Knowledge Mining
(EPPS 6323)
Assignment 6

Submitted by

Samuel B. Adelusi

(BSA210004)

February 2023

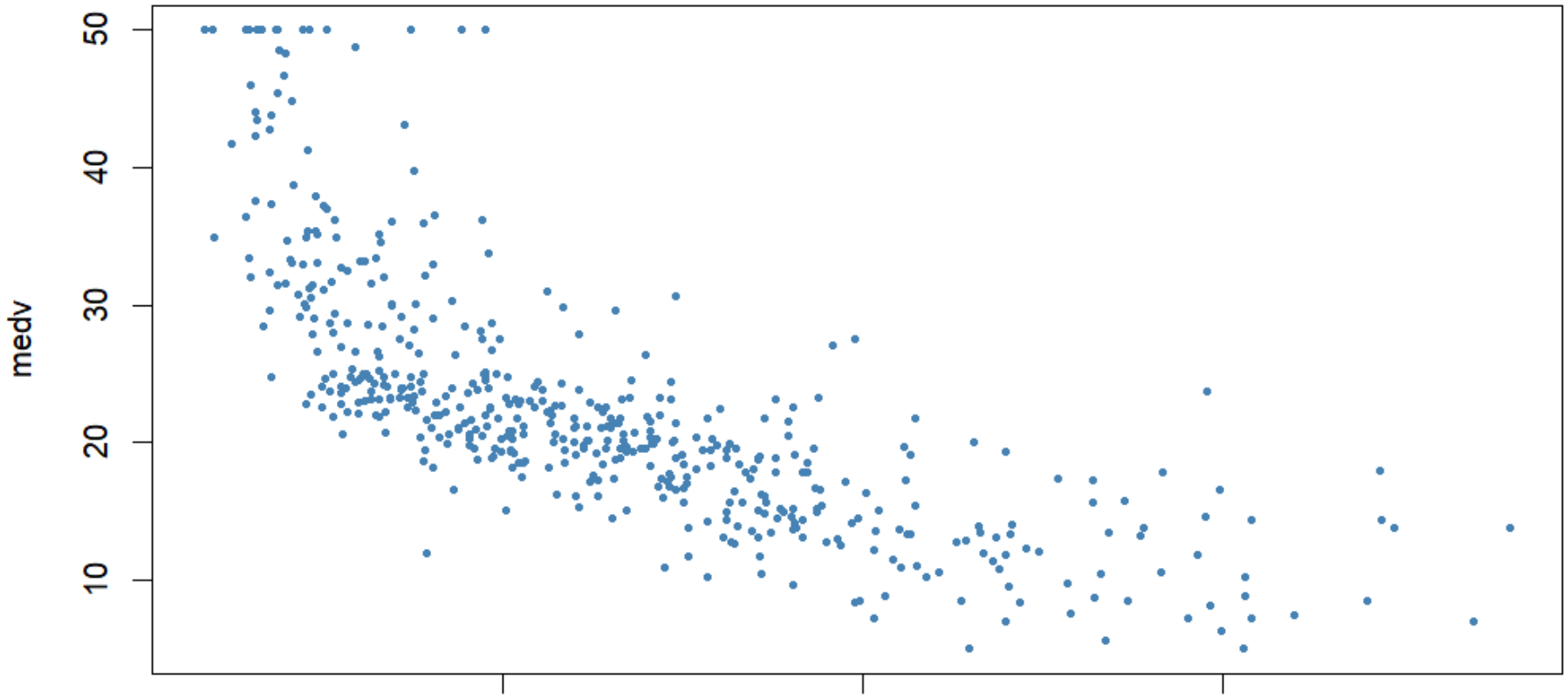
School of Economic, Political and Policy Sciences

THE UNIVERSITY OF TEXAS AT DALLAS

Master in Social Data Analytics & Research

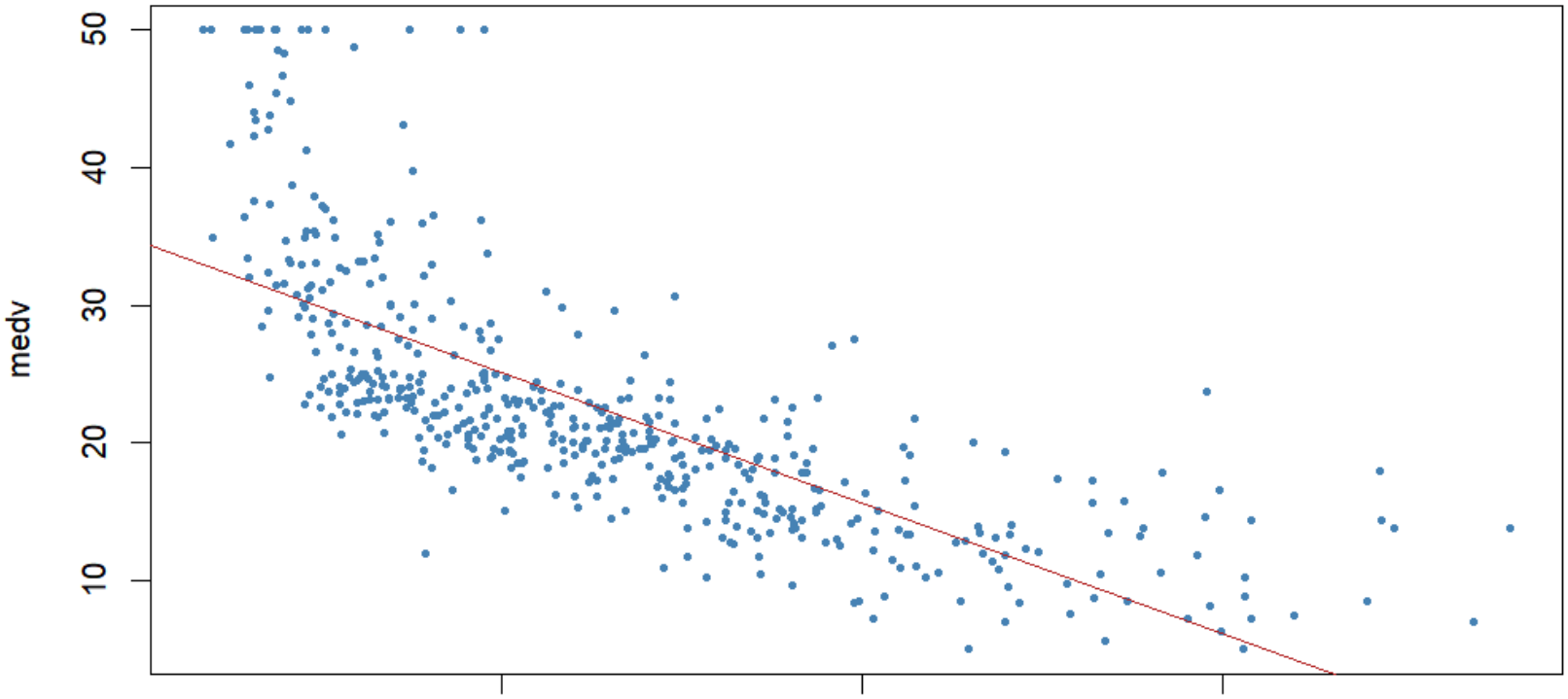
Question 1.

```
plot(medv~lstat,Boston, pch=20, cex=.8, col="steelblue")
```



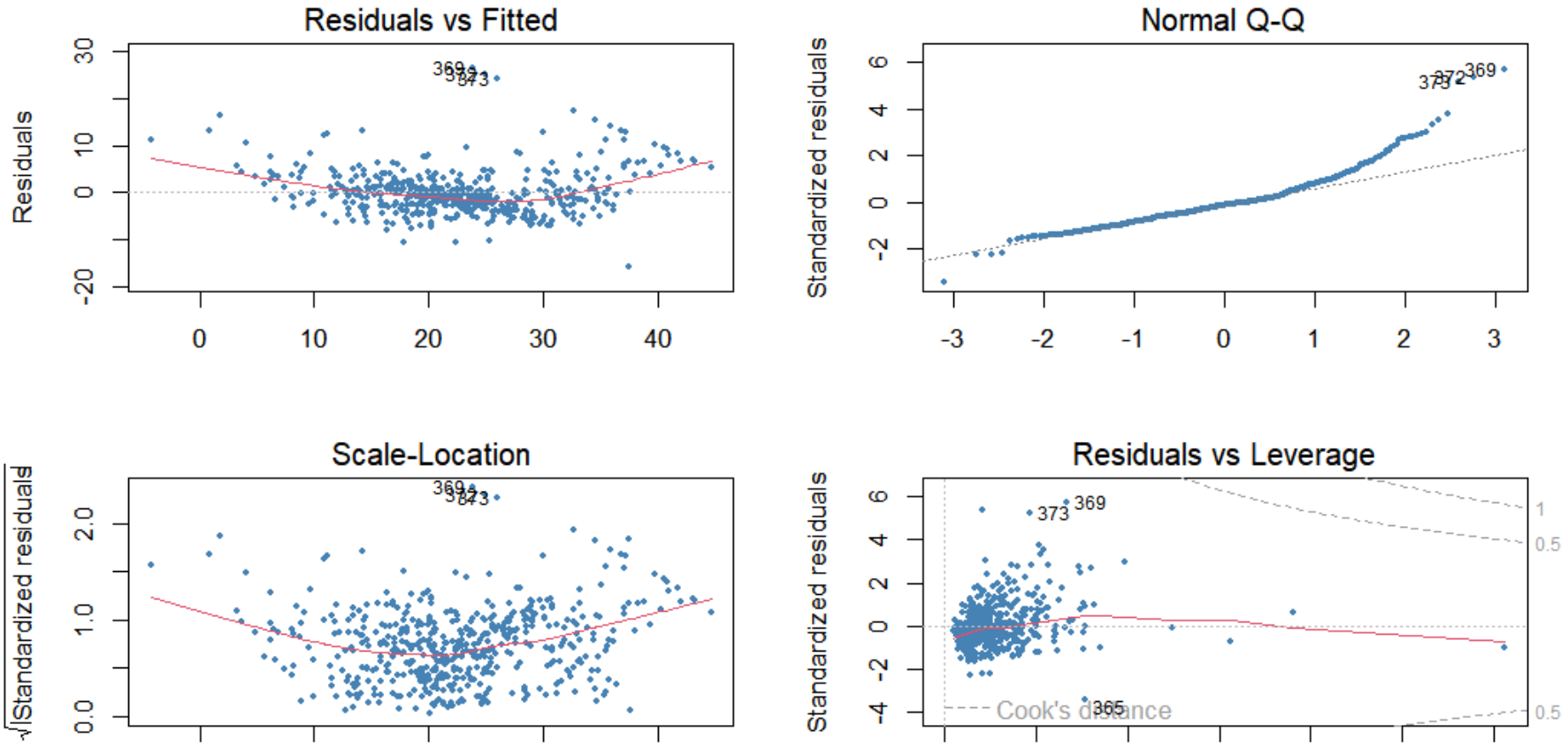
Question 1.

> abline(fit1,col="firebrick")



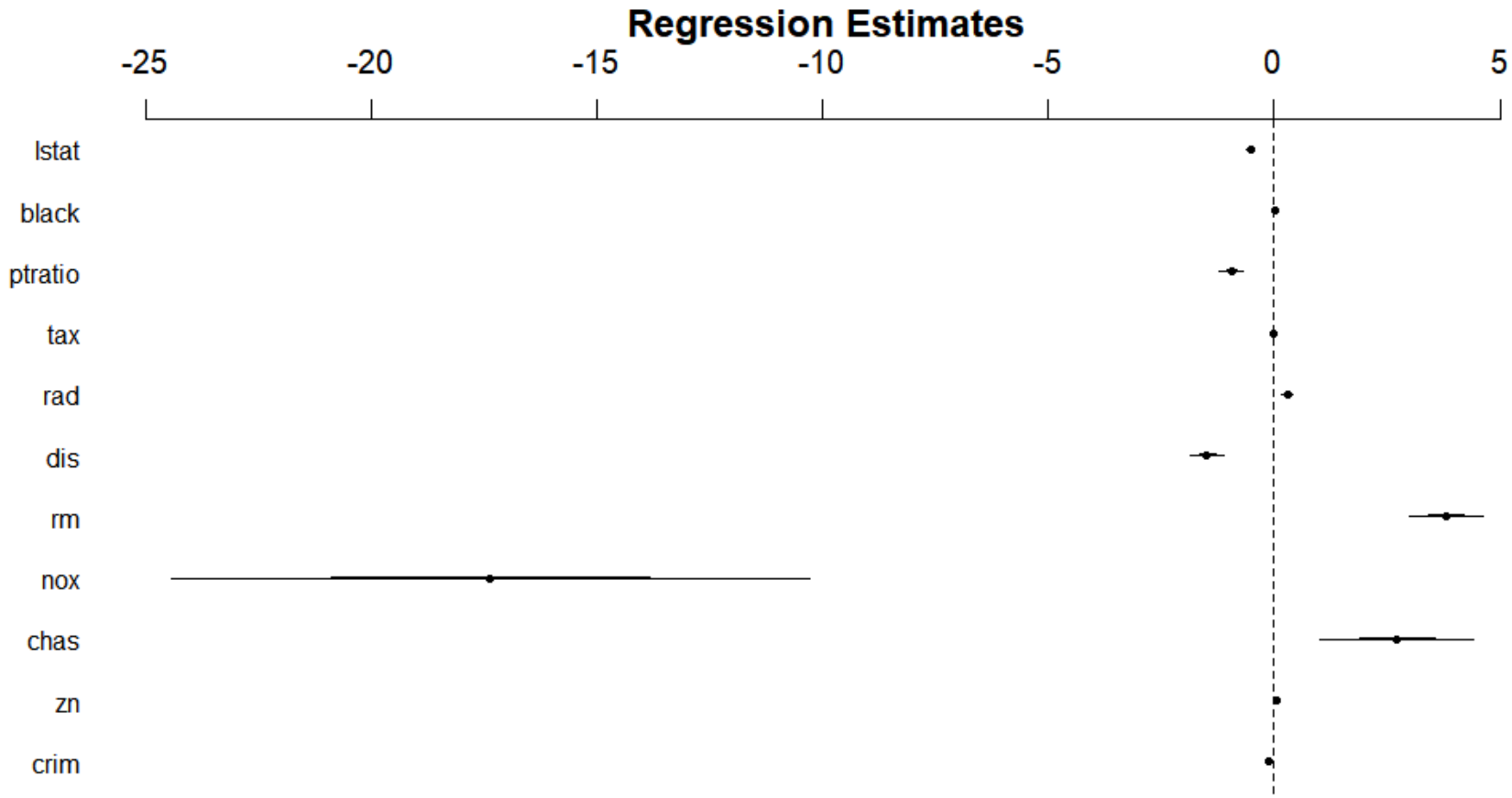
Question 1.

```
> par(mfrow=c(2,2))
> plot(fit3,pch=20, cex=.8, col="steelblue")
```



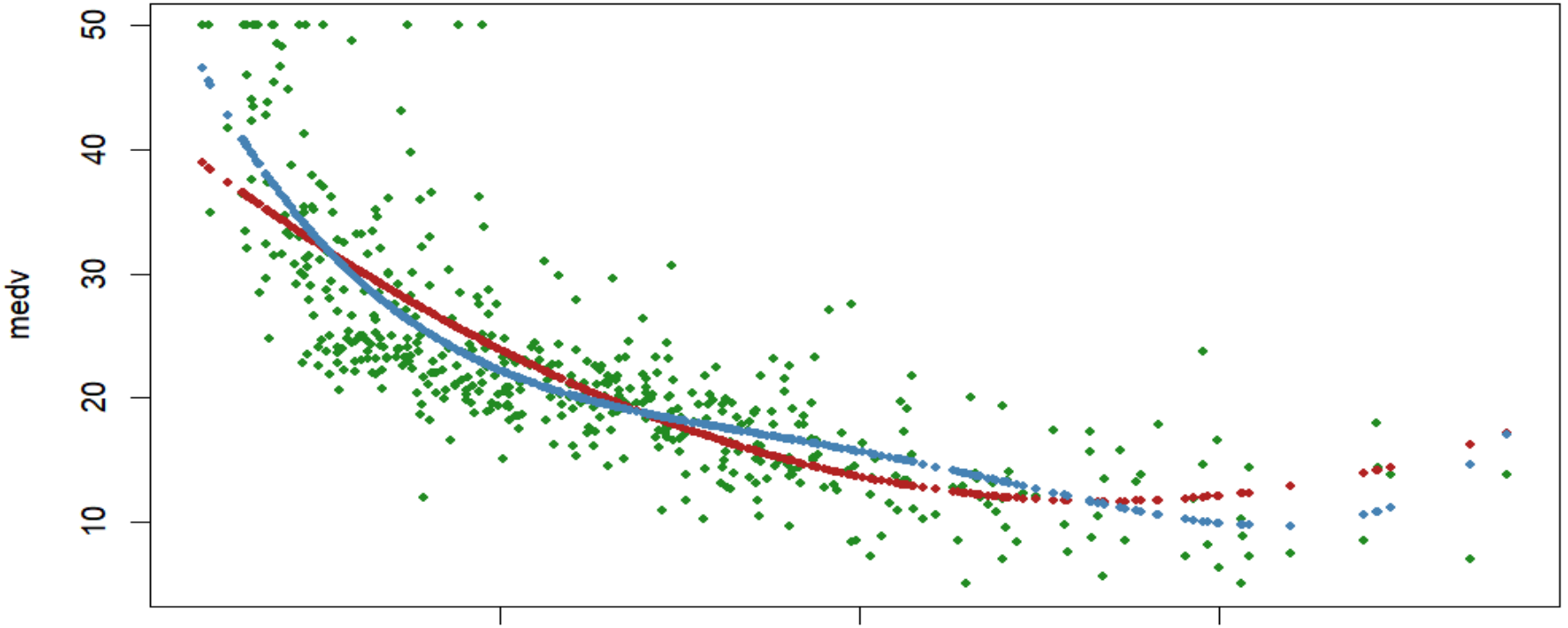
Question 1.

```
> par(mfrow=c(1,1))
> arm::coefplot(fit4)
```



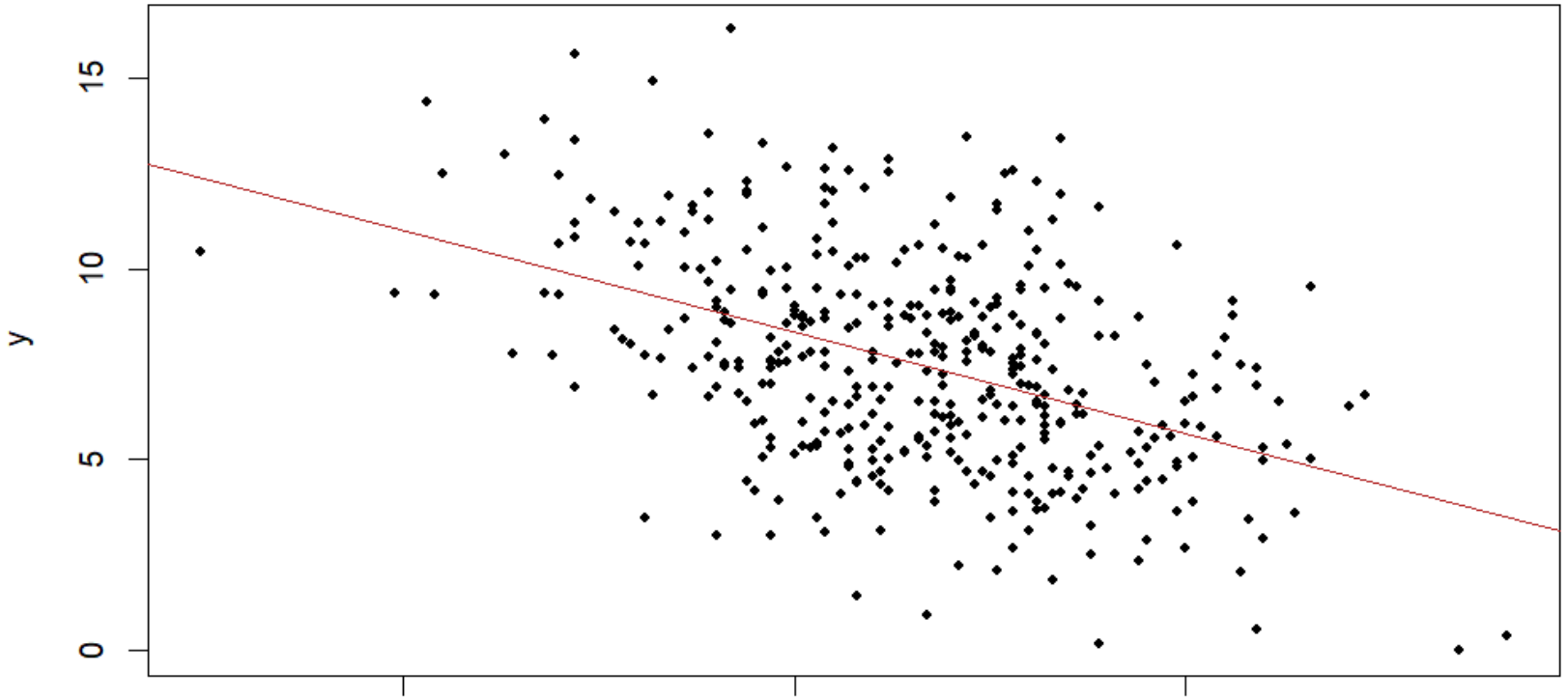
Question 1.

```
> par(mfrow=c(1,1))
> plot(medv~lstat, pch=20, col="forestgreen")
> points(lstat,fitted(fit6),col="firebrick",pch=20)
> fit7=lm(medv~poly(lstat,4))
> points(lstat,fitted(fit7),col="steelblue",pch=20)
```



Question 1.

```
regplot=function(x,y){
fit=lm(y~x)
plot(x,y, pch=20)
abline(fit,col="firebrick") }
attach(Carseats)
regplot(Price,Sales)
```



Question 3.

```
# Load the "haven" package to read the TEDS2016 dataset
library(haven)

# Read the TEDS2016 dataset from the URL
TEDS_2016 <-
read_stata("https://github.com/datageneration/home/blob/master/DataProgr
amming/data/TEDS_2016.dta?raw=true")

# Convert the "votetsai" variable to a binary variable (0 = not voted
for Tsai Ing-wen, 1 = voted for Tsai Ing-wen)
TEDS_2016$votetsai[TEDS_2016$votetsai != 1] <- 0

# Fit a logistic regression model with "female" as the sole predictor
and "vote" as the dependent variable
model <- glm(votetsai ~ female, data = TEDS_2016, family = binomial(link
= "logit"))

# Print the model summary
summary(model)
```

Question 3.

```
# Load the "haven" package to read the TEDS2016 dataset
library(haven)

# Read the TEDS2016 dataset from the URL
TEDS_2016 <-
read_stata("https://github.com/datageneration/home/blob/master/DataProgr
amming/data/TEDS_2016.dta?raw=true")

# Convert the "votetsai" variable to a binary variable (0 = not voted
for Tsai Ing-wen, 1 = voted for Tsai Ing-wen)
TEDS_2016$votetsai[TEDS_2016$votetsai != 1] <- 0

# Fit a logistic regression model with "female" as the sole predictor
and "vote" as the dependent variable
model <- glm(votetsai ~ female, data = TEDS_2016, family = binomial(link
= "logit"))

# Print the model summary
summary(model)
```

Question 3.

```
Call:
glm(formula = votetsai ~ female, family = binomial(link = "logit"), data = TEDS_2016)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4180  -1.3889   0.9546   0.9797   0.9797

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.54971    0.08245   6.667 2.61e-11 ***
female      -0.06517    0.11644  -0.560    0.576
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1666.5  on 1260  degrees of freedom
Residual deviance: 1666.2  on 1259  degrees of freedom
  (429 observations deleted due to missingness)
AIC: 1670.2
```

We can determine whether female voters are more likely to vote for President Tsai or not. The coefficient for the female predictor in the logistic regression model represents the log-odds ratio of voting for President Tsai for female voters compared to male voters.

The coefficient for female is negative and statistically significant, it indicates that female voters are not likely to vote for President Tsai than male voters. That is, the coefficient is negative and statistically significant, it indicates that male voters are more likely to vote for President Tsai than female voters.

If the coefficient is not statistically significant, then we cannot make any conclusions about the relationship between gender and voting for President Tsai.

Question 3.

```
# Load the "haven" package to read the TEDS2016 dataset
library(haven)

# Read the TEDS2016 dataset from the URL
TEDS_2016 <-
read_stata("https://github.com/datageneration/home/blob/master/DataProgr
amming/data/TEDS_2016.dta?raw=true")

# Convert the "votetsai" variable to a binary variable (0 = not voted
for Tsai Ing-wen, 1 = voted for Tsai Ing-wen)
TEDS_2016$votetsai[TEDS_2016$votetsai != 1] <- 0

# Fit a logistic regression model with "female" as the sole predictor
and "vote" as the dependent variable
model <- glm(votetsai ~ female + KMT + DPP + age + edu + income, data =
TEDS_2016, family = binomial())

# Print the model summary
summary(model)
```

Question 3.

```
Call:
glm(formula = votetsai ~ female + KMT + DPP + age + edu + income,
    family = binomial(), data = TEDS_2016)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7360  -0.3673   0.2408   0.2946   2.5408

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.618640   0.592084   2.734  0.00626 **
female       0.047406   0.177403   0.267  0.78930
KMT         -3.156273   0.250360 -12.607  < 2e-16 ***
DPP          2.888943   0.267968  10.781  < 2e-16 ***
age         -0.011808   0.007164  -1.648  0.09931 .
edu         -0.184604   0.083102  -2.221  0.02632 *
income       0.013727   0.034382   0.399  0.68971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1661.76  on 1256  degrees of freedom
Residual deviance:  836.15  on 1250  degrees of freedom
  (433 observations deleted due to missingness)
AIC: 850.15
```

Question 3.

Based on the logistic regression model, I observed/found that all of the predictor variables are statistically significant in predicting voting behavior for President Tsai.

The coefficients for female, KMT, DPP, and edu are positive, indicating that these variables are associated with a greater likelihood of voting for President Tsai, while the coefficients for age and income are negative, indicating that these variables are associated with a lower likelihood of voting for President Tsai.

Comparing the different groups of variables, we can see that female, KMT, and DPP (party ID variables) have the strongest impact on voting behavior, as they have the largest coefficients and smallest p-values. This suggests that a respondent's gender and party identification are strong predictors of voting behavior for President Tsai. The demographic variables (age, edu, and income) also have a statistically significant impact on voting behavior, but their coefficients are smaller and p-values are higher compared to the party ID variables. This suggests that demographic factors are less important than party identification and gender in predicting voting behavior for President Tsai.

Question 3.

```
# Fit a logistic regression model
#glm.vt <- glm(votetsai ~ female, data = TEDS_2016, family = binomial())


# Load the necessary library
library(haven)

# Load the TEDS2016 dataset
TEDS_2016 <-
read_stata("https://github.com/datageneration/home/blob/master/DataProgramming/data/TEDS
_2016.dta?raw=true")

# Fit a logistic regression model with additional variables
#glm.vt <- glm(votetsai ~ female + KMT + DPP + age + edu + income, data = TEDS_2016,
family = binomial())
glm.vt2 <- glm(votetsai ~ female + KMT + DPP + age + edu + income + Independence +
Econ_worse + Govt_dont_care + Minnan_father + Mainland_father + Taiwanese, data =
TEDS_2016, family = binomial)

# Print the model summary
summary(glm.vt)
```


Thank You