BANA 7052: Applied Linear Regression

Final exam (60 points)

**Instructions:** The final is take home and open book. It will be due by noon on Friday, December 7-th. Please upload your competed final to blackboard. This is **NOT** a group assignment and absolutely **NO** working together. Failure to do so will result in a zero for the final. There is a bonus question at the end worth five points.

## Part I: multiple choice

**Please clearly circle your multiple choice answers**

**Question 1:** In a multiple linear regression model, which of the following are considered as random?

1. The continuous response or dependent variable
2. The expected value of
3. The unknown parameter
4. The unknown error variance

**Question 2:** When developing a linear regression model, adding additional predictors to the model will

1. Always increase
2. Always increase adjusted
3. Sometimes decrease
4. Always decrease adjusted

**Question 3:** What is the best way to identify multicollinearity?

1. Including interaction terms in the linear regression model
2. Variance inflation factors (VIFs)
3. Residual diagnostic plots
4. Adjusted

**Question 4:** The definitions for AIC and BIC (or SBC) are:

where is the log likelihood, is the number of parameters, is the number of observations (i.e., the sample size). Which one of the following statements is correct?

1. AIC penalizes complex models more than BIC
2. One can only use AIC as a model selection criterion in stepwise regression
3. For the same model, if the AIC is less than the BIC then it is an indication that you have found a good model
4. When BIC is used as the model selection criterion, usually fewer independent variables are included in the model compared with the model selected using AIC

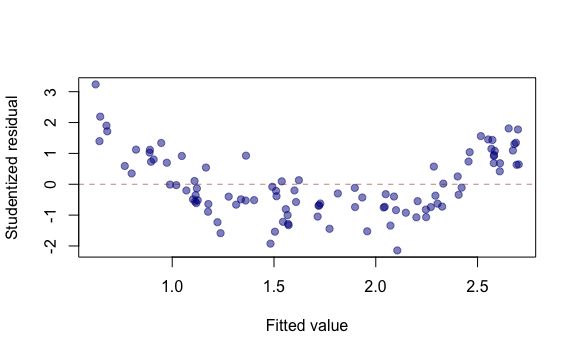
**Question 5:** In simple linear regression, a confidence interval for the mean response is narrower for values of that are:

1. Closer to the mean of the the ’s
2. Further from the mean of the ’s

**Question 6:** When building a linear regression model of the form for a data set, the presence of multicollinearity in the data may result in \_\_\_\_\_\_\_\_\_ standard errors for the coefficient estimates than if the data came from an orthogonal design (i.e., when and are uncorrelated).

1. Smaller
2. Larger

**Question 7:** What does the following residual versus fitted value plot suggest about a model between a single predictor and ?



1. Heteroscedasticity of the error term
2. A nonlinear between and
3. Non-normality of the error term
4. Satisfactory residuals

**Question 8:** When a linear regression model is being developed, adding additional variables to the model will

1. always decrease model SSE
2. always decrease model AIC
3. always increase model adjusted
4. always decrease model MSE

**Question 9:** The ordinary residuals refer to

**Question 10:** For a fitted simple linear regression model, which one of the following properties is **NOT** true?

1. The fitted regression line passes through the point
2. The residuals sum to zero:
3. The residuals, ,. are always independent

**Question 11:** The diagonal elements of hat matrix, also referred to as the *hat values* or *leverage values*, measures the influence of observation on the regression line when removing observation .

1. True
2. False

**Question 12:** In a regression study, a 95% confidence interval for was given as: . What does this confidence interval mean?

1. The interval contains the true with 95% probability
2. 95% of all possible ’s are in
3. 95% of the interval contains the true
4. If we were to repeat the experiment many times (i.e., repeatedly take a new sample of size and compute the same confidence interval), roughly 95% of the generated intervals would contain the true

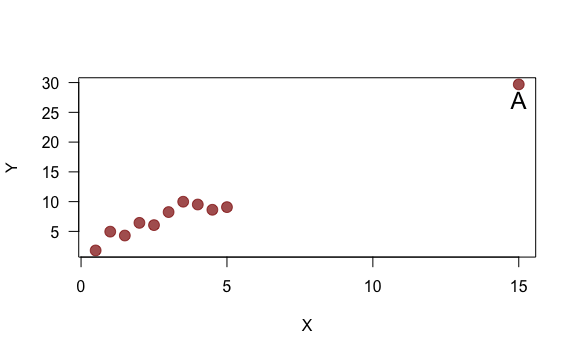
**Question 13:** In a regression study, a 95% confidence interval for was given as: . Which of following is correct?

1. The 90% interval will be wider than the 95% interval
2. The 99% interval will be wider than the 95% interval

**Question 14:** In simple linear regression, the confidence interval for the mean response is always narrowest around

1. True
2. False

**Question 15:** Point A in the far right is likely to be



1. High influential point of the regression line
2. High leverage point

**Question 16:** Suppose you have built a multiple linear regression model according to . If you were to test versus you would use

1. -test
2. -test based on an ANOVA
3. Partial -test

**Question 17:** In simple linear regression, assume we used a -test for , and we rejected the null hypothesis at the specified level. We would conclude that:

1. The linear regression model is useful compared with
2. There is no linear relationship between and
3. The relationship between and is quadratic
4. There is no relationship between and

**Question 18:** Which of the following cases lead to multicollinearity?

1. There are indicator variables being used as predictors
2. A predictor can be expressed as a linear combination of other predictors
3. The variances across all predictors are not the same
4. The predictors are not normally distributed and are positively skewed

**Question 19:** What does the last line of output “F-statistic….p-value: …” indicate in the following multiple regression output?

##   
## Call:  
## lm(formula = mpg ~ cyl + disp + hp + wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.0562 -1.4636 -0.4281 1.2854 5.8269   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.82854 2.75747 14.807 1.76e-14 \*\*\*  
## cyl -1.29332 0.65588 -1.972 0.058947 .   
## disp 0.01160 0.01173 0.989 0.331386   
## hp -0.02054 0.01215 -1.691 0.102379   
## wt -3.85390 1.01547 -3.795 0.000759 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.513 on 27 degrees of freedom  
## Multiple R-squared: 0.8486, Adjusted R-squared: 0.8262   
## F-statistic: 37.84 on 4 and 27 DF, p-value: 1.061e-10

1. All of the coefficients are significantly different from 0 at the level
2. None of the independent variables explains any of the variation in
3. At least one of the independent variables explains some of the variation in
4. The model explained 1.061e-10 of the variability in

**Question 20:** If one wishes to incorporate seasonal dummy variables for monthly data into a regression model, how many dummy variables should be in the model?

1. 12
2. 11
3. 10
4. 1

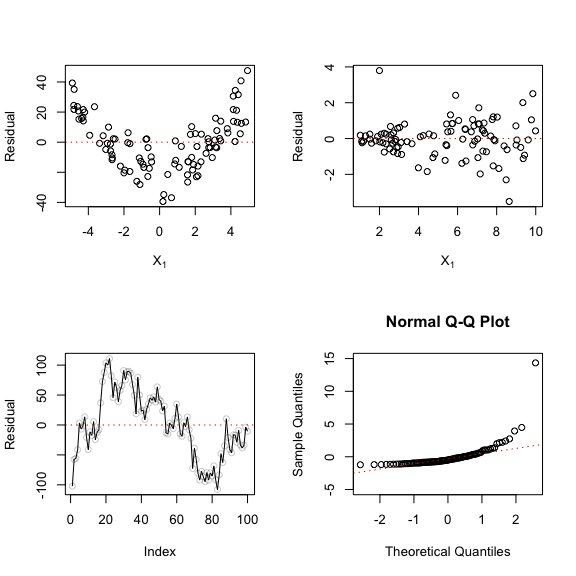
## Part II: short answer response

For the multiple linear regression model

**Question 21:** (2 pts.) Recall that the variance inflation factor for is defined as . Explain in one sentence what is. Explain in one sentence why it is better to use VIFs than pairwise correlations when trying to detect the presence of multicollinearity in a regression data set.

**Question 22:** (2 pts.) For variable selection criteria based on residuals, among , , , , and , which are “good” criteria for variable selection? Explain your reason in two sentences.

**Question 23:** (4 pts.) Please state one possible violation of standard model assumptions for each of the residual plots below (one sentence for each plot):



**Question 24:** (4 pts.) Suppose you want to build a linear regression for response variable weight ( ) using covariate height ( ) and gender ( for female and for male).

1. Suppose you want to allow the slope (of height) to be the same for different gender but different intercept, how would you build the linear regression model? Please specify the model in one line.
2. Suppose you want to allow both slope (of height) and intercept to be different for different gender, how would you build the linear regression model? Please specify the model in one line.

**Question 25:** (2 pts.) An engineer has stated: “Reduction of the number of candidate explanatory variables should always be done using the objective forward stepwise regression procedure.” Discuss.

**Question 26:** (4 pts.) A junior investment analyst used a polynomial regression model of relatively high order in a research seminar on municipal bonds and obtained an of 0.991 in the regression of net interest yield of bond ( ) on industrial diversity index of municipality ( ) for seven bond issues. A classmate, unimpressed, said: “You overfitted. Your curve follows the random effects in the data.”

1. Comment on the classmate’s criticism.
2. Might be more appropriate than as a descriptive measure here?

**Question 27:** (4 pts.) A student who used a regression model that included indicator variables was upset when receiving only the following output on the multiple regression printout: XTRANSPOSE X SINGULAR. What does this mean and what is a likely source of the error?

**Question 28:** (4 pts.) In a regression study of factors affecting learning time for a certain task (measured in minutes), gender of learner was included as a predictor variable () that was coded if male and if female. It was found that the estimated coefficient of was with a standard error of 3.8. An observer questioned whether the coding scheme for gender is fair because it results in a positive coefficient, leading to longer learning times for males than females. Comment.

**Question 29:** (2 pts.) A student stated: “Adding predictor variables to a regression model can never reduce , so we should include all available predictor variables in the model.” Comment.

**Question 30:** (2 pts.) The members of a health spa pay annual membership dues of $300 plus a charge of $2 for each visit to the spa. Let denote the dollar cost for the year for a member and the number of visits by the member during the year. Express the relation between and mathematically. Is it a functional relation or a statistical relation?

**Question 31:** (2 pts.) Evaluate the following statement: “For the least squares method to be fully valid, it is required that the distribution of be normal.”

**Question 32:** (2 pts.) A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures ( ) and sales ( ) for one of the team’s products:

* Estimated regression equation:
* Two-sided *p*-value for estimated slope:

The student stated: “The message I get here is that the more we spend on advertising this product, the fewer units we sell!” Comment.

**Question 33:** (2 pts.) What is a residual? Why are residuals important in regression analysis?

**Question 34:** (4 pts.) The following regression model is being considered in a water resources study:

State the reduced models for testing whether or not:

**Bonus:** (5 pts.) An analyst wanted to fit the regression model by the method of least squares when it is known that . How can the analyst obtain the desired fit using standard statistical software (e.g., R, Python, or SAS)? No need to run any code, just describe in general how you could accomplish this.