if exactly *two* predictors are used. Given that

$$\hat{\beta}_0 = 1.1,\ \hat{\beta}_1 = 2.2,\ \hat{\beta}_2 = 0.5,$$

are the least squares estimators when all three predictors are used, write the equation of the best fitting among the three *two -predictor* models. Explain *clearly* your argument.

(Unless stated, all matrices/vectors in 5.1)

(5.1) $e'e = (Y - X\beta^0)'(Y - X\beta^0)$

$= (\hat{e} + X\hat{\beta} - X\beta^0)'(\hat{e} + X\hat{\beta} - X\beta^0)$ ✓

$= \hat{e}'\hat{e} + \hat{e}'X(\hat{\beta} - \beta^0) + (\hat{\beta}' - \beta^{0\prime})X'\hat{e} + (\hat{\beta} - \beta^0)'(\hat{\beta} - \beta^0)$

scalar ⇒ No effect of overall transpose

[Using $X'X = I$] ✓

$= \hat{e}'\hat{e} + (\hat{\beta} - \beta^0)'(\hat{\beta} - \beta^0) + 2\hat{e}'X(\hat{\beta} - \beta^0)$

Least square estimator is $\perp$ to $X \Rightarrow \hat{e}'X = 0$

$\Rightarrow e'e = \hat{e}'\hat{e} + (\hat{\beta} - \beta^0)'(\hat{\beta} - \beta^0)$ ✓

4

(5.3) $1.1x_1 + 2.2x_2 + 0.5x_3$

$1.1(x_1 + 2x_2) + 0.5x_3$?

Best fitting should be

$1.1(x_1 + 2x_2) + 0.5x_3$

$\beta_0(x_1 + 2x_2) + \beta_2 x_3$

Correlation among covariates ↑ variance of estimated coefficients

if x_1 and x_2 are taken separately there will be ↓ correlation high.

Space for rough work or additional calculations

Space for rough work or additional calculations

$$V_i = \gamma_0 + \gamma_1 x$$

$$\downarrow \quad \downarrow \quad \uparrow$$

$$\hat{\gamma}_0 \quad \hat{\gamma}_1 \quad x^*$$

predict int for V_i

$$\_ \leq \arcsin\left(\sqrt{\sigma_i(x)}\right) \propto \leq \_$$

$$(\arcsin^{-1} \_)^2 \leq \qquad \leq (\arcsin^{-1} \_)^2$$

$$e'e = (Y - X\beta^0)'(Y - X\beta^0)$$

$$= (\hat{e} + X\hat{\beta} - X\beta^0)'(\hat{e} + X\hat{\beta} - X\beta^0)$$

$$= (\hat{e}' + \hat{\beta}'X' - \beta^{0\prime}X')(\hat{e} + X\hat{\beta} - X\beta^0)$$

$$= [\hat{e}' + (\hat{\beta}' - \beta^{0\prime})X'][\hat{e} + X(\hat{\beta} - \beta^0)]$$

$$= \hat{e}'\hat{e} + \hat{e}'X(\hat{\beta} - \beta^0) + (\hat{\beta}' - \beta^{0\prime})X'\hat{e}$$

$$+ (\hat{\beta}' - \beta^{0\prime})(\hat{\beta} - \beta^0) \quad (\because X'X = I)$$

$\hat{e}'X(\hat{\beta} - \beta^0)$ is scalar $\Rightarrow$ transpose no effect

$$\Rightarrow \hat{e}'\hat{e} + (\hat{\beta}' - \beta^{0\prime})'(\hat{\beta} - \beta^0) + 2\hat{e}'X(\hat{\beta} - \beta^0)$$