INFX 573: Problem Set 5 - Learning from Data

Shuyang Wu

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:

```
# Find and look at the birth data
#library(help="Sleuth3")
birthData <- ex0724
summary(birthData)
```

```
##       Year         Denmark        Netherlands         Canada
##  Min.   :1950   Min.   :0.5108   Min.   :0.5087   Min.   :0.5120
##  1st Qu.:1961   1st Qu.:0.5127   1st Qu.:0.5120   1st Qu.:0.5128
##  Median :1972   Median :0.5141   Median :0.5129   Median :0.5136
##  Mean   :1972   Mean   :0.5142   Mean   :0.5130   Mean   :0.5137
##  3rd Qu.:1983   3rd Qu.:0.5153   3rd Qu.:0.5139   3rd Qu.:0.5145
##  Max.   :1994   Max.   :0.5175   Max.   :0.5160   Max.   :0.5153
##                                                   NA's   :24
```

```
##       USA
##  Min.   :0.5120
##  1st Qu.:0.5122
##  Median :0.5126
##  Mean   :0.5126
##  3rd Qu.:0.5128
##  Max.   :0.5134
##  NA's   :24
```

```
#Denmark lm
fit.d <- lm(Denmark ~ Year, data = birthData)
summary(fit.d)
```

```
##
## Call:
## lm(formula = Denmark ~ Year, data = birthData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003225 -0.001339  0.000089  0.001119  0.003790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.987e-01  4.080e-02  14.673   <2e-16 ***
## Year        -4.289e-05  2.069e-05  -2.073   0.0442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001803 on 43 degrees of freedom
## Multiple R-squared:  0.09083,    Adjusted R-squared:  0.06968
## F-statistic: 4.296 on 1 and 43 DF,  p-value: 0.04424
```

```
#Netherlands lm
fit.n <- lm(Netherlands ~ Year, data = birthData)
summary(fit.n)
```

```
##
## Call:
## lm(formula = Netherlands ~ Year, data = birthData)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0031437 -0.0008246  0.0002819  0.0009287  0.0021478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.724e-01  2.792e-02   24.08  < 2e-16 ***
## Year        -8.084e-05  1.416e-05   -5.71 9.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001233 on 43 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.418
## F-statistic: 32.61 on 1 and 43 DF,  p-value: 9.637e-07
```

```r
#Canada lm
fit.c <- lm(Canada ~ Year, data = birthData)
summary(fit.c)
```

```
##
## Call:
## lm(formula = Canada ~ Year, data = birthData)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.494e-03 -6.161e-04 -8.312e-05  4.951e-04  1.284e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.338e-01  5.480e-02  13.390 3.98e-11 ***
## Year        -1.112e-04  2.768e-05  -4.017 0.000738 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000768 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.4592, Adjusted R-squared:  0.4307
## F-statistic: 16.13 on 1 and 19 DF,  p-value: 0.0007376
```

```r
#USA lm
fit.u <- lm(Denmark ~ Year, data = birthData)
summary(fit.u)
```

```
##
## Call:
## lm(formula = Denmark ~ Year, data = birthData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003225 -0.001339  0.000089  0.001119  0.003790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.987e-01  4.080e-02  14.673   <2e-16 ***
## Year        -4.289e-05  2.069e-05  -2.073   0.0442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001803 on 43 degrees of freedom
## Multiple R-squared:  0.09083,    Adjusted R-squared:  0.06968
## F-statistic: 4.296 on 1 and 43 DF,  p-value: 0.04424
```

(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. Denmark: malebirth = -4.289e-05 x Year + 5.987e-01 Netherlands: makebirth = -8.084e-05 x Year + 6.724e-01 Canada: malebirth = -1.112e-04 x Year + 7.338e-01 USA: malebirth = -4.289e-05 x Year + 5.987e-01

(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? Denmark: 4.296 on 1 and 43 DF, p-value: 0.04424 Netherlands: 32.61 on 1 and 43

DF, p-value: 9.637e-07 Canada: 16.13 on 1 and 19 DF, p-value: 0.0007376 USA: 4.296 on 1 and 43 DF, p-value: 0.04424 Given all p-values are smaller than 0.05, the null hypothesis that birth is not correlated with year can be rejected. The four negative slopes in the models suggests a negative association between year and birth, so it's suggestive that male birth rates are declining over the years.

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.

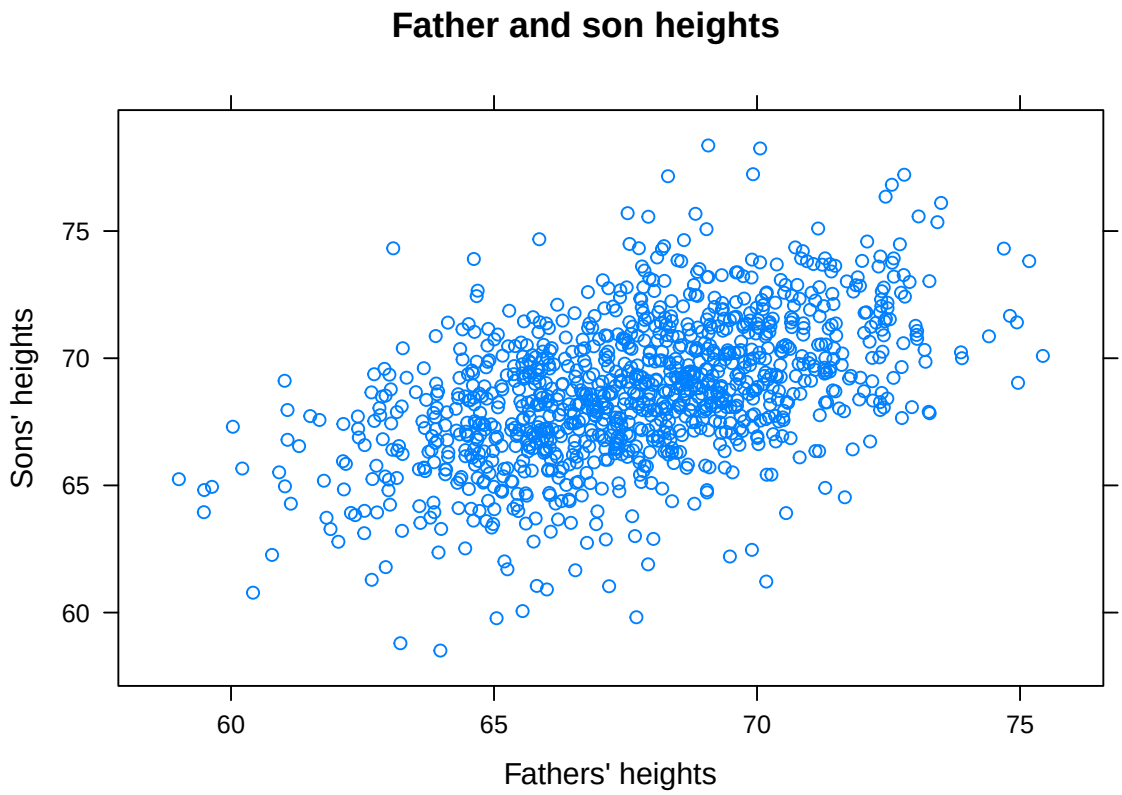
```
# exlore data
str(heightData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1078 obs. of  2 variables:
##  $ fheight: num  65 63.3 65 65.8 61.1 ...
##  $ sheight: num  59.8 63.2 63.3 62.8 64.3 ...
```

```
summary(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
```

```
# Visualize data
xyplot(sheight ~ fheight, data = heightData,
  xlab = "Fathers' heights",
  ylab = "Sons' heights",
  main = "Father and son heights"
)
```




```
cor(heightData$fheight, heightData$sheight)
```

```
## [1] 0.5013383
```

The dataset has two columns, fathers' heights and the corresponding sons' heights. fathers' heights have a mean of 67.77, sons' heights have a mean of 68.68, both distribution seems normal(no obvious skewness).The correlation coefficient being 0.501 implies a weak positive correlation between father and sons' heights.

```
# exlore data
fit.1 <- lm(sheight ~ fheight, data = heightData)
#b
summary(fit.1)
```

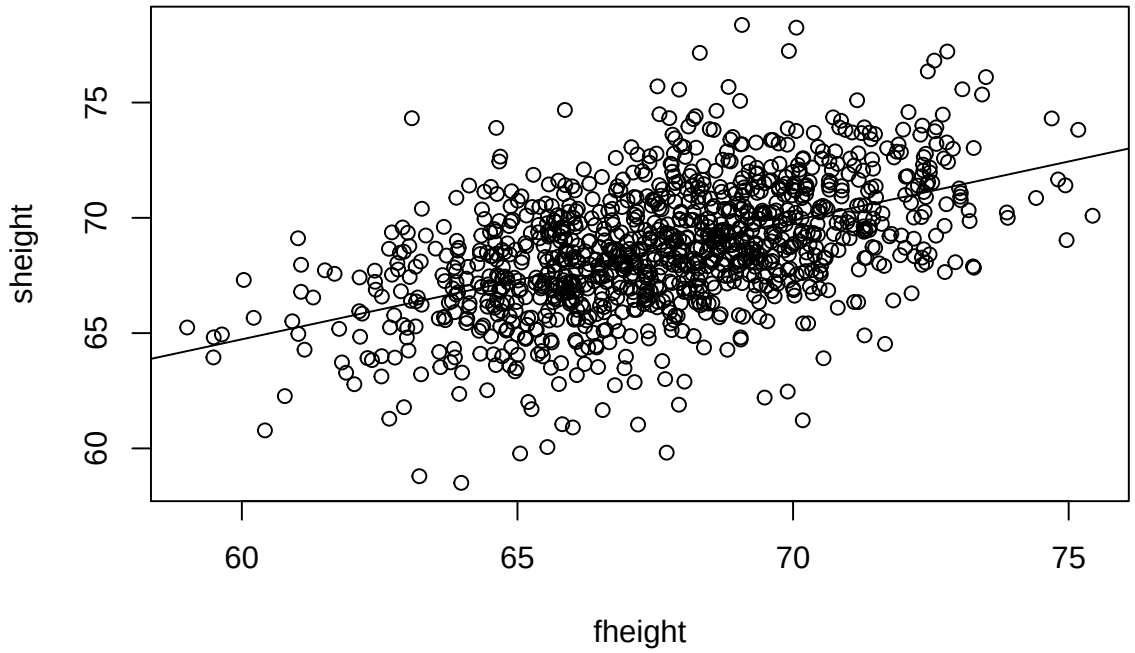
```
##
## Call:
## lm(formula = sheight ~ fheight, data = heightData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8772 -1.5144 -0.0079  1.6285  8.9685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.88660    1.83235   18.49   <2e-16 ***
## 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
```

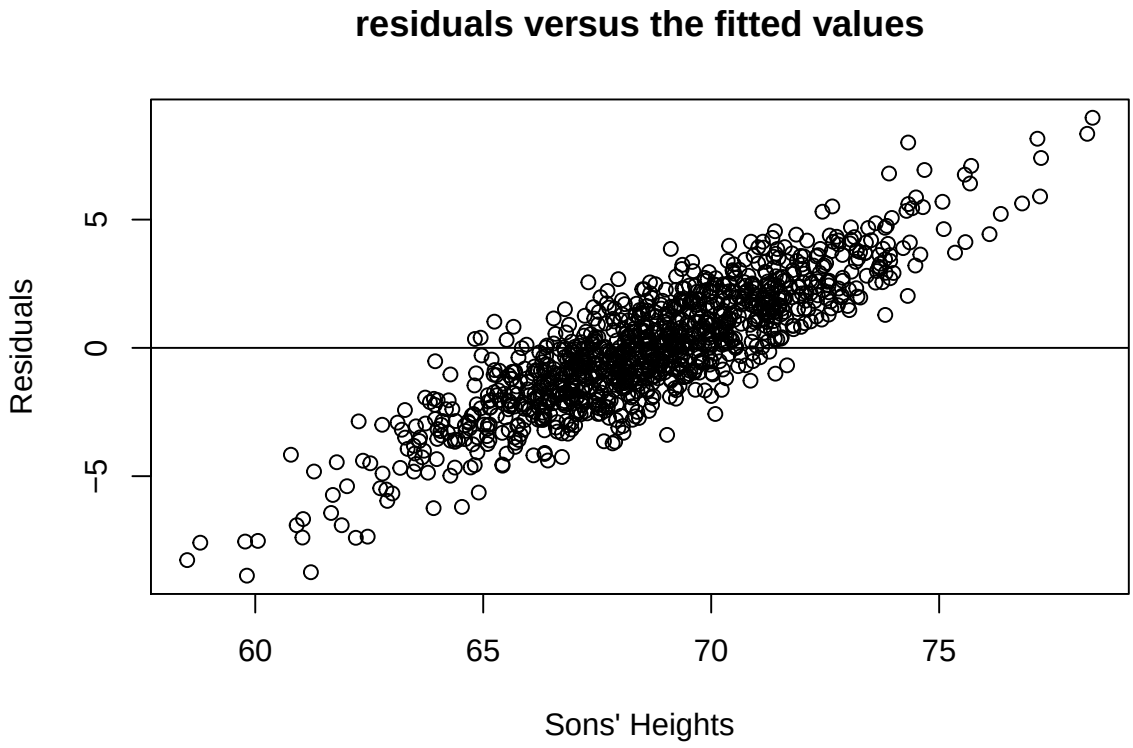
```
#c
confint(fit.1)
```

```
##                  2.5 %     97.5 %
## (Intercept) 30.2912126 37.4819961
## fheight      0.4610188  0.5671673
```

```
#d
plot(sheight ~ fheight, data = heightData)
abline(lm(sheight ~ fheight, data = heightData))
```




```
#e
fit.res = resid(fit.1)
plot(heightData$sheight, fit.res,
     ylab="Residuals", xlab="Sons' Heights",
     main="residuals versus the fitted values")
abline(0, 0)
```




```
#f
array.1 <- data.frame(fheight = c(50, 55, 70, 75, 90))
predict(fit.1, array.1, interval="predict")
```

```
##        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
```

(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. sheight = 0.51409 x fheight + 33.88660

(c) Find the 95% confidence intervals for the estimates. You may find the `confint()` command useful. slope falls into [0.4610188, 0.5671673], intercept falls into [30.2912126, 37.4819961]

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

(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?
Residuals range from -10 to 10, residual is smallest around mean sheight, larger with higher sheights and smaller with lower shights. My concern is that linear model might not best fit the data as it only seems to fit best around the mean, thus the real relationship might be non-linear.

(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. fit
50: 59.59126 55: 62.16172 70: 69.87312 75: 72.44358 90: 80.15498

3. **Extra Credit:**

   (a) What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model? The explanatory variables are assumed to have little or no multicollinearity. (Multicollinearity occurs when the independent variables are not independent from each other.)

   (b) Why can an R^2 close to one not be used as evidence that the simple linear regression model is appropriate? R squared == 1 indicates that the model explains all the variability of the response data around its mean. But it cannot determine whether the coefficient estimates and predictions are biased of if the model fits the data, so the residual plots should also be looked at to determine appropriateness.

   (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? The intercept (often labeled the constant) is the expected mean value of Y when all X=0. Height 0 does not have meaning in real life, thus the intercept does not imply anything about people with zero height. Simply linear regression is not meaningless as the intercept is just a constant that's being added to the model on each corresponding y value given x to better fit the data.

   (d) Suppose you had data on pairs (X, Y) which gave the scatterplot been below. How would you approach the analysis? There seems to be two trends in this scatterplot, I would do a clustering analysis trying to split the data into two groups and then do linear regression models for both groups.

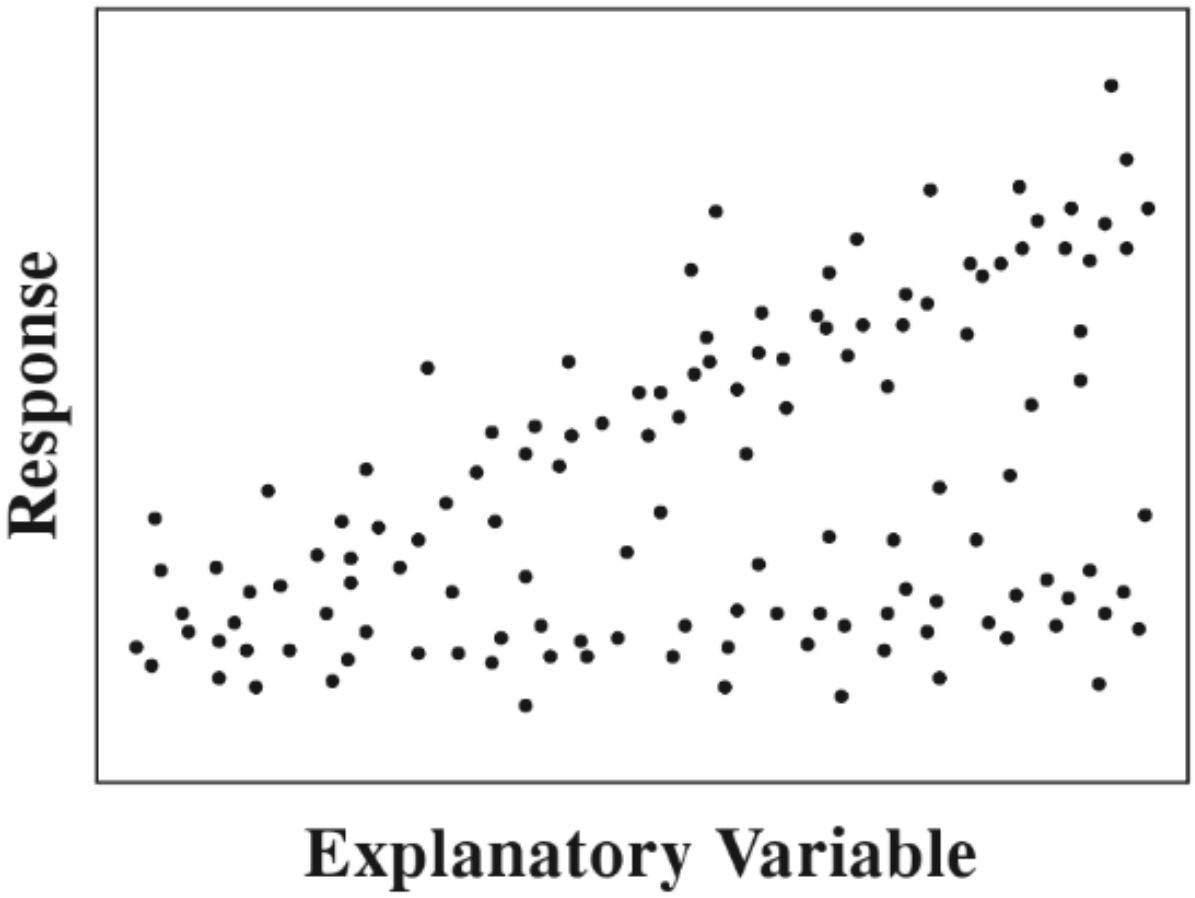


Figure 1: Scatterplot for Extra Credit (d).