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

STAT 510: Regression Analysis

Term 221, Final Exam
Saturday December 31, 2022, 12:30 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**12**	
2	**06**	
3	**03**	
4	**05**	
5	**04**	
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 in mathematical part. 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 at least 3 decimal points of your numerical answers.**

Code 1

Q1: (1x12 = 12 pts.) Multiple choice questions.

(1) In a multiple linear regression model, the estimator for σ^2 is $\frac{SSE}{n-k-1}$. What is the reason for choosing $(n-k-1)$ as divisor?

A. To make the resulting estimator unbiased.
B. To minimize the sampling variance.
C. To minimize the sum of squares of errors.
D. To maximize the likelihood function.
E. To minimize the error variance.

(2) The estimated slope of a simple linear regression model represents the

A. average value of Y when $X = 0$.
B. average change in Y for a unit change in X.
C. average difference between the values of Y and X.
D. percentage variation in Y explained by X.
E. standard deviation in response variable Y.

(3) Which of the following statements is true for the model selection criteria $AIC = n\ln\frac{SSE}{n} - 2(k+1)$ and $SBC = n\ln\frac{SSE}{n} - (k+1)\ln n$?

A. The penalty is higher for SBC when sample size is large.
B. The penalty is higher for AIC when sample size is large.
C. AIC is always larger than SBC.
D. SBC is always larger than AIC.
E. Both are equal for same number of predictors.

(4) Which one of the following is **not** true about all possible regressions?

A. Mallow's C_p is usually plotted against $(k+1)$.
B. Adjusted R^2 is always less than R^2.
C. $R^2_{\text{prediction}}$ determines how well the model predicts new observations.
D. Maximizing adjusted R^2 is equivalent to minimizing SSE.
E. Minimizing prediction error sum of squares $\left(\sum e^2_{(i)}\right)$ is equivalent to maximizing $R^2_{\text{prediction}}$.

(5) In a multiple linear regression model, the generalized F statistic $\frac{(T\hat{\beta}-c)'\left[T(X'X)^{-1}T'\right]^{-1}(T\hat{\beta}-c)/r}{SSE(\text{Full})/(n-k-1)}$ can be used for

A. testing the significance of single predictor
B. testing the significance of some predictors
C. testing the significance of all predictors
D. testing linear constraint(s) on the model
E. all of above

(6) What does "centering the predictors" do to the regression?

A. It removes all the multicollinearity
B. It fixes the non-constant variance
C. It fixed the non-normality
D. It removes the excess multicollinearity
E. It brings the variance of response variable close to one.

(7) Which one of the following is true about the link between regression and analysis of variance (ANOVA)?

A. Any linear regression problem can be treated as an ANOVA problem.
B. Completely randomized design with t treatments can be written as a regression model with $(t-1)$ predictors.
C. For any ANOVA problem, it is always recommended to use regression technique.
D. For writing a regression model for ANOVA problem, all predictors are numerical variables.
E. None of above.

(8) In multiple linear regression, which one of the following can be used to differentiate between interpolation and extrapolation?

A. Adjusted R^2
B. Partial F test
C. Variance Inflation Factor
D. Normal QQ plot
E. Hat matrix

(9) Identify which statements is **not** true about the variable selection in model building?

A. It is the most common corrective technique for multicollinearity.
B. Finding best regression through model building is a compromise between as many predictors as possible and as few predictors as possible.
C. Forward selection is preferred over the stepwise regression.
D. All possible regressions is one of the model building techniques.
E. Different model building techniques may produce different models.

(10) Which one of the following is true about the PRESS residuals $e_{(i)} = y_i - \hat{y}_{(i)}$?

A. It is always equal to e_i if calculated using the same model.
B. It is always smaller than e_i in absolute terms.
C. It always stays between 0 and 1.
D. It measures the impact of heteroskedasticity.
E. It is usually higher than e_i in absolute terms if the corresponding observation is outlier.

(11) Mallow's $C_p = \frac{SSE}{\hat{\sigma}^2} - n + 2(k+1)$ estimates

A. the likelihood of fitted values being equal to the actual response.
B. percentage of variation in y that is explained by the regression.
C. the standardized total mean square error of fitted values.
D. the bias of sum of squares of errors.
E. the total variance of error term.

(12) Circle the one that is **not** a possible source of multicollinearity?

A. Variable elimination
B. Data collection method
C. Constraint(s) on the model
D. Choice of the model
E. Over-defined model

Name: _____ ID #: _____

Q2: (3 + 3 = 6 pts.) Consider a multiple regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, $V - Cov(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$ and $\boldsymbol{V}_{n\times n}$ is not an identity matrix. Using square root matrix $\boldsymbol{K}$ such that $\boldsymbol{K'K} = \boldsymbol{KK} = \boldsymbol{V}$, define a new set of variables $\boldsymbol{y_1} = \boldsymbol{K^{-1}y}$, $\boldsymbol{X_1} = \boldsymbol{K^{-1}X}$ and $\boldsymbol{\epsilon_1} = \boldsymbol{K^{-1}\epsilon}$. The new model is given as $\boldsymbol{y_1} = \boldsymbol{X_1\beta} + \boldsymbol{\epsilon_1}$.

a) Prove that $V - Cov(\boldsymbol{\epsilon_1}) = \sigma^2\boldsymbol{I}$ where $\boldsymbol{I}_{n\times n}$ is identity matrix. Also, show that the variances of error terms in new model are constant and the covariances are zero.

b) Using the new model $\boldsymbol{y_1 = X_1\beta + \epsilon_1}$, derive an estimate of $\boldsymbol{\beta}$ and express it in terms of original variables $\boldsymbol{X}$ and $\boldsymbol{y}$.

Q3: (3 pts.) Suppose that the true regression model was $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ but you ignored the second variable (X_2) and fitted a simple linear regression model $y = \gamma_0 + \gamma_1 X_1 + \varepsilon$. The least square estimators are as follows:

$\hat{\gamma}_1 = \frac{S_{X_1 Y}}{S_{X_1 X_1}}$ and $\hat{\gamma}_0 = \bar{y} - \hat{\gamma}_1 \bar{X}_1$ where $S_{XY} = \sum_{i=1}^{n}(y_i - \bar{y})(X_{1i} - \bar{X}_1)$ and $S_{X_1 X_1} = \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2$.

Is the least square estimator of slope ($\hat{\gamma}_1$) unbiased estimator of β_1? If no, then find the bias of $\hat{\gamma}_1$.

Hint: $\text{Bias}(\hat{\gamma}_1) = \text{E}(\hat{\gamma}_1) - \beta_1$

[Blank page]

Report at least 3 decimal points of your numerical answers.

Name: _____ ID #: _____

Q5: Code _____

Data on the thrust of a jet turbine engine and six candidate regressors are given in the Excel sheet with $n = 32$.

(a) (2 pts.) Use all the 6 predictors and run a ridge regression with $K = 0.03$. The adjusted R^2 of the fitted ridge regression model is equal to _____.

(b) (3 pts.) Using the backwards elimination with $\alpha_{OUT} = 0.1$, find the best model. Also, calculate C_p and AIC for your **final model**.

Final model: $\hat{y} =$ _____

$C_p =$_____ , $AIC =$_____

Q6: Code _____

Download the Excel file for this question containing the data on two variables y and X. Fit a linear spline to these data to predict y using two knots i.e. $k_1 = 15$ and $k_2 = 29$. The line should be continuous at the knots.

Fit the model: $y = \beta_0 + \beta_1 X + \gamma_1 S_1 + \gamma_2 S_2 + \epsilon$ where

$S_1 = \begin{cases} x - 15, & x > 15 \\ 0, & x \leq 15 \end{cases}$ and $S_2 = \begin{cases} x - 29, & x > 29 \\ 0, & x \leq 29 \end{cases}$

(1 pt.) Fitted model: $\hat{y} =$ _____

(1 pt.) Also, predict y when $x = 25$. Predicted y when $x = 25$ is equal to _____

(2 pts.) A 99% prediction interval for y when $x = 25$ is given as [_____ , _____]

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

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \quad r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}, \quad SS_R = \hat{\beta}_1 S_{XY}, \quad \hat{\sigma} = MSE = \frac{SSE}{n-2},$$

$$R^2 = \frac{SS_R}{SS_T}, \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right), \quad \hat{\mu}_{Y|x_0} \sim N\left(\mu_{Y|x_0}, \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]\right)$$

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

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

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

$$\hat{\mu}_{y|x=x_0} \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MSE(\boldsymbol{x_0'(X'X)^{-1}x_0})}, \qquad \hat{y}_0 \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MSE(1 + \boldsymbol{x_0'(X'X)^{-1}x_0})}$$

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

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

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

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y}\ln y, & \lambda = 0 \end{cases}, \quad SS^* = SSE_{(\hat{\lambda})}\left(e^{\left(\chi^2_{\alpha,1}/n\right)}\right), \quad \hat{\alpha}_1 = \hat{\alpha}_0 + \frac{\hat{\gamma}}{\hat{\beta}_1}, \quad \widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y}$$

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

$$C_p = \frac{SSE}{\hat{\sigma}^2} - n + 2(k+1), \; AIC = n\ln\left(\frac{SSE}{n}\right) + 2(k+1), \; BIC = n\ln\left(\frac{SSE}{n}\right) + (k+1)\ln n$$

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