RMarkdown STAT 2218 F23 Problem Set 2 (Simple Linear Regression)

Jack Ambery

Due 10/13/2023

For Problems 1 through 3 you will use the data set MedSchool.csv. Each part is worth 10 points for a total of 100 points.

Data: MedSchool.csv used for Problems 1 to 3. Data collected from a random sample of students who applied to medical school in a given year. It contains the medical school admission status and information on GPA and standardized test scores for each subject.

* Subject: Code for different student subjects
* Acceptance: Indicator for whether student was accepted to med school or not: 1=accepted or 0=denied
* Sex: F=female or M=male
* SCIGPA: Grade Point Average in Science and Math classes
* GPA: Overall undergraduate grade point average
* Verbal: Verbal reasoning subscore on MCAT exam
* PhySci: Physical sciences subscore on MCAT exam
* Writing: Writing subscore on MCAT exam
* BioSci: Biological sciences subscore on MCAT exam
* MCAT: Score on the MCAT exam (sum of subscores above)
* NumApps: Number of medical schools applied to

# Input data file below.

Import data into RStudio using the base option, and then cut and paste YOUR file directory in the read.csv command below. It may be different from what I have listed below since your files are located in a different directory/folder. Then you can attach the file.

#Your file will be different! See instructions above.  
MedSchool <- read.csv("C:/Users/jacka/OneDrive - fairfield.edu/Fall 2023/Statistics II/datasets/MedSchool.csv")  
attach(MedSchool)

# 1. Problem 1 (50 pts)

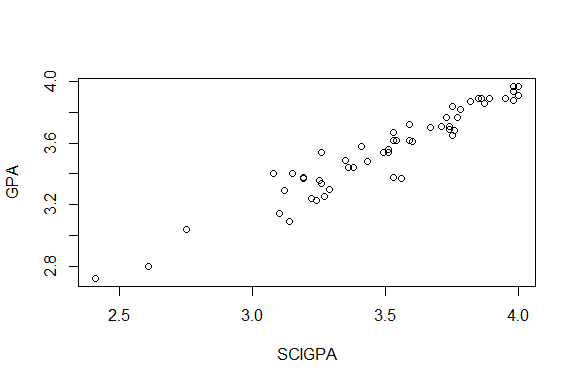
We wish to investigate whether the overall GPA of students can be predicted from the GPA in science and math.

## Problem 1a.

Make an appropriate graph and find the correlation. What do these show you about the relationship between overall GPA and the GPA in science and math?

## Problem 1a R Code

# This creates a scatterplot of GPA vs. SCIGPA  
plot(GPA~SCIGPA)



# cor(SCIGPA, GPA) gives us the correlation coefficient of SCIGPA and GPA  
cor(SCIGPA, GPA)

## [1] 0.9583393

## Problem 1a Analysis

Comment on the strength and direction of the relationship and whether a linear model is an appropriate one.

The scatter plot shows a strong, positive correlation between the GPA of students’ science and math GPAs and total GPAs. The given correlation coefficient of 0.96 also confirms that there is a very strong positive correlation between our two variables. These tell us that a linear model is appropriate and that it is likely that if someone’s math and science GPA is high, their total GPA is also high.

## Problem 1b

Is there evidence, at the 5% level, that overall GPA can be predicted from GPA in science and math? If the model is significant, give the least squares line to predict the overall GPA from GPA in science and math.

## Problem 1b R Code

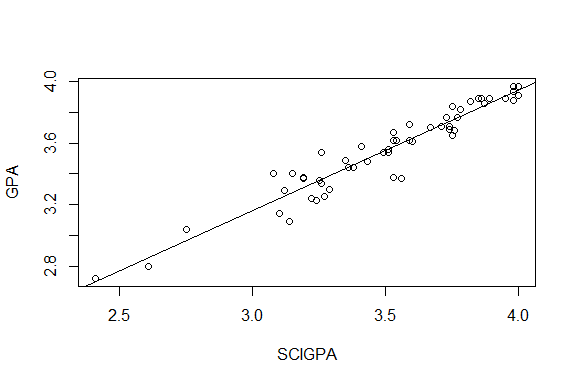
#create the linear model for GPA vs SCIGPA  
out = lm(GPA~SCIGPA)  
  
#give p-value and test statistic  
anova(out)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SCIGPA 1 4.0695 4.0695 596.62 < 2.2e-16 \*\*\*  
## Residuals 53 0.3615 0.0068   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#helps us find the least squares line  
summary(out)

##   
## Call:  
## lm(formula = GPA ~ SCIGPA)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.229907 -0.033059 0.003245 0.051159 0.176587   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.80758 0.11296 7.149 2.59e-09 \*\*\*  
## SCIGPA 0.78436 0.03211 24.426 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08259 on 53 degrees of freedom  
## Multiple R-squared: 0.9184, Adjusted R-squared: 0.9169   
## F-statistic: 596.6 on 1 and 53 DF, p-value: < 2.2e-16

#creates the scatter plot  
plot(GPA~SCIGPA)  
#writes least squares line on the scatter plot  
abline(out)



## Problem 1b Analysis

* State your null and alternative hypotheses
* Give the type of test you used, test statistic, dfs, and p-value
* Write a sentence summarizing your conclusion.
* If the model is significant, give the least squares model to predict the overall GPA from GPA in science and math.
* Refer to the relevant R output above.

To find out if overall GPA can be predicted from GPA in math and science, I used a linear regression hypothesis test.

H0: = 0; There is no linear relationship between SCIGPA and GPA. This is equivalent to = 0.

Ha: != 0; There is a linear relationship between SCIGPA and GPA. This is equivalent to != 0;

Results of the test:

F\* = 596.62; dfs = 1, 53; p-value = 2.2e-16

At the 5% level, there is evidence of a linear relationship between GPA in math and science and overall GPA. The p-value is well below our , 0.05.

Using the least squares model, we can predict overall GPA using the following formula:

= 0.808 + 0.784(SCIGPA)

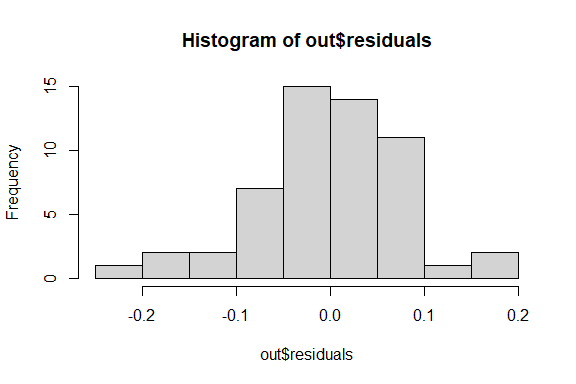
The least squares model tells us that for each increment of 1 in the GPA of science and math, overall GPA is predicted to increase by 0.784. A math and science GPA of 0 predicts an overall GPA of 0.808.

## Problem 1c

Check all of the assumptions for the hypothesis test and comment on whether each condition is or is not met. Should the test and model be used?

## Problem 1c R code

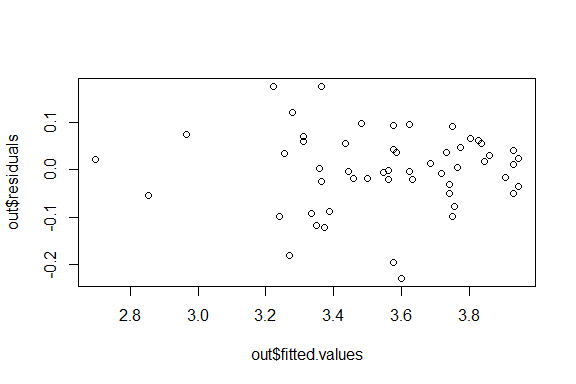
# 1. Linear model is appropriate  
# see graphs above  
  
# 2. Residuals or errors are Normal  
# histogram of the residuals  
hist(out$residuals)



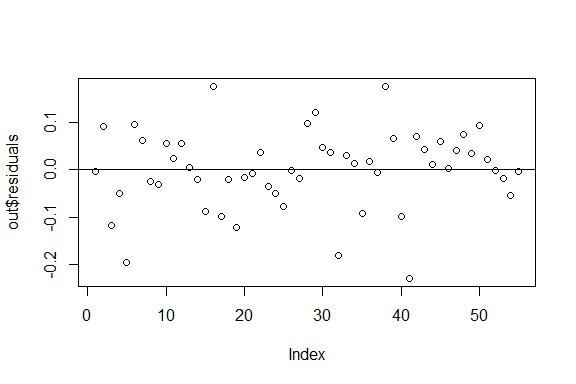
# shapiro test run on the residuals  
shapiro.test(out$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: out$residuals  
## W = 0.96905, p-value = 0.1668

# 3. Errors have constant variance:  
  
plot(out$residuals~out$fitted.values)



# 4. Errors are independent:   
  
plot(out$residuals)  
  
#reference line  
abline(0,0)



## Problem 1c Analysis

Explain whether each condition is met or not. Should the test and model be used?

1. Linear model is appropriate:

The scatter plot generated in part 1b shows that there is a strong linear trend between overall GPA and math and science GPA.

1. Residuals or errors are Normal:

It is safe to assume the residuals are normally distributed because the histogram of the residuals generated above appears to be normal.

A Shapiro-Wilk test can also confirm a normal distribution. Since the Shapiro-Wilk test above yields a high p-value of 0.167, it is safe to assume the residuals are normally distributed.

1. Errors have constant variance:

The scatter plot above shows that the residuals have constant variance. Since the points in the scatter plot are randomly distributed about the x-axis and there are no irregular groupings like V or U patterns, this test supports a linear model is appropriate.

1. Errors are independent:

Since the scatter plot of the residuals above displays points that are randomly scattered about the x-axis, this test supports that a linear model is appropriate. The are no dense groupings or patterns in the scatter plot.

Since all assumptions are met, it is appropriate to use a linear model.

## Problem 1d

Find a 90% interval for the predicted overall GPA for a student who had a 3.5 GPA in science and math.

## Problem 1d Code

#creating data frame with new observation x0  
newX = data.frame(SCIGPA = 3.5)  
  
#prediction interval for an individual response  
predict(out, newX, interval = 'prediction', level = 0.90)

## fit lwr upr  
## 1 3.552845 3.41333 3.692359

## Problem 1d Analysis

State what type of interval you used, give the interval and interpret the results.

To determine the predicted overall GPA for a student who had a 3.5 GPA in science and math with 90% confidence, I used a prediction interval for an individual response. Below is the calculated interval.

3.413 <= <= 3.692

With 90% confidence, we can expect that the overall GPA for a student with a 3.5 math and science GPA is between 3.413 and 3.692.

## Problem 1e

Find a 95% interval for the predicted average overall GPA for all students who have a 3.3 GPA in science and math.

## Problem 1e Code

#creating data frame with new observation x0  
newX = data.frame(SCIGPA = 3.3)  
  
#confidence interval for mean response  
predict(out, newX, interval = 'confidence', level = 0.95)

## fit lwr upr  
## 1 3.395973 3.37017 3.421775

## Problem 1e Analysis

State what type of interval you used, give the interval and interpret the results.

To determine the predicted overall GPA for all students who had a 3.3 GPA in science and math with 95% confidence, I used a confidence interval for mean response. Below is the calculated interval.

3.37 <= <= 3.422

With 95% confidence, we can expect that the overall GPA for all students with a 3.3 math and science GPA is between 3.37 and 3.422.

# 2. Problem 2 (20 points)

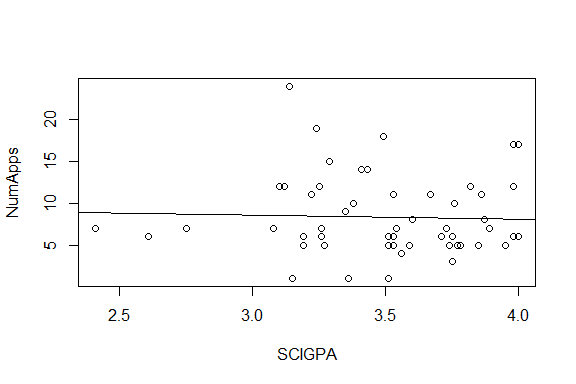
In this problem we will focus on science and math GPA and the number of medical schools students apply to. Specifically, we wish to investigate whether the number of med schools students apply to can be predicted from science and math GPA.

## Problem 2a.

Make an appropriate graph and find the correlation. What do these show you about the relationship between the number of med schools students apply to and science and math GPA?

## Problem 2a R Code

# This creates a scatter plot of NumApps vs. SCIGPA  
plot(NumApps~SCIGPA)  
abline(lm(NumApps~SCIGPA))



# cor(NumApps, GPA) gives us the correlation coefficient of SCIGPA and NumApps  
cor(SCIGPA, NumApps)

## [1] -0.03556831

## Problem 2a Analysis

Comment on the strength and direction of the relationship and whether a linear model is an appropriate one.

Looking at the scatter plot of the number of med schools applied to versus the math and science GPAs of students, it does not appear that there is much of any correlation at all. The trendline has a slightly negative direction, but I do not think this is significant given there is almost no correlation.

The calculated correlation coefficient is also -0.036. Since this value is so close to zero, this also supports no correlation between the number of applications and math and science GPA.

A linear model does not appear to be appropriate.

## Problem 2b

Is there evidence, at the 5% level, that the number of med schools a student applies to is linearly related to science and math GPA? If the model is significant, give the least squares line to predict number of med schools a student applies to from science and math GPA.

## Problem 2b R Code

#create the linear model for NumApps vs SCIGPA  
out = lm(NumApps~SCIGPA)  
  
#give p-value and test statistic  
anova(out)

## Analysis of Variance Table  
##   
## Response: NumApps  
## Df Sum Sq Mean Sq F value Pr(>F)  
## SCIGPA 1 1.54 1.5393 0.0671 0.7966  
## Residuals 53 1215.19 22.9281

#helps us find the least squares line  
summary(out)

##   
## Call:  
## lm(formula = NumApps ~ SCIGPA)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.533 -3.231 -1.726 2.764 15.462   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 10.0523 6.5492 1.535 0.131  
## SCIGPA -0.4824 1.8618 -0.259 0.797  
##   
## Residual standard error: 4.788 on 53 degrees of freedom  
## Multiple R-squared: 0.001265, Adjusted R-squared: -0.01758   
## F-statistic: 0.06714 on 1 and 53 DF, p-value: 0.7966

## Problem 2b Analysis

* State your null and alternative hypotheses
* Give the type of test you used, test statistic, dfs, and p-value
* Write a sentence summarizing your conclusion.
* If the model is significant, give the least squares model to predict number of med schools applied to from the science and math GPA.
* Refer to the relevant R output above.

To find out if the number of med schools applied to can be predicted from GPA in math and science, I used a linear regression hypothesis test.

H0: = 0; There is no linear relationship between math and science GPA and number of med schools applied to. This is equivalent to = 0.

Ha: != 0; There is a linear relationship between math and science GPA and number of med schools applied to. This is equivalent to != 0;

Results of the test:

F\* = 0.0671; dfs = 1, 53; p-value = 0.797

At the 5% level, there is not evidence of a linear relationship between the number of med schools applied to and math and science GPA. At 0.797, the p-value is well over our , 0.05.

Since there is no linear relationship, our linear model is not significant. The least squares model will not help us predict any values since this is the case.

# 3. Problem 3 (30 points)

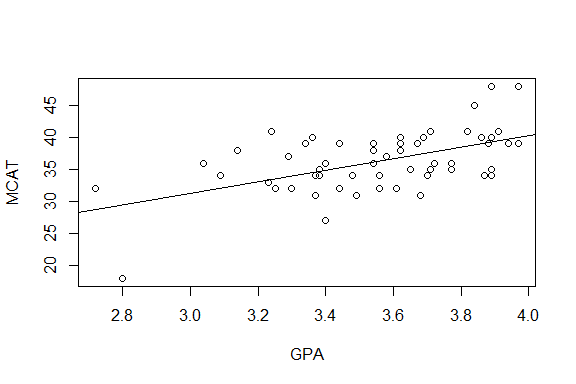
We wish to investigate whether students’ MCAT scores can be predicted from the overall GPA of the student.

## Problem 3a.

Make an appropriate graph and find the correlation. What do these show you about the relationship between MCAT score and overall GPA?

## Problem 3a R Code

# This creates a scatter plot of MCAT score vs. GPA  
plot(MCAT~GPA)  
abline(lm(MCAT~GPA))



# cor(GPA, MCAT) gives us the correlation coefficient of GPA and MCAT score  
cor(GPA, MCAT)

## [1] 0.5414202

## Problem 3a Analysis

Comment on the strength and direction of the relationship and whether a linear model is an appropriate one.

The scatter plot of MCAT score vs. GPA does seem to suggest a linear relationship of moderate strength.

The calculated correlation coefficient also suggested a moderate correlation with a value of 0.541.

## Problem 3b

Is there evidence, at the 5% level, that MCAT score can be predicted from overall GPA? If the model is significant, give the least squares line to the MCAT score from overall GPA.

## Problem 3b R Code

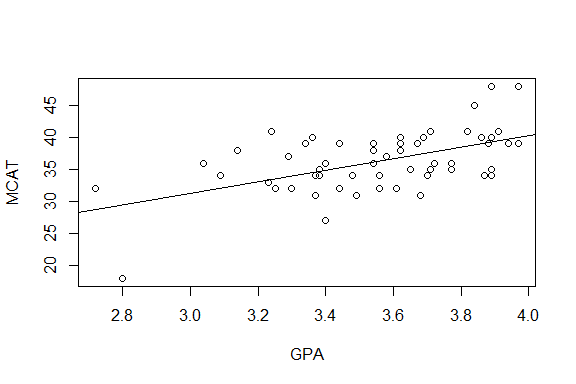
#create the linear model for MCAT vs GPA  
out = lm(MCAT~GPA)  
  
#give p-value and test statistic  
anova(out)

## Analysis of Variance Table  
##   
## Response: MCAT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GPA 1 367.27 367.27 21.979 1.969e-05 \*\*\*  
## Residuals 53 885.64 16.71   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#helps us find the least squares line  
summary(out)

##   
## Call:  
## lm(formula = MCAT ~ GPA)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.4148 -2.5168 -0.1519 2.6653 8.6616   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.923 6.922 0.567 0.573   
## GPA 9.104 1.942 4.688 1.97e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.088 on 53 degrees of freedom  
## Multiple R-squared: 0.2931, Adjusted R-squared: 0.2798   
## F-statistic: 21.98 on 1 and 53 DF, p-value: 1.969e-05

#creates the scatter plot  
plot(MCAT~GPA)  
#writes least squares line on the scatter plot  
abline(out)



## Problem 3b Analysis

* State your null and alternative hypotheses
* Give the type of test you used, test statistic, dfs, and p-value
* Write a sentence summarizing your conclusion
* If the model is significant, give the least squares model to predict MCAT score from overall GPA.
* Refer to the relevant R output above.

To find out if MCAT score can be predicted from GPA, I used a linear regression hypothesis test.

H0: = 0; There is no linear relationship between MCAT score and GPA. This is equivalent to = 0.

Ha: != 0; There is a linear relationship between MCAT score and GPA. This is equivalent to != 0;

Results of the test:

F\* = 21.979; dfs = 1, 53; p-value = 1.969e-05

At the 5% level, there is evidence of a linear relationship between MCAT score and GPA. The p-value is well below our , 0.05.

Using the least squares model, we can predict MCAT score using the following formula:

= 3.92 + 9.104(GPA)

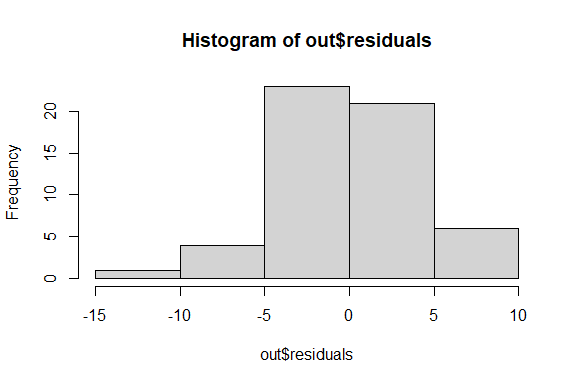
The least squares model tells us that for each increment of 1 in the GPA, MCAT score is predicted to increase by 9.104. A GPA of 0 predicts an MCAT score of 3.92.

## Problem 3c

Check all of the assumptions for the hypothesis test and comment on whether each condition is or is not met. Should the test and model be used?

## Problem 3c R code

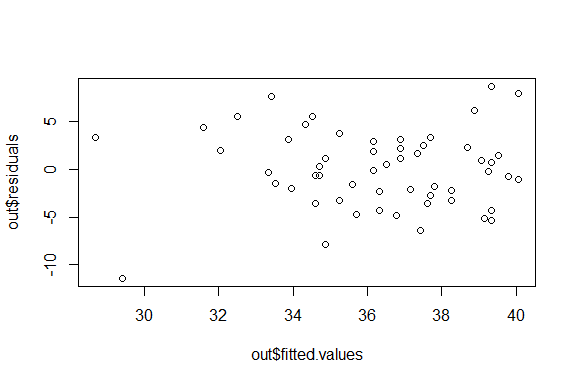
# 1. Linear model is appropriate  
# see graphs above  
  
# 2. Residuals or errors are Normal  
# histogram of the residuals  
hist(out$residuals)



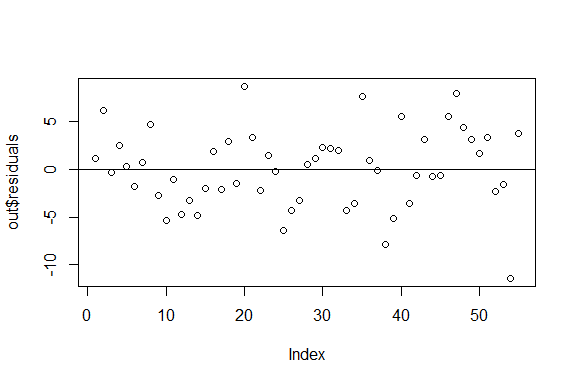
# shapiro test run on the residuals  
shapiro.test(out$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: out$residuals  
## W = 0.99256, p-value = 0.9817

# 3. Errors have constant variance:  
  
plot(out$residuals~out$fitted.values)



# 4. Errors are independent:   
  
plot(out$residuals)  
  
#reference line  
abline(0,0)



## Problem 3c Analysis

Explain whether each condition is met. Should the test and model be used?

1. Linear model is appropriate:

The scatter plot generated in part 3b shows that there is a moderate linear trend between MCAT score and GPA. This supports the use of a linear model.

1. Residuals or errors are Normal:

It is safe to assume the residuals are normally distributed because the histogram of the residuals generated above appears to be normal.

A Shapiro-Wilk test can also confirm a normal distribution. Since the Shapiro-Wilk test above yields a high p-value of 0.982, it is safe to assume the residuals are normally distributed.

1. Errors have constant variance:

The scatter plot above shows that the residuals have constant variance. Since the points in the scatter plot are randomly distributed about the x-axis and there are no irregular groupings like V or U patterns, this test supports a linear model is appropriate.

1. Errors are independent:

Since the scatter plot of the residuals above displays points that are randomly scattered about the x-axis, this test supports that a linear model is appropriate. The are no dense groupings or patterns in the scatter plot.

Since all assumptions are met, it is appropriate to use a linear model.

End of Problem Set