ST 314 Data Analysis 08

Topics:

- Describing Scatterplots and Bivariate Regression Analysis
- Lessons Covered: 39 - 45
- Textbook Chapter (Optional): 12 and 13.1

Grading:

- Points are listed next to each question and should total 25 points.
- Grading will be based on the content of the data analysis as well as the overall appearance of the document.
- Late assignments will not be graded.

Deadlines:

- Final Submission: **Monday, December 7th**. All submissions must be PDF files.

Instructions:

- Clearly label and **type answers** to the questions on the proceeding pages in Word, Google Docs, or other word processing software.
- Insert diagrams or plots as a picture in an appropriate location.
- Math Formulas need to be typed with Math Type, LateX, or clearly using key board symbols such as +, -, *, /, sqrt() and ^
- Submit assignment to the Gradescope link as a PDF. Indicate the pages to the individual questions and also verify the correct document has been uploaded. Failing to follow this direction may result in point deductions.

Allowances:

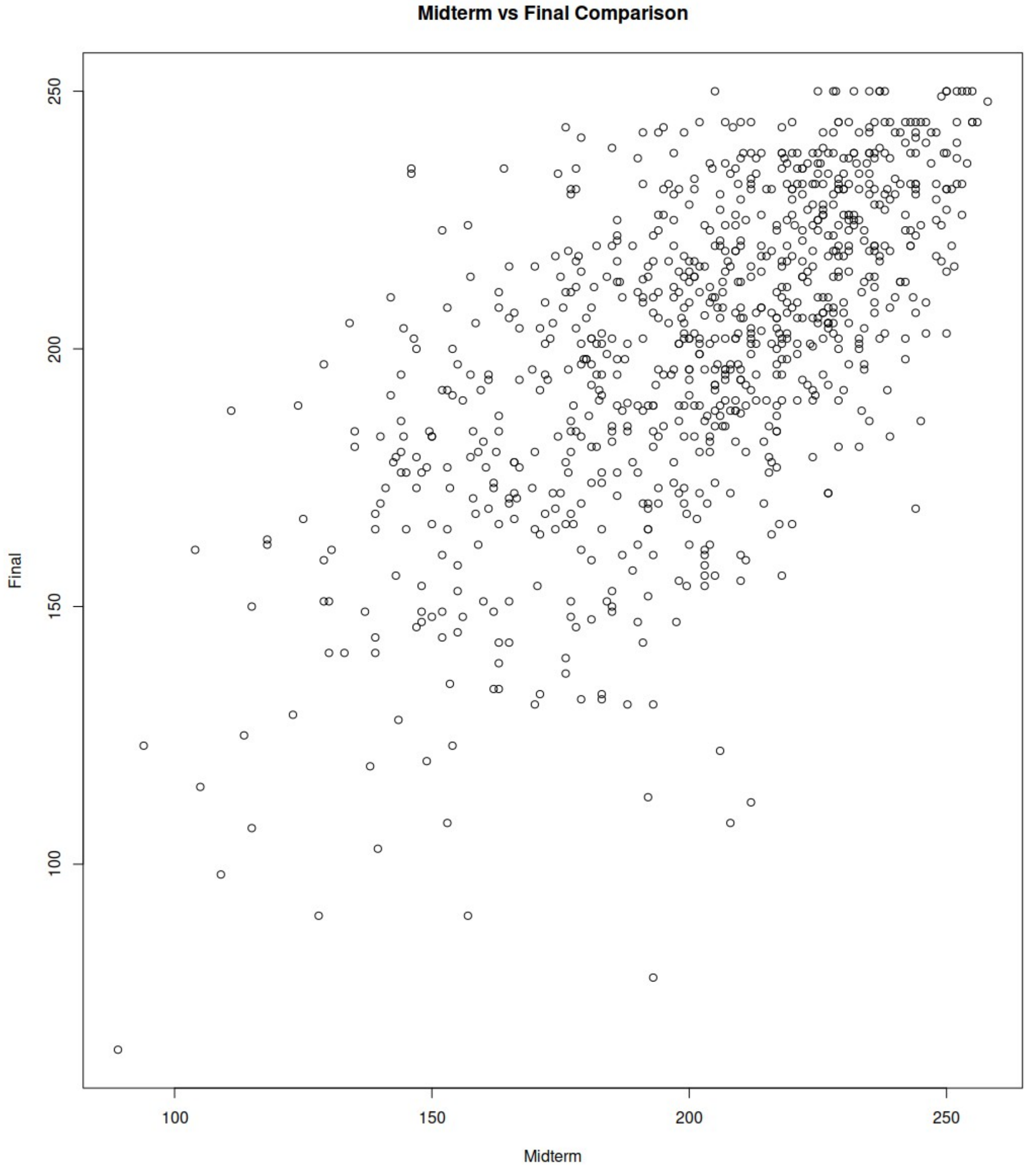
- Any resources listed or posted in our class.
- You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
- Outside Resources need to be cited. Websites such as Chegg, Coursehero, Koofers are discouraged, but if used need to be cited and within the boundaries of academic honesty.

ST 314 Data Analysis 08

The dataset *ExamDataW19.csv* represents the midterm and final exam grades for students in the ST314 online and campus courses for the last three consecutive terms. Use this data to complete a regression analysis in R and answer the following questions.

Part 1. (10 points) Describing the relationship between your two variables

a. **(6 points)** Graphically: Make a scatterplot of the relationship between ST314 student midterm grades and final exam grades. Paste your plot. Describe in context the relationship from the scatterplot. Include strength, direction, form and outliers (if any).



The graph of midterm grades on the X axis against final exam grades on the Y axis has a moderatel positive correlation. It appears that those that scored well on either test also did so on the other test, though there are a couple students that scored highly on the midterm, and low on the final, and visa versa. There are a couple outlier students that it might be worth removing, for example the student that scored (80, 60) and the student that scored (190, 80).

b. **(4 points)** Numerically: Calculate the correlation coefficient r. Describe in context the strength of the relationship based on your value.

The Correlation coefficient, r is 0.636, which indicates a moderate correlation between a score on the midterm exam and a similar score on the final.

This coefficient tells us that scores of a certain type (high, low, average) are correlated with a similar score or outcome on the other test, for example it is likely that a student scores poorly on both exams or well on both exams, but not as likely that they score well on one exam and poorly on the other.

ST 314 Data Analysis 08

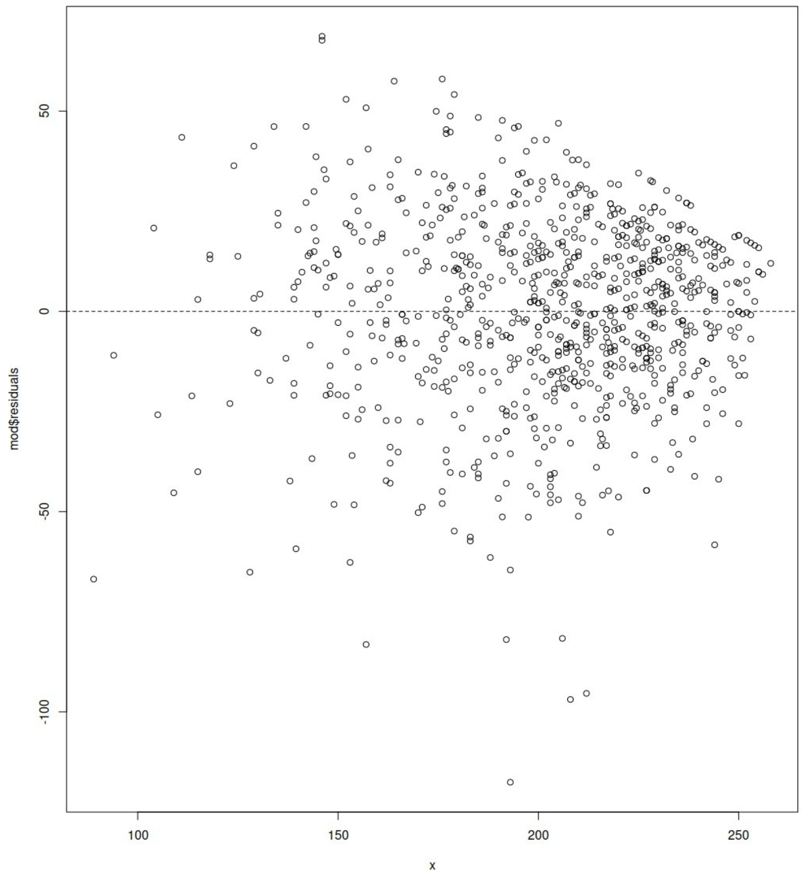
Part 2. (15 points) Calculate the Least Square Regression Line (Model) and Check Conditions for Inference

a. **(4 points)** Using R, calculate the least squares regression line that predicts final exam scores from midterm exam scores for ST314 students. Paste the R output for the model summary. State the least squares regression line (model).

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-117.571  -13.617    1.962   16.059   68.671 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 75.49269    5.28281   14.29   <2e-16 ***
x            0.62217    0.02587   24.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.69 on 850 degrees of freedom
Multiple R-squared:  0.4049,	Adjusted R-squared:  0.4042 
F-statistic: 578.4 on 1 and 850 DF,  p-value: < 2.2e-16
```

$\hat{y} = 75.49269 + 0.62217\,x$

b. **(6 points)** Plot the residuals for the model. Include a reference line at 0. Paste your plot. Check the linearity, normality and constant variation conditions using the residual plot. State why each condition is met or why it is not met.



Conditions for interpreting residual plot

- The relationship is linear in the population.
  - Yes, there is no ‘U’ shape.
- The response varies normally about the population regression line.
  - There is about equal scatter above and below the regression line
- Observations are independent.
  - Observations are independent by design.

- The standard deviation of the responses is the same for all values of x.
    - Failure: Funnel shape indicates that this is not true.

c. **(5 points)** Using your linear model, predict what the final score would be for a student who received an 85% grade on the midterm, on average.

$$\hat{y}=75.49269+0.6221\,x$$
$$\hat{y}(0.85)=75.49269+0.6221*0.85$$
$$\hat{y}(0.85)=76.021\%$$

ST 314 Data Analysis 08

Optional Questions:
These questions are good practice for using simple linear regression. However, you don't need to complete them prior to turning in the assignment.

Part 3. Is your model a good fit? Use your R output from the model in Part 2a. From the output, is there statistical evidence midterm exam score is a significant predictor of final exam score? Use a significance level of 0.05.

a. **(1 point)** State the null and alternative hypothesis for the individual t test on the slope.
b. **(1 point)** State the test statistic, degrees of freedom and p-value from the output.
c. **(2 points)** Make a conclusion. Include context, a statement in terms of the alternative and whether to reject the null based on the level of significance.
d. **(3 points)** Calculate the 95% confidence interval for the slope. Interpret the point and interval estimate for β_1.

Part 4. Prediction: A common goal in regression is to use the estimated model for prediction.

a. **(2 point)** Using your model from part 2a, provide a final exam score prediction for students that earned a midterm exam score of 200. Show work by hand! You may validate with code provided in R.
b. **(1 point)** The following intervals are the 95% confidence and prediction intervals for a midterm value of 200. Which is the confidence interval? Which is the prediction interval? How do you know?

   (153.4, 246.5)   (198.3, 201.5)

c. **(2 point)** Write an appropriate interpretation of both the intervals. Include context.
d. **(1 point)** Based on your midterm exam score, calculate your predicted final exam score based on the model.
e. **(1 point)** Do you think the predicted value is a reasonable score to assume for your final exam grade? Why or why not? Consider the data and the strength of the model. Consider other factors that might influence your grade.