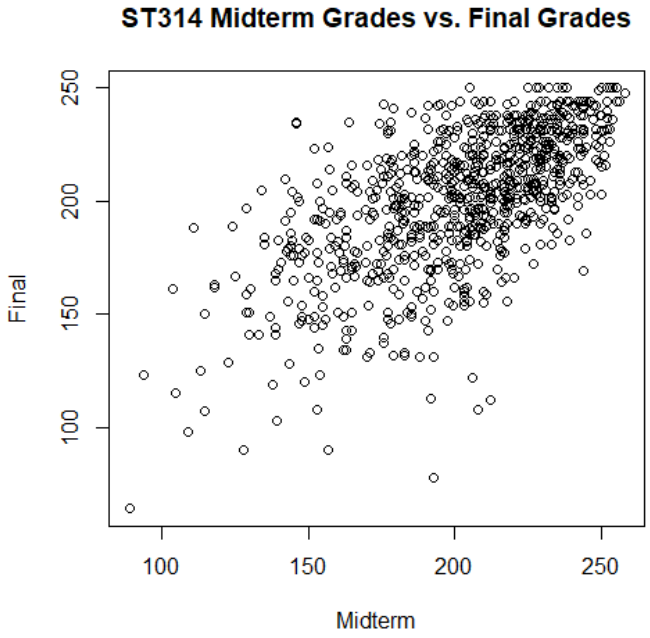
3/4/2019

Braeden Kuether

Data Analysis #8

Part 1 –

a) There is a positive correlation between midterm and final grades among ST314 students, meaning that students who did well on the midterm generally did well on the final too. The strength of the scatter plot is somewhat strong as the spread among data points is not that large. The form of the plot is linear, and there are few outliers if any. Most notably, a few students who were around the 200 point mark for the midterm who scored lower to much lower than 150 points on the final.



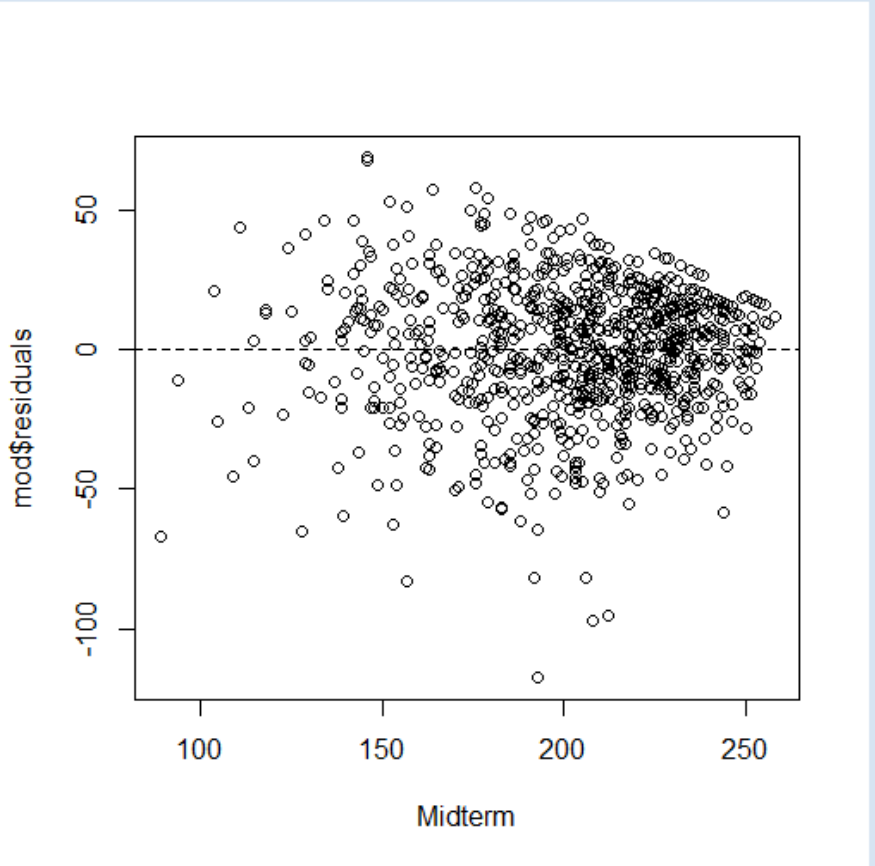
b) The correlation coefficient is r = 0.6363. This means that the relationship is not strong, but is noticeable and relevant. Generally correlation coefficients of 0.7 and higher are considered “strong” correlations.

Part 2 –

a) The least squares regression line is $\hat{y} = 0.62217x + 75.49269$

```
> summary(mod)

Call:
lm(formula = Final ~ Midterm)

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 ***
Midterm      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
```

b)



Seeing as there is no curvature or distinct pattern in the residuals, I would say that it is in fact linear. Additionally, there are approximately a similar amount of data points above and below the zero line making it normal. Lastly, I would say that the variation about the regression line is not perfectly constant as values higher on the x axis tend to be more condensed than those at lower x values, creating a somewhat "funnel shape".

Part 3 –

a)

Null – H0: B1 = 0

Alternative – Ha: B1 /= 0

b) t = 24.05 , DOF = 850 , p = 2 x 10^(-16)

c) There is convincing evidence that midterm and final scores for students in ST314 are related. The 95% confidence interval estimates that this correlation is in between 0.57 and 0.67 meaning that the null hypothesis of 0 is not holding. The null hypothesis is rejected with a significance level of 0.05 where the t-stat is equal to 24.05, the degrees of freedom is 850 and the p value is equal to 2 x 10^(-16). Therefore, there is a correlation between midterm and final grades.

d) B1 = 0.62217, SEb1 = 0.02587, t* = 1.963

Therefore, the 95% confidence interval would be:

0.62217 +/- 1.963 * 0.02587 , which is equal to and interval of (0.57139, 0.67295).

The point estimate B1 would be the middle value of 95% confidence interval for the slope, and the interval estimate tells us with a 95% confidence level that the slope of the equation lies in that interval.

Part 4 –

a) The point estimate would be: $\hat{y}$ = 0.62217(200) + 75.49269 = 199.92
   This means that we would expect for a student who got a 200 on the midterm to get a 199.92 on the final.

b) The first interval is much larger and therefore the prediction interval because prediction intervals are significantly larger than confidence intervals. The second interval is the prediction interval. This is because the prediction interval looks at the range that the student might get on the next exam, which could be very large, while the confidence interval looks at the true mean for all students with that test score.

c) The interpretation of the confidence interval would be that we can conclude with 95% confidence that the true average final score of students who received a 200 on the midterm will be within that range. The interpretation of the prediction interval would be that we can conclude with 95% confidence that a student with a midterm score of 200 will receive a final score within that range.

d) $\hat{y} = 0.62217(181) + 75.49269 = 188.105$

e) I'd say this is a reasonable score to assume I will receive on the final because I didn't really study for the first one and I hate this class, so I probably won't work up the motivation to study for the final. Considering the strength of the model, I would say it's reasonable to assume that score on my final, although the prediction interval is quite large, so it's hard to say with any real confidence if I will land close to that number.