Department of Mathematics and Statistics
Semester 152

STAT510 Final Exam Thursday May 12, 2016

Name: _____ ID #: _____

Question	Marks	Marks Obtained
1	14	
2	2	
3	4	
4	4	
5	4	
6	2	
7	8	
8	6	
Multiple Choice	16	
Total	60	

1) A model to predict the sale price of a house based on the 3 predictor variables:
   - X_1= Area of the living space in the house,
   - X_2 = Number of bedrooms,
   - X_3 = Area of the land.

Model 1

Source	SS	Df
Model	$2.49*10^{11}$	3
X_1	$1.18*10^{11}$	1
X_2	$0.69*10^{11}$	1
X_3	$0.62*10^{11}$	1
Error	$7.81*10^{11}$	63
Total	$10.30*10^{11}$	66

Variable	Coefficient	Standard Error
Constant	290558.1	75398.80
X_1	6187.99	2508.56
X_2	-36018.68	14665.17
X_3	2.51	1.12

Model 2

Source	SS	Df
Model	$2.49*10^{11}$	3
X_2	$4.02*10^{7}$	1
X_1	$1.87*10^{11}$	1
X_3	$0.62*10^{11}$	1
Error	$7.81*10^{11}$	63
Total	$10.30*10^{11}$	66

Variable	Coefficient	Standard Error
Constant	290558.1	75398.80
X_2	-36018.68	14665.17
X_1	6187.99	2508.56
X_3	2.51	1.12

a) Write down the estimated regression equation for the full model.

b) Test the hypothesis of the significance of the regression. Write the null hypothesis, alternative, the value of the test statistic, the decision.

c) Give values for the following:
   i) $SSR(X_2)$

   ii) $SSR(X_3|X_1, X_2)$

   iii) $SSR(X_1|X_2)$

d) In the context of the problem, explain the hypothesis $H_0: \beta_3 = 0$ given that X_1 andX_2 are in the model.

e) What is the p-value of the test of the hypothesis above?

f) What is the p-value for testing hypothesis $H_0: \beta_2 = 0$ given that X_2 andX_3 are in the model?

g) If you wanted to fit a simple linear regression using bedrooms as the only predictor, would the result be that the number of bedrooms is a significant predictor of Sale Price? Explain using information from the outputs above.

2) In examining case diagnostics in multiple regression, under what circumstance is it acceptable to remove a case that is clearly a Y outlier?

3) In the scatter plot below

Consider the two points P and Q.

a) Which point has a large studentized residual?

b) Which point has a small studentized residual?

c) Which point has small leverage?

d) Which point has a high leverage?

4) Suppose you have four predictor variables X_1, X_2, X_3, and X_4. You run a forward selection procedure and the variables are entered as follows: X_2, then X_4, then X_1, then X_3. You also run an all subsets regression analysis using R^2 as the criterion for the best model for each possible number of predictors, would the same models result from this analysis as from the forward stepwise procedure? In other words, would "all subsets regression" definitely identify the following as the best models for 1, 2, 3, and 4 variables?

Model	Best Model	Yes/No
One variable	$E(Y) = \beta_0 + \beta_1 X_2$	
Two Variables	$E(Y) = \beta_0 + \beta_1 X_2 + \beta_2 X_4$	
Three Variables	$E(Y) = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1$	
Four Variables	$E(Y) = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1 + \beta_4 X_3$	

5) A company is concerned that employees in a certain job at its Riyadh office are not being given raises at the same rate as employees in the same job at its Khobar office. Using a random sample of employees from each city, a regression model is fit with

Y = employee salary

X_1 = length of time employee has oreked for the company

X_2 = 1 if employee is in Riyadh, 0 if employee is in Khobar

New employees, who have $X_1 = 0$, all start at the same salary, so the complay is not interested in fitting a model with different intercept, only with different slopes.

a) Write the full and reduced models for determining whether the slopes are different for employees in the two cities.

6) A researcher once attempted to estimate an asset demand equation that included the following explanatory variables: current wealth W_t , wealth in the previous period W_{t-1}, and the change in wealth $\Delta W = W_t - W_{t-1}$. What problem did this researcher encounter? What should have been done to solve this problem?

7) Consider the model $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i; i = 1, \cdots, n$ where $x_{1i} = \begin{cases} 1 & i = 1, \cdots, n_1 \\ 0 & i = n_1 + 1, \cdots, n \end{cases}$

a) Write down the X'X matrix

b) Find the vector of estimators $\boldsymbol{b}$

c) Find *SSE* and *SSR*.

8) Derive the weighted least squares normal equations for fitting a simple linear regression function when $\sigma_i^2 = kX_i$ where k is a proportionality constant.

Multiple Choice Questions

9) In linear regression analysis with usual assumptions, which of the following quantities is the same for all individual units in the analysis?
   a) h_{ii}
   b) $S(Y_i)$
   c) $S(e_i)$
   d) $S(\hat{Y}_i)$

10) Which of the following is not a valid use of a regression line
   a) To estimate the average value of Y at a specified value of X
   b) To predict the value of Y for an individual, given that individual's X value.
   c) To estimate the change in Y for a one unit change in X
   d) To determine if a change in X causes a change in Y.

11) Which of the following is the best way to determine whether or not there is a statistical significant linear relationship between two quantitative variables?
   a) Fit a regression line from a sample and see if the sample slope is 0.
   b) Compute the correlation coefficient and see if it is greater in magnitude than 0.5.
   c) Fit a regression line from a sample and test the null hypothesis that the slope is 0.
   d) Fit a regression line from a sample and test the null hypothesis that the intercept is 0.

12) In a regression model with $p-1$ predictor variable chosen from a set of $P-1$ possible predictor variable, which of the following indicated that bias is not a problem with the model?
   a) Mallow's $C_p \leq p$
   b) Mallow's $C_p \leq P$
   c) Mallow's $C_p > p$
   d) Mallow's $C_p \leq P$

13) Which of the following diagnostic measures is based on Y values only?
   a) Cook's Distance
   b) Studentized Deleted Residual
   c) DFFITS
   d) None of the above.

14) Which of the following is the most appropriate method to test $H_0: \beta_k = 0 \; vs \; H_a: \beta_k > 0$?
   a) A t test
   b) An F test
   c) A test of a full versus reduced model
   d) All the above

15) When choosing between regression models it is always preferable to choose the one with
    a) The highest R^2
    b) The least number of independent variables
    c) The highest F statistic
    d) All the above but there is no set procedure to determine the most preferable regression model

16) A regression line is fit using 15 observations. The confidence interval for *y* when *x* = 23 is found to be (28.5, 31.6). If more observations are taken and the values are found to be similar to those already recorded, this confidence interval would most likely
    a) Be bigger because there are more observations
    b) Be shorter because there are more observations
    c) Change but it is not clear how because we need to know how far 23 is from $\bar{x}$
    d) Stay the same

17) If the F statistic is statistically significant, then
    a) All the t statistics will be statistically significant
    b) Some of the t statistics will be statistically significant
    c) Generally all the t statistics will be statically significant
    d) None of the above

18) Collinearity
    a) Occurs when we can determine the value of the dependent variable form a set of independent variables
    b) Occurs when we can determine the value of one independent variable from a set of other independent variables
    c) Occurs if there are too many independent variables in regression
    d) Is one of the many problems that may occur in multiple regression which violate assumptions about the residuals

19) The correlation coefficient measures
    a) How well the plot of the observations of the sample fit a line
    b) How the independent and dependent variables vary together
    c) The strength of the statistical significance of the regression line
    d) The variation of the observed values

20) If a multiple regression yields an F statistic = -0.12. It means that
   a) It is unlikely that the multiple regression is statistically significant
   b) The coefficient of determination is also negative
   c) The F statistic has been calculated incorrectly
   d) At least one of the $\beta's$ may equal 0.

21) A confidence interval for the mean response is narrower when made for values of *x* that are
   a) Closer to the mean of the *x*'s
   b) Further from the mean of the *x*'s
   c) Closer to the mean of the *y*'s
   d) Further from the mean of the *y*'s

22) A prediction interval for a new response is narrower when made for values of *x* that are
   a) Closer to the mean of the *x*'s
   b) Further from the mean of the *x*'s
   c) Closer to the mean of the *y*'s
   d) Further from the mean of the *y*'s

23) In a regression model with a dummy variable without interaction there can be
   a) More than one slope and more than one intercept
   b) More than one slope but only one intercept
   c) Only one slope but more than one intercept
   d) Only one slope and one intercept

24) Studies have shown a high positive correlation between the number of firefighters dispatched to combat a fire and the financial damages resulting from it. A politician commented that the fire chief should stop sending so many firefighters since they are clearly destroying the place. This is an example of
   a) Extrapolation
   b) Dummy variables
   c) Misuse of causality
   d) Multicollinearity