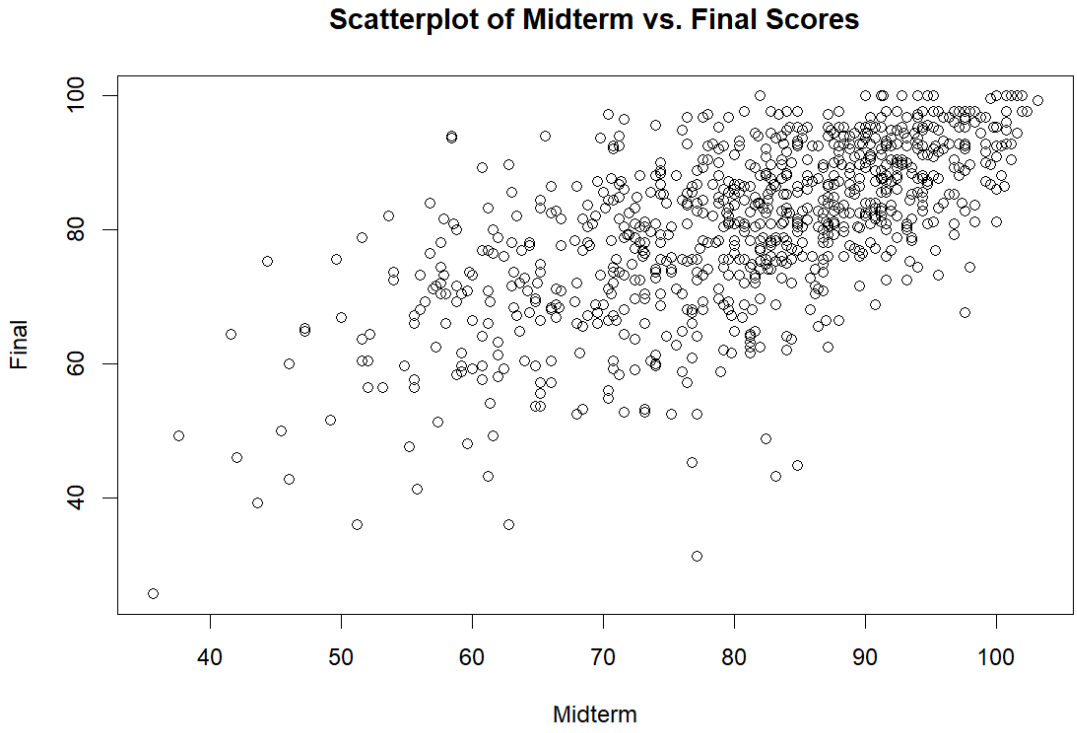
ST 314 Data Analysis 07

The dataset *ExamDataW24.csv* represents the midterm and final exam grades for students in the ST314 online and campus courses for three previous consecutive terms. Use these data to complete a regression analysis in R and answer the following questions.

Part 1. Describing the relationship between your two variables

a. **(2 points)** Graphically: Make a scatterplot of the relationship between ST314 student midterm grades (explanatory variable) and final exam grades (response variable). Include your plot in your submission. Describe in context the relationship from the scatterplot. Include strength, direction, form and outliers (if any).



The scatterplot of midterm versus final scores shows a positive relationship, meaning that higher midterm scores point to higher final scores. The strength of this relationship is moderate to strong, as the points are clustered around an upward trend line. The relationship appears linear, a straight line could describe the trend. There are a few outliers, such as some low midterm scores paired with high final scores and vice versa.

b. **(2 points)** Numerically: Calculate the correlation coefficient r. Report the value of r and describe in context the strength of the relationship based on your value.
The calculated correlation coefficient r is 0.636. This value indicates a moderate to strong positive linear relationship between midterm and final scores. In context, this means that higher midterm scores are associated with higher final scores.

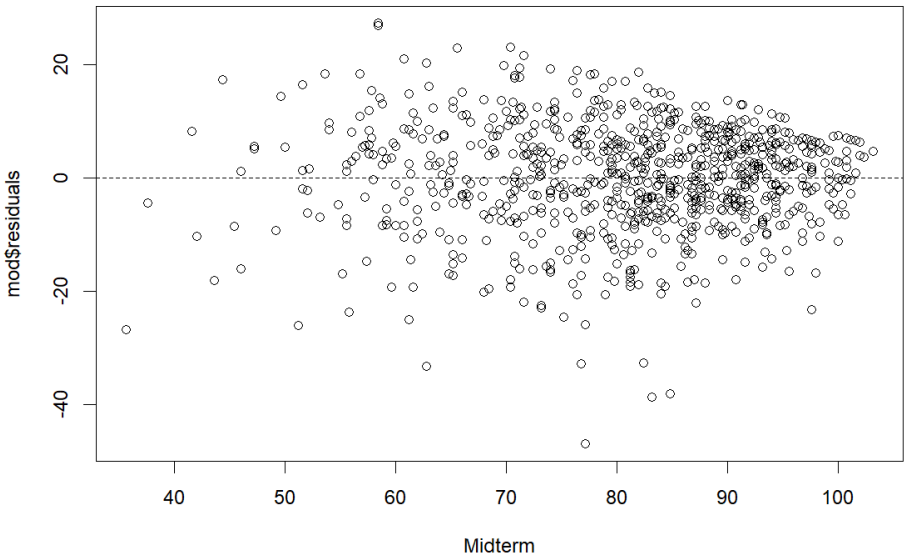
Part 2. Calculate the Least Square Regression Line (Model) and Check Conditions for Inference

a. **(3 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. Separately, state the least squares regression line (model) using statistical notation.

```
> mod = lm(Final~Midterm)
> summary(mod) #note here the line has been named mod

Call:
lm(formula = Final ~ Midterm)

Residuals:
    Min      1Q  Median      3Q     Max
-47.029  -5.447   0.785   6.424  27.468

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.19708    2.11312   14.29   <2e-16 ***
Midterm      0.62217    0.02587   24.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.477 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
```

The least squares regression line can be stated as: $\hat{Y} = 30.19708 + 0.62217X$
Where $\hat{Y}$ is the predicted final score, and X is the midterm score.

b. **(4 points)** Plot the residuals for the model. Include a reference line at 0. Include your plot in your submission. Check the linearity, normality and constant variation conditions using the residual plot. State why each condition is met or why it is not met.



Linearity is met, the residuals are randomly scattered around zero without a clear pattern. Normality is met based on the symmetry and spread of the residuals. Constant Variation is mostly met, with some minor deviations that do not violate the assumption.

c. **(4 points)** Using your linear model, predict what the final score would be for a student who received a 84 grade on the midterm, on average.
For a student who got an 84 on the midterm, they are expected to get a 82.46 on the final.

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. **(2 points)** State the null and alternative hypothesis for the individual t test on the slope.
Null: The slope is equal to 0. (Midterm can not predict the score of the final)
Alternative: The slope is not equal to 0 (Midterm can predict the score of the final)

b. **(2 points)** State the test statistic, degrees of freedom and p-value from the output.
Test statistic: 24.05
Degrees of freedom: 850
p-value: < 2.2e-16

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.
Given the p-value is less than the significance level of 0.05, we reject the null hypothesis. This means there is statistically significant evidence that the midterm exam score is a predictor of the final exam score.

d. **(2 points)** Calculate the 95% confidence interval for the slope. Interpret the point and interval estimate for β_1.
The 95% confidence interval for the slope is from 0.5714 to 0.6729. The point estimate for the slope is 0.6222. This means that for each point scored on the midterm, the final exam score is predicted to increase by 0.6222 points.

Gradescope Page Matching (2 points)

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page "Submitting Assignments in Gradescope". Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item "Gradescope Page Matching".