IMT 573: Problem Set 7

Regression

YOUR NAME HERE

Due: August 01, 2021

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Download the `07_ps_regression.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `07_ps_regression.Rmd`.
2. Replace the "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. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```r
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps7_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

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

```r
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
library(knitr) # this will keep code on the page!
library(ggplot2)
library(tidyverse)
library(caret)
library(leaps)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Problem 1: Housing Values in Suburbs of Boston

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. This data is much used in data science and statistics to demonstrate regression problems; and while it has a lot of advantages it will comes with concerns. Load this data and use it to answer the following questions.

(a) Briefly describe where these data come from and why they were collected. Be sure to mention any concerns you have about these data. The source of the data is: Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102. Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

They were collected to create a regression. My concern for this data is that it is very outdated and is likely unusable outside of data wrangling/visualization practice.

```
boston <- Boston
boston <- subset(boston, select = c("crim", "indus", "rm", "age", "dis", "rad","ptratio", "medv"))
```

(b) Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.

(d) Consider this data in context, what is the response variable of interest? The median value indicates median home value in thousands. We want to see if crime, industrial area, age of buildings, location, and education play a factor in median home values.

```
m1 <- lm(medv ~ crim, data = boston)
summary(m1)
```

(e) For each predictor, fit a simple linear regression model to predict the response. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

```
##
## Call:
## lm(formula = medv ~ crim, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## crim        -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
m2 <- lm(medv ~ indus, data = boston)
summary(m2)
```

```
##
## Call:
## lm(formula = medv ~ indus, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.017  -4.917  -1.457   3.180  32.943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.75490    0.68345   43.54   <2e-16 ***
## indus       -0.64849    0.05226  -12.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.234,  Adjusted R-squared:  0.2325
## F-statistic:   154 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
m3 <- lm(medv ~ rm, data = boston)
summary(m3)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
m4 <- lm(medv ~ age, data = boston)
summary(m4)
```

```
##
## Call:
## lm(formula = medv ~ age, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## age         -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
m5 <- lm(medv ~ dis, data = boston)
summary(m5)
```

```
##
## Call:
## lm(formula = medv ~ dis, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.016  -5.556  -1.865   2.288  30.377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3901     0.8174  22.499  < 2e-16 ***
## dis           1.0916     0.1884   5.795 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 504 degrees of freedom
## Multiple R-squared:  0.06246,    Adjusted R-squared:  0.0606
## F-statistic: 33.58 on 1 and 504 DF,  p-value: 1.207e-08
```

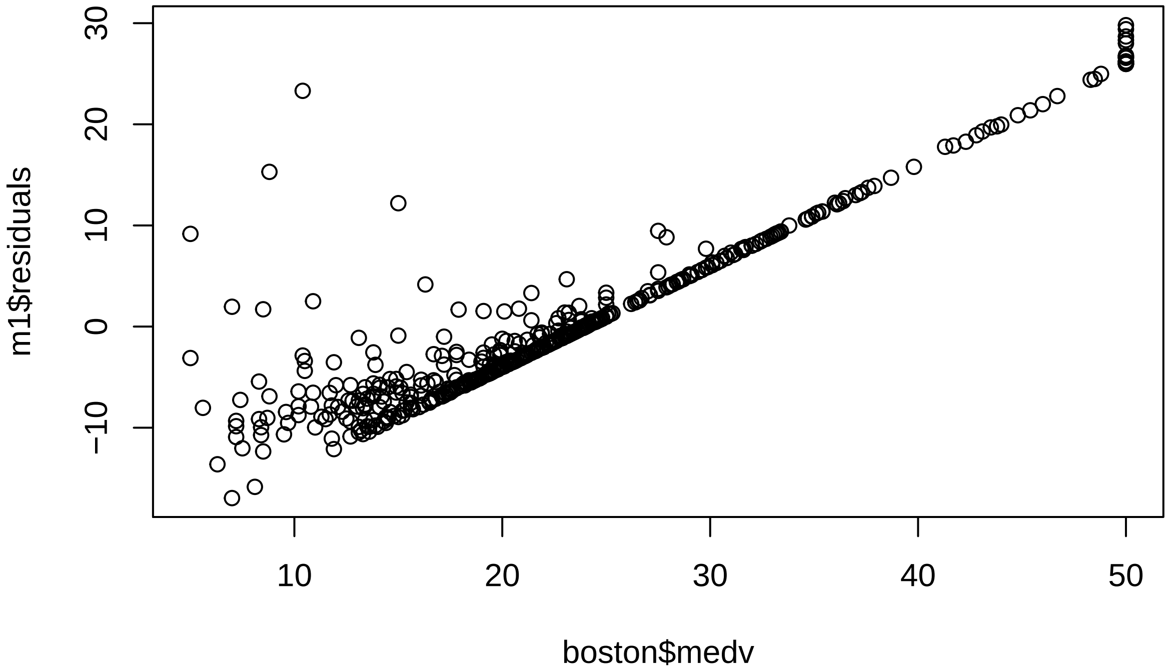
```r
m6 <- lm(medv ~ rad, data = boston)
summary(m6)
```

```
##
## Call:
## lm(formula = medv ~ rad, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.770  -5.199  -1.967   3.321  33.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.38213    0.56176  46.964   <2e-16 ***
## rad         -0.40310    0.04349  -9.269   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.509 on 504 degrees of freedom
## Multiple R-squared:  0.1456, Adjusted R-squared:  0.1439
## F-statistic: 85.91 on 1 and 504 DF,  p-value: < 2.2e-16
```

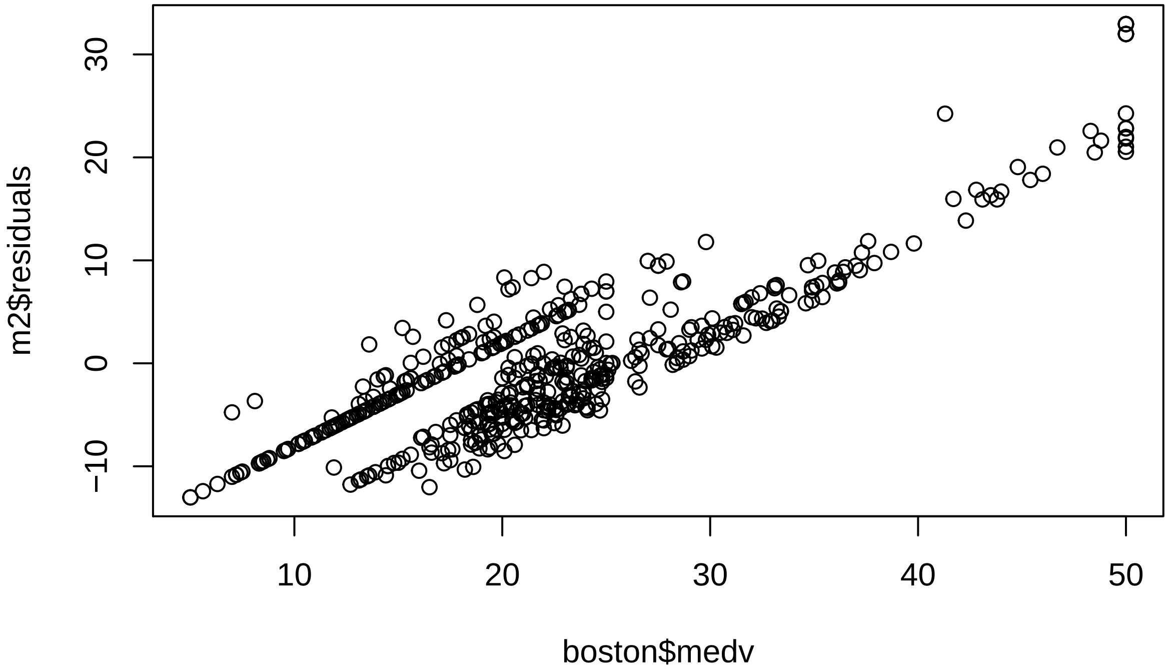
```
m7 <- lm(medv ~ ptratio, data = boston)
summary(m7)
```

```
##
## Call:
## lm(formula = medv ~ ptratio, data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.345      3.029   20.58   <2e-16 ***
## ptratio       -2.157      0.163  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.931 on 504 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2564
## F-statistic: 175.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

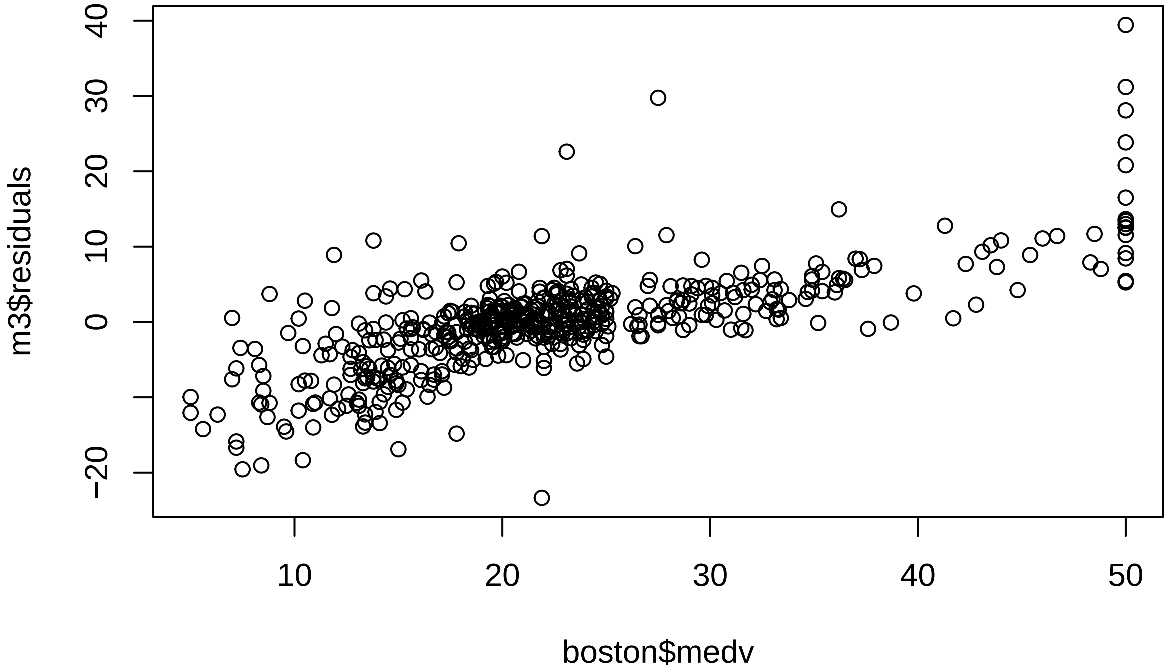
```
plot(m1$residuals ~ boston$medv)
```



```
plot(m2$residuals ~ boston$medv)
```



```
plot(m3$residuals ~ boston$medv)
```



```r
model <- lm(medv ~ crim + indus + rm + age + dis + rad + ptratio, data = boston)
summary(model)
```

(f) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```
##
## Call:
## lm(formula = medv ~ crim + indus + rm + age + dis + rad + ptratio,
##     data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.456  -2.658  -0.634   1.944  39.172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.22473    4.37559   2.108 0.035510 *
## crim        -0.17807    0.03773  -4.720 3.07e-06 ***
## indus       -0.23459    0.06099  -3.846 0.000135 ***
## rm           6.74193    0.40566  16.620  < 2e-16 ***
## age         -0.07096    0.01386  -5.121 4.34e-07 ***
## dis         -1.09386    0.20288  -5.392 1.08e-07 ***
## rad          0.01063    0.04432   0.240 0.810495
## ptratio     -0.91521    0.13739  -6.662 7.19e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.616 on 498 degrees of freedom
## Multiple R-squared:  0.6323, Adjusted R-squared:  0.6271
## F-statistic: 122.3 on 7 and 498 DF,  p-value: < 2.2e-16
```

We can reject a null hypothesis for industry and rad.

(g) How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response. The data correlates well with the plots I previously created.

(h) Is there evidence of a non-linear association between any of the predictors and the response? To answer this question, for each predictor X fit a model of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

There is a strong non-linear association with the predictors: crime, age of buildings, and teacher ratio

```
step.model <- stepAIC(model, direction = "both",
                      trace = FALSE)
summary(step.model)
```

(i) Consider performing a stepwise model selection procedure to determine the bets fit model. Discuss your results. How is this model different from the model in (4)?

```
##
## Call:
## lm(formula = medv ~ crim + indus + rm + age + dis + ptratio,
##     data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.392  -2.622  -0.640   1.963  39.282
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.98666    4.25755   2.111   0.0353 *
## crim        -0.17355    0.03266  -5.315 1.61e-07 ***
## indus       -0.23028    0.05823  -3.955 8.77e-05 ***
## rm           6.75644    0.40075  16.860  < 2e-16 ***
## age         -0.07089    0.01384  -5.122 4.32e-07 ***
## dis         -1.09652    0.20239  -5.418 9.39e-08 ***
## ptratio     -0.90493    0.13042  -6.939 1.23e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.611 on 499 degrees of freedom
## Multiple R-squared:  0.6322, Adjusted R-squared:  0.6278
## F-statistic:   143 on 6 and 499 DF,  p-value: < 2.2e-16
```

This model has lower prediction errors, therefore predicting the best fit model.

(j) Evaluate the statistical assumptions in your regression analysis from (7) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

Problem 2: A Critical Perspective to the Boston Housing Data

(a) When were these data collected? Did you note this in you descriptive above? Did the date surprise you? Yes, I noted above that the data is from 1978. This date did not surpirse me as I can't even imagine buying a tree house for $25,000 today let alone a house in Boston.

(b) Amidst data features like number of rooms and access to highways are features like crime rate, and percentage Black per town. Whether intentional or not, someone looking at this data might infer a link between crime and race just due to the variables present; or even worse might use the data to support harsh policing policies based on race. Suppose for a moment we have a modern version of this dataset; the "Seattle Housing Data." Discuss, in a few paragraphs, how this hypothetical dataset could be used (1) in a harmful way, and (2) in a beneficial way for society. This data could be used in a harmful way by increasing police presence in areas of low income and diversity, leading to oppression. Beneficial ways of using these types of data would be indentifying equity areas that are not growing as fast as other neighborhoods, and diverting economic resources towards those areas.