Name: Section:

The final is due **Wednesday, August 17th, 2016 at 11:59PM Central Time** and must be turned in by that time to receive any credit for the exam. There is a space to submit your completed exam under Unit 15, in 15.4. There will be **no extensions** on this deadline so be sure to plan ahead. If there is an emergency which prevents you from finishing the exam, you should email me, but as I most likely would not give an extension, you should also turn in what you have at the time.

I would prefer you return the exam as one Word or PDF document. If you cannot use one of these formats, please contact me before the due date and let me know what format you would like to use to ensure that I will be able to open and grade your submitted document.

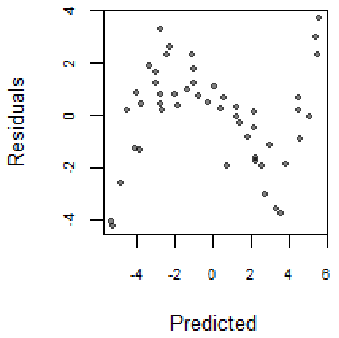
You may not discuss the final (or any class material) with other students or anyone else during the exam**. All questions should be directed to me** (Chelsea: cdallen@smu.edu); the graders and tutor will not answer questions during the exam.

Turning in this exam is an agreement that you have adhered to the SMU Honor Code and that you have neither given nor received assistance in completing this exam.

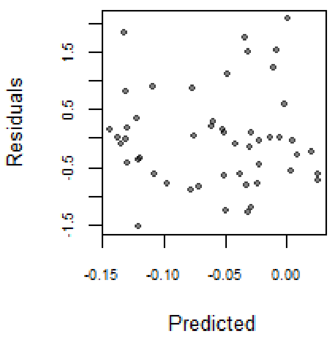
Multiple Choice/Multiple Answer (6 points each)

In this section, select the option(s) that best answer(s) the question. On "select all that apply" questions, it is still possible for there to be only one correct answer. There is partial credit available on "select all that apply" questions. No explanation is necessary for your answers.

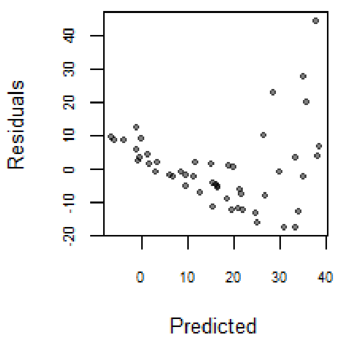
a



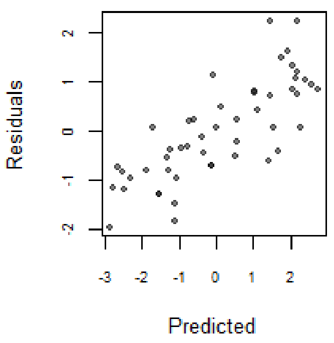
b



c



d



1. The plots on the right are of the residuals versus the predicted values of a model.
Which of these plots would be the best residuals for a linear regression model? Note: for each option, the label is above the plot.
Select only ONE answer.
   a. [top left]
   b. [top right]
   c. [bottom left]
   d. [bottom right]

2. While using one-way ANOVA on 4 groups, say we want to specifically compare the means of groups 3 and 4 and come up with a confidence interval for $\mu_3 - \mu_4$. Should we use a contrast or a t-test? Why?
Select only ONE answer.

a. We should use a t-test; the t-test is always best way to compare two groups' means.
b. We should use a t-test; a contrast will be biased by the other groups.
c. We should use a contrast; there is no way to do this with a t-test.
d. We should use a contrast; a t-test will not use the best estimate of variance.

3. In multiple linear regression, multicollinearity is when some of your explanatory variables are strongly correlated with each other. Why is this undesirable?
Select ALL that apply.

a. It can give misleading coefficients.
b. It decreases the adjusted-R^2 of the model.
c. We will not be able to fit the model at all if the variables are correlated.
d. It makes it difficult to separate the different effects of the correlated variables.

4. Which of the following is/are always true for a simple linear regression model?
[Note: if one of the options is only sometimes true, then it should not be selected. Only select things that are true of every simple linear regression model.]
Select ALL that apply.

a. The residuals will sum to 0.
b. Any outliers should be deleted from the dataset.
c. Negative coefficients are a sign that the model does not fit well or is not correct.
d. The regression line will go through the point $(\bar{x}, \bar{y})$ [the mean of x and the mean of y]
e. If the simple linear regression F-test is not significant, the two variables are not related.
f. At every x-value, the confidence interval for the mean is narrower than the prediction interval.

5. Say we do a t-test at the $\alpha = 0.10$ level for the below hypotheses and get a p-value of 0.073.

$$H_0: \mu = 18$$
$$H_a: \mu > 18$$

Which of the following is/are correct conclusions for this test?
Select ALL that apply.

a. We have sufficient evidence (p = 0.073) to reject the alternative hypothesis. The mean is 18.
b. With a p-value of 0.073, we have sufficient evidence to reject the null in favor of the alternative hypothesis that the mean is greater than 18.
c. With a p-value of 0.073, we do not have sufficient evidence to reject the null hypothesis. The mean is greater than 18.
d. We reject the null hypothesis. There is sufficient evidence (p = 0.073) to conclude that the mean is greater than 18.
e. With a p-value of 0.073, there is not sufficient evidence to reject the null hypothesis that the mean is 18.
f. We fail to reject the null hypothesis. There is not sufficient evidence (p = 0.073) to say that mean is greater than 18.

6. The null hypothesis for the overall F-test for a multiple regression with 3 explanatory variables is

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

Which is the alternate hypothesis?

Select ALL that apply.

a. $H_a: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$

b. $H_a: \beta_1 \neq 0\ \ or\ \ \beta_2 \neq 0\ \ or\ \ \beta_3 \neq 0$

c. $H_a: \beta_1 \neq 0\ \ and\ \ \beta_2 \neq 0\ \ and\ \ \beta_3 \neq 0$

d. $H_a: at\ least\ one\ of\ the\ slopes\ (\beta's)\ is\ not\ 0$

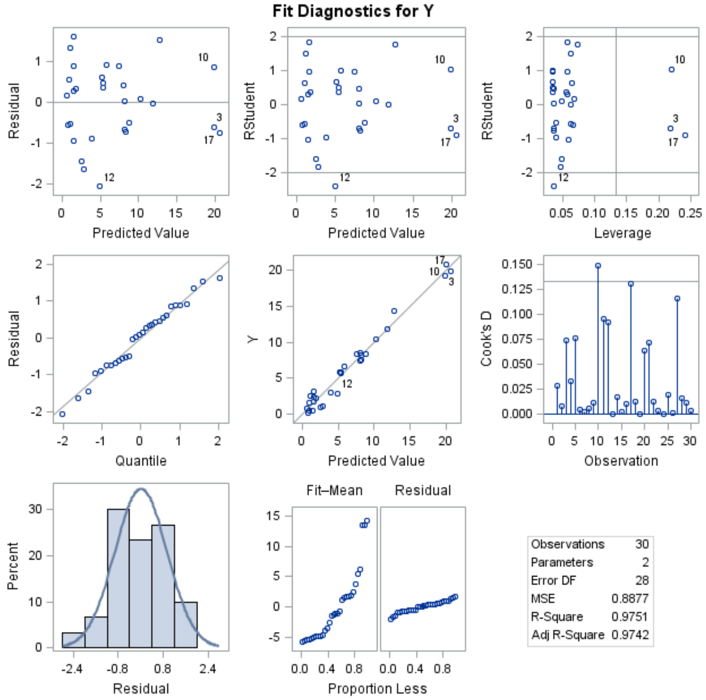
Short Answer (8 points each)

Answer the questions in complete sentences. 1-2 sentences should be enough to answer each question.

7. (This question is two parts, each worth 4 points.)

a. What is a Type I Error?

b. The probability of making a Type I Error is the same as the significance level, α, which we control. Why don't we just make α really small to decrease the chance of making a Type I Error? (Hint: it's because there is a trade-off. What is that trade off?)



Observations	30
Parameters	2
Error DF	28
MSE	0.8877
R-Square	0.9751
Adj R-Square	0.9742

8. The residual plots on the right are for a simple linear regression model. Do you think linear regression is appropriate for this data? List what the assumptions are for simple linear regression and how you know each one is or is not met, based on the plots. You can assume the assumption of independence is met.

9. Why do we need to use adjusted-R^2 as opposed to R^2 to measure whether a model is "better" with or without a certain variable?

Analysis (5 points for each part)

The questions below all use two datasets, which is in the same zip file as this final. Use the dataset and your software of choice (and whatever hand calculations you may feel are necessary) to answer the following questions. For each question, include your code or a description of your calculations and a copy of any tables you refer to. **Please note that I do NOT want a print-out of the data (or the datalines if you choose to put your dataset in that way) or tables or plots that you do not refer to in your answers.**

10. Say whenever I go on a trip I write down how many nights I will be gone and the weight of my luggage after I'm finished packing. This data is in Travel.csv. This dataset contains the following variables:

- Nights: The number of nights the trip is for
- Weight: The weight, in pounds, of the suitcase

a. Use simple linear regression to model the weight of the suitcase based on the number of nights of the trip. Fit the model

$$\widehat{Weight} = b_0 + b_1 \cdot Nights$$

and give the equation. (Assume that all the assumptions of linear regression are met. You do not need to check them.)

b. Interpret the slope and intercept of the model in context.

c. Say I am packing for another trip that will be 4 nights. Give an 95% interval predicting how much my bag will weigh and interpret that interval in the context of the problem.

d. Say I am packing for a long trip that will be 21 nights. Do you think this model will be an accurate predictor of my suitcase weight? Why or why not? (You do not need to actually predict the weight, unless you feel it strengthens your argument.)

e. **Extra Credit:** Perform a lack-of-fit test (to determine if we need higher order terms). Give a conclusion for the test including the p-value. Do we need higher order terms? (Regardless, you do not need to add anything to the model; only determine if you *would* add something to the model.)

11. Later it occurs to me that there might be a difference between my business and personal trips. So, I go back and classify my trips with an indicator variable. The new dataset, TravelCat.csv, now has three variables:

- Nights: The number of nights the trip is for
- Business: An indicator variable that is 1 if the trip is for business and 0 if the trip is personal
- Weight: The weight, in pounds, of the suitcase

a. Use multiple regression to model the weight of the suitcase based on the number of nights of the trip AND the type of trip (business or not). In other words, fit the model

$$\widehat{Weight} = b_0 + b_1 \cdot Nights + b_2 \cdot Business$$

Give the equation.

b. Based on your equation, would you expect a business or a personal trip to have a heavier bag if the lengths of the trip were the same? How do you know?

c. Using the equation from part a, give the two separate equations for personal trips and for business trips.

d. Do you think adding in the variable for type of trip improves the model? (There is more than one way to answer this question. I am looking for whether your justification is reasonable.)

e. **Extra Credit**: We should also consider the possibility of an interaction. Fit the model

$$\widehat{Weight} = b_0 + b_1 \cdot Nights + b_2 \cdot Business + b_3 \cdot Business \cdot Nights$$

i. Give the equation and also give the two separate equations for personal trips and for business trips. Please be sure to label your equations so I know which is which.

ii. Do you think the model is “better” with the interaction? Justify your answer (again, there is not only one right way to do this; I am looking for whether your justification makes sense.).