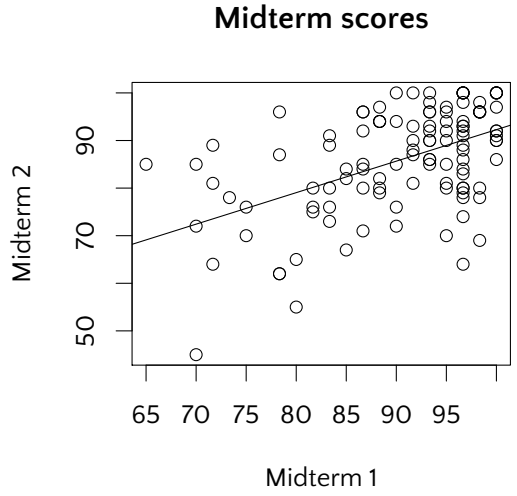
STAT 323 Bonus Assignment

Instructor: Claudia Marie Mahler
Name: Yifeng Pan
UCID: 30063828

Summer 2019

1 A math professor is interested in the relationship between students' performance on the first midterm in his class and their performance on the second midterm in his class. He gathers the midterm 1 and midterm 2 data for his $n = 112$ students (this data is called "Midterm1" and "Midterm2" respectively in the "Bonus Assignment Data" R file). Note that the scores for both midterm 1 and midterm 2 are out of 100 points.

a Creat a scatterplot of the data. Describe the general relationship (trend) of the midterm 1 and midterm 2 scores.



b What is the correlation between midterm 1 grades and midterm 2 grades? Interpret this correlation.

$r \approx 0.5048535$

c State the least-squares estimate of the model predicting midterm 2 scores from midterm 1 scores.

$m_2 \approx 0.6619698 m_1 + 26.0843211$

d Interpret the slope of this regression equation in the context of the data.

For every 1 point increase on a student's score on midterm 1, there will be a 0.6620 point increase on the student's score on midterm 2.

e Predict the midterm 2 score of a student who scored 60 on midterm 1.

$m_2 \approx 0.6619698 m_1 + 26.0843211$
$= 0.6619698(60) + 26.0843211 \approx 65.80$

f Predict the midterm 2 score of a student who scored 92 on midterm 1.

$m_2 \approx 0.6619698 m_1 + 26.0843211$
$= 0.6619698(92) + 26.0843211 \approx 86.99$

g Suppose a student scored a 80 on midterm 1. Calculate the residual for this student if they also scored an 80 on midterm 2.

$\epsilon = m_2 - \hat{m}_2 \approx 80 - (0.6619698 m_1 + 26.0843211)$
$= 80 - (0.6619698(80) + 26.0843211) \approx 0.9581$

h Assuming the normality condition is met, what would you expect the normal probability plot of the residuals to look like? Describe it in words or draw a picture.

See figure 1 on page 2.
The historgram is close to normal. On the normal probabily plot, the dots are close to the lines.

i Conduct a hypothesis test to determine if midterm 2 scores can be expressed as a linear function of midterm 1 scores. Please state all steps of the hypothesis testing process (hypotheses, test statistic, p-value, conclusion). Use $\alpha = 0.05$.

$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$
$t = \frac{\beta_1 - \beta_{1_0}}{SE_{\beta_1}} \approx \frac{0.662 - 0}{0.107917} \approx 6.134$
p-value $= 2*(1-pt(t, 110)) \approx 1.377*10^{-8} < 0.05 = \alpha$
Therefore there is a linear dependency of the score on midterm 1 on midterm 2.

j Construct a 95% confidence interval for the slope. Does this confidence interval suggest that the slope is significantly different from zero?

Confidence interval: (0.4481033, 0.8758362).
Yes, it does not contain 0.

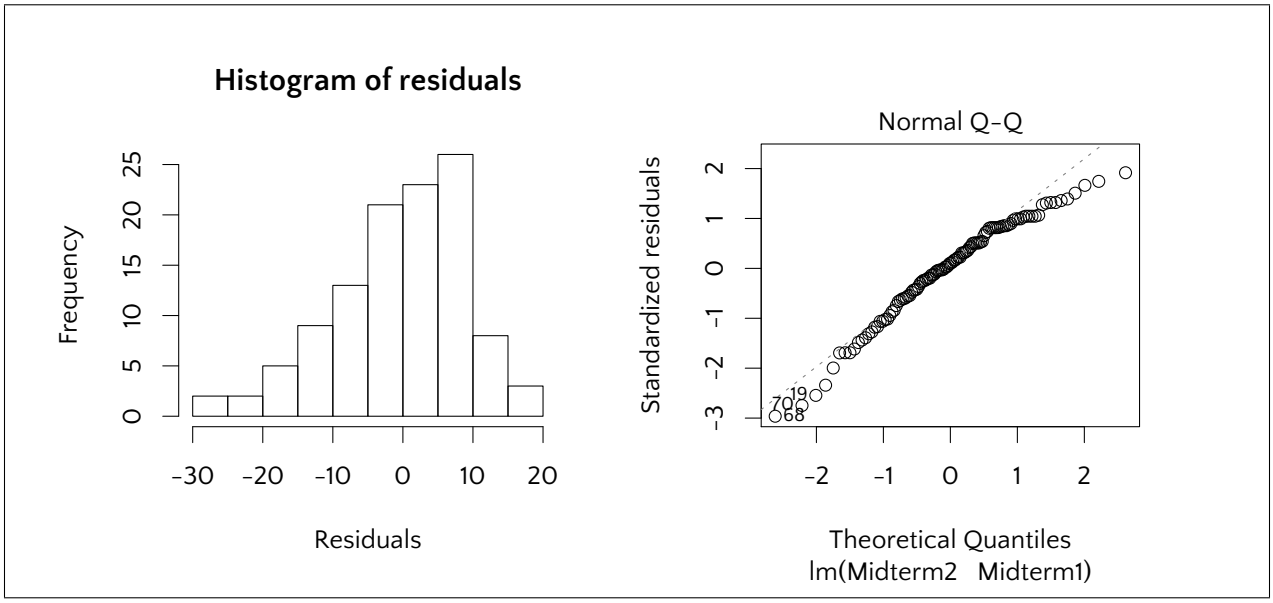


Figure 1:

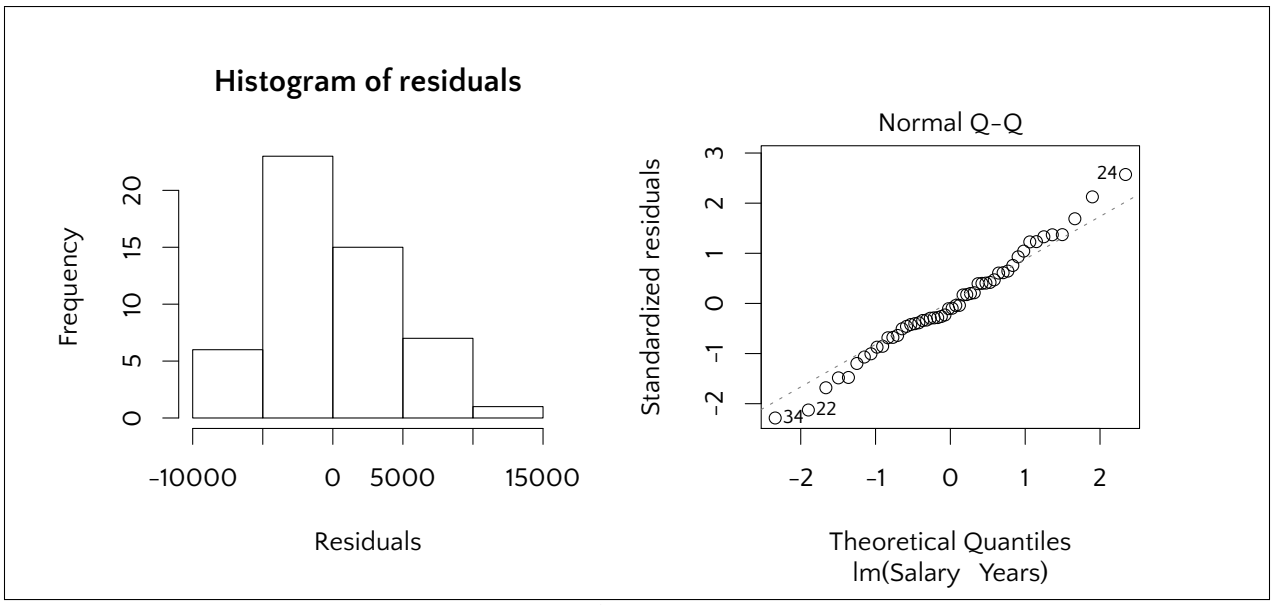


Figure 2:

2 Is academic salary related to the number of years since the highest degree earned? A random sample of 52 tenure-track professors was taken. Each professor's academic yearly salary (in dollars) was recorded, as well as the time (in years) since their highest degree had been earned (this data is called "Salary" and "Years" respectively in the "Bonus Assignment Data" R file).

a What is the value of the coefficient of determination? Interpret its meaning in this context.

$r^2 \approx 0.4554$

b Write down the least-squares estimate of the model predicting yearly salary from years since highest degree earned.

$s \approx 390.6y + 17500$

c Interpret the slope of this regression equation.

For every 1 year increase the amount of time since highest degree earned, there will be a 390.6 dollar increase in salary.

d Interpret the y-intercept of this regression equation. Does the y-intercept have a "practical" (meaningful) interpretation?

It means the starting salary from highest degree earned is 17500 dollars.
Yes

e Predict the yearly salary for a tenure-track professor for whom there has been 29 years since his/her highest degree was earned.

$s \approx 390.6451y + 17502.2574$
$= 390.6451(29) + 17502.2574 \approx 28830$

f A tenure-track professor for whom there has been 29 years since his/her highest degree was earned is actually earning $\$22,450$ yearly. Compute this professor's residual.

$\epsilon = s - \hat{s} \approx 22450 - (390.6451(29) + 17502.2574)$
≈ -6381

g Examine the normality plot. What condition/assumption does the normality plot check for? Does it appear to be met in this case?

See figure 2 on page 2.
The normality plot checks if the data approximates a normal distro. In this case, it does, as the dots are close enough to the line. And the histogram is close enough to a bell curve.

3 R has a built-in dataset called trees that contains observations for $n = 31$ black cherry trees. The first column, "Girth," contains the girth of the trees (in inches); the second column, "Height," contains the height of the trees (in feet); the third column, "Volume," contains the volume of timber (in cubic feet).

a State the least-squares estimate of the model predicting diameter from height.

(Note: Diameter is mislabeled as girth in the dataset.)
d and h are in inches: $d \approx 0.02131h - 6.188$

b Predict the diameters of trees of the following heights: 60 feet, 73 feet, and 85 feet.

In inches.
Height of 60 feet:
$d \approx 0.02131226(60 * 12) - 6.18839451 \approx 9.156$
Height of 73 feet:
$d \approx 0.02131226(73 * 12) - 6.18839451 \approx 12.48$
Height of 85 feet:
$d \approx 0.02131226(85 * 12) - 6.18839451 \approx 15.55$

c Interpret the meaning of the slope in the context of the data.

For every 1 inch increase the height, there will be a 0.02131 inch increase in diameter.

d Suppose you wanted to conduct a hypothesis test for a positive linear relationship between height and diameter. State the appropriate hypotheses and test statistic.

$H_0 : \beta_1 \leq 0, H_a : \beta_1 > 0$
$t = \frac{\beta_1 - \beta_{1_0}}{SE_{\beta_1}} \approx \frac{0.02131226 - 0}{0.006513191} \approx 3.272$

e State the probability expression of the p-value as well as the p-value itself. What conclusion would we reach with respect to our hypotheses in part d. using $\alpha = 0.05$?

p-value $= P(T_{df} > 3.272\ldots) = 1 - pt(t, df) \approx 0.1379\% < 5\% = \alpha$
Therefore there is a linear dependency of height on diameter.

f If you wanted to report a 95% interval for the diameter of a single tree that was 80 feet tall, would you report a confidence interval or prediction interval? Why?

Prediction interval, as it predicts the dependent variable from a given independent variable
Prediction interval: (13.08276, 15.45999) in inches.

g What proportion of σ^2 in diameter is explained by its linear relationship with height?

$r^2 \approx 0.2697$

h Examine the residual plot. What condition/assumption does the residual plot check for? Does it appear to be met in this case?

It checks for consistancy in variance. In this case, yes, the points are in a random pattern.