KING FAHD UNIVERSITY OF PETROLEUM & MINERALS, DHAHRAN, SAUDI ARABIA
DEPARTMENT OF MATHEMATICS

STAT 510: Regression Analysis

Term 221, Second Major Exam
Monday November 21, 2022, 06:00 PM

Name: ______________________________ ID #: __________________

Question No	Full Marks	Marks Obtained
1	**06**	
2	**06**	
3	**02**	
4	**03**	
5	**07**	
6	**06**	
Total	**30**	

Instructions:

1. Formula sheet will be provided to you in exam. You are not allowed to bring, with you, formula sheet or any other printed/written paper.
2. Mobiles are not allowed in exam. If you have your **mobile** with you, **turn it off** and put it **on the table** so that it is visible to proctor.
3. Show all the calculation steps. There are points for the steps so if your miss them, you lose points. For multiple choice type questions, showing calculation steps is not required.
4. Derive every result that you use in your solution, unless mentioned otherwise.
5. Anything bold in a question indicates that it is a vector or matrix.
6. Report your answers up to at least 4 decimal points.

[Blank Page I]

Answer Q1 – Q4 using the multiple linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where the OLS estimates are given as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'$.

Q1: (6 pts.) Mathematically show that the correlation between the two residuals e_i and e_j can be written entirely in terms of elements of hat matrix $\forall\ i \neq j$.
Note: You can use the following results without deriving: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$

Q2: (6 pts.) Mathematically show that the Cook's distance $D_i = \frac{[\hat{\beta}-\hat{\beta}_{(i)}]' X' X [\hat{\beta}-\hat{\beta}_{(i)}]}{(k+1)MSE}$ is equal to $\frac{r_i^2}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right)$. Also show that $D_i = \frac{\sum_{i=1}^{n}\left(\hat{y}_i-\hat{y}_{(i)}\right)^2}{(k+1)MSE}$.

Note: You can use the following results without deriving: $r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$, $\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}\right] = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i}{1-h_{ii}}$

Q2 continues….

Q3: (2 pts.) Suppose a scientist has the data on response variable y and predictors X_1, X_2 and X_3. She believes that the true relationship between the response variable is intrinsically linear and given as:

$$y = \left[\ln\left(\beta_0 + \frac{\beta_1}{X_1} + \beta_2 \ln X_2 + \beta_2 X_3\right)\right]^2$$

Transform the variables such that the relationship becomes linear. The new variables are

$y' =$________________, $X_1' =$________________, $X_2' =$________________, $X_3' =$________________

Q4: (3 pts.) An analyst fits model 1: $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$ to a sample of size n and obtains the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_4 X_{4i}$ through the method of ordinary least squares (OLS). The residuals of model 1 are defined as $e_i = (y_i - \hat{y}_i)$.

After obtaining the fitted values ($\hat{y}_i$) from model 1, the analyst fits model 2: $y_i = a + b\hat{y}_i + \varepsilon_i$. In the first exam, we proved that $\hat{a} = 0$, $\hat{b} = 1$ and hence the fitted values of model 2 are equal to the fitted values of model 1 i.e. $\hat{y}_i^* = \hat{y}_i$.

Mathematically show that the coefficient of determination for model 2 is equal to the coefficient of determination for model 1.

[Blank Page II]

[Blank Page III]

Q5: Code _____ Data represent a sample of n = 43 college male measured at ten different heights. There are multiple weight observations at most of the heights, which are measured to the nearest inch. Fit a regression model regression keeping Weight as the response variable and Height as a predictor. Test for the possible lack of fit using F test.

(1 pt.) H0: ____

(1 pt.) H1: ____

(1 pt.) F = __________________________ (1 pt.) p-value = _______________________

(1 pt.) Decision:
(A) Reject H_0
(B) Accept H_0
(C) Fail to reject H_0
(D) Fail to accept H_0

(2 pt.) Conclusion:

Q6: Code _____ A scientist has the data on response variable y and predictors X_1, X_2 and X_3. He believes that the true relationship between the response variable is $y = \left[\ln\left(\beta_0 + \frac{\beta_1}{X_1} + \beta_2 \ln X_2 + \beta_3 X_3\right)\right]^2$. Transform the variables such that the relationship becomes linear. The new variables are

$$y' = e^{\sqrt{y}}, \quad X_1' = \frac{1}{X_1}, \quad X_2' = \ln X_2, \quad X_3' = X_3$$

(a) (1 pt.) Fit a linear regression model on the transformed variables. The transformed fitted model is given as:

$\hat{y}' =$ __________ + __________X_1' + __________X_2' + __________X_3'

(b) (3 pts.) Predict the original response y when $x_{10} = 66, x_{20} = 7, x_{30} = 12$. Also construct a 90% prediction interval for original y when $x_{10} = 66, x_{20} = 7, x_{30} = 12$.

$\hat{y}_{X=x_0}$ =___________ Critical value = ______________

Lower Prediction Limit = ___________ Upper Prediction Limit = ___________

(c) (2 pts.) Is the prediction done in part (b) interpolation or extrapolation? Justify your answer with a valid procedure.

Good Luck

$$S_{XX}=\sum x^2-\frac{1}{n}(\sum x)^2,\quad S_{YY}=SST=\sum y^2-\frac{1}{n}(\sum y)^2,\quad S_{XY}=\sum xy-\frac{1}{n}(\sum y)(\sum x)$$

$$\hat\beta_1=\frac{S_{XY}}{S_{XX}},\quad \hat\beta_0=\bar y-\hat\beta_1\bar x,\quad r=\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}},\quad SS_R=\hat\beta_1 S_{XY},\quad \hat\sigma=MSE=\frac{SSE}{n-2},\quad R^2=\frac{SS_R}{SS_T},$$

$$\hat\beta_0\sim N\left(\beta_0,\sigma^2\left[\frac{1}{n}+\frac{\bar x^2}{S_{xx}}\right]\right),\quad \hat\beta_1\sim N\left(\beta_1,\frac{\sigma^2}{S_{xx}}\right),\quad \hat\mu_{y|x=x_0}\pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_0-\bar x)^2}{S_{xx}}\right)}$$

$$\hat{\boldsymbol\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},\ \ SST=\boldsymbol{y}'\boldsymbol{y}-\frac{(\sum y_i)^2}{n},\ \ SSE=\boldsymbol{y}'\boldsymbol{y}-\hat{\boldsymbol\beta}'\boldsymbol{X}'\boldsymbol{y},\ \ V-Cov(\hat{\boldsymbol\beta})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$V-Cov(\boldsymbol{e})=\sigma^2(\boldsymbol{I}-\boldsymbol{H}),\quad MSE=\frac{SSE}{n-k-1},\quad R^2_{adj}=1-\frac{SSE/(n-k-1)}{SST/(n-1)},\quad VIF_j=\frac{1}{1-R_j^2},$$

$$T_{n-k-1}=\frac{\hat\beta_j-\beta_{j0}}{s.e.(\hat\beta_j)},\quad F=\frac{SSR/k}{SSE/(n-k-1)},\quad F=\frac{(\boldsymbol{T}\hat{\boldsymbol\beta}-\boldsymbol{c})'\left[\boldsymbol{T}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{T}'\right]^{-1}(\boldsymbol{T}\hat{\boldsymbol\beta}-\boldsymbol{c})/r}{SSE(FM)/(n-k-1)},\quad \boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

$$\hat\mu_{y|\boldsymbol{x}=\boldsymbol{x_0}}\pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MSE(\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0})},\qquad \hat y_0\pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MSE(1+\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0})}$$

$$w_{ij}=\frac{x_{ij}-\bar x_j}{\sqrt{s_{jj}}},\ y_i^0=\frac{y_i-\bar y}{\sqrt{s_{yy}}},\ s_{jj}=\sum_{i=1}^n(x_{ij}-\bar x_j)^2,\ s_{yy}=\sum_{i=1}^n(y_i-\bar y)^2\ \forall\ i=1,2,\dots,n \text{ and } j=1,2,\dots,k$$

$$d_i=\frac{e_i}{\sqrt{MSE}},\qquad r_i=\frac{e_i}{\sqrt{MSE(1-h_{ii})}},\qquad t_i=\frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}} \text{ where } S_{(i)}^2=\frac{(n-k-1)MSE-\frac{e_i^2}{(1-h_{ii})}}{n-k-2}$$

$$\sum_{i=1}^m\sum_{j=1}^{n_i}(y_{ij}-\hat y_i)^2=\sum_{i=1}^m\sum_{j=1}^{n_i}(y_{ij}-\bar y_i)^2+\sum_{i=1}^m n_i(\bar y_i-\hat y_i)^2,\quad F=\frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)}$$

$$y^{(\lambda)}=\begin{cases}\frac{y^\lambda-1}{\lambda\dot y^{\lambda-1}}, & \lambda\neq 0\\ \dot y\ln y, & \lambda=0\end{cases},\quad SS^*=SSE_{(\hat\lambda)}\left(e^{\left(\chi^2_{\alpha,1}/n\right)}\right),\quad \hat\alpha_1=\hat\alpha_0+\frac{\hat\gamma}{\hat\beta_1}$$

$$\hat{\boldsymbol\beta}_{WLS}=(\boldsymbol{X_1}'\boldsymbol{X_1})^{-1}\boldsymbol{X_1}'\boldsymbol{y_1},\quad \boldsymbol{X_1}=\sqrt{w}\boldsymbol{X},\quad \boldsymbol{y_1}=\sqrt{w}\boldsymbol{y},\quad |e_i|=a+b\hat y_i+\varepsilon_i,\quad w_i=1/\left(\widehat{|e_i|}\right)^2$$

Measure of influence	**Critical value(s)**
$D_i=\frac{(\hat{\boldsymbol\beta}_{(i)}-\hat{\boldsymbol\beta})'X'X(\hat{\boldsymbol\beta}_{(i)}-\hat{\boldsymbol\beta})}{MSE(k+1)}=\left(\frac{r_i^2}{k+1}\right)\frac{h_{ii}}{(1-h_{ii})}$	1
$DFBETAS_{j,i}=\frac{\hat\beta_j-\hat\beta_{j(i)}}{\sqrt{S_{(i)}^2\times C_{jj}}}=\frac{r_{j,i}}{\sqrt{\boldsymbol{r_j}'\boldsymbol{r_j}}}\frac{t_i}{\sqrt{1-h_{ii}}}$ where $\boldsymbol{r_j}'$ is the j^{th} row of $\boldsymbol{R}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$	$2/\sqrt{n}$
$DFFITS_i=\frac{\hat y_i-\hat y_{(i)}}{\sqrt{S_{(i)}^2\times h_{ii}}}=t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$	$2\sqrt{\frac{k+1}{n}}$
$COVRATIO_i=\frac{\left\lvert\left(X_{(i)}'X_{(i)}\right)^{-1}S_{(i)}^2\right\rvert}{\lvert(X'X)^{-1}MSE\rvert}=\left(\frac{S_{(i)}^2}{MSE}\right)^{k+1}\left(\frac{1}{1-h_{ii}}\right)$	$1\pm 3\left(\frac{k+1}{n}\right)$