MATH 4322 Homework 2 - Solutions

Instructor: Dr. Cathy Poliak

Fall 2022

Problem 1

The following output is based on predicting `sales` based on three media budgets, `TV`, `radio`, and `newspaper`.

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

a. Give the estimated model to predict sales.

Answer

$$\widehat{\text{sales}} = 2.9389 + 0.0458 \times \text{TV} + 0.1885 \times \text{radio} - 0.0001 \times \text{newspaper}$$

b. Describe the null hypothesis to which the p-values given in the `Coefficients` table correspond. Explain this in terms of the `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

Answer For each of these:
H_0: TV is not needed in the model if radio and newspaper are in the model $t = 32.809$, p-value ≈ 0.
H_0: radio is not needed in the model if TV and newspaper are in the model $t = 21.893$, p-value ≈ 0.
H_0: newspaper is not needed in the model if TV and radio are in the model. $t = -0.177$, p-value $= 0.86$.

c. Are there any variables that may not be significant in predicting `sales`?

Answer Yes, since the p-value is large (greater than 0.05) for `newspaper` this variable might not be needed in the model.

Problem 2

Based on the previous problem,the following is the output from the full model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq   F value Pr(>F)
## TV          1 3314.6  3314.6 1166.7308 <2e-16 ***
## radio       1 1545.6  1545.6  544.0501 <2e-16 ***
## newspaper   1    0.1     0.1    0.0312 0.8599
## Residuals 196  556.8     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is based on the model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## TV          1 3314.6  3314.6 1172.50 < 2.2e-16 ***
## radio       1 1545.6  1545.6  546.74 < 2.2e-16 ***
## Residuals 197  556.9     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is based on the model $\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \epsilon$

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## TV          1 3314.6  3314.6  312.14 < 2.2e-16 ***
## Residuals 198 2102.5    10.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Determine the `AIC` for all three models.

Answer

Model 1: $\text{AIC} = 2(4) + 200 \times ln\left(\frac{556.8}{200}\right) = 212.7777485$
Model 2: $\text{AIC} = 2(3) + 200 \times ln\left(\frac{556.9}{200}\right) = 210.8136648$
Model 3: $\text{AIC} = 2(2) + 200 \times ln\left(\frac{2102.5}{200}\right) = 474.5130051$

b) Determine the C_p for all three models.

Answer

Model 1: $C_p = \frac{556.8}{2.8} + 2(4) - 200 = 6.8571429$
Model 2: $C_p = \frac{556.9}{2.8} + 2(3) - 200 = 4.8928571$
Model 3: $C_p = \frac{2102.5}{2.8} + 2(2) - 200 = 554.8928571$

c) Determine the adjusted R^2 for all three models.

Answer

SST = 3314.6+1545.6+.1+556.8 = 5417.1
The SST is the same for all of the models.
Model 1: $R^2 = 1 - \frac{556.8/(200-3-1)}{5417.1/199} = 0.8956$
Model 2: $R^2 = 1 - \frac{556.9/(200-2-1)}{5417.1/199} = 0.8943$
Model 3: $R^2 = 1 - \frac{2102.5/(200-1-1)}{5417.1/199} = 0.6099$

d) Determine the `RSE` for all three models.

Answer
Model 1: $\text{RSE} = \sqrt{\frac{556.8}{196}} = 1.6855$
Model 2: $\text{RSE} = \sqrt{\frac{556.9}{197}} = 1.6813$
Model 3: $\text{RSE} = \sqrt{\frac{2102.5}{198}} = 3.2586$

e) Which model best fits to predict `sales` based on these statistics?

Answer
The AIC, C_p and RSE are all the smallest with Model 2. The adjusted R^2 is slightly larger for Model 1 but not by much. Thus the best of the three models is:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Problem 3

Suppose we have a data set with five predictors, X_1 =GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$,$\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Answer
The predicted model is:

$$\hat{\text{salary}} = \begin{cases} 85 + 10 \times \text{GPA} + 0.07 \times \text{IQ} + 0.01 \times \text{GPA} \times \text{IQ} & \text{if Female} \\ 50 + 20 \times \text{GPA} + 0.07 \times \text{IQ} + 0.01 \times \text{GPA} \times \text{IQ} & \text{if Male} \end{cases}$$

i. This is false because the y-intercept is higher for a female.

ii. This is false, because of the interaction term, as the GPA increases, the starting salary for a male will become higher.

iii. This is true, a higher GPA for a male will allow the starting salary to be higher.

iv. This is false, a lower GPA for a female will allow the starting salary to be higher for females.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Answer
predicted salary $= 85 + 10 \times 4.0 + 0.07 \times 110 + 0.01 \times 4.0 \times 110 = 137.1$
or \$137,100.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Answer
This is probably true, we need to determine this with a t-test.

Problem 4

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots, p$ predictors. Answer true or false to the following statements.

(a) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection. **True**

(b) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by backward stepwise selection. **True**

(c) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection. **False**

(d) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by backward stepwise selection. **False**

(e) The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k + 1)$ - variable model identified by best subset selection. **False**

Problem 5

This question involves the use of simple linear regression on the *Auto* data set. This can be found in the `ISLR2` package in `R`.

(a) Use the `lm()` function to perform a simple linear regression with *mpg* as the response and *horsepower* (*hp*) as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

   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 or negative?
   iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals? Give an interpretation of these intervals.

Answer

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.2.1
```

```
data(Auto)
auto.lm = lm(mpg ~ horsepower,data = Auto)
summary(auto.lm)
```

```
##
## 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
```

i. There is appears to be a relationship between horsepower and mpg.

ii. This seems to be a somewhat strong relationship as the $R^2 = 0.6059$.

iii. This is a negative relationship.

iv. See the output below:

```r
predict(auto.lm, newdata = data.frame(horsepower = 98),interval = "p")
```

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

```r
predict(auto.lm, newdata = data.frame(horsepower = 98),interval = "c")
```

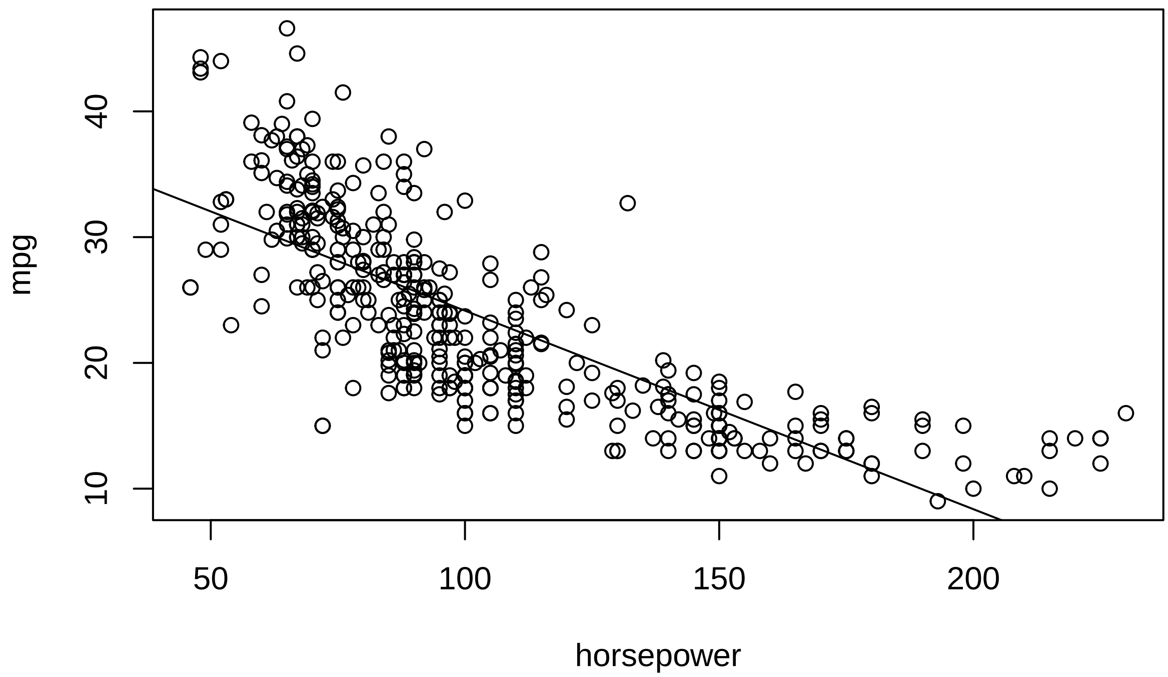
```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

The predicted `mpg` is 24.46708.
The prediction interval is [14.8094, 34.12476], this means for **one** automobile that has a horsepower of 98, we are 95% confident that the mpg is between 14.8094 and 34.12476.
The confidence interval is [23.97308, 24.96108], this means for **all** of the automobiles that have a horsepower of 98, we are 95% confident that the **mean** mpg will be between 23.97308 and 24.96108.

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

Answer

```r
attach(Auto)
plot(horsepower,mpg)
abline(auto.lm)
```

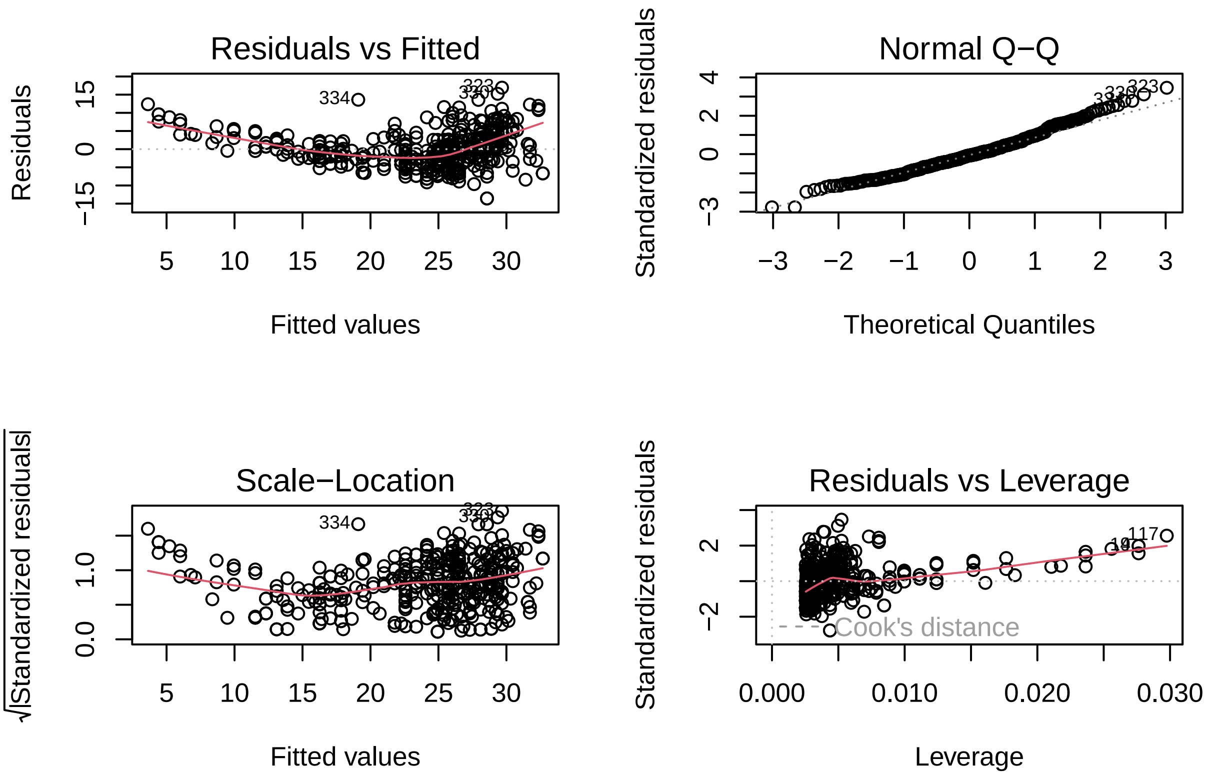



```r
detach(Auto)
```

(c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Answer

```r
par(mfrow = c(2,2))
plot(auto.lm)
```



This may not be a linear relationship.

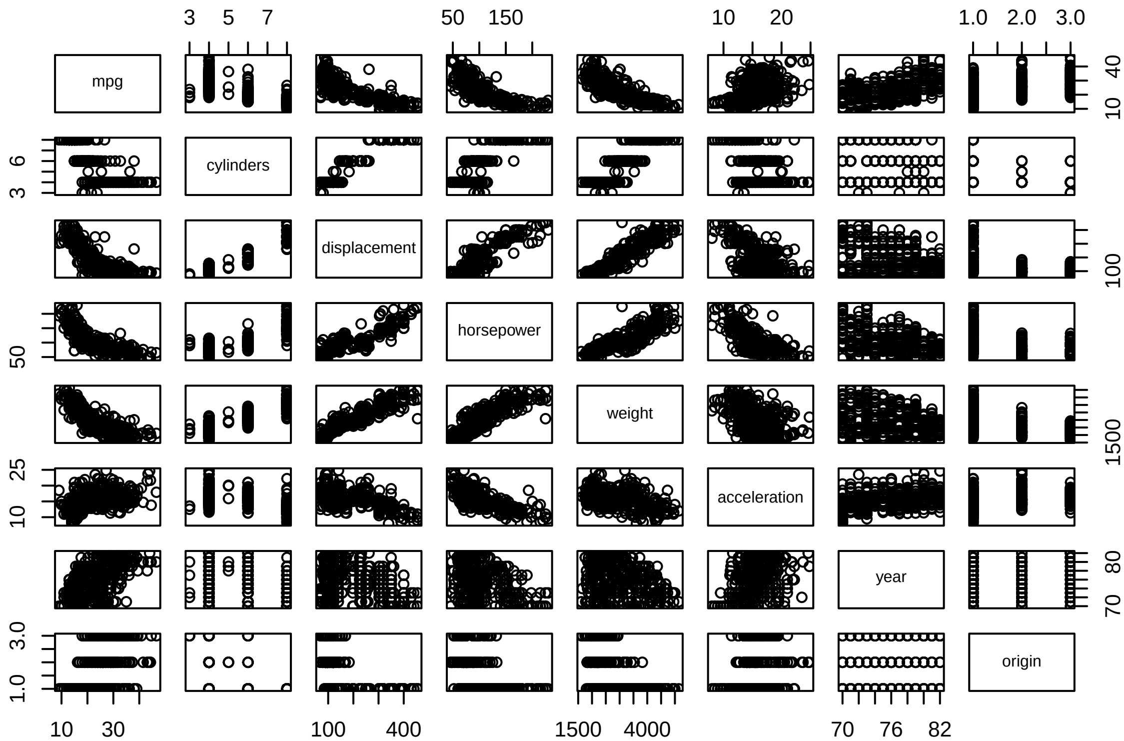
Problem 6

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.

Answer

```
pairs(~mpg+cylinders+displacement+horsepower+weight+acceleration+year+origin,data = Auto)
```



(b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the name variable, `cor()` which is qualitative.

```
round(cor(Auto[,1:7]),3)
```

```
##                 mpg cylinders displacement horsepower weight acceleration
## mpg           1.000    -0.778       -0.805     -0.778 -0.832        0.423
## cylinders    -0.778     1.000        0.951      0.843  0.898       -0.505
## displacement -0.805     0.951        1.000      0.897  0.933       -0.544
## horsepower   -0.778     0.843        0.897      1.000  0.865       -0.689
## weight       -0.832     0.898        0.933      0.865  1.000       -0.417
## acceleration  0.423    -0.505       -0.544     -0.689 -0.417        1.000
## year          0.581    -0.346       -0.370     -0.416 -0.309        0.290
##                year
## mpg           0.581
## cylinders    -0.346
## displacement -0.370
## horsepower   -0.416
## weight       -0.309
## acceleration  0.290
## year          1.000
```

(c) Use the `lm()` function to perform a multiple linear regression with *mpg* as the response and all other variables except name as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

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?
Answer

```
auto.new = Auto[,-9]
auto.new$origin = as.factor(auto.new$origin)
auto.new$cylinders = as.factor(auto.new$cylinders)
auto.lm = lm(mpg~.,data = auto.new)
summary(auto.lm)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = auto.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6797 -1.9373 -0.0678  1.6711 12.7756
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.208e+01  4.541e+00  -4.862 1.70e-06 ***
## cylinders4    6.722e+00  1.654e+00   4.064 5.85e-05 ***
## cylinders5    7.078e+00  2.516e+00   2.813  0.00516 **
## cylinders6    3.351e+00  1.824e+00   1.837  0.06701 .
## cylinders8    5.099e+00  2.109e+00   2.418  0.01607 *
## displacement  1.870e-02  7.222e-03   2.590  0.00997 **
## horsepower   -3.490e-02  1.323e-02  -2.639  0.00866 **
## weight       -5.780e-03  6.315e-04  -9.154  < 2e-16 ***
## acceleration  2.598e-02  9.304e-02   0.279  0.78021
## year          7.370e-01  4.892e-02  15.064  < 2e-16 ***
## origin2       1.764e+00  5.513e-01   3.200  0.00149 **
## origin3       2.617e+00  5.272e-01   4.964 1.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 380 degrees of freedom
## Multiple R-squared:  0.8469, Adjusted R-squared:  0.8425
## F-statistic: 191.1 on 11 and 380 DF,  p-value: < 2.2e-16
```

Comments:
i. Test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_6 = 0$ against H_a : at least one of the β_j is not zero. $p - value \approx 0$. Thus there is at least one predictor associated with *mpg*.
ii. For testing each one predictor separately, $H_0 : \beta_j = 0$ it appears that only *acceleration* does not have a statistically significant to *mpg*.
iii. The coefficient for the *year* is 0.073 so for each additional year, the mpg is predicted on average to increase by 0.073 keeping all of the other variables constant.

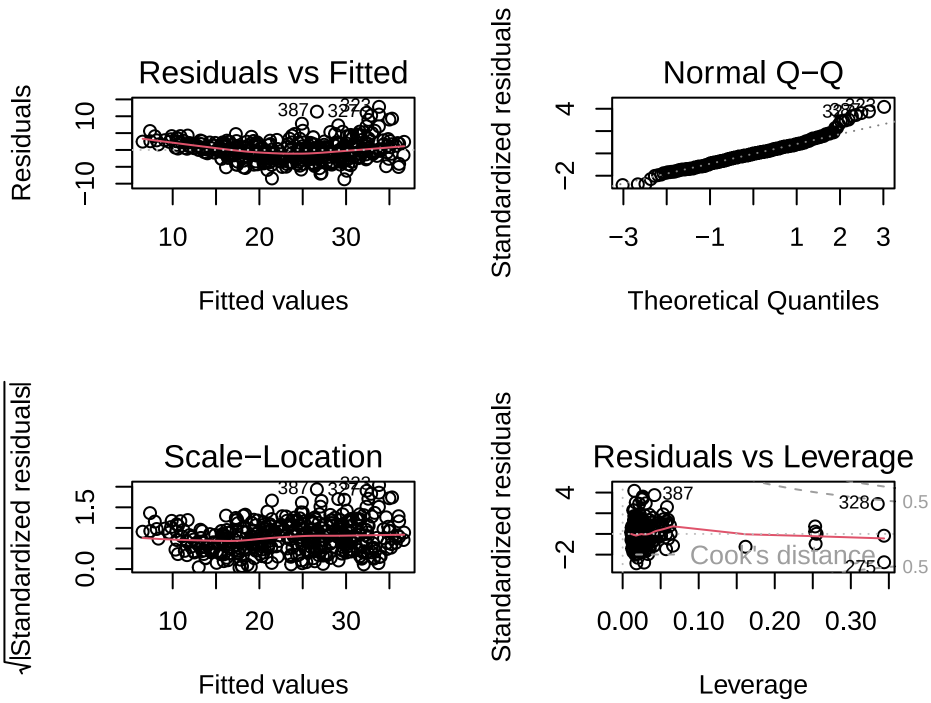
(d) Use the `plot()` function to produce diagnostic plots of the linear regression fit based on the predictors that appear to have a statistically significant relationship to the response. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

Answer
Take out acceleration:

```
auto.new2 = auto.new[,-6]
auto.lm2 = lm(mpg~., data = auto.new2)
summary(auto.lm2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = auto.new2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7037 -1.9501 -0.0552  1.7105 12.7932
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.162e+01  4.231e+00  -5.111 5.09e-07 ***
## cylinders4    6.784e+00  1.637e+00   4.144 4.20e-05 ***
## cylinders5    7.147e+00  2.501e+00   2.857 0.004510 **
## cylinders6    3.403e+00  1.813e+00   1.877 0.061262 .
## cylinders8    5.137e+00  2.102e+00   2.444 0.014983 *
## displacement  1.848e-02  7.169e-03   2.578 0.010312 *
## horsepower   -3.706e-02  1.071e-02  -3.459 0.000604 ***
## weight       -5.696e-03  5.535e-04 -10.291  < 2e-16 ***
## year          7.358e-01  4.868e-02  15.114  < 2e-16 ***
## origin2       1.763e+00  5.506e-01   3.203 0.001476 **
## origin3       2.621e+00  5.264e-01   4.979 9.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.094 on 381 degrees of freedom
## Multiple R-squared:  0.8469, Adjusted R-squared:  0.8429
## F-statistic: 210.7 on 10 and 381 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(auto.lm2)
```



These plots show some outliers observation numbers: 387,323, 327
High leverage: 387, 328, 275
It appears that the linearity fit is good.

(e) Use the * and/or : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

Answer

```
auto.int = lm(mpg ~ cylinders + displacement*horsepower + horsepower*weight + year + origin, data = aut
summary(auto.int)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement * horsepower + horsepower *
##     weight + year + origin, data = auto.new2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7565 -1.4899 -0.0843  1.4168 12.0178
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -7.583e+00  4.316e+00  -1.757 0.079734 .
## cylinders4              5.856e+00  1.516e+00   3.863 0.000132 ***
## cylinders5              7.464e+00  2.297e+00   3.250 0.001259 **
## cylinders6              5.197e+00  1.728e+00   3.008 0.002803 **
## cylinders8              6.455e+00  2.042e+00   3.161 0.001700 **
## displacement           -2.243e-02  1.660e-02  -1.351 0.177530
## horsepower             -1.842e-01  2.162e-02  -8.521 3.79e-16 ***
## weight                 -7.717e-03  1.513e-03  -5.099 5.41e-07 ***
```

```
## year                      7.523e-01  4.523e-02  16.635  < 2e-16 ***
## origin2                   1.056e+00  5.251e-01   2.011 0.045084 *
## origin3                   1.695e+00  4.971e-01   3.411 0.000718 ***
## displacement:horsepower  1.968e-04  9.529e-05   2.066 0.039544 *
## horsepower:weight         2.768e-05  1.047e-05   2.644 0.008533 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.84 on 379 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8676
## F-statistic: 214.4 on 12 and 379 DF,  p-value: < 2.2e-16
```

It appears that there might be interaction effects with horsepower and displacement also horsepower and weight. However, when we add these interaction terms, the displacement is no longer significant.

```
auto.lm3 = lm(mpg ~ cylinders + displacement +  sqrt(horsepower) + weight + origin, data = auto.new2)
summary(auto.lm3)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + sqrt(horsepower) +
##     weight + origin, data = auto.new2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.994 -2.235 -0.542  1.758 15.765
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      44.0684287  3.1692690  13.905  < 2e-16 ***
## cylinders4        7.8227761  2.0337518   3.846  0.00014 ***
## cylinders5        9.9647779  3.0964663   3.218  0.00140 **
## cylinders6        4.0868709  2.2409849   1.824  0.06898 .
## cylinders8        6.2616424  2.6039750   2.405  0.01666 *
## displacement      0.0063803  0.0085501   0.746  0.45599
## sqrt(horsepower) -1.7726717  0.2759663  -6.424 3.96e-10 ***
## weight           -0.0037309  0.0006861  -5.438 9.65e-08 ***
## origin2           0.0051860  0.6652473   0.008  0.99378
## origin3           2.6162364  0.6490513   4.031 6.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 382 degrees of freedom
## Multiple R-squared:  0.7629, Adjusted R-squared:  0.7573
## F-statistic: 136.6 on 9 and 382 DF,  p-value: < 2.2e-16
```

When I transform some of the variables, the R^2 actually gets lower. This percent of variation in *mpg* that can be explained is lower with these transformations. So it might not be best to use them. Just the original model without *acceleration*.

(f) Try a few different transformations of the variables, such as *log(X)*, $\sqrt{X}$, X^2. Comment on your findings.

Problem 7

This problem involves the `Boston` data set, from the `ISLR2` package. 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.

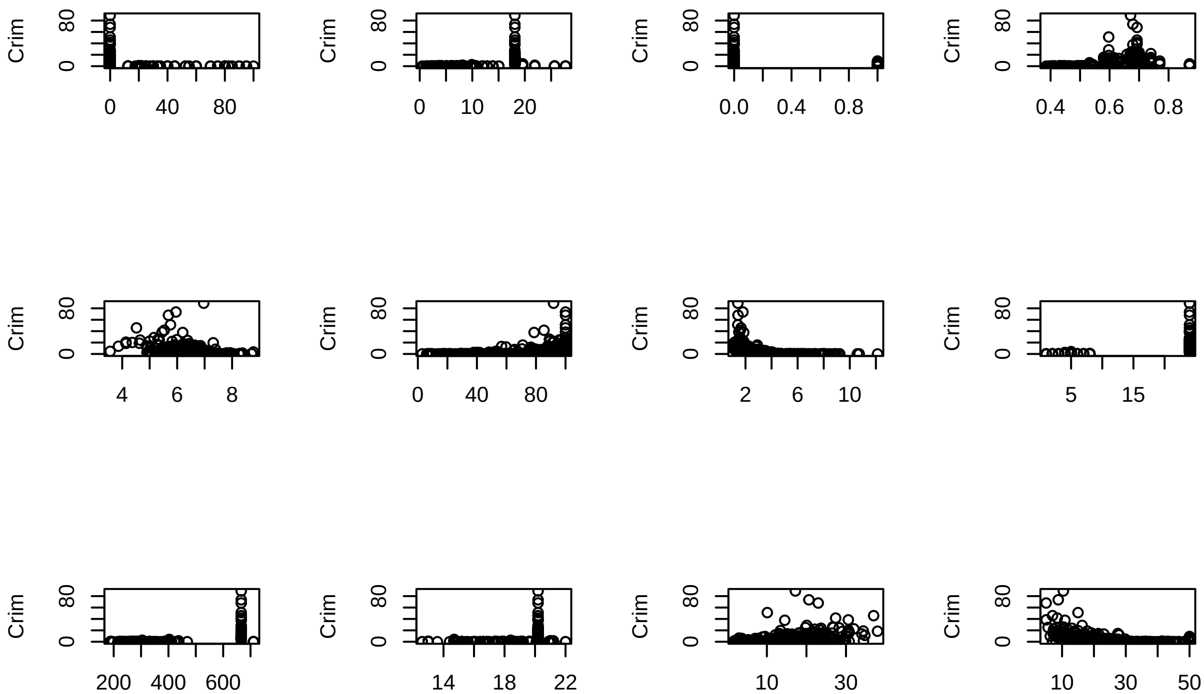
Answer

```
library(ISLR2)
b.boston = NA
f.boston = NA
p.boston = NA
for (i in 1:ncol(Boston)-1) {
  lm.fit = lm(Boston$crim~Boston[,i+1])
  b.boston[i] = lm.fit$coef[2]
  f.boston = summary(lm.fit)$fstatistic
  p.boston[i] = pf(f.boston[1],f.boston[2],f.boston[3],lower.tail = F)
}
cbind(colnames(Boston[,-1]),b.boston,p.boston)
```

```
##                 b.boston             p.boston
##  [1,] "zn"      "-0.073934977404123" "5.50647210767939e-06"
##  [2,] "indus"   "0.509776331104228"  "1.45034893302726e-21"
##  [3,] "chas"    "-1.89277655080378"  "0.209434501535199"
##  [4,] "nox"     "31.2485312011229"   "3.75173926035698e-23"
##  [5,] "rm"      "-2.68405122411395"  "6.34670298468782e-07"
##  [6,] "age"     "0.107786227139533"  "2.85486935024409e-16"
##  [7,] "dis"     "-1.5509016824101"   "8.51994876692653e-19"
##  [8,] "rad"     "0.617910927327201"  "2.69384439818606e-56"
##  [9,] "tax"     "0.0297422528227653" "2.35712683525675e-47"
## [10,] "ptratio" "1.15198278707059"   "2.94292244735986e-11"
## [11,] "lstat"   "0.548804782062398"  "2.65427723147327e-27"
## [12,] "medv"    "-0.363159922257603" "1.17398708219434e-19"
```

The only one that does not seem significant for crime per capita is if the suburb bounds the Charles River or not. All of the others seem significant.

```
par(mfrow = c(3,4))
for (i in 1:12) {
  plot(Boston[,i+1],Boston$crim,xlab = "", ylab = "Crim")
}
```



(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 $H_0 : \beta_j = 0$?

```
lm.fit = lm(crim ~ ., data = Boston)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.534 -2.248 -0.348  1.087 73.923
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.7783938  7.0818258   1.946 0.052271 .
## zn           0.0457100  0.0187903   2.433 0.015344 *
## indus       -0.0583501  0.0836351  -0.698 0.485709
## chas        -0.8253776  1.1833963  -0.697 0.485841
## nox         -9.9575865  5.2898242  -1.882 0.060370 .
## rm           0.6289107  0.6070924   1.036 0.300738
## age         -0.0008483  0.0179482  -0.047 0.962323
## dis         -1.0122467  0.2824676  -3.584 0.000373 ***
## rad          0.6124653  0.0875358   6.997 8.59e-12 ***
## tax         -0.0037756  0.0051723  -0.730 0.465757
## ptratio     -0.3040728  0.1863598  -1.632 0.103393
```

```
## lstat        0.1388006  0.0757213   1.833 0.067398 .
## medv        -0.2200564  0.0598240  -3.678 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.46 on 493 degrees of freedom
## Multiple R-squared:  0.4493, Adjusted R-squared:  0.4359
## F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

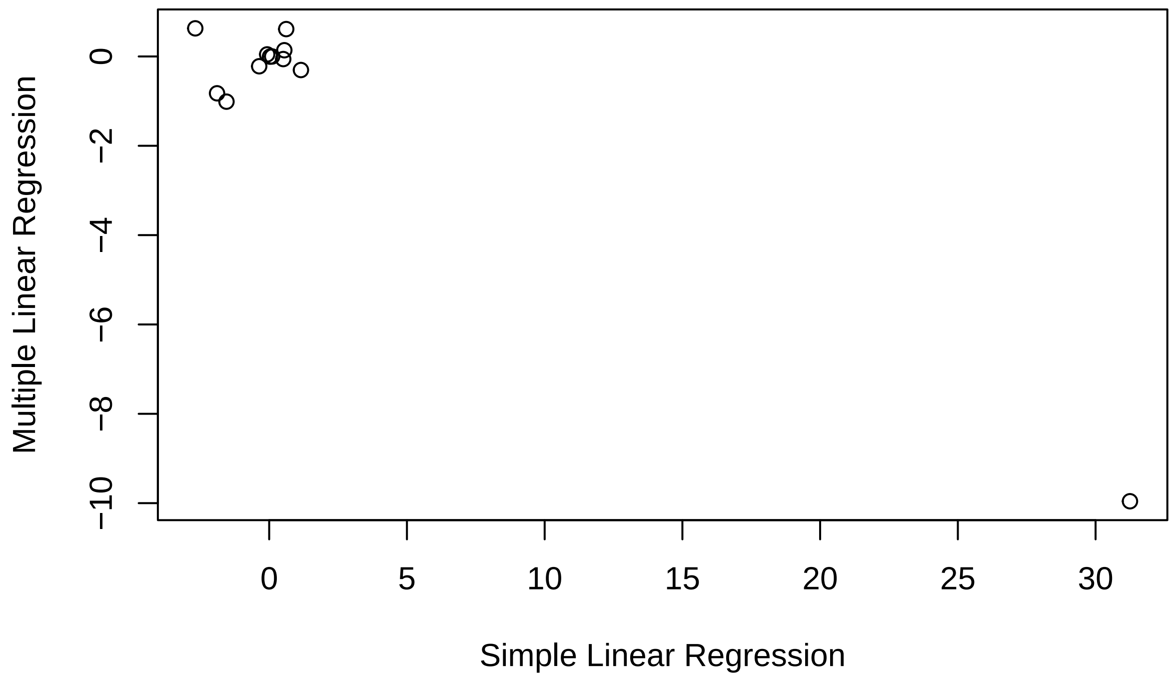
```
summary(lm(crim ~ zn + nox + dis + rad + ptratio + lstat + medv, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ zn + nox + dis + rad + ptratio + lstat +
##     medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.655 -2.143 -0.319  1.050 74.740
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.46682    6.02424   2.899 0.003904 **
## zn            0.04497    0.01803   2.494 0.012951 *
## nox         -12.45782    4.77637  -2.608 0.009375 **
## dis          -0.94255    0.26270  -3.588 0.000366 ***
## rad           0.56152    0.04813  11.667  < 2e-16 ***
## ptratio      -0.34703    0.18288  -1.898 0.058322 .
## lstat         0.11479    0.06945   1.653 0.098997 .
## medv         -0.19026    0.05369  -3.543 0.000432 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.452 on 498 degrees of freedom
## Multiple R-squared:  0.4452, Adjusted R-squared:  0.4374
## F-statistic: 57.08 on 7 and 498 DF,  p-value: < 2.2e-16
```

It appears that only `zn`, `nox`, `dis`, `rad`, `ptratio`, `lstat`, and `medv` are significant in predicting `crim`.

(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.

```
plot(b.boston,lm.fit$coefficients[-1],xlab = "Simple Linear Regression", ylab = "Multiple Linear Regres
```



(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

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

I will only do the predictors that are significant in the full model, `zn`, `nox`, `dis`, `rad`, `ptratio`, `lstat`, and `medv` are significant in predicting `crim`.

```
library(stargazer)
lm.zn = lm(crim ~ poly(zn,3),data = Boston)
lm.nox = lm(crim ~ poly(nox,3),data = Boston)
lm.dis = lm(crim ~ poly(dis,3),data = Boston)
lm.rad = lm(crim ~ poly(rad,3),data = Boston)
lm.ptratio = lm(crim ~ poly(ptratio,3), data = Boston)
lm.lstat = lm(crim ~ poly(lstat,3),data = Boston)
lm.medv = lm(crim ~ poly(medv,3), data = Boston)
stargazer(lm.zn,lm.nox,lm.dis,lm.rad,lm.ptratio,lm.lstat,lm.medv,
          title = "Polynomial Models",
          font.size = "scriptsize")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Sep 19, 2022 - 10:24:46 AM

It appears that only the `medv` predictor might be better for a non-linear association. The others have a very small R^2.

Table 1: Polynomial Models

	Dependent variable:						
	crim						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
poly(zn, 3)1	-38.750^{***}						
	(8.372)						
poly(zn, 3)2	23.940^{***}						
	(8.372)						
poly(zn, 3)3	-10.072						
	(8.372)						
poly(nox, 3)1		81.372^{***}					
		(7.234)					
poly(nox, 3)2		-28.829^{***}					
		(7.234)					
poly(nox, 3)3		-60.362^{***}					
		(7.234)					
poly(dis, 3)1			-73.389^{***}				
			(7.331)				
poly(dis, 3)2			56.373^{***}				
			(7.331)				
poly(dis, 3)3			-42.622^{***}				
			(7.331)				
poly(rad, 3)1				120.907^{***}			
				(6.682)			
poly(rad, 3)2				17.492^{***}			
				(6.682)			
poly(rad, 3)3				4.698			
				(6.682)			
poly(ptratio, 3)1					56.045^{***}		
					(8.122)		
poly(ptratio, 3)2					24.775^{***}		
					(8.122)		
poly(ptratio, 3)3					-22.280^{***}		
					(8.122)		
poly(lstat, 3)1						88.070^{***}	
						(7.629)	
poly(lstat, 3)2						15.888^{**}	
						(7.629)	
poly(lstat, 3)3						-11.574	
						(7.629)	
poly(medv, 3)1							-75.058^{***}
							(6.569)
poly(medv, 3)2							88.086^{***}
							(6.569)
poly(medv, 3)3							-48.033^{***}
							(6.569)
Constant	3.614^{***}	3.614^{***}	3.614^{***}	3.614^{***}	3.614^{***}	3.614^{***}	3.614^{***}
	(0.372)	(0.322)	(0.326)	(0.297)	(0.361)	(0.339)	(0.292)
Observations	506	506	506	506	506	506	506
R^2	0.058	0.297	0.278	0.400	0.114	0.218	0.420
Adjusted R^2	0.053	0.293	0.274	0.396	0.108	0.213	0.417
Residual Std. Error (df = 502)	8.372	7.234	7.331	6.682	8.122	7.629	6.569
F Statistic (df = 3; 502)	10.349^{***}	70.687^{***}	64.374^{***}	111.573^{***}	21.484^{***}	46.629^{***}	121.272^{***}
Note:							$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Problem 8

This problem focuses on the **collinearity** problem.

(a) Perform the following commands in R:

```
set.seed (1)
x1=runif (100)
x2 =0.5* x1+rnorm (100) /10
y=2+2* x1 +0.3* x2+rnorm (100)
```

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

Answer
The linear model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
The regression coefficients are: $\beta_0 = 2$, $\beta_1 = 2$ and $\beta_2 = 0.3$.

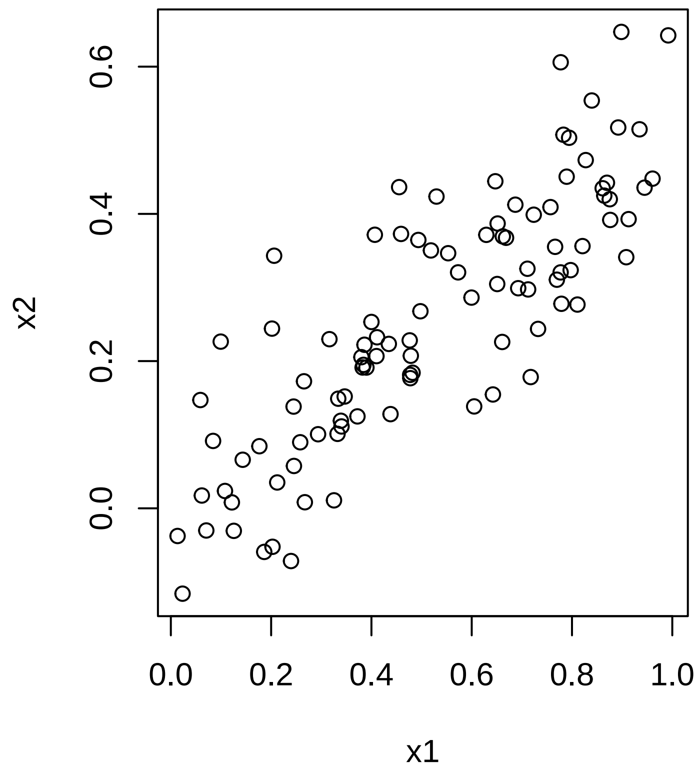
(b) What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

Answer

```
cor(x1,x2)
```

```
## [1] 0.8351212
```

```
plot(x1,x2)
```



(c) Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

Answer

```
summary(lm(y ~ x1 + x2))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

For testing $H_0 : \beta_1 = \beta_2 = 0$ against H_a : at least one β_j is not zero. We get a p-value close to zero. So at least one of the variables $x1$, $x2$ is related to y.

$\hat{\beta}_0 = 2.1305$
$\hat{\beta}_1 = 1.4396$
$\hat{\beta}_2 = 1.0097$

From the actual values of β_0, β_1, and β_2. This estimate is close for β_0 and somewhat to β_1 but not for β_2.

For testing $H_0 : \beta_1 = 0$ we reject that hypothesis with a p-value $= 0.0487$.
For testing $H_0 : \beta_2 = 0$ we fail to reject the null hypothesis with a p-value $= 0.3754$.

(d) Now fit a least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

Answer

```r
summary(lm(y~x1))
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

$\hat{\beta}_0$ and $\hat{\beta}_1$ are close to the original coefficients.

If we test $H_0 : \beta_1 = 0$ we would reject the null hypothesis.

(e) Now fit a least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

Answer

```r
summary(lm(y~x2))
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

This shows that x2 is associated with y by rejecting $H_0 : \beta_2 = 0$.

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

Answer

What (c) says is that if $x1$ is in the model to predict y, then we do not need $x2$. Which is true because $x2$ was calculated based on $x1$. So it does not really contradict each other.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1=c(x1 , 0.1)
x2=c(x2 , 0.8)
y=c(y,6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

Answer

```
summary(lm(y ~ x1 + x2))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```
summary(lm(y ~ x1))
```

```
##
## Call:
```

```
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

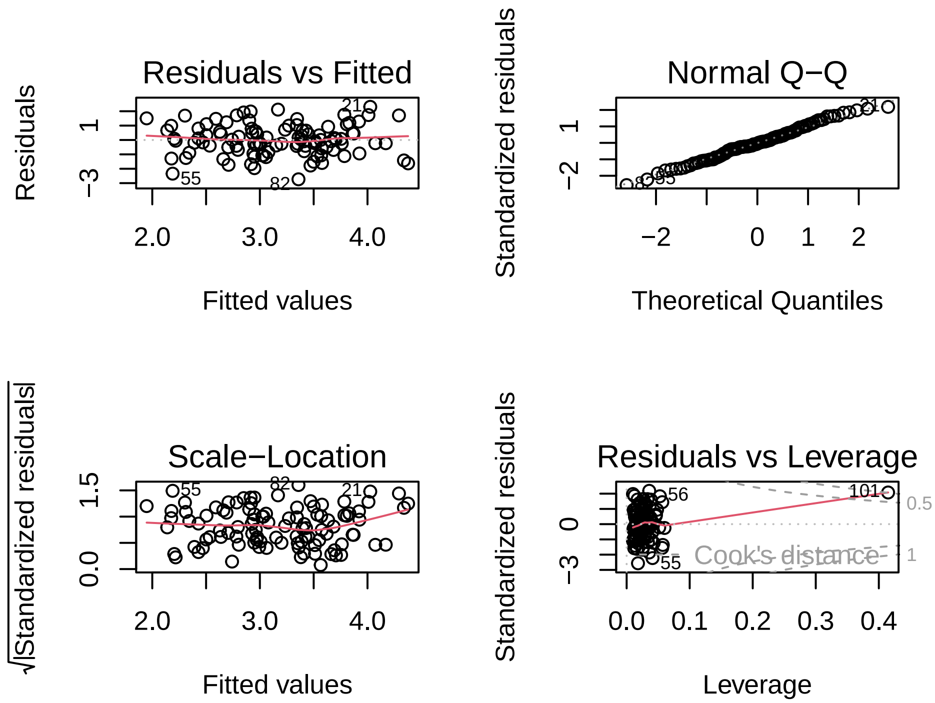
```
summary(lm(y ~ x2))
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

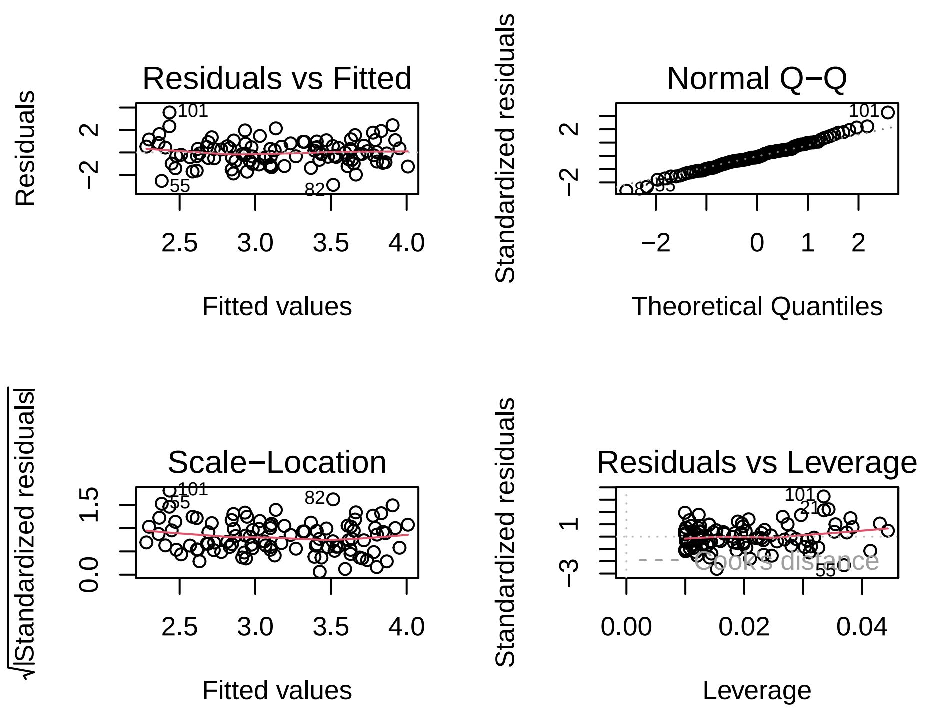
This does change the estimates of β_1 and β_2.

Plots

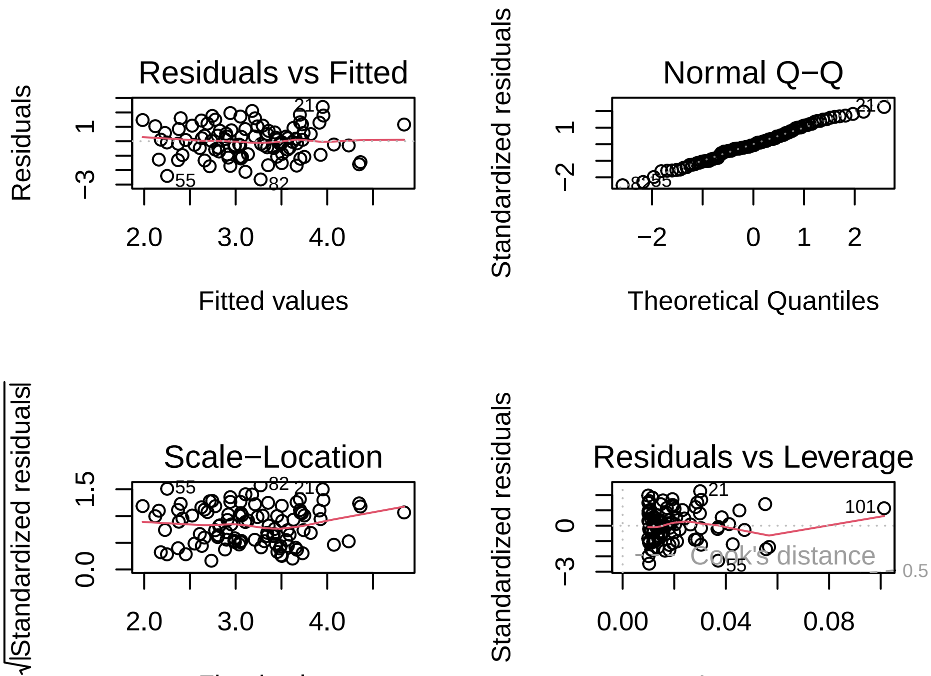
```
par(mfrow = c(2,2))
plot(lm(y ~ x1 + x2))
```




```
plot(lm(y ~ x1))
```




```
plot(lm(y ~ x2))
```



This extra point has high leverage.

Problem 9

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

(b) Generate a response vector Y of length n = 100 according to the model

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

where β_0, β_1, β_2, and β_3 are constants of your choice.

(c) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both X and Y.

(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

Answer

(a) Generating X and ϵ.

```
set.seed(1)
X = rnorm(100)
e = rnorm(100)
```

(b) Generate Y. Let $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_2 = -0.75$ and $\beta_3 = 5$.

```r
Y = 2 + 0.5*X -0.75*X^2 + 5*X^3 + e
```

(c) Use regsubsets

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.2.1
```

```r
new.data = data.frame(cbind(Y,X))
fit.y = regsubsets(Y ~ poly(X,10),data = new.data)
(fit.res = summary(fit.y))
```

```
## Subset selection object
## Call: regsubsets.formula(Y ~ poly(X, 10), data = new.data)
## 10 Variables  (and intercept)
##                Forced in Forced out
## poly(X, 10)1      FALSE      FALSE
## poly(X, 10)2      FALSE      FALSE
## poly(X, 10)3      FALSE      FALSE
## poly(X, 10)4      FALSE      FALSE
## poly(X, 10)5      FALSE      FALSE
## poly(X, 10)6      FALSE      FALSE
## poly(X, 10)7      FALSE      FALSE
## poly(X, 10)8      FALSE      FALSE
## poly(X, 10)9      FALSE      FALSE
## poly(X, 10)10     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)4 poly(X, 10)5
## 1  ( 1 ) "*"          " "          " "          " "          " "
## 2  ( 1 ) "*"          " "          "*"          " "          " "
## 3  ( 1 ) "*"          "*"          "*"          " "          " "
## 4  ( 1 ) "*"          "*"          "*"          " "          "*"
## 5  ( 1 ) "*"          "*"          "*"          "*"          "*"
## 6  ( 1 ) "*"          "*"          "*"          "*"          "*"
## 7  ( 1 ) "*"          "*"          "*"          "*"          "*"
## 8  ( 1 ) "*"          "*"          "*"          "*"          "*"
##          poly(X, 10)6 poly(X, 10)7 poly(X, 10)8 poly(X, 10)9 poly(X, 10)10
## 1  ( 1 ) " "          " "          " "          " "          " "
## 2  ( 1 ) " "          " "          " "          " "          " "
## 3  ( 1 ) " "          " "          " "          " "          " "
## 4  ( 1 ) " "          " "          " "          " "          " "
## 5  ( 1 ) " "          " "          " "          " "          " "
## 6  ( 1 ) " "          " "          " "          " "          "*"
## 7  ( 1 ) " "          "*"          " "          " "          "*"
## 8  ( 1 ) " "          "*"          " "          "*"          "*"
```

```r
fit.stat = cbind(fit.res$adjr2,fit.res$cp,fit.res$bic)
colnames(fit.stat) = c("Adjr2","Cp","BIC")
print(fit.stat)
```

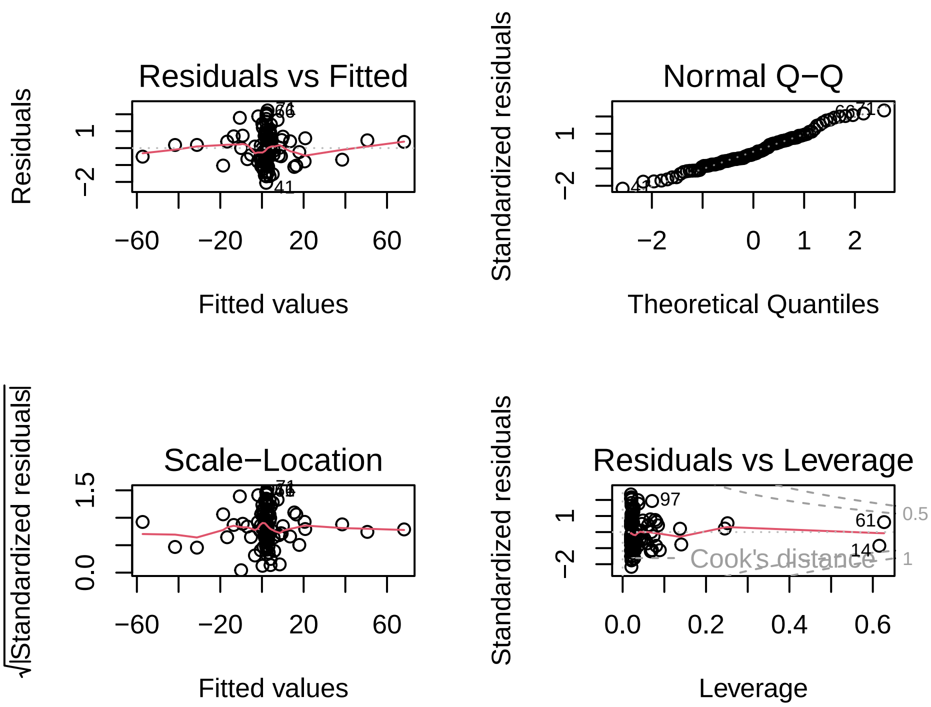
```
##          Adjr2          Cp       BIC
## [1,] 0.6795785 6009.726765 -105.6167
## [2,] 0.9931301   35.572292 -486.2859
## [3,] 0.9949543    2.185943 -513.5775
## [4,] 0.9950267    1.866261 -511.4660
```

```
## [5,] 0.9950654    2.193128 -508.6989
## [6,] 0.9950653    3.235128 -505.1616
## [7,] 0.9950181    5.119994 -500.6855
## [8,] 0.9949686    7.027330 -496.1844
```

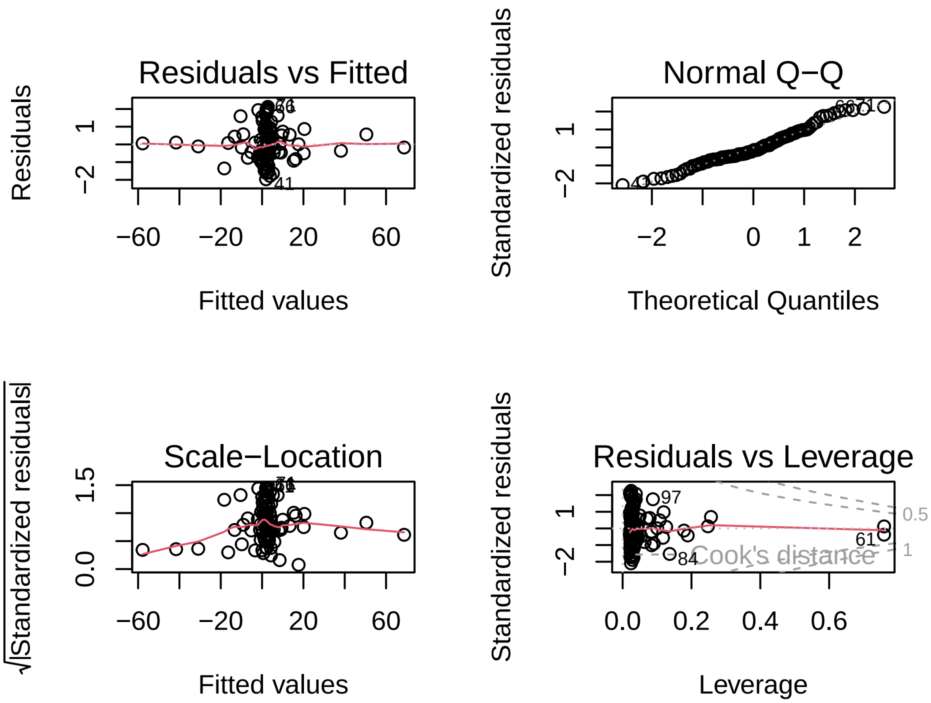
The model with the 4th degree appears to be the best subset.

Plots:

```
par(mfrow = c(2,2))
plot(lm(Y ~ poly(X,4)))
```




```
plot(lm(Y ~ poly(X,5)))
```



(d) Using stepwise selections

```
step(lm(Y ~ poly(X,10)), direction = "backward")
```

```
## Start:  AIC=4.64
## Y ~ poly(X, 10)
##
##                  Df Sum of Sq     RSS    AIC
## <none>                           84.1   4.64
## - poly(X, 10)  10     18098 18181.5 522.30
##
## Call:
## lm(formula = Y ~ poly(X, 10))
##
## Coefficients:
##   (Intercept)   poly(X, 10)1   poly(X, 10)2   poly(X, 10)3   poly(X, 10)4
##        2.4619       111.4209         5.7812        75.1300         1.2571
##  poly(X, 10)5   poly(X, 10)6   poly(X, 10)7   poly(X, 10)8   poly(X, 10)9
##        1.4802         0.1190        -0.3298        -0.1079        -0.2958
## poly(X, 10)10
##       -0.9512
```

```
step(lm(Y ~ poly(X,10)), direction = "forward")
```

```
## Start:  AIC=4.64
## Y ~ poly(X, 10)
##
## Call:
## lm(formula = Y ~ poly(X, 10))
##
```

```
## Coefficients:
##   (Intercept)   poly(X, 10)1   poly(X, 10)2   poly(X, 10)3   poly(X, 10)4
##        2.4619       111.4209         5.7812        75.1300         1.2571
##  poly(X, 10)5   poly(X, 10)6   poly(X, 10)7   poly(X, 10)8   poly(X, 10)9
##        1.4802         0.1190        -0.3298        -0.1079        -0.2958
## poly(X, 10)10
##       -0.9512
```

This shows that all of the terms is used in the regression