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

STAT 310: Regression Analysis

Term 221, Final Exam
Sunday December 25, 2022, 07:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**15**	
2	**04**	
3	**03**	
4	**02**	
5	**06**	
6	**05**	
Total	**35**	

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 4 decimal points of your numerical answers.**

Code 1

[Blank page]

Q1: (1x15 = 15 pts.) Multiple choice questions.

(1) In multiple linear regression, which test is used for testing the significance of whole model?

A. T-test
B. F-test
C. Z-test
D. Chi-square test
E. Durbin-Watson test

(2) How do we accommodate a categorical predictor in the regression line?

A. Using unit length scaling
B. Using unit normal scaling
C. Using indicator variable(s)
D. Using ridge regression
E. Using weighted least squares method

(3) Why is the number of indicator variables to be entered into the regression model always equal to the number of categories (c) minus 1?

A. To avoid the situation of perfect multicollinearity
B. To control for other variables in the model
C. To increase the R-squared value
D. To fix the violation of constant variance assumption
E. To reduce the effect of influential observation

(4) Weighted least squares method is

A. used when the assumption of independence is violated
B. a special case of ridge regression
C. used when the assumption of normality is violated
D. used to find the optimal transformation on response variable
E. a special case of generalized least squares

(5) In a linear regression model $y = \beta_0 + \beta_1 X + \gamma_1 I + \epsilon$ with X as a continuous predictor and I as an indicator variable, how do we interpret the value of γ_1?

A. the difference between the $y-$ intercepts
B. the difference between β_0 and β_1
C. the difference between the two R-square values
D. the difference between the slopes
E. average change in y due to a unit change in X

(6) Ill-conditioning of $\boldsymbol{X'X}$ matrix can occurs due to

A. heteroskedasticity
B. narrow range of X variables
C. non-constant variance
D. large value of intercept
E. small value of intercept

(7) In multiple linear regression, which one of the following in an indication of the presence of multicollinearity?

A. high coefficient of determination of the model
B. large values on the diagonal elements of hat matrix
C. large values on the diagonal of $(\boldsymbol{W'W})^{-1}$ matrix
D. strong correlation between the response and predictor variables
E. large values on the diagonal of $\boldsymbol{X'y}$ matrix

(8) With reference to variable selection and model building techniques, which one of the following is true?

A. A predictor can be removed from the model in forward selection method
B. All insignificant predictors are removed in the first step of backwards elimination
C. All insignificant predictors are added in the first step of forward selection
D. Stepwise regression is likely to give a model with more predictors as compared to forward selection
E. Stepwise regression is likely to give a model with less predictors as compared to backwards elimination

(9) In case of polynomial regression, which one of the following techniques can be used to reduce the multicollinearity?

A. Normalizing
B. Squaring
C. Interaction
D. Centering
E. Ill-conditioning

(10) A lack of fit test can only be applied in the presence of

A. replicates
B. multicollinearity
C. normality
D. heteroskedasticity
E. at least 2 predictors

(11) Ridge regression is applied when

A. there is heteroskedasticity
B. the error variance is not constant
C. the error term is not normally distributed
D. there is multicollinearity
E. there are not enough data

(12) With reference to variable selection and model building techniques, which one of the following is **not** true?

A. Mallow's C_p is plotted against $k + 1$
B. MSE of the model can decrease by adding new predictor(s)
C. The model with larger AIC is better
D. SSE of the model cannot decrease by adding new predictor(s)
E. Maximizing R^2 is equivalent to minimizing SSE

(13) In a simple linear regression model $y = \beta_0 + \beta_1 X + \epsilon$ for predicting the house price (y) using the area of the house (X) as a predictor, which one of the following is true?

A. β_0 is the average house price with no area
B. β_0 is the average change in house price due to one unit increase in covered area
C. β_0 is the average house price of all the houses in sample
D. β_0 is the average house price of all the houses in population
E. β_0 has no practical interpretation

(14) What is the problem associated with cubic spline model: $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \gamma_1 K^1 + \gamma_2 K^2 + \gamma_3 K^3 + \epsilon$ where $K = \begin{cases} 0 & \text{if } x \leq k \\ x - k & \text{if } x > k \end{cases}$ and the knot is at point k.

A. The line is not continuous
B. The line is not straight
C. The line is not linear
D. The line is not smooth
E. None of above

(15) A large value of $COVRATIO_i = \frac{\left| s_i^2 \left(\boldsymbol{X}'_{(i)} \boldsymbol{X}_{(i)} \right)^{-1} \right|}{|\text{MSE}(\boldsymbol{X}'\boldsymbol{X})^{-1}|}$ indicates that the i^{th} observation is

A. downgrading the precision of estimates
B. improving the precision of estimates
C. influencing the predicted value in a positive way
D. influencing the predicted value in a negative way
E. influencing the overall $\boldsymbol{\beta}$ vector

Name: _____ ID #: _____

Q2: (4 pts.) Consider a multiple linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\beta}$ is the vector of $(k+1)$ unknown parameters. Derive ordinary least squares estimates of $\boldsymbol{\beta}$.

Q3: (3 pts.) For model selection through "All Possible Regressions", show that maximizing the adjusted coefficient of determination $R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{SSE}{SST}\right)$ is equivalent to minimizing the MSE.

Q4: (2 pts.) A study is conducted to determine the effects of company size and the presence or absence of a safety program on the number of hours lost due to work-related accidents. A total of 28 companies are selected for the study. The variables are as follows:

$$y = \text{lost work hours}$$

$$X_1 = \text{number of employees}$$

$$X_2 = \begin{cases} 1 & \text{if safety program is used} \\ 0 & \text{if no safety program is used} \end{cases}$$

Fit a linear regression model for predicting the lost work hours based on the number of employees and the presence or absence of a safety program. This model should incorporate two lines i.e. first when safety program is used and the second when no safety program is used. Moreover, both the lines should have different intercepts and different slopes. Write down the model to be fitted for the said purpose.

Name: _____ ID #: _____

Q5: Code _____

(6 pts.) Data on the thrust of a jet turbine engine and six candidate regressors are given in the Excel sheet with $n = 32$. Using the stepwise regression with $\alpha_{IN} = \alpha_{OUT} = 0.1$, find the final model. Write down all the details of each step. Calculate $R^2_{\text{prediction}}$, C_p and AIC for the **final model**.

Final model: $\hat{y} =$

$R^2_{\text{prediction}} =$_____ , $C_p =$_____ , $AIC =$_____

Q6: Code _____ (1+2+2 = 5 pts.) A study is conducted to determine the effects of company size and the presence or absence of a safety program on the number of hours lost due to work-related accidents. A total of 28 companies are selected for the study. The variables are as follows:

$$y = \text{lost work hours}$$

$$X_1 = \text{number of employees}$$

$$X_2 = \begin{cases} 1 & \text{if safety program is used} \\ 0 & \text{if no safety program is used} \end{cases}$$

Fit a linear regression model for predicting the lost work hours based on the number of employees and the presence or absence of a safety program. This model should incorporate two lines i.e. first when safety program is used and the second when no safety program is used. Moreover, both the lines should have different intercepts and different slopes i.e.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12}(X_1 X_2) + \epsilon$$

(a) The fitted model is:

$$\hat{y} = \_\_\_\_\_ + \_\_\_\_\_ X_1 + \_\_\_\_\_ X_2 + \_\_\_\_\_ (X_1 X_2)$$

(b) Referring to the full model from part a), should we force both the lines to have equal slopes? **Justify your answer statistically.**

(c) Referring to the model from part a), check using the Box-Cox method if there is transformation (y^{λ}) needed on the response variable. Use $\alpha = 0.05$ for creating an interval estimate for λ.

Optimal λ = _____

95% confidence interval for λ: (_____ , _____)

Is the transformation needed? (i) Yes (ii) No

λ	SSE
-1	
0	
0.5	
0.9	
1	
1.1	
1.25	
1.5	
2	