Chapter 3 Linear Regression

Jishen Yin

2020/5/3

```
knitr::opts_chunk$set(echo = TRUE)
library(ISLR)
library(MASS)
library(tidyverse)
library(GGally)
```

Problem 8

This question involves the use of simple linear regression on the `Auto` data set.

(a) Perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Comment on the summary.

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive and negative?

```
data(Auto)
model <- lm(mpg ~ horsepower, data = Auto)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

They have linear relationship. R^2 is 0.6059, which indicates that their relationship is not strong. The coefficient is negative, indicating a negative relationship between `mpg` and `horsepower`.

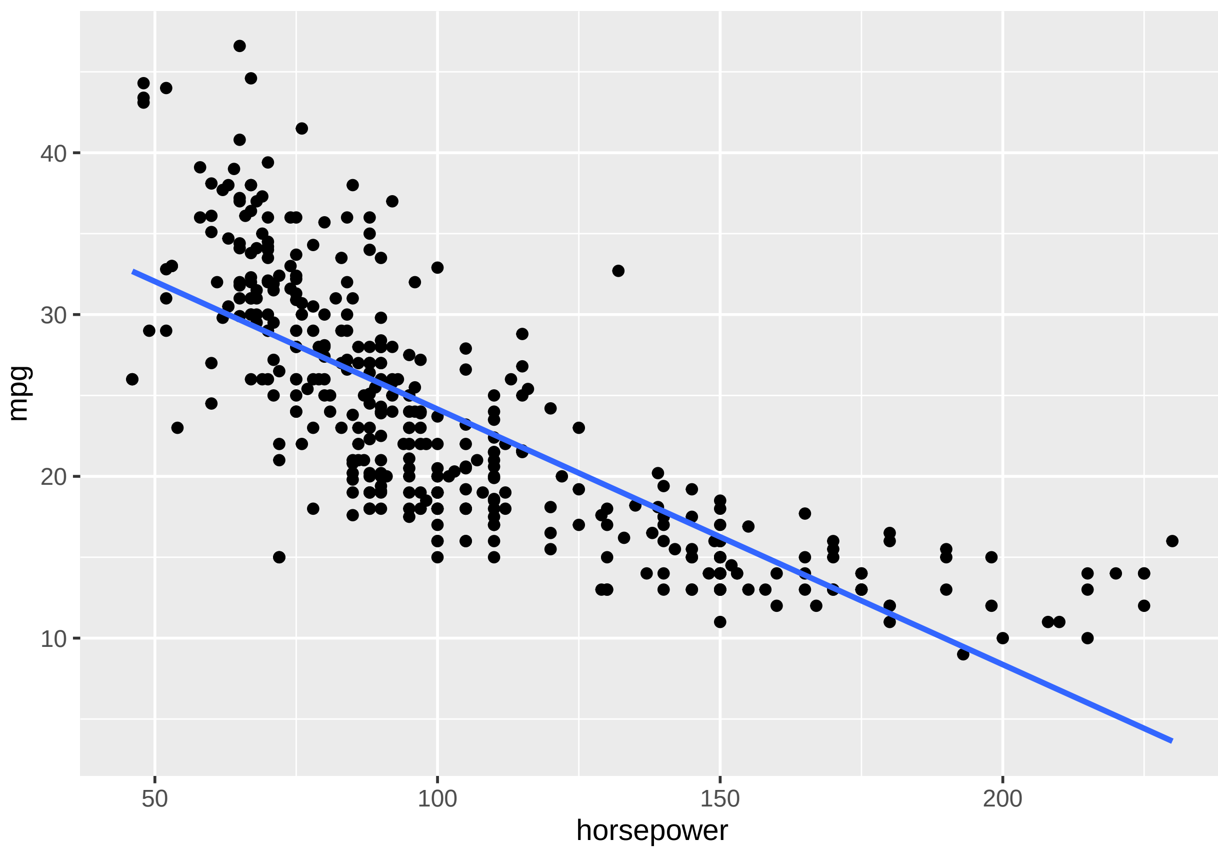
iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

```
predict.lm(model, data.frame(horsepower=98), interval = "prediction", level = 0.95)
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

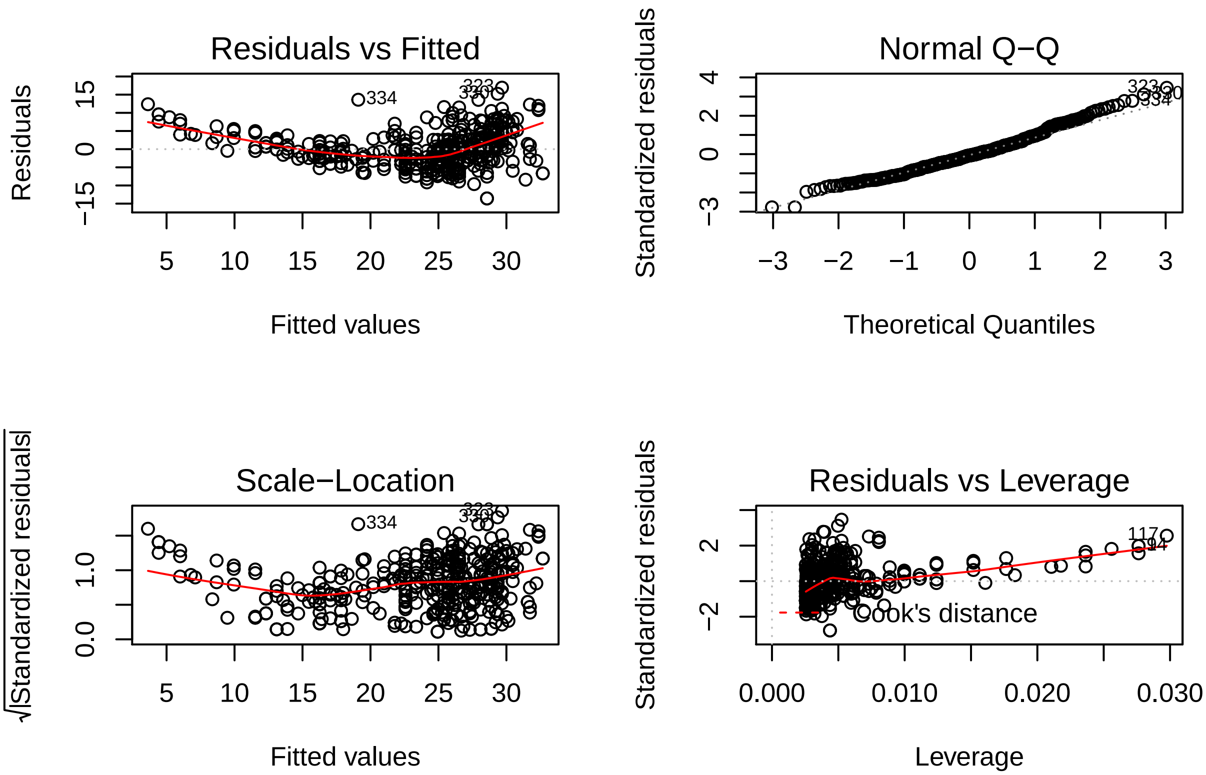
(b) Plot the response and the predictor. Display the least squares regression line.

```
ggplot(data = Auto) +
  geom_point(aes(x = horsepower, y = mpg)) +
  geom_smooth(aes(x = horsepower, y = mpg), method = "lm", se = FALSE)
```



(c) Produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```
par(mfrow=c(2, 2))
plot(model)
```



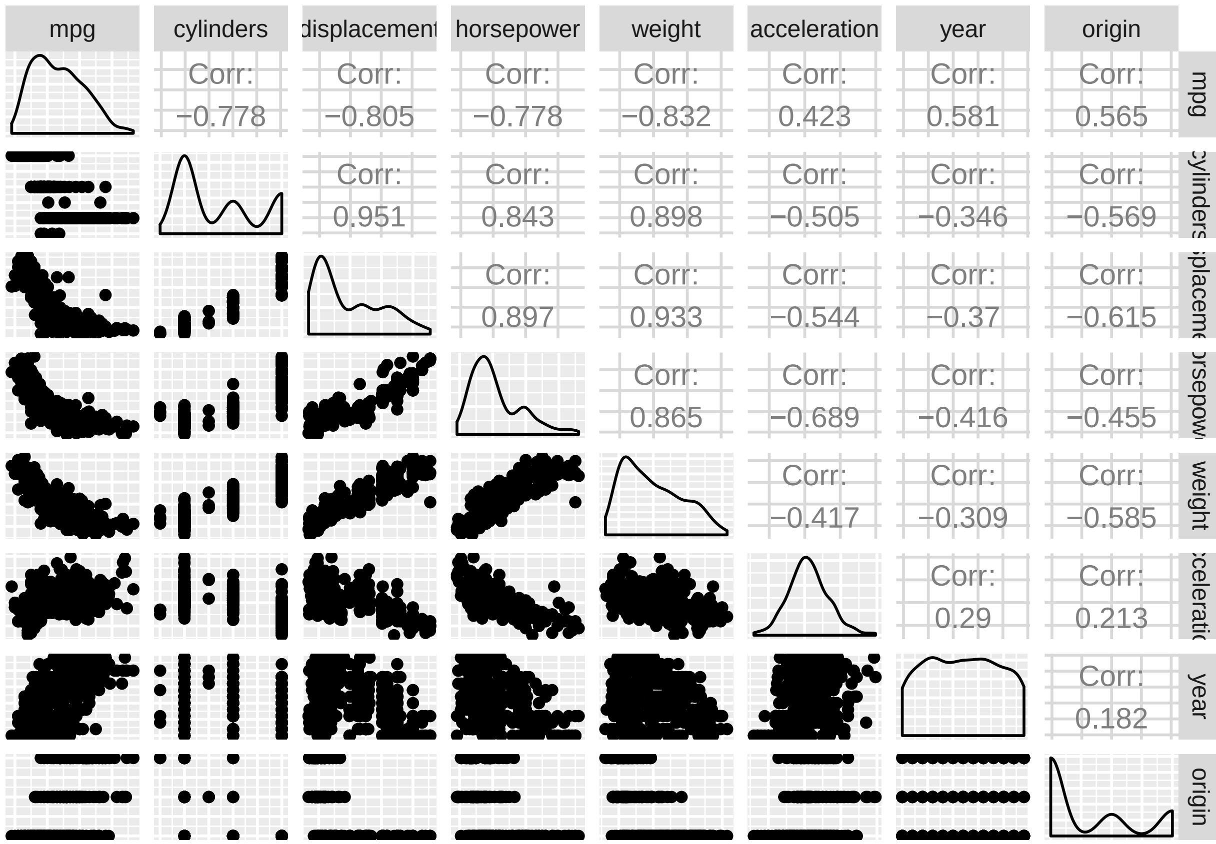
The variance is not constant among the predictor. Number 323, 330, 334 are potential outliers. Number 94, 117 are potential influential points.

Problem 9

This question involves the use of multiple linear regression on the `Auto` data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
ggpairs(Auto[, 1:8], axisLabels = "none")
```



(b) Compute the matrix of correlations between the variables.

```r
cor(Auto[, 1:8])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c) Perform a multiple linear regression with `mpg` as the response and all other variables except `name`. Comment on the summary.

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the `year` variable suggest?

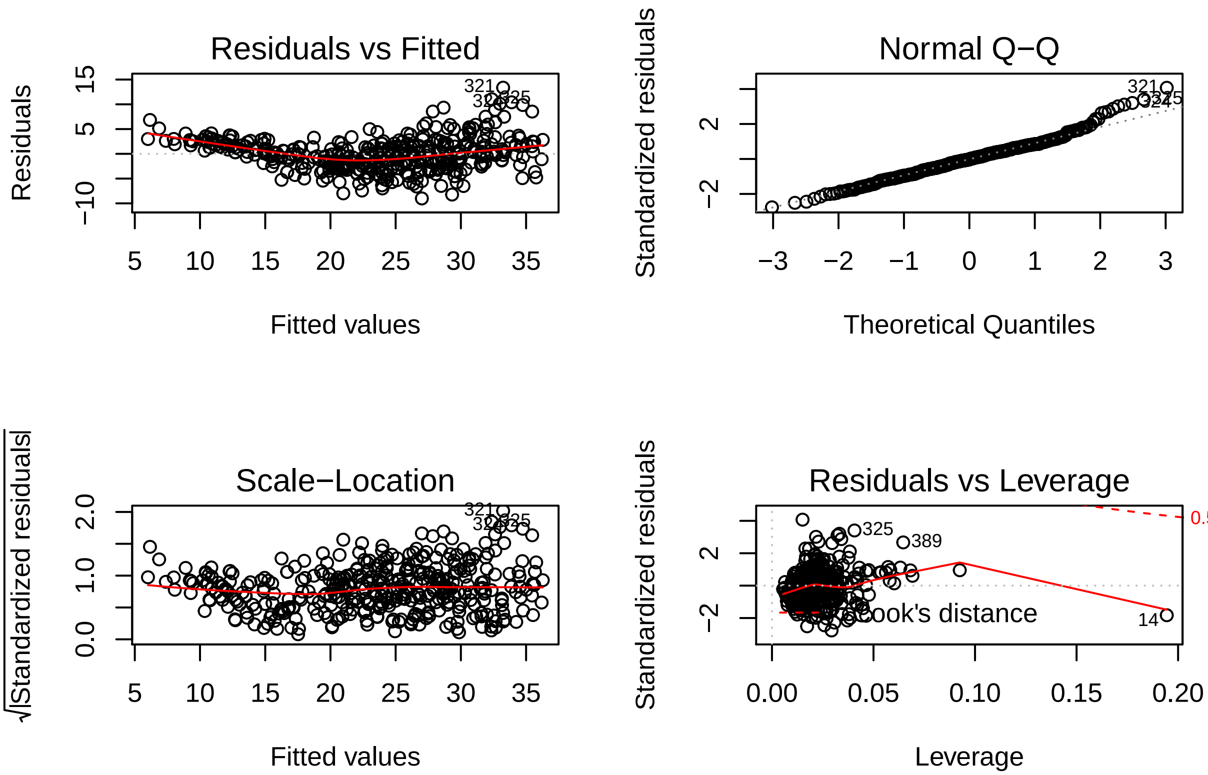
```
model <- lm(mpg ~ .-name, data = mutate(Auto, origin = as.factor(origin)))
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = mutate(Auto, origin = as.factor(origin)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders    -4.897e-01  3.212e-01  -1.524 0.128215
## displacement  2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower   -1.818e-02  1.371e-02  -1.326 0.185488
## weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration  7.910e-02  9.822e-02   0.805 0.421101
## year          7.770e-01  5.178e-02  15.005  < 2e-16 ***
## origin2       2.630e+00  5.664e-01   4.643 4.72e-06 ***
## origin3       2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

`displacement`, `weight`, `year` and `origin` have significant relationship to the response. On average, `mpg` will increase 0.7 after one year.

(d) Produce diagnostic plots of the linear regression fit. Comment on any problems.

```
par(mfrow = c(2, 2))
plot(model)
```



(e) Fit linear regression models with interaction effects.

```
model2 <- lm(mpg ~ (.-name)^2, data = mutate(Auto, origin = as.factor(origin)))
bic <- step(model2, direction = "both", k = log(nrow(Auto)), trace = 0)
summary(bic)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + horsepower + weight + acceleration +
##     year + origin + displacement:weight + horsepower:year + acceleration:origin,
##     data = mutate(Auto, origin = as.factor(origin)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9276 -1.4649 -0.0121  1.2379 11.4958
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -5.589e+01  9.584e+00  -5.832 1.17e-08 ***
## displacement         -5.525e-02  8.711e-03  -6.342 6.46e-10 ***
## horsepower            5.101e-01  8.691e-02   5.869 9.56e-09 ***
## weight               -8.298e-03  7.197e-04 -11.529  < 2e-16 ***
## acceleration         -3.271e-01  1.029e-01  -3.179   0.0016 **
## year                  1.507e+00  1.195e-01  12.603  < 2e-16 ***
## origin2              -1.061e+01  2.379e+00  -4.461 1.08e-05 ***
```

```
## origin3              -5.193e+00  3.065e+00  -1.695   0.0910 .
## displacement:weight   1.692e-05  2.157e-06   7.846 4.40e-14 ***
## horsepower:year      -7.556e-03  1.188e-03  -6.361 5.76e-10 ***
## acceleration:origin2  7.185e-01  1.374e-01   5.228 2.84e-07 ***
## acceleration:origin3  3.825e-01  1.883e-01   2.031   0.0429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 380 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8812
## F-statistic: 264.6 on 11 and 380 DF,  p-value: < 2.2e-16
```

Problem 10

This question should be answered using the `Carseats` data set.

(a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`.

```r
data(Carseats)

reg <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(reg)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```r
small <- lm(Sales ~ Price + US, data = Carseats)
summary(small)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(f) How well do the models in (a) and (e) fit the data?

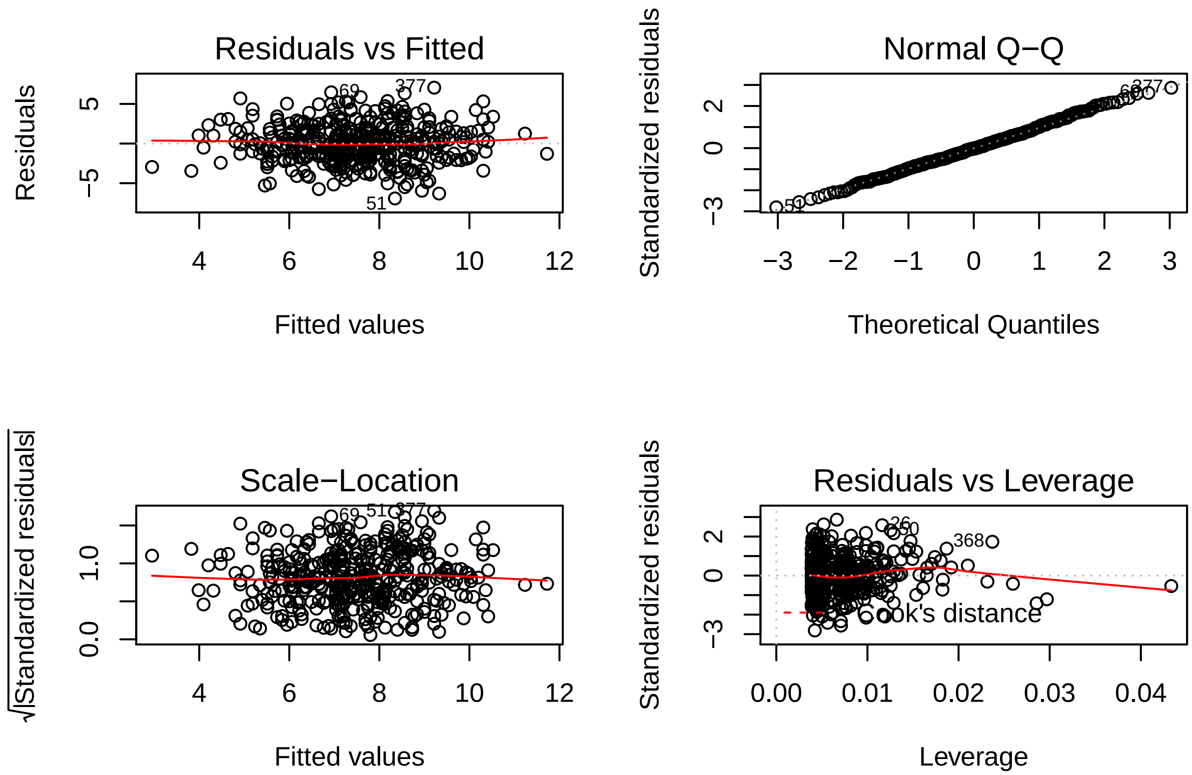
(g) Using the model from (e), obtain 95% confidence intervals for the coefficients.

```r
confint(small)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

```r
par(mfrow = c(2, 2))
plot(small)
```



Problem 15

This problem involves the `Boston` data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. 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.

```
data(Boston)
```

```
l <- lapply(colnames(Boston)[2:14], function(x){
  formul <- formula(paste("crim ~", x))
  model <- lm(formul, data = Boston)
  return(data.frame(varname = x,
                    sim_coef = model$coefficients[2],
                    sim_r2 = summary(model)$r.squared))
})

df <- do.call(rbind, l)
```

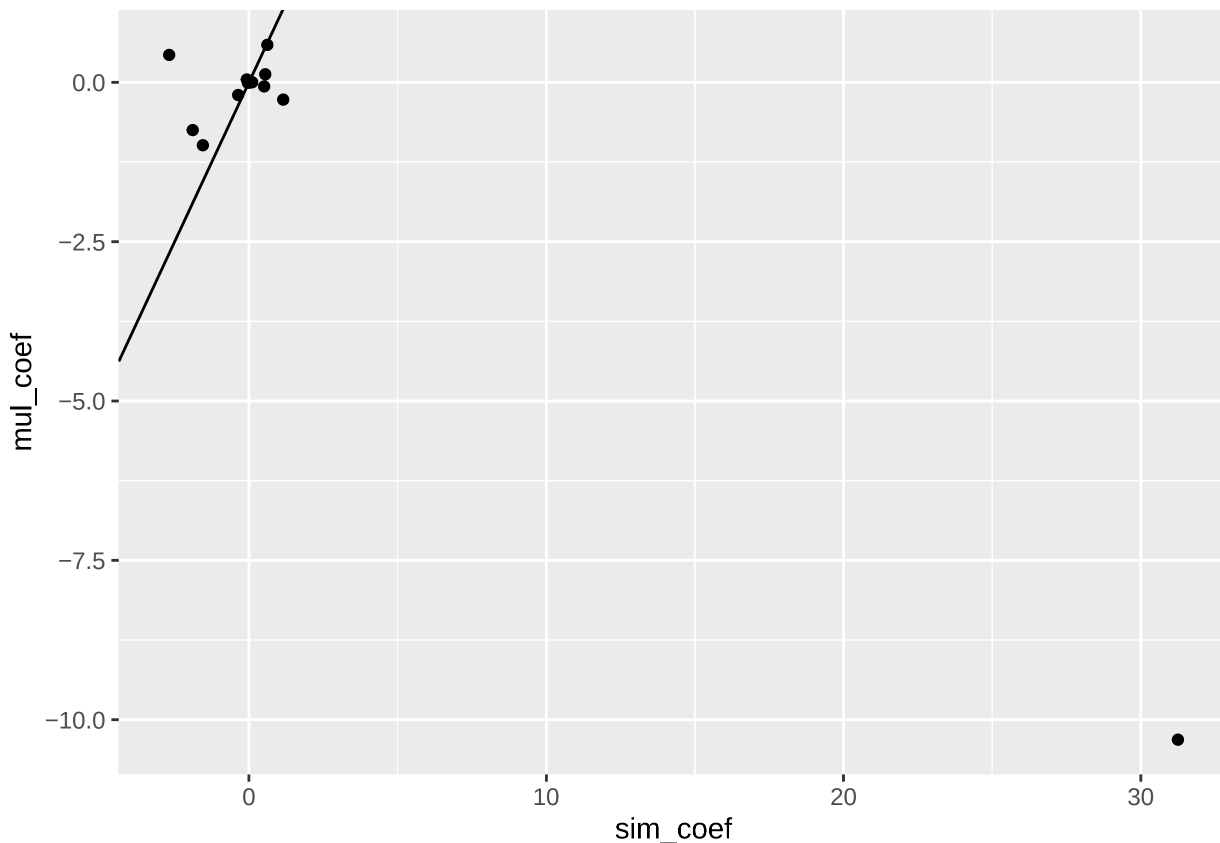
(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```r
full <- lm(crim ~ ., data = Boston)
df["mul_coef"] <- full$coefficients[2:14]
summary(full)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

```r
ggplot(data = df) +
  geom_point(aes(x = sim_coef, y = mul_coef)) +
  geom_abline(slope = 1, intercept = 0)
```



(d) Is there evidence of 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

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

```
df["mulno_r2"] <- sapply(colnames(Boston)[2:14], function(x){
  formul <- formula(paste("crim ~", x, "+I(", x, "^2)+I(", x, "^3)"))
  model <- lm(formul, data = Boston)
  return(summary(model)$r.squared)
})
```