Problem Set 3

Divya Krishnan

Monday, Nov 30, 2015

Univariate and Multivariate Regression

```
# Explicitly added the package of coefplot for plotting regression coefficients
library("coefplot")
# Adding libraries for Birth data
library("Sleuth3")
# Adding library for height data
library("UsingR")
# Adding library for boston data
library("MASS")
```

Simple Linear Regression 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:

```
# Extracting birth data
birthData <- ex0724
```

(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 = 0.5987 + (-4.289e-05) * Year

Netherlands = 0.6724 + (-8.084e-05) * Year
Canada = 0.7338 + (-1.112e-04) * Year
USA = 0.6201 + (-5.429e-05) * Year

```
# Fitting linear model for each country
fitDenmark<-lm(Denmark ~ Year,data=birthData)
fitNetherlands<-lm(Netherlands ~ Year,data=birthData)
fitCanada<-lm(Canada ~ Year,data=birthData)
fitUSA<-lm(USA ~ Year,data=birthData)
# Linear model
fitDenmark
```

```
##
## Call:
## lm(formula = Denmark ~ Year, data = birthData)
```

```
##
## Coefficients:
## (Intercept)         Year
##   5.987e-01   -4.289e-05
```

```
fitNetherlands
```

```
##
## Call:
## lm(formula = Netherlands ~ Year, data = birthData)
##
## Coefficients:
## (Intercept)         Year
##   6.724e-01   -8.084e-05
```

```
fitCanada
```

```
##
## Call:
## lm(formula = Canada ~ Year, data = birthData)
##
## Coefficients:
## (Intercept)         Year
##   0.7337857   -0.0001112
```

```
fitUSA
```

```
##
## Call:
## lm(formula = USA ~ Year, data = birthData)
##
## Coefficients:
## (Intercept)         Year
##   6.201e-01   -5.429e-05
```

(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?

The t-statistic for the countries are as follows (obtained from the summary) -

Denmark -2.073
Netherlands -5.71
Canada -4.017
USA -5.779

The t statistic gives results of the t-test with the null hypothesis that the beta-j coefficient is 0. The t-statistic for beta1 coefficient has the null hypothesis that the true linear model has slope zero. All the t-statistic values for beta1(as written above) indicate that they are in the region of rejection (greater than 1.96) for the two-sided t test, and hence we can reject the null hypothesis that the slopes of the regression lines are zero. This establishes a statistically significant association between year and the proportion of male births. Hence, looking at the negative regression coefficient(beta1) estimates, we can conclude that the proportion of births that are male is truly declining over this period.

```
# Summary of linear models, which provides the t-statistic as well
summary(fitDenmark)
```

```
##
## 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
```

```
summary(fitNetherlands)
```

```
##
## 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
```

```
summary(fitCanada)
```

```
##
## 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
```

```
summary(fitUSA)
```

```
##
## Call:
## lm(formula = USA ~ Year, data = birthData)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.343e-04 -1.800e-04 -1.714e-05  2.571e-04  3.743e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.201e-01  1.860e-02  33.340  < 2e-16 ***
## Year        -5.429e-05  9.393e-06  -5.779 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002607 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6183
## F-statistic:  33.4 on 1 and 19 DF,  p-value: 1.439e-05
```

Analysis and Prediction using Regression 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:

```
# Extracting height data
heightData<-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.

The statistical summaries of the variable indicate that the mean height for fathers is 67.69 inches and sons is 68.68. Most of the values in the dataset are in the range of (58,79). The histogram of the father's height and son's height suggests that each distribution is similar to a normal distribution. Scatterplot of the father's and son's height indicates a linear relationship between the two variables. The pearson correlation of father's and sons' height is 0.5 indicating a positive relationship.

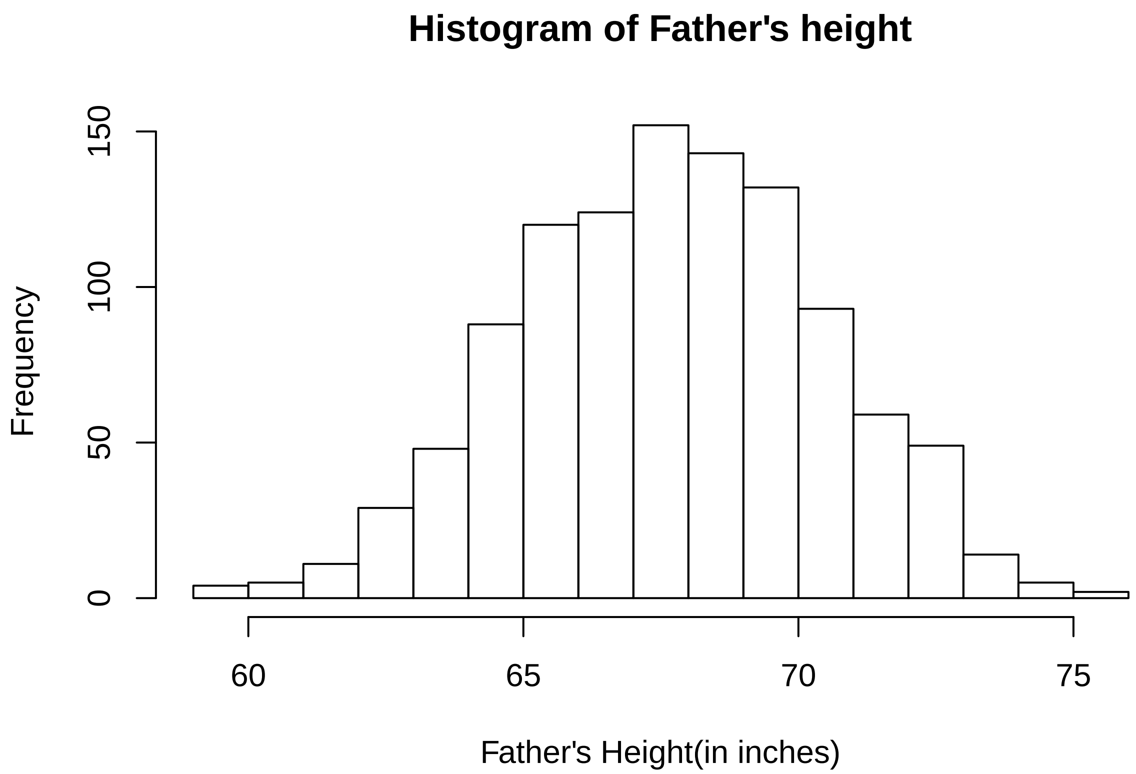
```
# Exploring the height data
str(heightData)
```

```
## '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 ...
```

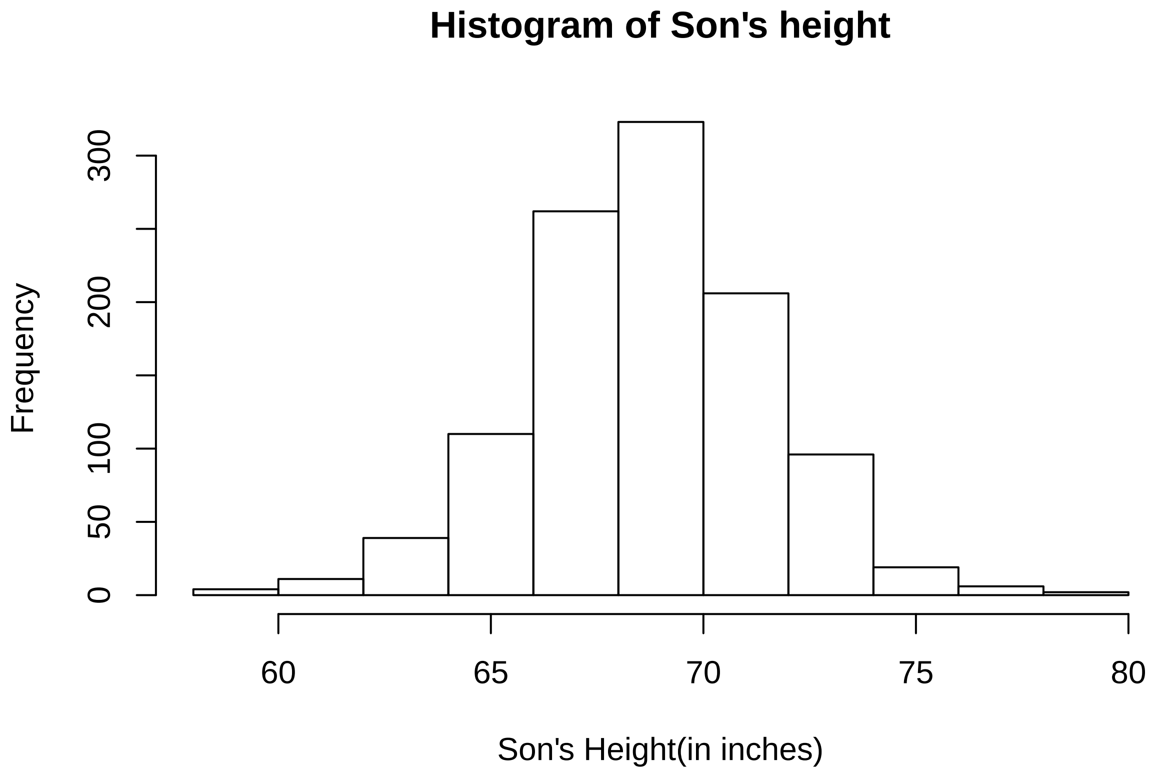
```
summary(heightData)
```

```
##     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
```

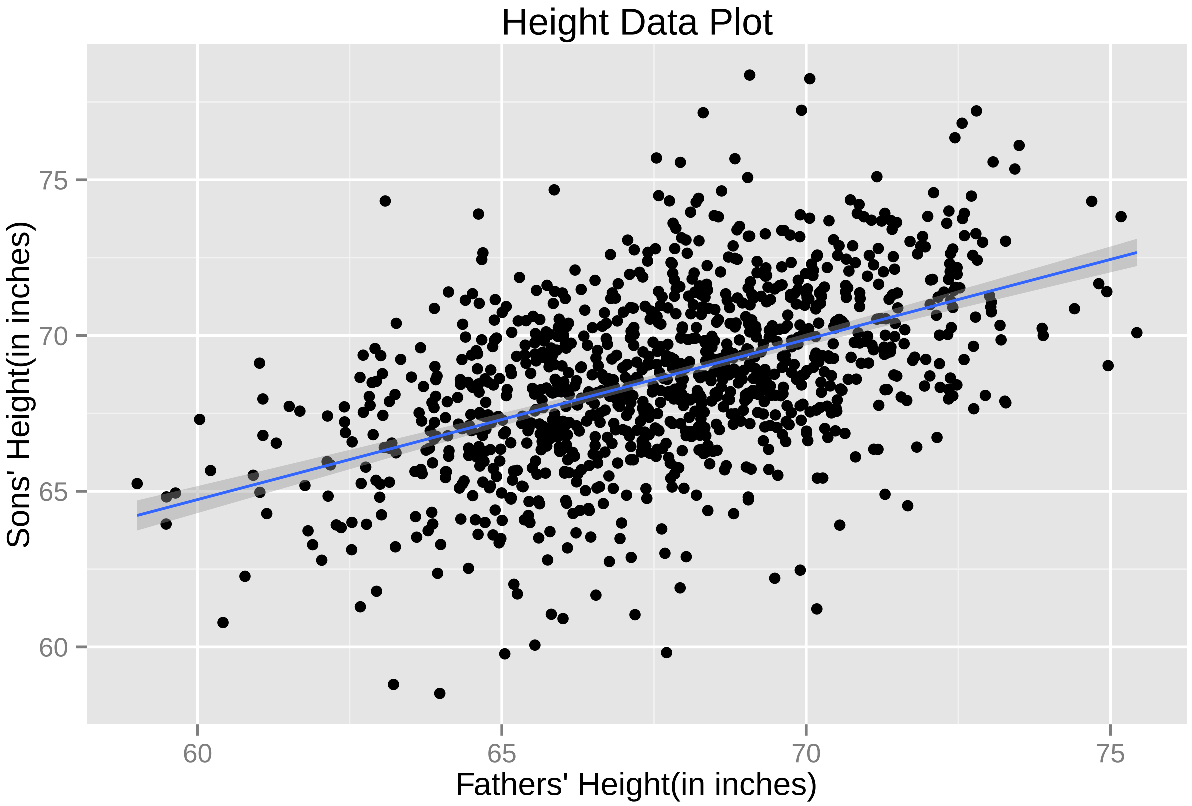
```
# Creating histogram for each of the variables
hist(heightData$fheight,xlab="Father's Height(in inches)",main="Histogram of Father's height")
```



```
hist(heightData$sheight,xlab="Son's Height(in inches)",main="Histogram of Son's height")
```



```r
# Plotting father's height with sons's height
g<- ggplot(heightData, aes(x=fheight, y=sheight))
g<- g + geom_point()
g<- g + geom_smooth(method="lm")
g<- g + labs(x="Fathers' Height(in inches)", y="Sons' Height(in inches)",title="Height Data Plot")
g
```



```
# Finding pearson correlation of between father and son heights
cor(heightData$fheight,heightData$sheight)
```

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

(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, y-hat-sheight = Beta-hat-0 + Beta-hat-i * fheight filling in estimated coefficient values and interpret the coefficient estimates.

sheight(estimate) = 33.8866 + 0.5141 * fheight

The above model can be intepreted as - For every 1 inch of increase in father's height, the son's height increases by about 0.5 inches.

```
# Fitting linear model for height data
fitHeight<-lm(sheight ~ fheight,data=heightData)
# Summary of linear model
summary(fitHeight)
```

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

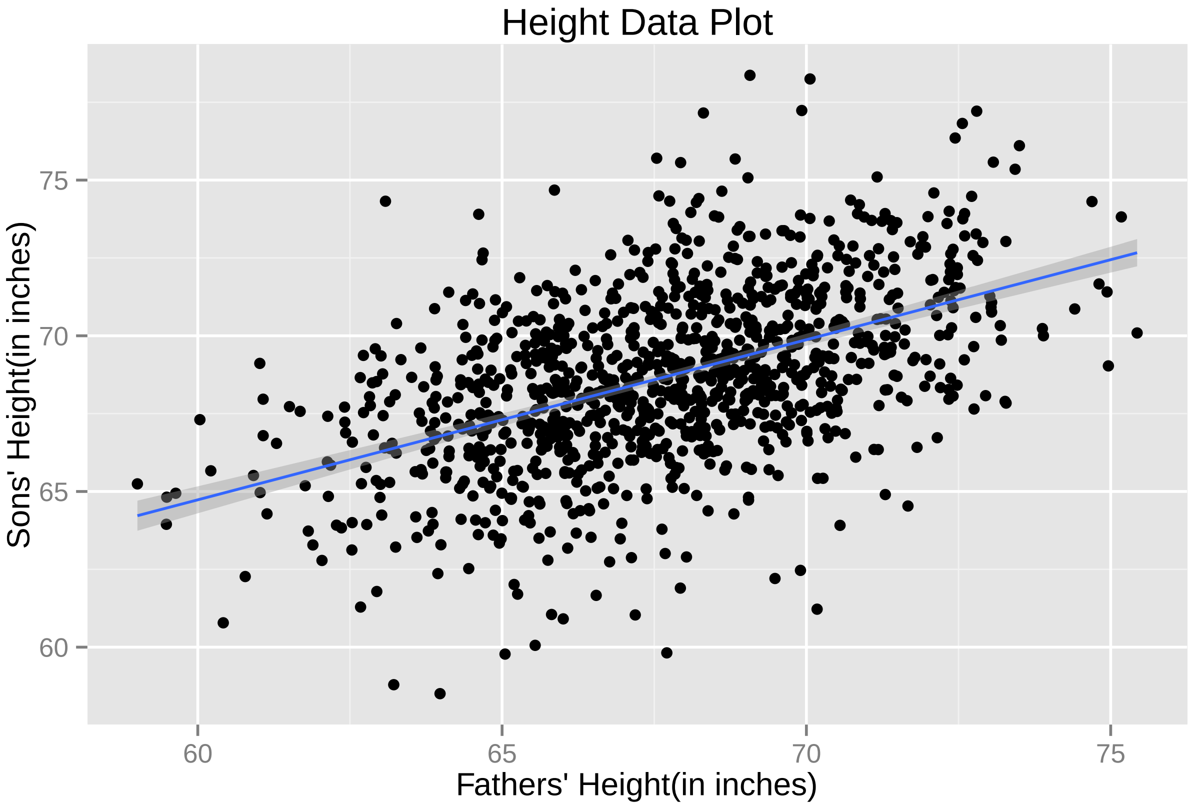
The 95% confidence interval for the estimates of intercept is - (30.2912,37.4820)
The 95% confidence interval for the estimates of the slope is - (0.4610,0.5672)

```
# Confidence interval for the estimates
confint(fitHeight)
```

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

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

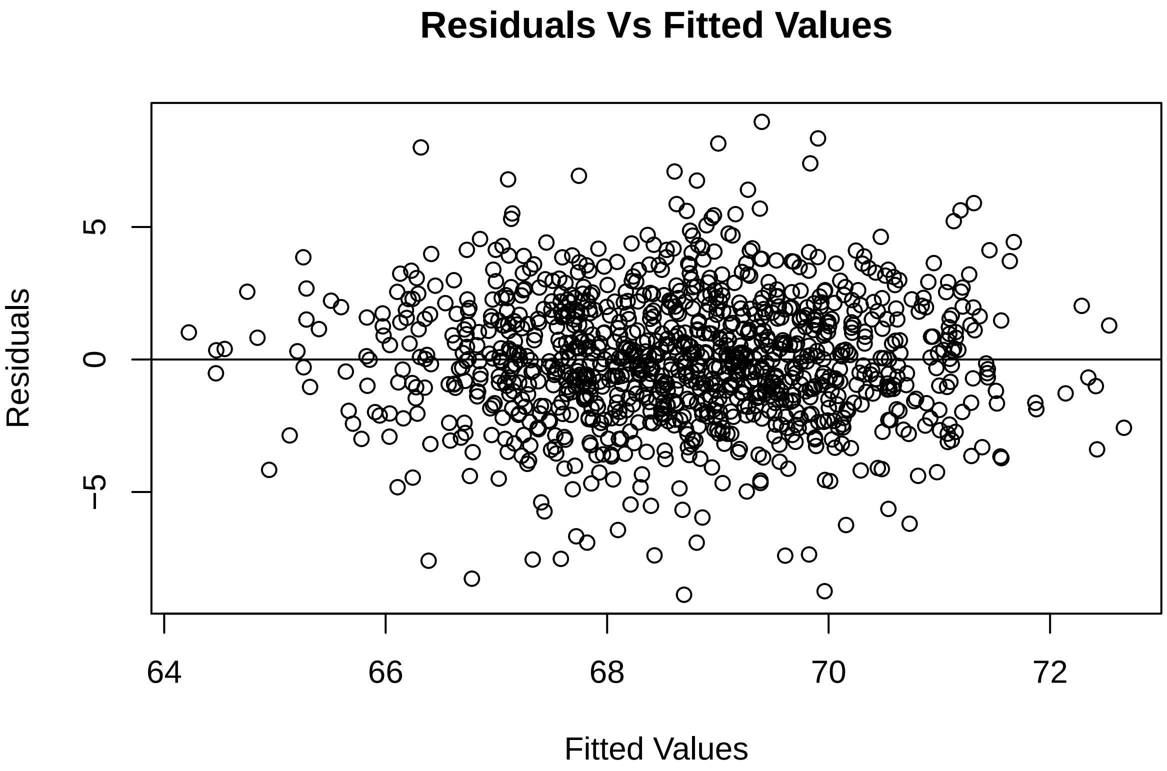
```
# Plotting father's height with sons's height
g<- ggplot(heightData, aes(x=fheight, y=sheight))
g<- g + geom_point()
g<- g + geom_smooth(method="lm")
g<- g + labs(x="Fathers' Height(in inches)", y="Sons' Height(in inches)"
             ,title="Height Data Plot")
g
```



(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?

There is no pattern observed in the visualization between residuals and the fitted values.The residual plot helps in identifying non-linear relationships between the predictors and the response. Since, the residual plot does not show in strong pattern, we can safely use a linear model to understand the relationship between father's height and son's height.

```
# Computing the fitted values using the linear model equation
fittedValues<-33.8866 + (0.5141 * heightData$fheight)
# Extracting the residuals
heightResiduals<-as.vector(fitHeight$residuals)
# Plotting residuals Vs fitted values
plot(heightResiduals ~ fittedValues,xlab="Fitted Values",
     ylab="Residuals",main="Residuals Vs Fitted Values")
# Plotting the linear regression line between residuals and fitted values
abline(lm(heightResiduals ~ fittedValues))
```



```r
# summary of the linear regression of residuals and fitted values
summary(lm(heightResiduals ~ fittedValues))
```

```
##
## Call:
## lm(formula = heightResiduals ~ fittedValues)
##
## 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)  -6.756e-14  3.615e+00       0        1
## fittedValues  9.781e-16  5.261e-02       0        1
##
## Residual standard error: 2.437 on 1076 degrees of freedom
## Multiple R-squared:  3.059e-31,  Adjusted R-squared:  -0.0009294
## F-statistic: 3.291e-28 on 1 and 1076 DF,  p-value: 1
```

(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.

Prediction for the new data -

Father's height(in inches) - Son's Height(in inches)

50 - 59.5913

55 - 62.1617

70 - 69.8731

75 - 72.4436

90 - 80.1550

```r
# Using predict function to predict the son's height for new father's height data
round(predict(fitHeight, newdata = data.frame(fheight = c(50,55,70,75,90))),4)
```

```
##       1       2       3       4       5 
## 59.5913 62.1617 69.8731 72.4436 80.1550
```

Analysis and Prediction using Multiple Regression 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. Load this data and use it to answer the following questions.

(a) Describe the data and variables that are part of the Boston dataset.

The data contains median house value for 506 neighborhoods in Boston, MA. It has the following attributes -

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's

References - https://archive.ics.uci.edu/ml/datasets/Housing

```r
# Extracting boston data set
boston<-get("Boston")
str(boston)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

(b) Consider this data what is the response variable of interest?

The response variable of interest could be the median value of houses, names medv. It could be very useful for home buyers and sellers to predict the value of a house, based on certain other known parameters.

(c) 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.

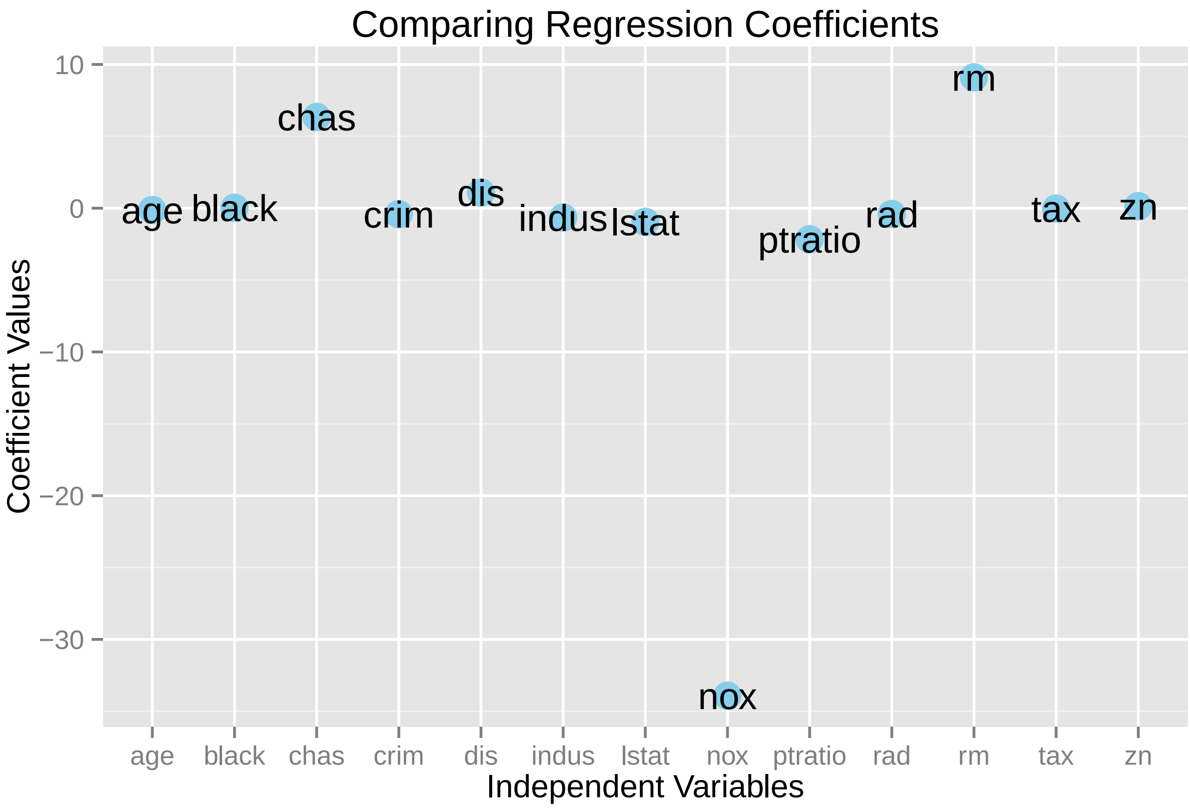
All the models show a statistically significant association between the independent variables and the response or dependent variable.
The independent variables that have the strongest association with the response variable is nox, rm,chas,ptration,dis and lstat, in decreasing order. Nox and ptratio have negative regression coefficient estimates whereas rm,chas and dis have positive regression coefficient estimates.

```r
# Storing the coefficient matrix of each univariate linear regression
fit<-list()
# Loop for computing linear regression for each independent variable
for(i in 1:(ncol(boston)-1))
{
  # Extracting the coefficients of the linear regression
  fit[[i]]<-summary(lm(boston$medv ~ boston[[i]]))$coefficients
}
# Renaming the elements of the list with the variable names
names(fit)<-colnames(boston)[1:13]
# Rounding the coefficients
lapply(fit,round,digits=2)
```

```
## $crim
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24.03       0.41   58.74        0
## boston[[i]]    -0.42       0.04   -9.46        0
```

```
##
## $zn
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20.92       0.42   49.25        0
## boston[[i]]     0.14       0.02    8.68        0
##
## $indus
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    29.75       0.68   43.54        0
## boston[[i]]    -0.65       0.05  -12.41        0
##
## $chas
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.09       0.42    52.9        0
## boston[[i]]     6.35       1.59     4.0        0
##
## $nox
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.35       1.81   22.83        0
## boston[[i]]   -33.92       3.20  -10.61        0
##
## $rm
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -34.67       2.65  -13.08        0
## boston[[i]]     9.10       0.42   21.72        0
##
## $age
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.98       1.00   31.01        0
## boston[[i]]    -0.12       0.01   -9.14        0
##
## $dis
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18.39       0.82   22.50        0
## boston[[i]]     1.09       0.19    5.79        0
##
## $rad
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26.38       0.56   46.96        0
## boston[[i]]    -0.40       0.04   -9.27        0
##
## $tax
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    32.97       0.95   34.77        0
## boston[[i]]    -0.03       0.00  -11.91        0
##
## $ptratio
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    62.34       3.03   20.58        0
## boston[[i]]    -2.16       0.16  -13.23        0
##
## $black
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.55       1.56    6.77        0
## boston[[i]]     0.03       0.00    7.94        0
##
## $lstat
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.55       0.56   61.42        0
## boston[[i]]    -0.95       0.04  -24.53        0
```

```
# Creating a list of coefficient estimates
coefRes<-data.frame(varName=as.character(),value=as.numeric(),stringsAsFactors = FALSE)
for(i in 1:(ncol(boston)-1))
{
  # Storing independent variable names
  var<-colnames(boston[i])
  coefRes[i,1]<-var
  # Extracting the beta1 estimates from the model
  value<-round(fit[[i]][2,1],4)
  coefRes[i,2]<-value
}
# Plotting the value of regression coefficients for each independent variable
g<-ggplot(coefRes,aes(x=varName,y=value))
g<-g + geom_point(color="skyblue",size=5)
g<-g + geom_text(aes(label=varName), size=5)
g<-g + labs(title="Comparing Regression Coefficients",
            x="Independent Variables", y="Coefficient Values")
g
```



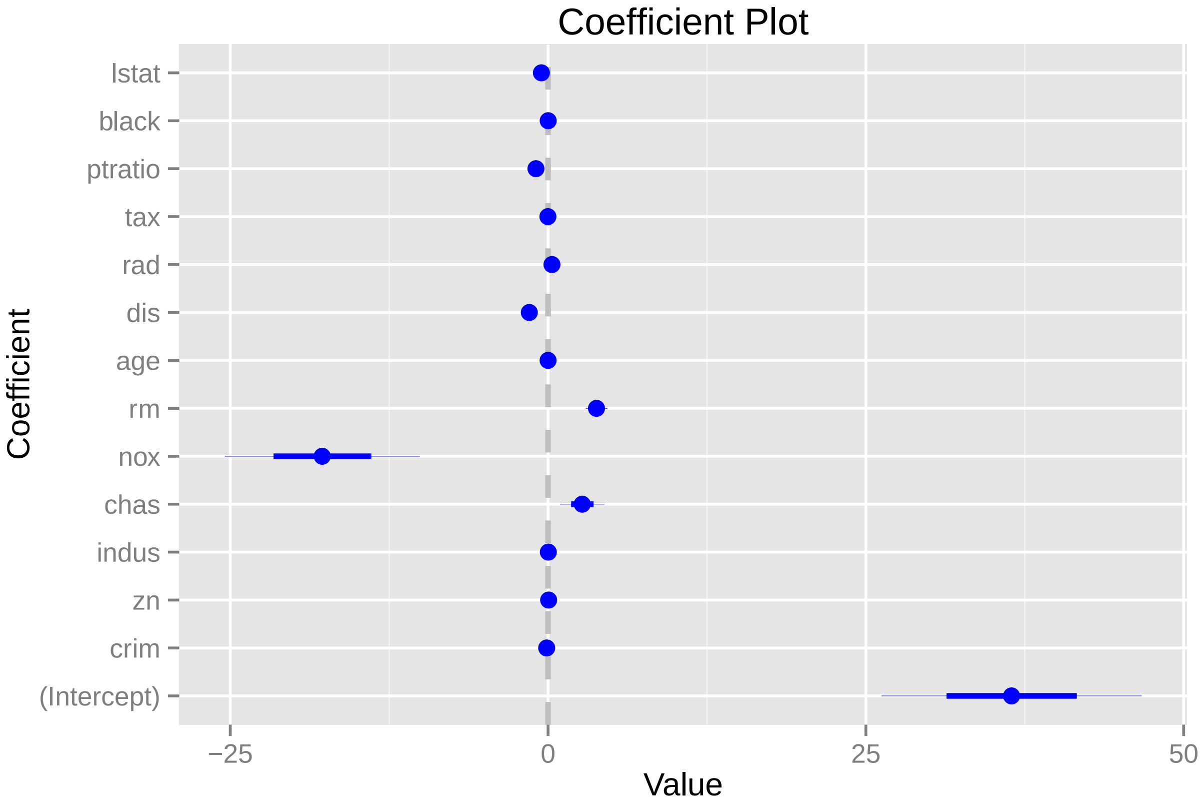
(d) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : Beta-j = 0?

The summary shows that only variables indus and age are not statistically significant. Hence, for these two variables we fail to reject the null hypothesis that beta-j is equal to zero. This means that the association between these variables and the medv is not statistically significant.
The summary and the coefplot of the multiple regression shows that nox, rm, chase, dis and ptratio have the strongest associations with the response variable. nox, dis and ptratio have negative beta1 estimates whereas chas and rm have positive beta1 estimates.

```r
# Variable for storing multiple regression model
model<-"medv ~"
# Loop to concatenate all the variables of the boston data set
for(i in 1:(ncol(boston)-1))
{
  # concatenating each independent variable into the model
  if(i<13)
  {
    # Extracting the name of the independent variables
    var<-paste0(colnames(boston)[i]," +")
    model<-paste(model,var,sep=" ")
  }else
  {
    # Excluding the + sign at the end of the last variable
    var<-paste0(colnames(boston)[i])
    model<-paste(model,var,sep=" ")
  }
}

# Performing multiple regression model
fitmlm<-lm(model,data=boston)
# Summary of multiple regression
summary(fitmlm)
```

```
##
## Call:
## lm(formula = model, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
# Plot for understanding the various regression coefficients
coefplot(fitmlm)
```



(e) How do your results from (c) compare to your results from (d)? Create a plot displaying the univariate regression coefficients from (c) on the x-axis and the multiple regression coefficients from part (d) on the y-axis. Use this visualization to support your response.

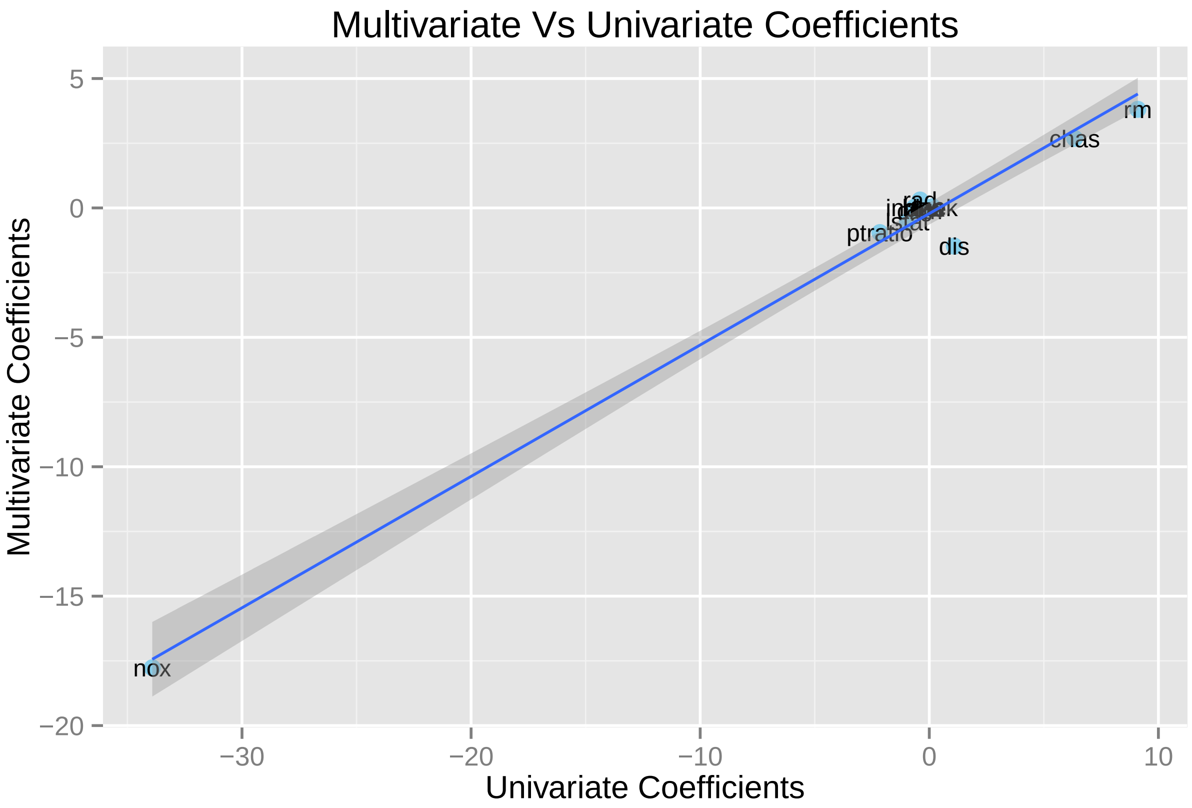
The plot between multivariate and univariate coefficients shows that although most variables estimates for both regression is close, variables like nox,dis,ptratio, chas,rm and lstat have differences. The linear regression of the multivariate and univariate coefficients shows that the slope is not equal to 1, which means there is definitely a difference between the coefficients. The linear model says that for every 1 unit increase in

the univariate coefficients, the multivariate coefficients increase by 0.51 units.
The correlation matrix shows that there is some colinearity in the data. Comparing the results of the univariate and multivariate coefficient, we can observe that the greatest absolute difference in the coefficients is for nox, rm, chas, dis and ptratio in decreasing order. These are also the independent variables that have strong relationships with the response variable. Multiple regression adjusts for colinearity in the data and hence the difference between the univariate and multivariate coefficients. The plot shows the absolute difference between the beta1 estimates of the two egression coefficients.

```
# Creating a correlation matrix to understand colinearity in the data
round(cor(as.matrix(boston[,1:13])),1)
```

```
##          crim   zn indus chas  nox   rm  age  dis  rad  tax ptratio black
## crim      1.0 -0.2   0.4 -0.1  0.4 -0.2  0.4 -0.4  0.6  0.6     0.3  -0.4
## zn       -0.2  1.0  -0.5  0.0 -0.5  0.3 -0.6  0.7 -0.3 -0.3    -0.4   0.2
## indus     0.4 -0.5   1.0  0.1  0.8 -0.4  0.6 -0.7  0.6  0.7     0.4  -0.4
## chas     -0.1  0.0   0.1  1.0  0.1  0.1  0.1 -0.1  0.0  0.0    -0.1   0.0
## nox       0.4 -0.5   0.8  0.1  1.0 -0.3  0.7 -0.8  0.6  0.7     0.2  -0.4
## rm       -0.2  0.3  -0.4  0.1 -0.3  1.0 -0.2  0.2 -0.2 -0.3    -0.4   0.1
## age       0.4 -0.6   0.6  0.1  0.7 -0.2  1.0 -0.7  0.5  0.5     0.3  -0.3
## dis      -0.4  0.7  -0.7 -0.1 -0.8  0.2 -0.7  1.0 -0.5 -0.5    -0.2   0.3
## rad       0.6 -0.3   0.6  0.0  0.6 -0.2  0.5 -0.5  1.0  0.9     0.5  -0.4
## tax       0.6 -0.3   0.7  0.0  0.7 -0.3  0.5 -0.5  0.9  1.0     0.5  -0.4
## ptratio   0.3 -0.4   0.4 -0.1  0.2 -0.4  0.3 -0.2  0.5  0.5     1.0  -0.2
## black    -0.4  0.2  -0.4  0.0 -0.4  0.1 -0.3  0.3 -0.4 -0.4    -0.2   1.0
## lstat     0.5 -0.4   0.6 -0.1  0.6 -0.6  0.6 -0.5  0.5  0.5     0.4  -0.4
##          lstat
## crim       0.5
## zn        -0.4
## indus      0.6
## chas      -0.1
## nox        0.6
## rm        -0.6
## age        0.6
## dis       -0.5
## rad        0.5
## tax        0.5
## ptratio    0.4
## black     -0.4
## lstat      1.0
```

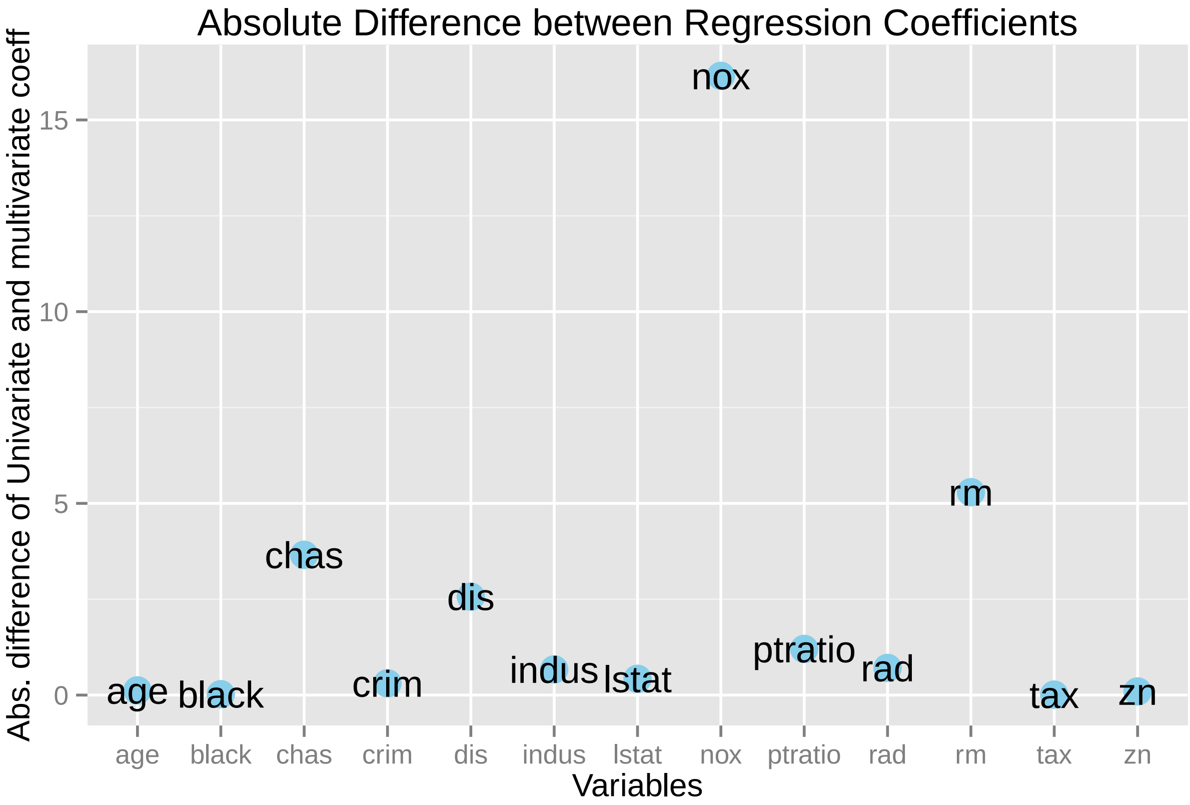
```
# Vectors for storing univariate, multivariate and
# the difference in the coefficients
univariateCoef<-vector()
multivariateCoef<-vector()
diffCoef<-vector()
# Extracting multivariate regression coefficients
tempY<-as.data.frame(fitmlm$coefficients)
colnames(tempY)<-"coefValue"
for(i in 1:(ncol(boston)-1))
{
  # Extracting the univariate regression coefficients
  tempX<-as.data.frame(fit[i])
  # Extracting the beta1 estimates for univariate regression
  univariateCoef[i]<-tempX[2,1]
  # Extracting the beta1 estimates for multivariate regression
  multivariateCoef[i]<-tempY$coefValue[i+1]
}
# Extracting the independent variables
var=colnames(boston)[1:13]
# Finding the difference between univariate and multivariate beta1 coefficients
diffCoef<-abs(multivariateCoef-univariateCoef)
# Creating a data frame of the beta1 estimates
lmComparison<-data.frame(var,univariateCoef,multivariateCoef,diffCoef)

# Plotting univariate and multivariate regression coefficients
g<-ggplot(data=lmComparison,aes(x=univariateCoef,y=multivariateCoef))
g<-g + geom_point(color="skyblue",size=3)
g<-g + geom_text(aes(label=var), size=3)
g<-g + geom_smooth(method="lm")
g<-g + labs(title="Multivariate Vs Univariate Coefficients",
            x="Univariate Coefficients", y="Multivariate Coefficients")
g
```



```
# Linear regression of multivariate and univariate coefficients
fitCoeff<-lm(multivariateCoef ~ univariateCoef)
summary(fitCoeff)
```

```
##
## Call:
## lm(formula = multivariateCoef ~ univariateCoef)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8122 -0.3178  0.2099  0.3204  0.7284
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.21763    0.19292  -1.128    0.283
## univariateCoef  0.50773    0.01944  26.124 2.99e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6855 on 11 degrees of freedom
## Multiple R-squared:  0.9841, Adjusted R-squared:  0.9827
## F-statistic: 682.5 on 1 and 11 DF,  p-value: 2.994e-11
```

```
# Plot the absolute difference of the regression estimate coefficients(beta1)
g<-ggplot(lmComparison,aes(x=var,y=diffCoef))
g<-g + geom_point(color="skyblue",size=5)
g<-g + geom_text(aes(label=var), size=5)
g<-g + labs(title="Absolute Difference between Regression Coefficients",
            x="Variables", y="Abs. difference of Univariate and multivariate coeff")
g
```



(f) 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 = Beta0 + Beta1 X + Beta2 X2 + Beta3 X3 + Epsilon

From the polynomial regression, it can be observed that some of the variables such as nox, rm and dis have a non-linear regression, which is statistically significant(p value<0.05). For variables chas, ptratio, tax, age and black, the non-linear relationships are not statistically significant (p>0.05). For variables crim, zn, indus, nox, rm, dis, rad and lstat have evidence that there is a statistically significant non-linear relationship between the independent variable and the response variable.nox variable seems to only have a statistically significant(p value=0.04) association between the cube of the variable and the response variable.

Except tax, black, ptratio and chas, there is evidence of non-linear relationship between the independent variables and the response variable. chas variable only has two values, so it does not have any scope of polynomial regression. nox is the only variable which has the bigger beta-j estimate for squared value than the beta1 estimate.

```
# Polynomial regression
fitPolynomial<-list()
# Loop for polynomial regression of each independent variable
for(i in 1:(ncol(boston)-1))
{
  # coefficient estimates of the polynomial regression
  fitPolynomial[[i]]<-summary(lm(boston$medv ~ boston[[i]] +
                                   I(boston[[i]]^2) + I(boston[[i]]^3),
                                    data=boston))$coefficients
}
# Renaming the list elements with the variable names
names(fitPolynomial)<-colnames(boston)[1:13]
# Rounding the coefficient estimates
lapply(fitPolynomial,round,digits=2)
```

```
## $crim
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         25.19       0.44   57.85     0.00
## boston[[i]]         -1.14       0.14   -7.87     0.00
## I(boston[[i]]^2)     0.02       0.01    3.49     0.00
## I(boston[[i]]^3)     0.00       0.00   -2.24     0.03
##
## $zn
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         20.45       0.44   46.91        0
## boston[[i]]          0.64       0.11    5.82        0
## I(boston[[i]]^2)    -0.02       0.00   -4.31        0
## I(boston[[i]]^3)     0.00       0.00    3.98        0
##
## $indus
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         37.08       1.66   22.29     0.00
## boston[[i]]         -2.81       0.51   -5.51     0.00
## I(boston[[i]]^2)     0.14       0.04    3.38     0.00
## I(boston[[i]]^3)     0.00       0.00   -2.37     0.02
##
## $chas
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.09       0.42    52.9        0
## boston[[i]]     6.35       1.59     4.0        0
##
## $nox
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -22.49      38.52   -0.58     0.56
## boston[[i]]        315.10     195.10    1.62     0.11
## I(boston[[i]]^2)  -615.83     320.48   -1.92     0.06
## I(boston[[i]]^3)   350.19     170.92    2.05     0.04
##
## $rm
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        241.31      47.33    5.10        0
## boston[[i]]       -109.39      22.97   -4.76        0
## I(boston[[i]]^2)    16.49       3.68    4.49        0
## I(boston[[i]]^3)    -0.74       0.19   -3.83        0
##
## $age
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         28.93       2.99    9.67     0.00
## boston[[i]]         -0.12       0.20   -0.61     0.54
## I(boston[[i]]^2)     0.00       0.00    0.60     0.55
## I(boston[[i]]^3)     0.00       0.00   -1.02     0.31
##
## $dis
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.04       2.91    2.42     0.02
## boston[[i]]          8.59       2.07    4.16     0.00
## I(boston[[i]]^2)    -1.25       0.41   -3.03     0.00
## I(boston[[i]]^3)     0.06       0.02    2.31     0.02
##
## $rad
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         30.25       2.57   11.78        0
## boston[[i]]         -3.80       1.31   -2.91        0
## I(boston[[i]]^2)     0.62       0.19    3.31        0
## I(boston[[i]]^3)    -0.02       0.01   -3.51        0
##
## $tax
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         52.22      13.97    3.74     0.00
## boston[[i]]         -0.16       0.11   -1.44     0.15
## I(boston[[i]]^2)     0.00       0.00    1.05     0.29
## I(boston[[i]]^3)     0.00       0.00   -0.93     0.35
##
## $ptratio
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        312.29     152.49    2.05     0.04
## boston[[i]]        -48.69      26.88   -1.81     0.07
## I(boston[[i]]^2)     2.84       1.56    1.82     0.07
## I(boston[[i]]^3)    -0.06       0.03   -1.89     0.06
##
## $black
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         12.60       2.52    5.01     0.00
## boston[[i]]         -0.02       0.06   -0.28     0.78
## I(boston[[i]]^2)     0.00       0.00    0.62     0.53
## I(boston[[i]]^3)     0.00       0.00   -0.47     0.64
##
## $lstat
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         48.65       1.43   33.91        0
## boston[[i]]         -3.87       0.33  -11.76        0
## I(boston[[i]]^2)     0.15       0.02    6.98        0
## I(boston[[i]]^3)     0.00       0.00   -5.01        0
```

Miscellaneous

(a) What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model?

Although the assumption for the linear model is that the relationship between explanatory variable and response variable should be linear, there is no particular assumption about the distribution of the explanatory variable.

(b) Why can an R2 close to one not be used as evidence that the simple linear regression model is appropriate?

It is possible that an r-squared value close to 1 is obtained even when one of the assumptions for the linear model is not satisfied. Hence, having an r-squared value close to one is not evidence that the simple linear model is appropriate.

(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?

Practically there cannot be a height of 0 units. Also, linear model helps us in predicting and understanding the relationship between height and weight according to the given data set. When we consider height of 0 units, we are extrapolating the regression line, which may not be correct. Hence, the regression does not imply that males of height 0 weigh 5 kgs on average. This however does not imply that the regression model is meaningless. Regression is very useful for inference about the relationship between variables and it is also useful for predicting. But we should be cautious about infering or predicting outside of the range of the explanatory variables.

(d) Suppose you had data on pairs (X; Y) which gave the scatterplot with cluster of points in two different directions. How would you approach the analysis?

The scatterplot shows that there could be other confounding variables that are influencing the relationship between the explanatory and the response variable. It could be that the explanatory variable has a different association with response variable for a subset of data having ‘f’ characteristics and a different association for the subset having ‘m’ charcateristics. For example, the females in the explanatory variable have a different association with the response variable than the males in the data set. Scatterplot shows two clusters behaving differently in the explanatory variable. Hence, we could consider other characteristics of the explanatory variable for better analysis.