Jordan Mayer
STAT 350
Lab 09
April 19, 2018

1. Code

```
#####
# Jordan Mayer
# STAT 350
# Lab 09
# April 19, 2018
#####

### setup ###
setwd("C:/Users/jordan/Google Drive/Courses Spring 2018/STAT 350/STAT 350
Labs/Lab 09 - Linear Regression")
# set working directory
library(ggplot2)  # set up ggplot2 for plotting
graphics.off()  # close any open figures
USData <- read.table("US_Data.txt", header=TRUE, sep="\t")  # get US Data
US_clean <- USData[complete.cases(USData),]  # clean US Data
# analyze only "typical counties"
m <- mean(US_clean$MedianIncome)
s <- sd(US_clean$MedianIncome)
US_typ <- subset(US_clean, m-2*s < US_clean$MedianIncome &
                   US_clean$MedianIncome < m+2*s)

### Part B ###

# create scatterplot
scatter <- ggplot(US_typ, aes(x=MedianIncome, y=TestScore)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle('Relationship between Test Score and Income') +
  xlab('Median Income') +
  ylab('Test Score')
ggsave(scatter, filename='scatter.jpg', width=6, height=6)

# check correlation
cor(US_typ$MedianIncome, US_typ$TestScore)

# perform linear regression and display results
US_lm <- lm(TestScore ~ MedianIncome, data=US_typ)
summary(US_lm)

# create scatterplot of residuals
res_scatter <- ggplot(data.frame(residuals = US_lm$residuals,
                                 MedianIncome = US_typ$MedianIncome),
                      aes(x = MedianIncome, y = residuals)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle('Residuals from Income/Test Score Analysis') +
  xlab('Median Income') +
  ylab('Residuals')
ggsave(res_scatter, filename='res_scatter.jpg',width=6,height=6)

# check normality of residuals using histogram and normal probability
# plot
```

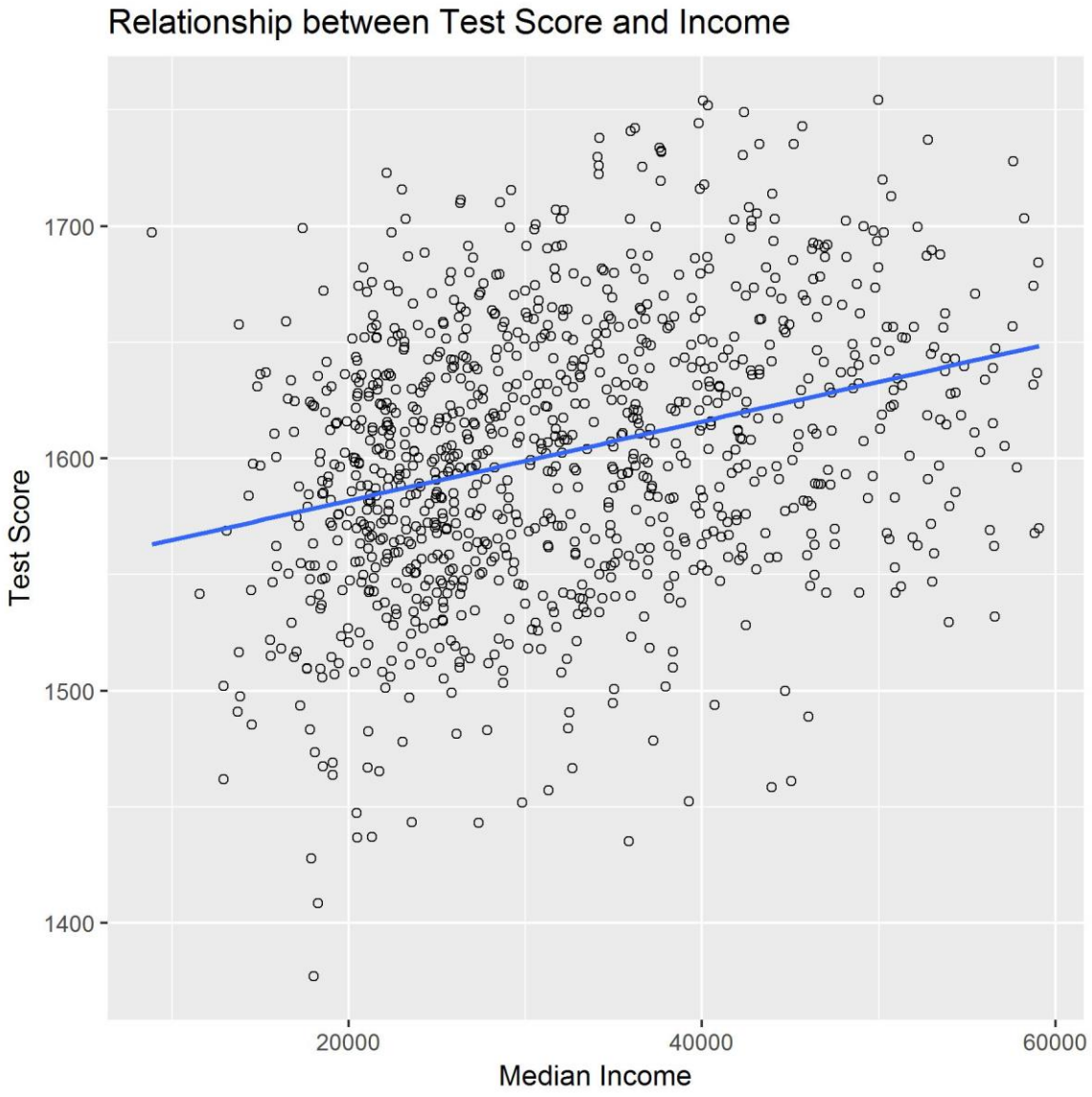
Jordan Mayer
STAT 350
Lab 09
April 19, 2018

```
attach(US_lm)
res_hist <- ggplot(data.frame(residuals=residuals),aes(residuals))+
  geom_histogram(aes(y=..density..),
                 bins=sqrt(length(residuals))+2,
                 fill='grey',col='black')+
  geom_density(col='red',lwd=1)+
  stat_function(fun=dnorm,args=list(mean=mean(residuals),
                                    sd=sd(residuals)),
                col='blue',lwd=1)+
  ggtitle('Residuals from Income/Test Score Analysis')+
  xlab('Data')+
  ylab('Proportion')
ggsave(res_hist, filename='res_hist.jpg',width=6,height=6)
res_qq <- ggplot(data.frame(residuals),aes(sample=residuals))+
  stat_qq()+
  geom_abline(slope=sd(residuals),intercept=mean(residuals))+
  ggtitle('Residuals from Income/Test Score Analysis')+
  xlab('Theoretical')+
  ylab('Sample')
ggsave(res_qq, filename='res_qq.jpg',width=6,height=6)
detach(US_lm)

# check two-sided 99% confidence interval for slope and intercept
confint(US_lm, level=0.99)
```

Jordan Mayer
STAT 350
Lab 09
April 19, 2018

2. Scatterplot



Jordan Mayer
STAT 350
Lab 09
April 19, 2018

3. Scatterplot Analysis

This scatterplot appears to have the following characteristics:

- Form: linear
- Direction: positive
- Strength: relatively weak

There are a few noticeable outliers on the left side of the plot. These are y-axis outliers specifically: one data point far above the general trend and two far below the general trend.

This relationship does appear to be approximately linear. The strength appears rather weak, though this may be influenced by the axis scales; further analysis should make the nature of the relationship (if any) more clear.

4. Correlation

```
> # check correlation
> cor(US_typ$MedianIncome, US_typ$TestScore)
[1] 0.3044976
```

The correlation coefficient between Test Score and Median Income is 0.3044976.

This suggests a weak, yet significant, association between Test Score and Median Income.

5. Correlation and Scatterplot

This correlation appears to be a good numerical summary of the pattern in the scatterplot, suggesting a weak positive relation between Test Score and Median Income. Here, the use of correlation is appropriate because both variables (Test Score and Median Income) are quantitative, the relationship is linear, and there are few outliers. If either variable were categorical, if the relationship were nonlinear, or if there were many extreme outliers, correlation would not be a useful numerical summary.

6. Estimated Linear Regression Line Equation

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.548e+03  5.537e+00  279.56   <2e-16 ***
MedianIncome 1.703e-03  1.647e-04   10.34   <2e-16 ***
```

The estimated equation for the linear regression line is:

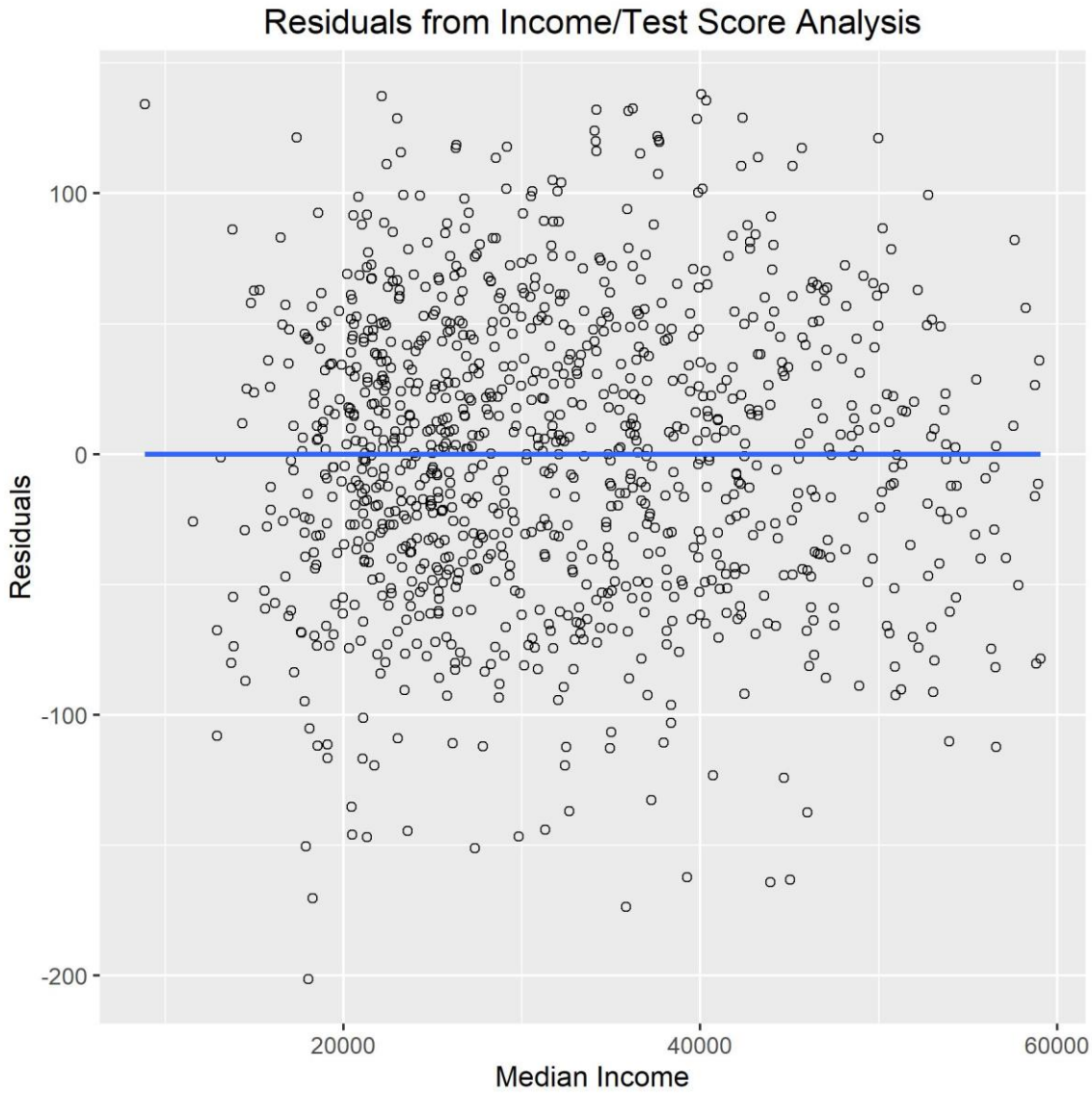
TestScore = 1548 + 0.001703 MedianIncome

```
Residual standard error: 55.9 on 1046 degrees of freedom
Multiple R-squared:  0.09272,   Adjusted R-squared:  0.09185
F-statistic: 106.9 on 1 and 1046 DF,  p-value: < 2.2e-16
```

$r^2 = 0.09272$

Jordan Mayer
STAT 350
Lab 09
April 19, 2018

7. Residuals Scatterplot



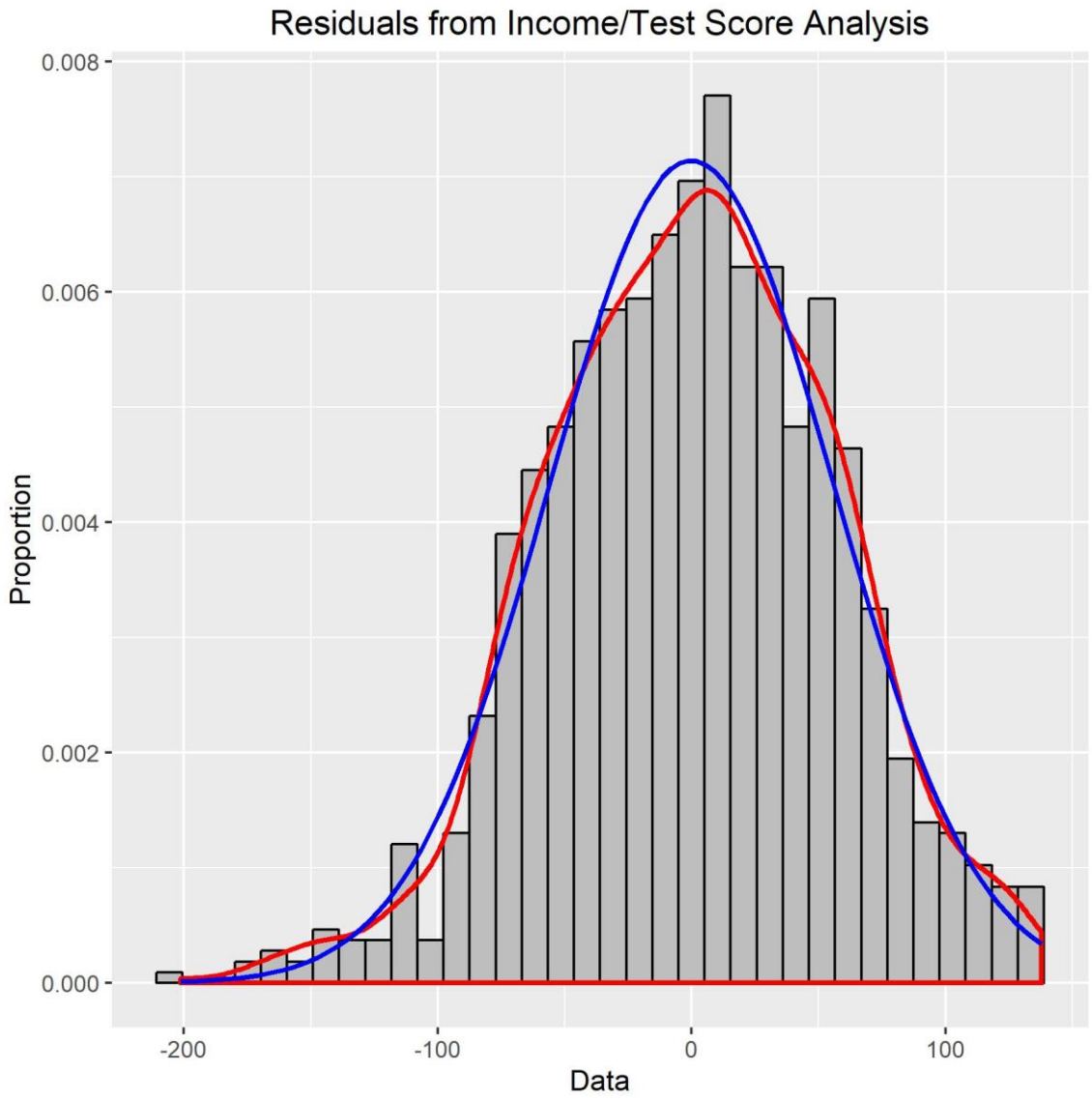
There appears to be no pattern to this data, indicating a linear association. There are a couple outliers near the left side of the plot, but even these are not very extreme. The standard deviation appears to be approximately constant.

This, there is nothing unusual to report and the conclusions from the residual plot are consistent with those from the scatterplot in part (3).

Jordan Mayer
STAT 350
Lab 09
April 19, 2018

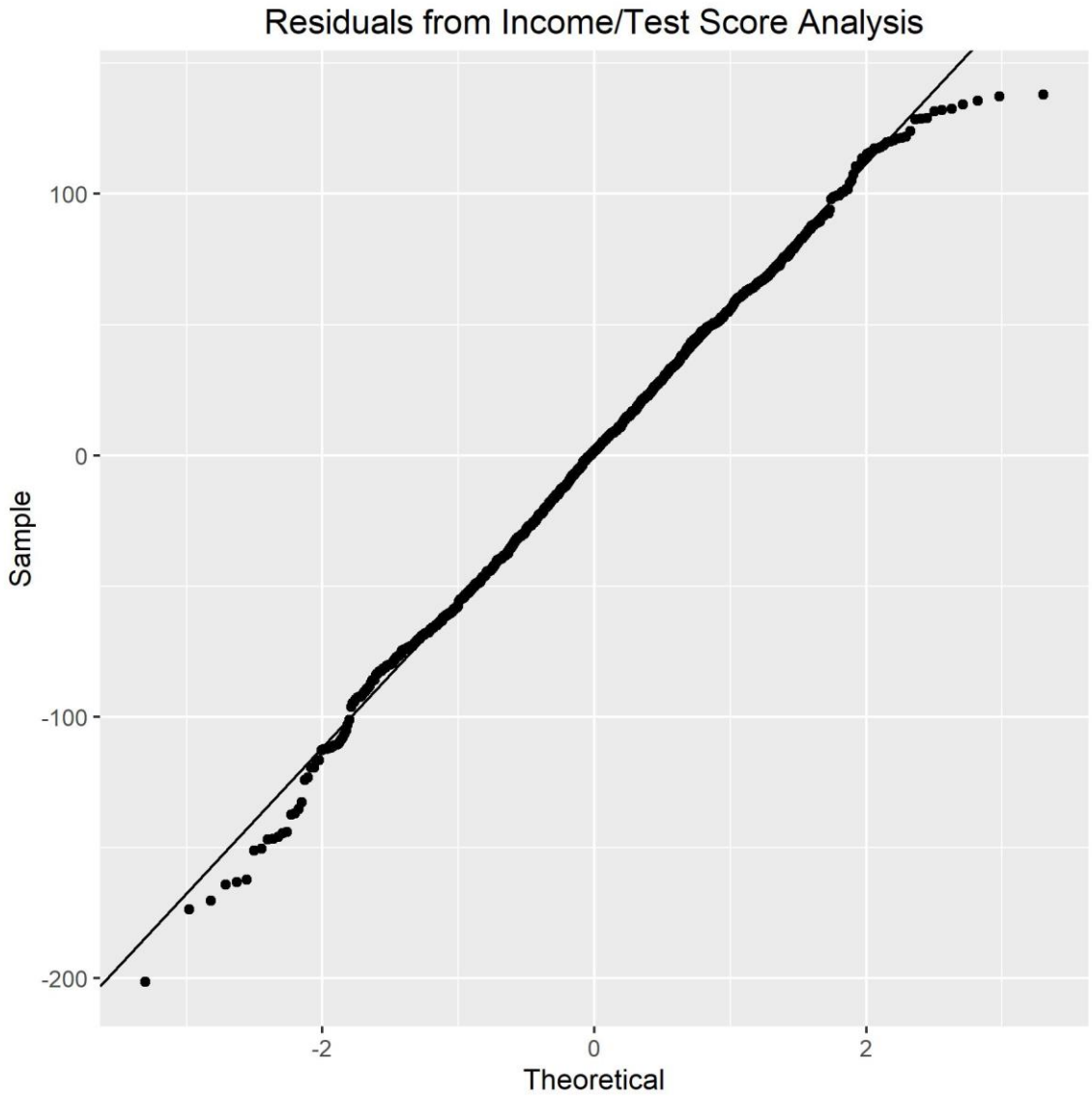
8. Normality of Residuals

The normality of the residuals can be checked using two plots: a histogram and a normal probability plot.



The histogram does indeed reflect a normal distribution in the residuals. This is clear from the general shape and from the similarity of the blue and red lines, which represent the density curve and the normal approximation, respectively.

Jordan Mayer
STAT 350
Lab 09
April 19, 2018



The normal probability plot also reflects a normal distribution, with most data points lying very close to the line of normality.

Both of these plots tell us that the distribution of the residuals is approximately normal.

Jordan Mayer
STAT 350
Lab 09
April 19, 2018

9. Linear Regression Assumptions

Performing linear regression requires four assumptions:

1. The data is collected from simple random samples (SRS).
   We cannot test this assumption, and must simply assume that it is true.
2. The relationship between the two variables is linear.
   We confirmed this through the original scatterplot and through the correlation value.
3. The standard deviation of the residuals is constant.
   We confirmed this through the scatterplot of the residuals.
4. The residuals are normally distributed.
   We confirmed this through our histogram and normal probability plot of the residuals.

Thus, all assumptions for linear regression analysis are reasonable, except the SRS assumption, which we cannot prove or disprove based on the available information.

10. 99% Confidence Intervals

```
> # check two-sided 99% confidence interval for slope and intercept
> confint(US_lm, level=0.99)
                    0.5 %       99.5 %
(Intercept)  1.533572e+03 1.562147e+03
MedianIncome 1.277939e-03 2.128047e-03
```

Slope:

99% CI: (0.001278, 0.002128)

We are 99% confident that the interval between 0.001278 and 0.002128 captures the slope of the linear regression line. In practical terms, the slope corresponds to the amount by which Test Score increases with each dollar by which Median Income increases. The fact that this value is small indicates that an increase in Median Income leads to a relatively small increase in Test Score.

Intercept:

99% CI: (1534, 1562)

We are 99% confident that the interval between 1534 and 1562 captures the intercept of the linear regression line. In practical terms, this value would correspond to the expected test score for a county with a median income of $0. In this situation, we should not be interested in the intercept – we would never expect a real US county to have a median income of $0.

Jordan Mayer
STAT 350
Lab 09
April 19, 2018

11. Evidence of Association

1. Definition of terms

β_1: the slope of the linear regression line, i.e. the amount by which Test Score increases for each dollar increase in Median Income.

2. Hypotheses

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

3. Test statistic (t_{ts} and F_{ts}), degree of freedom (DF), p-value (p)

```
> summary(US_lm)

Call:
lm(formula = TestScore ~ MedianIncome, data = US_typ)

Residuals:
     Min       1Q   Median       3Q      Max
-201.400  -39.003    1.657   39.675  137.918

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.548e+03  5.537e+00  279.56   <2e-16 ***
MedianIncome 1.703e-03  1.647e-04   10.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.9 on 1046 degrees of freedom
Multiple R-squared:  0.09272,   Adjusted R-squared:  0.09185
F-statistic: 106.9 on 1 and 1046 DF,  p-value: < 2.2e-16
```

$t_{ts} = 10.34$

$F_{ts} = 106.9$

$DF = 98$

$p < 2*10^{-16}$

4. Conclusion

$\alpha = 0.01$

➔ $p << \alpha$ → reject H_0

The data provides strong evidence ($p < 2*10^{-16}$) to the claim that there is an association between Test Score and Median Income.

12. CIs and Hypothesis Test

The results of (10) and (11) are similar. The 99% confidence interval for the slope of the linear regression line is entirely above zero, meaning we are 99% confident that an interval greater than zero captures the slope, and therefore that there is an association between Test Score and Median Income. Our hypothesis test produces the same conclusion, with an extremely low p-value. It is worth noting that (11) did make it clearer how strong the evidence is for an association between Test Score and Median Income.

Jordan Mayer
STAT 350
Lab 09
April 19, 2018

13. Summary

a) Appropriateness of Model

It would be inappropriate to apply this model, given the very low r^2 value ($r^2 = 0.09272 << 1$).

b) Relationship

There is a weak positive relationship between Median Income and Test Score. There is very strong evidence ($p < 2*10^{-16}$) that an association exists between Median Income and Test Score, but the slope of the linear regression line is not very large, and this regression line itself has a very small r^2 value.

c) Causality

Although our model itself should not be trusted, we do have strong evidence of a positive association between Median Income and Test Score. While correlation does not necessarily correspond to causation, it is reasonable to suppose that a causal relationship does exist here. Children from families with higher incomes are likely to have more resources available to prepare for a standardized test, such as preparatory books and courses. Thus, it would make sense if having a higher Median Income causes a county to have a higher average Test Score.