INFX 573: Problem Set 5 - Learning from Data

Pierre Augustamar

Due: Tuesday, November 8, 2016

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset5.Rmd` file from Canvas. Open `problemset5.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset5.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps5.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse);
library(Sleuth3) # Contains data for problemset
library(UsingR) # Contains data for problemset
library(MASS) # Modern applied statistics functions
```

1. Davis et al. (1998) collected data on the proportion of births that were male in Denmark, the Netherlands, Canada, and the United States for selected years. Davis et al. argue that the proportion of male births is declining in these countries. We will explore this hypothesis. You can obtain this data as follows:

   (a) Use the `lm` function in **R** to fit four (one per country) simple linear regression models of the yearly proportion of males births as a function of the year and obtain the least squares fits. Write down the estimated linear model for each country.

   ```
   #linear model of the relationship between year's proportion and male's birth in Denmark
   denmark.lm = lm(Denmark ~ Year, data=ex0724)
   coefficients(denmark.lm) #Get coefficient result
   ```

   ```
   ##   (Intercept)          Year
   ##  5.987233e-01 -4.288538e-05
   ```

- Denmark's males births = 5.987233e-01 -4.288538e-05*yearly proportions

```
#linear model of the relationship between year's proportion and male's birth in the Netherands
netherlands.lm = lm(Netherlands ~ Year, data=ex0724)
coefficients(netherlands.lm) #Get Coefficient result
```

```
##   (Intercept)          Year
##  6.723984e-01 -8.084321e-05
```

- Netherlands' males births = 6.723984e-01 -8.084321e-05*yearly proportions

```
#linear model of the relationship between year's proportion and male's birth in Canada
canada.lm = lm(Canada ~ Year, data=ex0724)
coefficients(canada.lm) #Get coefficient result
```

```
##   (Intercept)          Year
##  0.7337857143 -0.0001111688
```

- Canada's males births = 0.7337857143 -0.0001111688*yearly proportions

```
#linear model of the relationship between year's proportion and male's birth in the US
usa.lm = lm(USA ~ Year, data=ex0724)
coefficients(usa.lm) #Get the coefficent results
```

```
##   (Intercept)          Year
##  6.200857e-01 -5.428571e-05
```

- USA's males births = 6.200857e-01 -5.428571e-05*yearly proportions

(b) Obtain the t-statistic for the test that the slopes of the regression lines are zero, for each of the four countries. Is there evidence that the proportion of births that are male is truly declining over this period?

```
usa.summary = summary(usa.lm) #generate summary model for USA's data
canada.summary = summary(canada.lm) #generate summary model for Canada's data
netherlands.summary = summary(netherlands.lm) #generate summary model for Netherlands' data
denmark.summary = summary(denmark.lm) #generate summary model for Denmark's data
```

Response

Given the following hypothesis. * H0: There is no evidence that the proportion of male births is declining
* Ha: There is evidence that the proportion of male births is declining
Using the following url, http://www.socscistatistics.com/pvalues/tdistribution.aspx, to calculate the p-value for a significance level = 0.05 and a two-tailed hypothesis.

Coefficient - t-value or t-statistic for Denmark

t-value = 14.673. Since the t-value is far away from 0 indicates that we could reject the null hypothesis. Also, the P-Value is .000126. The p-value result is significant at $p < 0.05$. Thus, there is evidence that the proportion of male births is declining in Denmark.

Coefficient - t-value or t-statistic for Netherlands

t-value = 24.08. Since the t-value is far away from 0 indicates that we could reject the null hypothesis. Also, the calculated p-value is 0.00018. The p-value result is significant at $p < 0.05$. Thus, there is evidence that the proportion of male births is declining in Netherlands.

Coefficient - t-value or t-statistic for Canada

t-value = 13.39. Since the t-value is far away from 0 indicates that we could reject the null hypothesis. Also, the calculated p-value is 0.00018. the p-value result is significant at $p < 0.05$. Thus, there is evidence that the proportion of male births is declining in Canada.

Coefficient - t-value or t-statistic for USA

t-value = 33.340. Since the t-value is far away from 0 indicates that we could reject the null hypothesis. Also, the p.value is less than 0.00001. Thus, because at a significance level less than 5%, there is evidence that the proportion of male births is declining in the US.

2. Regression was originally used by Francis Galton to study the relationship between parents and children. One relationship he considered was height. Can we predict a man's height based on the height of his father? This is the question we will explore in this problem. You can obtain data similar to that used by Galton as follows:

```
# Import and look at the height data
heightData <- tbl_df(get("father.son"))
```

(a) Perform an exploratory analysis of the dataset. Describe what you find. At a minimum you should produce statistical summaries of the variables, a visualization of the relationship of interest in this problem, and a statistical summary of that relationship.

```
dim(heightData) # obtain dimension of the data frame
```

```
## [1] 1078    2
```

```
describe(heightData) #describe in more details the height data frame
```

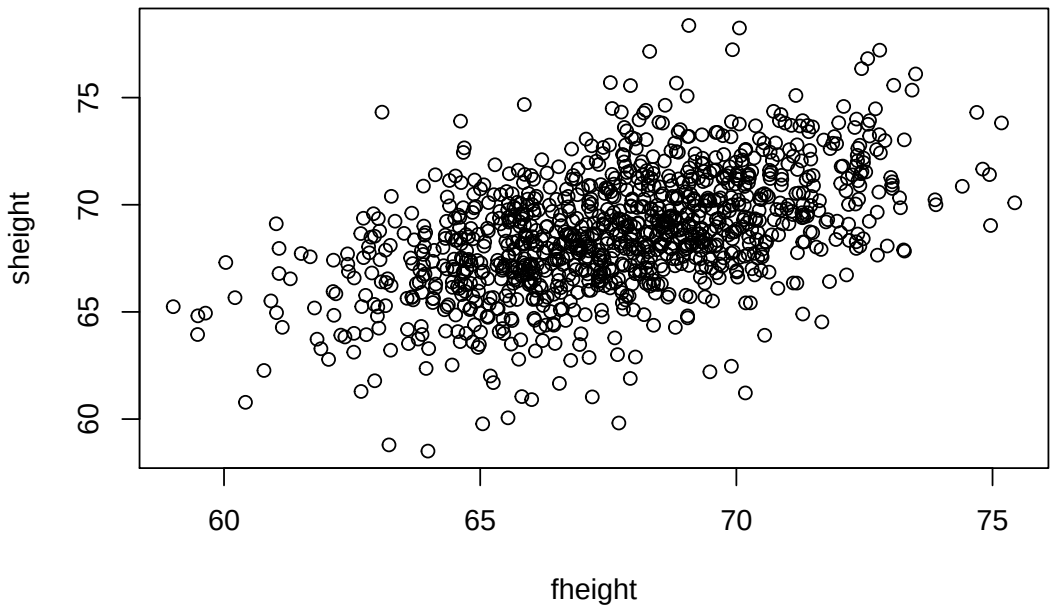
```
## heightData
##
##  2  Variables      1078  Observations
## ---------------------------------------------------------------------------
## fheight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1078        0     1078        1    67.69     3.11    63.06    64.30
##      .25      .50      .75      .90      .95
##    65.79    67.77    69.60    71.29    72.34
##
## lowest : 59.00800 59.48391 59.48980 59.63777 60.03436
## highest: 74.80830 74.94058 74.96203 75.17469 75.43393
## ---------------------------------------------------------------------------
## sheight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1078        0     1076        1    68.68    3.138    64.00    65.23
##      .25      .50      .75      .90      .95
##    66.93    68.62    70.47    72.25    73.29
##
## lowest : 58.50708 58.79456 59.77827 59.81693 60.05859
## highest: 77.15504 77.21338 77.23474 78.24760 78.36479
## ---------------------------------------------------------------------------
```

```
summary(heightData) #summarize heightData
```

```
##     fheight         sheight
##  Min.   :59.01   Min.   :58.51
##  1st Qu.:65.79   1st Qu.:66.93
##  Median :67.77   Median :68.62
##  Mean   :67.69   Mean   :68.68
##  3rd Qu.:69.60   3rd Qu.:70.47
##  Max.   :75.43   Max.   :78.36
```

The data shows that fathers have an average height = 67.69, the minimum height is 59.00 and the tallest is 75.43. On the other hand sons' have an average height = 68.68, the minimum height = 58.50, and the tallest is 78.36. Also, there were no missing records or no records that are equal to NA. All the fathers's entries have distinct heights, on the other hand, two records in the son's column have identical heights. Finally, there were 1078 records with two variables.

```
#visualization
plot(heightData) # scatter plot between father's height and son's height
```



The graph shows that there is a strong positive correlation son's height and father's height. Also, it is important to note that there are few outliers but nothing that affected the upward trend. I can easily fit an uphill line depicting a positive linear trend.

(b) Use the `lm` function in R to fit a simple linear regression model to predict son's height as a function of father's height. Write down the model,

$$\hat{y}_{\texttt{sheight}} = \hat{\beta}_0 + \hat{\beta}_i \times \texttt{fheight}$$

filling in estimated coefficient values and interpret the coefficient estimates.

```r
#generate the linear regression model
height.lm = lm(heightData$sheight ~ heightData$fheight, data = heightData)
summary(height.lm) #generate summary of the linear regression model
```

```
##
## Call:
## lm(formula = heightData$sheight ~ heightData$fheight, data = heightData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8772 -1.5144 -0.0079  1.6285  8.9685
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.88660    1.83235   18.49   <2e-16 ***
## heightData$fheight  0.51409    0.02705   19.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.437 on 1076 degrees of freedom
## Multiple R-squared:  0.2513, Adjusted R-squared:  0.2506
## F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Linear regression model

Son's height = 33.88 + 0.51*father's height

Residuals

The distribution of the residuals appears to be strongly symmetrical with a mean under zero. Thus, most of the data are strongly concentrated towards specific trends and that there will be a minimal number of outliers.

Coefficient Estimate

The coefficient estimate shows that on average a son's grows an additional 33 inches from his father. The slope which is the effect of a father's height on his son growth. The slope in this model shows that for a son's increase in height, it goes up by 0.51409 inches.

Coefficient - Standard Error

The standard error shows that a father's height can vary by 0.02705 inches.

Coefficient - t-value

The t-value at 18.49 for the intercept and 19.01 for the slope tells is far away for 0. Thus, it would signify that we could reject the null hypothesis. In this scenario, we could say that a relationship exists between a father's height and the expected height of his son.

Coefficient - Pr(>|t|)

The p-values in this example are very close to zero. Thus, it is a sign that we can reject the null hypothesis and conclude that there is a relationship between a father's height and the resulted height of his son. Also, the three stars represent a highly significant p-value.

Residual Standard Error

The residual standard error was calculated with 1076 degrees for freedom. Also, it shows that the actual height for a son to catch up to his father's height can deviate from the true regression line by about 2.437 inches, on average. Thus, given the mean height of a son's growth is about 33.88 and that the residual standard error is 2.473, then the percentage error is 7.19%. The distribution of the residuals appear to be strongly symmetrical with a mean under zero. Thus, most of the data are strongly concentrated towards a specific trends and that there will be a minimal number of outliers.

Coefficient Estimate

The coefficient estimate shows that on average a son's grows an additional 33 inches from his father. The slope which is the effect of a father's height on his son growth. The slope in this model shows that for a son's ncrease in height, it goes up by 0.51409 inch.

Coefficient - Standard Error

The standard error shows that a father's height can vary by 0.02705 inches.

Coefficient - t-value

The t-value at 18.49 for the intercept and 19.01 for the slope tells is far away for 0. Thus, it would signify that we could reject the null hypothesis. In this scenario, we could say that a relationship exists between a father's height and the expected height of his son.

Coefficient - Pr(>|t|)

The p-values in this example is very close to zero. Thus, it is a sign that we can reject the null hypothesis and conclude that there is a relationship between a father's height and the resulted height of his son. Also, the three stars represent a highly significant p-value.

Residual Standard Error

The residual standard error was calculated with a 1076 degrees for feedom. Also, it shows that the actual height for a son to to catch up to his father's height can deviate from the true regression line by about 2.437 inches, on average. Thus, given the mean height of a son's growth is about 33.88 and that the residual standard error is 2.473, then the percentage error is 7.19%.

(c) Find the 95% confidence intervals for the estimates. You may find the `confint()` command useful.

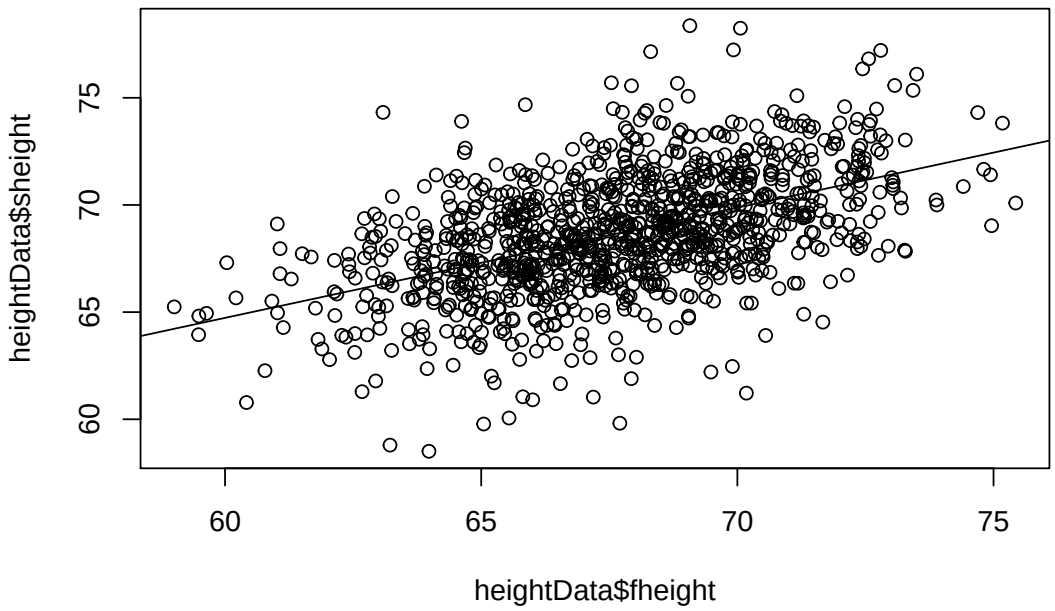
```
#generare confidence interval at 95%
confint(height.lm, level=0.9)
```

```
##                             5 %       95 %
## (Intercept)          30.8700534 36.9031553
## heightData$fheight    0.4695635  0.5586226
```

The 95% confidence intervals of the estimates has a lower bound of 30.87 and upper bound of 36.90 for son's height or inetercept. And, it has a a lower bound of 0.46, and upper bound of 0.55 for the father's height.

(d) Produce a visualization of the data and the least squares regression line.

```
#generate scatter plots
plot(heightData$fheight, heightData$sheight)
abline(height.lm)
```

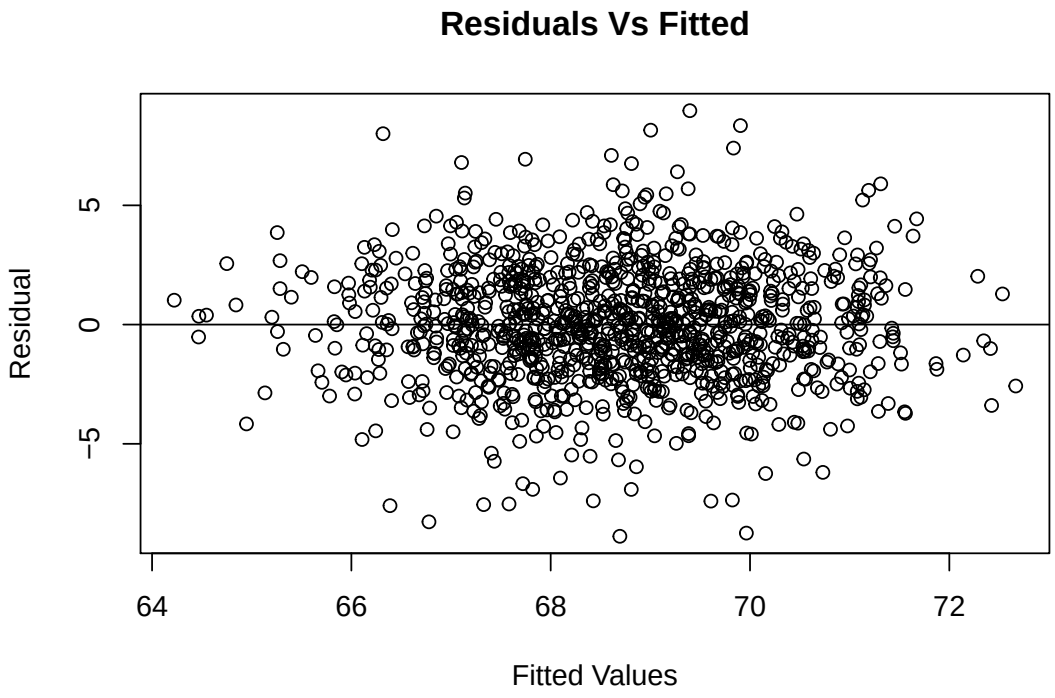


The graph shows a positive linear trend with an intercept around 64 inches.

(e) Produce a visualization of the residuals versus the fitted values. (You can inspect the elements of the linear model object in R using `names()`). Discuss what you see. Do you have any concerns about the linear model?

```
height.rs = resid(height.lm) #get the residuals value
height.fitted = fitted.values(height.lm) #get fitted values

#plot the residuals against the fitted values of the variable sheight
plot(height.fitted, height.rs, ylab="Residual", xlab="Fitted Values", main="Residuals Vs Fitte
abline(0,0)
```



The points in the graph appear to be randomly scattered around zero. It shows no obvious pattern between residuals and the fitted values. Ideally, we want to have a graph with no pattern as this will indicates a good fit for a linear model. Thus, I don't anticipate any concerns with the linear model.

(f) Using the model you fit in part (b) predict the height was 5 males whose father are 50, 55, 70, 75, and 90 inches respectively. You may find the `predict()` function helpful.

```
x <- heightData$sheight # set the son's height to x
y <- heightData$fheight # set the father's height to y
fitHeight <- lm(x ~ y, heightData) # create a model
#predict height for father's that are either 50, 55, 70, 75 and 90 inches.
predict(fitHeight, data.frame(y=c(50,55,70,75,90)), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 59.59126 54.71685 64.46566
## 2 62.16172 57.33140 66.99204
## 3 69.87312 65.08839 74.65785
## 4 72.44358 67.64470 77.24246
## 5 80.15498 75.22740 85.08255
```

The results tells us that 95 percent prediction for:

- father's heigh of 50 inches will have a son between 54.71 and 64.46 inches.
- father's height of 55 inches will have a son between 57.33 and 66.99 inches.
- father's height of 70 inches will have a son between 65.08 and 74.65 inches.
- father's height of 75 inches will have a son between 67.64 and 77.24 inches.
- father's height of 90 inches will have a son between 75.22 and 85.08 inches.

3. **Extra Credit:**

(a) What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model? ###Response * Explanatory Variable is the indepent variable. * Explanatory variable is also known as the predictor variable. * The effect of the Explanatory Variable on the expected value of the dependent variable is additive. * Linear relationship between Explanatory Variable (independent) and Response (dependent) * Explanatory Variable may be caused by, or directly influenced by the other variables *

(b) Why can an R^2 close to one not be used as evidence that the simple linear regression model is appropriate? Because R^2 tends to vary based on the noise level. If noise are added or removed through useless covarates to the model then the R^2 will vary and could improve the R^2.

(c) Consider a regression of weight on height for a sample of adult males. Suppose the intercept is 5 kg. Does this imply that males of height 0 weigh 5 kg, on average? Would this imply that the simple linear regression model is meaningless?

Response - Does this imply that males of height 0 weigh 5 kg, on average?

No, I cannot infer that males of height = 0 weigh 5kg because height = 0 is not part of the given value.

Response - Would this imply that the simple linear regression model is meaningless?

No, we do not have sufficient data to come to such conclusion. It may not be applicable for this case, but applicable for other situations.

(d) Suppose you had data on pairs (X, Y) which gave the scatterplot been below. How would you approach the analysis?

Response

I would would analyze the scatterplot based on the direction, form, strength, and outliers.

Direction

I would check whether the positive values of the independent variable are associated with positive larger values of the dependent variable which is known as a positive trend. In the similar case, I would also check if the positive independent variable is associated with smaller values of the dependent variable which is known as a negative trend. Based on this analysis, looking at the graph from left to right we can see that this chart tends to show a positive trending.

Form

The form would tell us whether there is an association or not, and whether this association is a straight line or a curve. A linear form would show a straight line, while a non-linear would show a curve of some point. When looking at the graph, there is both a sense of an uphill linear form as well as some upward curve. Since there is not a true particular linear form, then we would conclude this is a non-linear form.

Strength

The graph tells us how tightly clustered the points are around the form. Looking at the chart, I would say that there is almost near zero correlation because there are so many scattered points. These points show the different form of patterns. We see both the possibility of an upward positive linear trend as well as a curvy positive line. Thus, because of the discrepancy, we could not clearly set in a particular pattern.

Outliers

Outliers are data points that are away from the trending line. In this graph, depending on what we are trying to analyze we could remove a set of data to show the strength of the positive linear trend. Also, we could remove a few sets of outliers to show the possible curve.

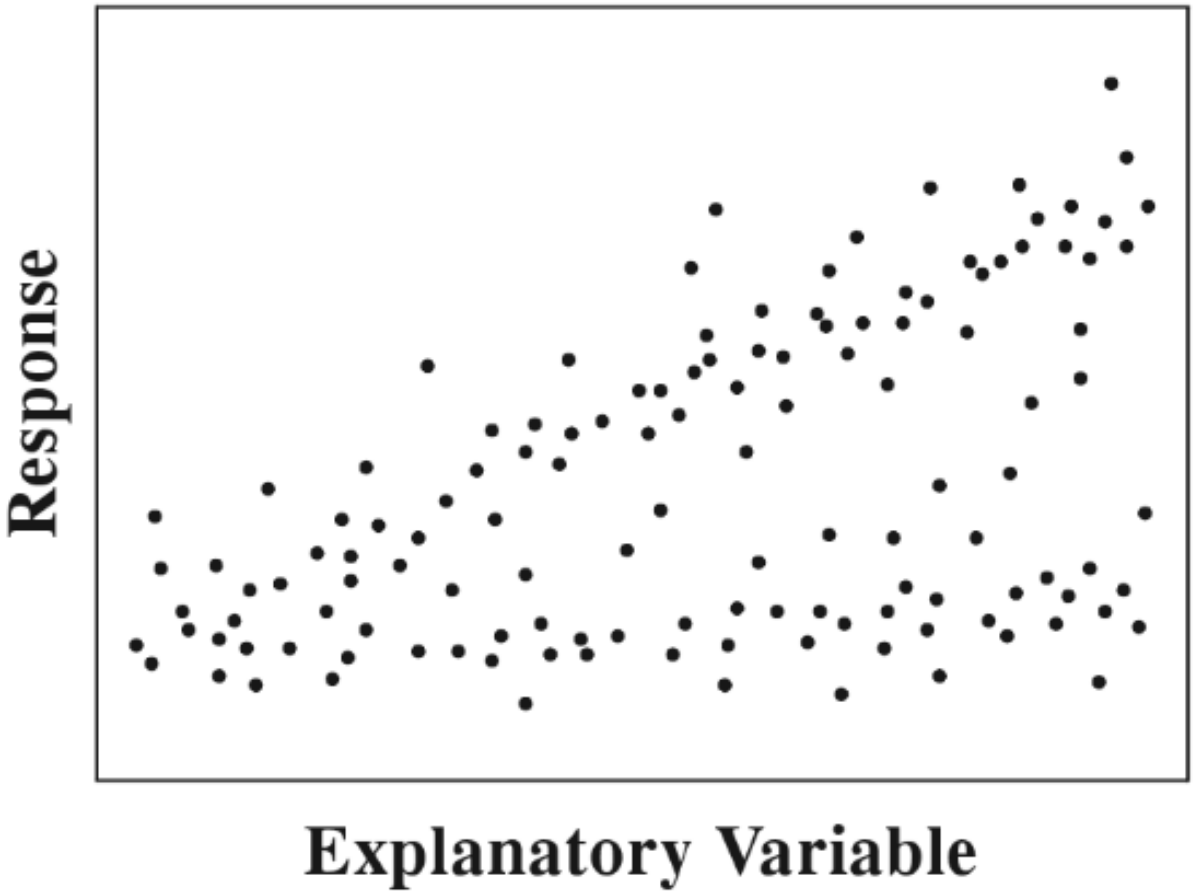


Figure 1: Scatterplot for Extra Credit (d).