Knowledge Mining
(EPPS 6323)
Assignment 7

Submitted by
Samuel B. Adelusi
(BSA210004)

February 2023

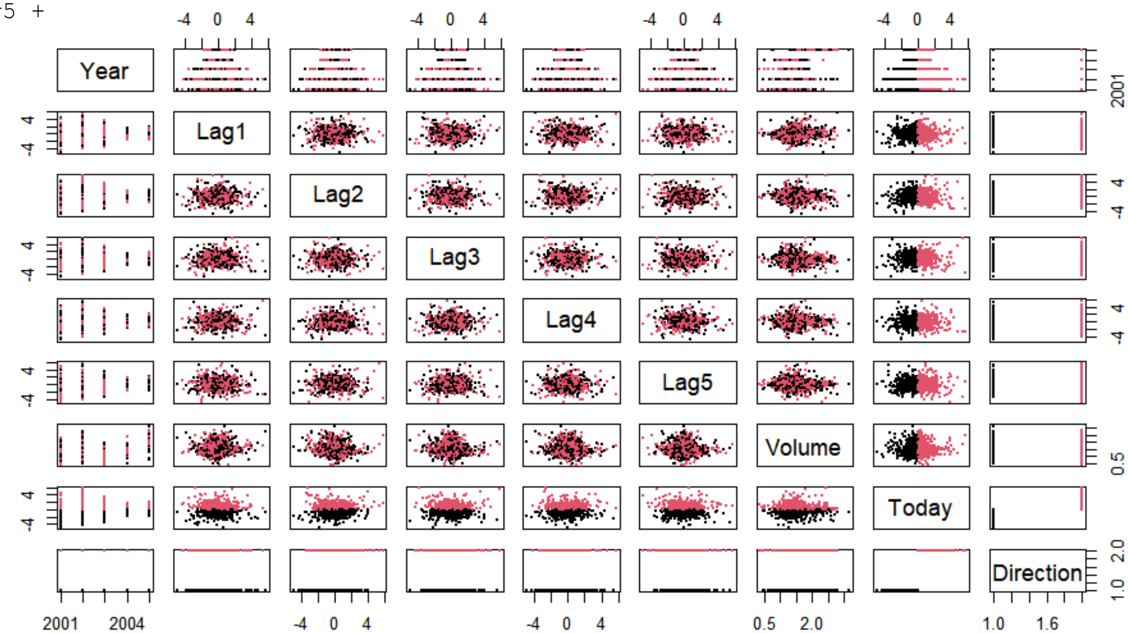
School of Economic, Political and Policy Sciences

THE UNIVERSITY OF TEXAS AT DALLAS

Question 1.

```
> summary(Smarket)
      Year           Lag1                Lag2                Lag3                Lag4                Lag5              Volume           Today
Direction
 Min.   :2001   Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.92200   Min.   :0.3561   Min.   :-4.922000   Down:602
 1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500   1st Qu.:-0.640000   1st Qu.:-0.640000   1st Qu.:-0.64000   1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648
 Median :2003   Median : 0.039000   Median : 0.039000   Median : 0.038500   Median : 0.038500   Median : 0.03850   Median :1.4229   Median : 0.038500
 Mean   :2003   Mean   : 0.003834   Mean   : 0.003919   Mean   : 0.001716   Mean   : 0.001636   Mean   : 0.00561   Mean   :1.4783   Mean   : 0.003138
 3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.59700   3rd Qu.:1.6417   3rd Qu.: 0.596750
 Max.   :2005   Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.73300   Max.   :3.1525   Max.   : 5.733000
>
```

```
> glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,
+             data=Smarket,family=binomial)
> summary(glm.fit)

Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial, data = Smarket)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.446   -1.203   1.065    1.145    1.326

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.126000   0.240736  -0.523    0.601
Lag1        -0.073074   0.050167  -1.457    0.145
Lag2        -0.042301   0.050086  -0.845    0.398
Lag3         0.011085   0.049939   0.222    0.824
Lag4         0.009359   0.049974   0.187    0.851
Lag5         0.010313   0.049511   0.208    0.835
Volume       0.135441   0.158360   0.855    0.392

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1731.2  on 1249  degrees of freedom
Residual deviance: 1727.6  on 1243  degrees of freedom
AIC: 1741.6
```



Question 2a.

The requirements of LDA (Linear Discriminant Analysis) are:

1. **Normality assumption**: The predictors (features) are normally distributed in each class.

2. **Homogeneity of variances (homoscedasticity)**: The variance of the predictors is the same in each class.

3. **Independence assumption:** The predictors are independent of each other within and among classes.

Question 2b.

Linear Discriminant Analysis (LDA) and Logistic Regression (LR) are both classification techniques, but they have different approaches and assumptions.

LDA is a generative model that assumes that the data for each class is normally distributed with a mean vector and covariance matrix. It calculates these parameters separately for each class and then uses Bayes' theorem to compute the probability of each class given a set of input features. LDA assumes that the variance of each class is the same, and that the classes have equal prior probabilities.

Logistic Regression, on the other hand, is a discriminative model that directly models the probability of each class given a set of input features. It assumes that the data is distributed according to a logistic distribution and uses maximum likelihood estimation to estimate the parameters of the logistic function. Logistic Regression does not make any assumptions about the distribution of the input features, but it assumes that the relationship between the input features and the output variable is linear.

In summary, LDA assumes that the data is normally distributed and estimates the probability of each class based on the input features, while Logistic Regression directly models the probability of each class given the input features and does not make any assumptions about the distribution of the input features.

Logistic Regression (LR) is for determining categorical variables, while Linear Discriminant Analysis (LDA) is for modeling continuous variables.

Question 2b.

LDA (Linear Discriminant Analysis) and Logistic Regression are both supervised learning algorithms that can be used for classification problems, but they have some differences:

- Assumption about the distribution of predictors: LDA assumes that the predictors are normally distributed within each class, while logistic regression makes no assumption about the distribution of predictors.

- Number of classes: LDA is a multiclass classification algorithm, meaning it can be used when there are more than two classes to predict. Logistic regression is typically used for binary classification problems.

- Output: LDA outputs a linear combination of the predictors that best separates the classes, while logistic regression outputs the probability of the outcome (class) given the predictor values.

- Interpretation: LDA provides information on how the predictors contribute to the classification of the different classes, while logistic regression provides information on the direction and strength of the relationship between the predictors and the outcome.

Question 2c.

ROC stands for Receiver Operating Characteristic. It is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied.

It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR is also known as sensitivity, recall or probability of detection, while FPR is also known as the probability of false alarm. The area under the ROC curve (AUC) is a common metric used to evaluate the overall performance of a classifier, where an AUC of 1 represents perfect performance and an AUC of 0.5 represents a random guess. A classifier with an AUC above 0.5 is better than random guessing.

ROC (Receiver Operating Characteristic) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is a performance metric for binary classification problems where the model predicts the probability of the positive class, and a threshold is chosen to determine the final classification decision. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

The area under the ROC curve (AUC) is also a commonly used performance metric that summarizes the overall performance of the classifier, with higher values indicating better discrimination ability. AUC values range from 0.5 to 1.0, with a value of 0.5 indicating random guessing and a value of 1.0 indicating perfect discrimination.

Question 2d. What is sensitivity and specificity?

Sensitivity and specificity are two common measures used to evaluate the performance of binary classification models.

Sensitivity, also known as true positive rate, is the proportion of actual positives (i.e., the proportion of samples from the positive class) that are correctly identified as such by the model. In other words, sensitivity measures how well the model can detect the positive class.
Sensitivity measures the proportion of true positive cases that are correctly identified by the classifier, that is, the proportion of cases where the actual value is positive (or "true") that are correctly predicted as positive by the classifier.

Specificity, also known as true negative rate, is the proportion of actual negatives (i.e., the proportion of samples from the negative class) that are correctly identified as such by the model. In other words, specificity measures how well the model can detect the negative class.
Specificity, on the other hand, measures the proportion of true negative cases that are correctly identified by the classifier, that is, the proportion of cases where the actual value is negative (or "false") that are correctly predicted as negative by the classifier.

In practice, sensitivity and specificity are often traded off against each other, and there is a trade-off between maximizing both measures at the same time. For example, by adjusting the threshold for classification, we can increase sensitivity at the cost of specificity or vice versa.
In general, a classifier with high sensitivity and high specificity is preferred, as it indicates that the classifier is able to accurately identify both positive and negative cases.

Question 2d. Which is more important in your opinion?

True positives may be more acceptable than true negatives.

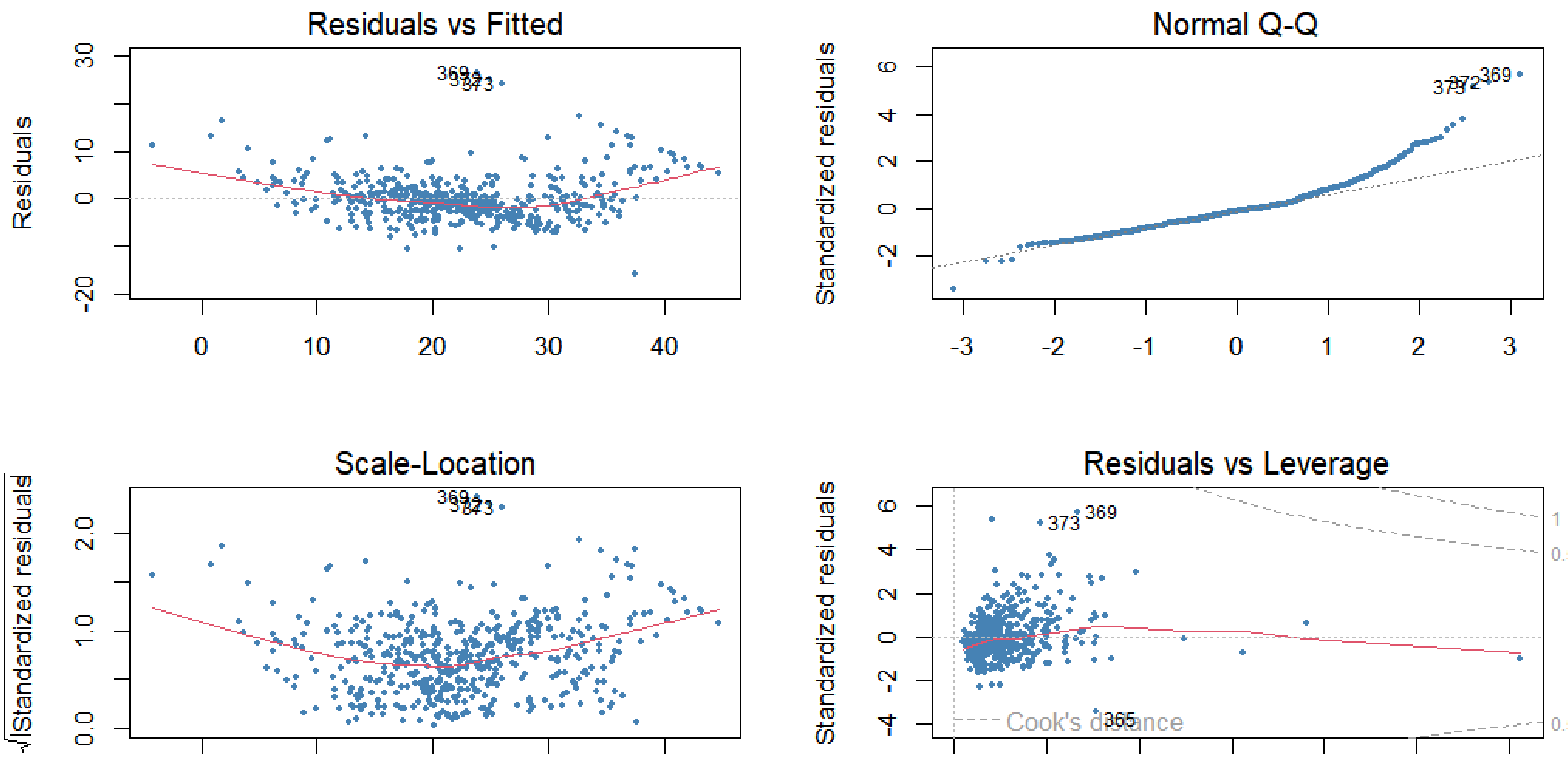
Sensitivity, also known as true positive rate, is the most important in my opinion.

In any other situations, true negative (specificity) may be more critical.

It ultimately depends on the nature and consequences of the decision being made based on the prediction.

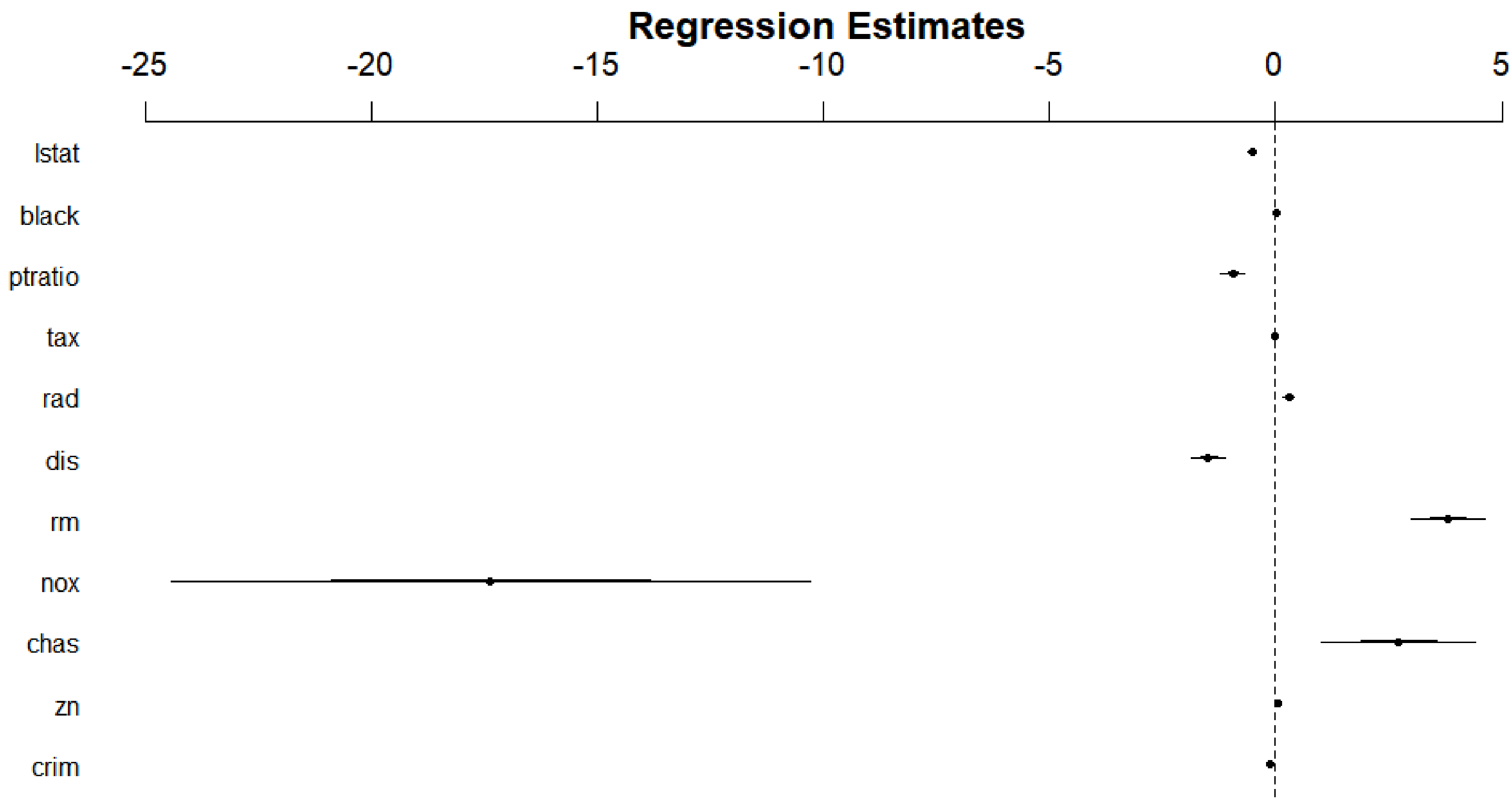
Question 1.

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



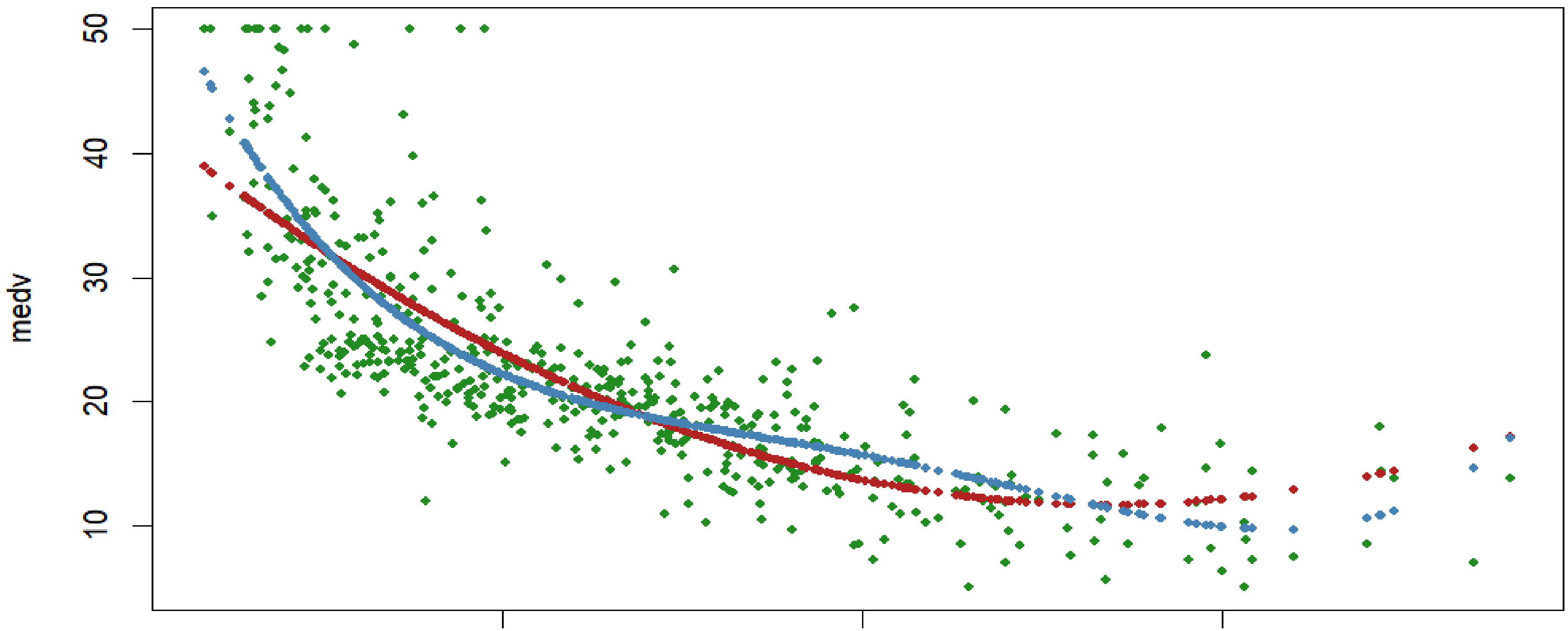
Question 1.

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



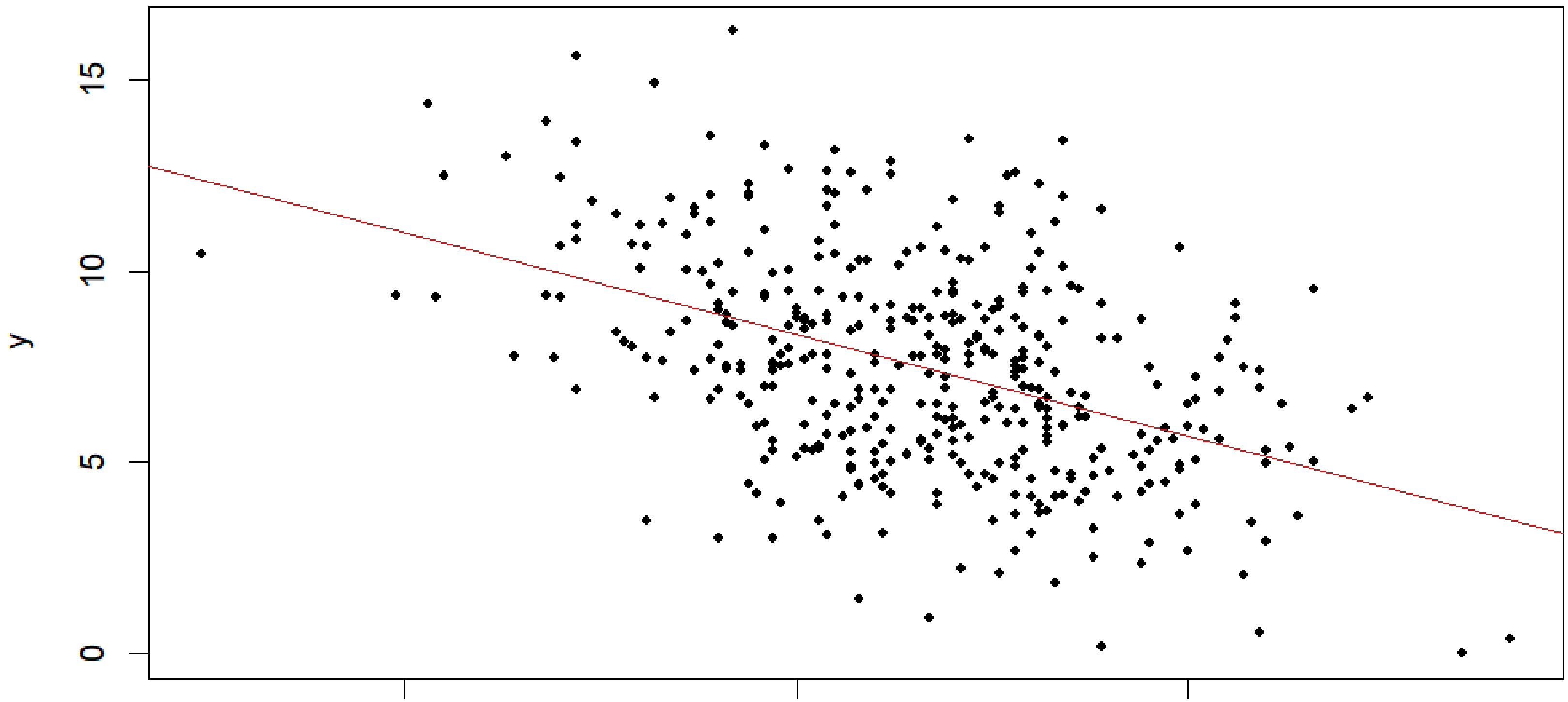
Question 1.

```
> 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