MATH 4322 Homework 2 Solutions

Instructor: Dr. Cathy Poliak

9/24/2021

Problem 1

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.

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

(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.

Model:

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

(a) Thus (iii) is correct if the GPA is high.

(b) Predicted salary: $85 + 10 \times 4 + 0.07 \times 110 + 0.01 \times 4 \times 110 = 137.1$

(c) This is **False**. The coefficients do not determine if there is an association between the predictor and response. You need to do a hypothesis test $H_0 : \beta_4 = 0$ if the p-value is *large* (greater than 0.05 usually) then we say we do not need that interaction term.

Problem 2

We perform stepwise, 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. True or False:

(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.

(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.

(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.

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

Answer

(a) True
(b) True
(c) False
(d) False
(e) False

Problem 3

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

(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.

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

(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

(a) Result from R:

```
library(ISLR)
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
```

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

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

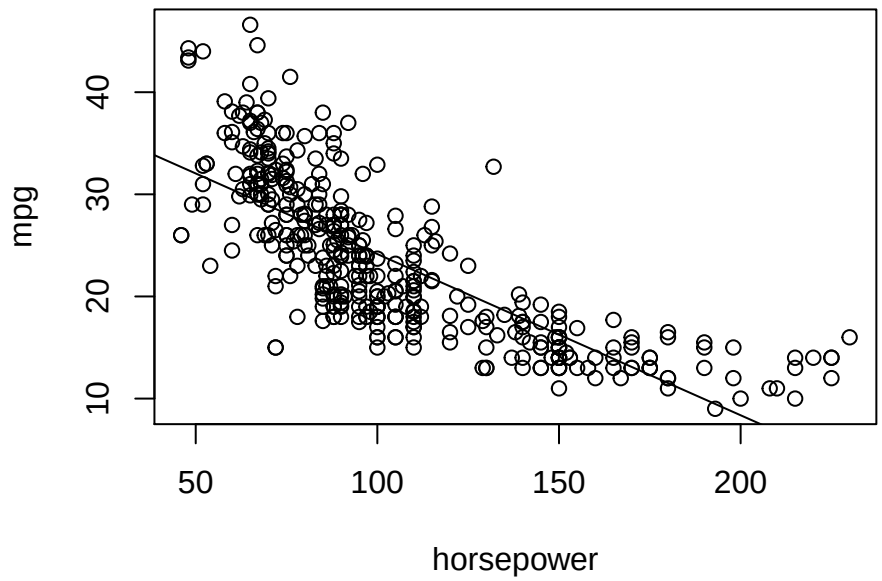
```
predict(auto.lm,newdata = data.frame(horsepower = 98),interval = "p")
```

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

i. By testing $H_0 : \beta_1 = 0$ we get a $p - value \approx 0$. Thus there is very strong evidence of a relationship between horsepower and mpg.
ii. Also correlation is -0.7784 somewhat strong association.
iii. Since the coefficient and the correlation is negative, this says that as the horsepower increases, the mpg will decrease.
iv. The predicted mpg for horespower of 98 is 24.467 mpg.
The 95% confidence interval is, (23.973, 24.961), for an automobiles that have a horsepower of 98 we are 95% confident that the mean(average) mpg will be between 23.973 and 24.961 mpg.
The 95% prediction interval is, (14.8094, 34.12476), for one automobile that has a horsepower of 98 we are 95% confident that the mpg for that one automobile is between 14.8094 and 34.12476.

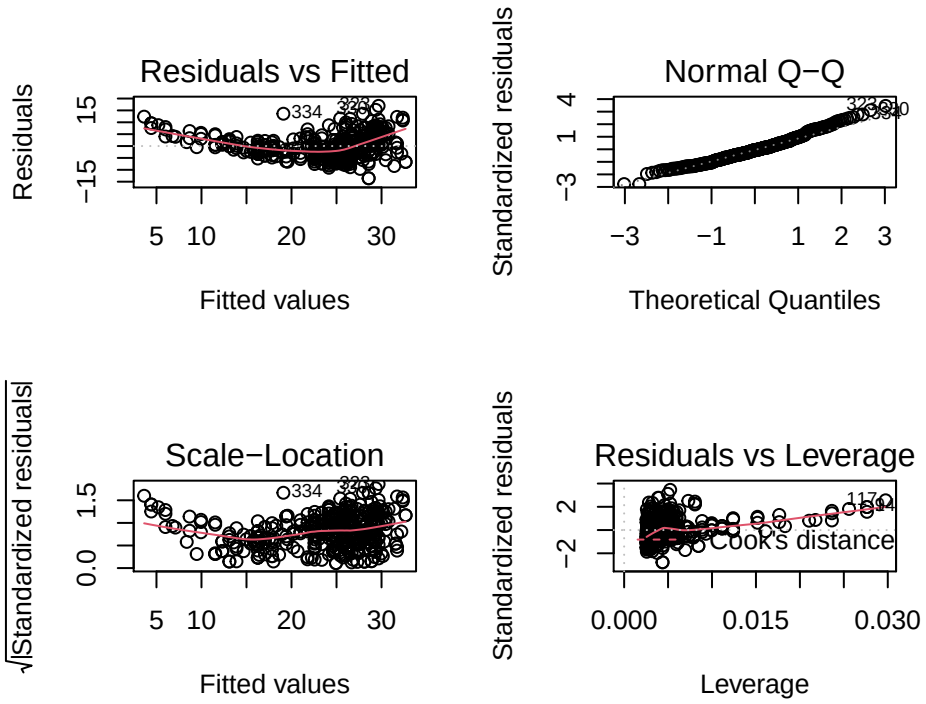
(b) Plot

```
plot(Auto$horsepower,Auto$mpg,xlab = "horsepower",ylab = "mpg")
abline(auto.lm)
```



(c) Diagnostic Plots

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



It appears that the relationship may not be linear when looking at the residual plot and the scatterplot.

Problem 4

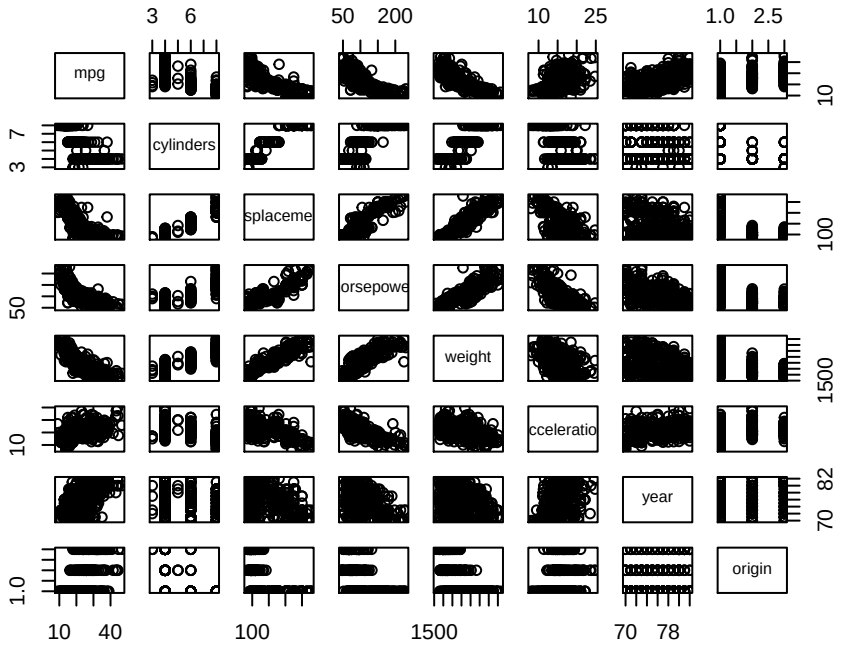
This question involves the use of multiple linear regression on the *Auto* data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.
(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.
(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?
(d) Use the `plot()` function to produce diagnostic plots of the linear regression fit based on the predictors that appear to have a statistically signifianct 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?
(e) Use the * and/or : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?
(f) Try a few different transformations of the variables, such as *log(X)*, $\sqrt{X}$, X^2. Comment on your findings.

Answer

(a) Plot of matrix:

```
pairs(Auto[,-9])
```



(b) Correlation matrix

```
round(cor(Auto[,-9]),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
## origin        0.565    -0.569       -0.615     -0.455 -0.585        0.213
##                year origin
## mpg           0.581  0.565
## cylinders    -0.346 -0.569
## displacement -0.370 -0.615
## horsepower   -0.416 -0.455
## weight       -0.309 -0.585
## acceleration  0.290  0.213
## year          1.000  0.182
## origin        0.182  1.000
```

(c) Regression:

```
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 signficant to *mpg*.
iii. The coefficient for the *year* is 0.073 so for each additional year, the mpg is predicted on average to increas by 0.073 keeping all of the other variables constant.

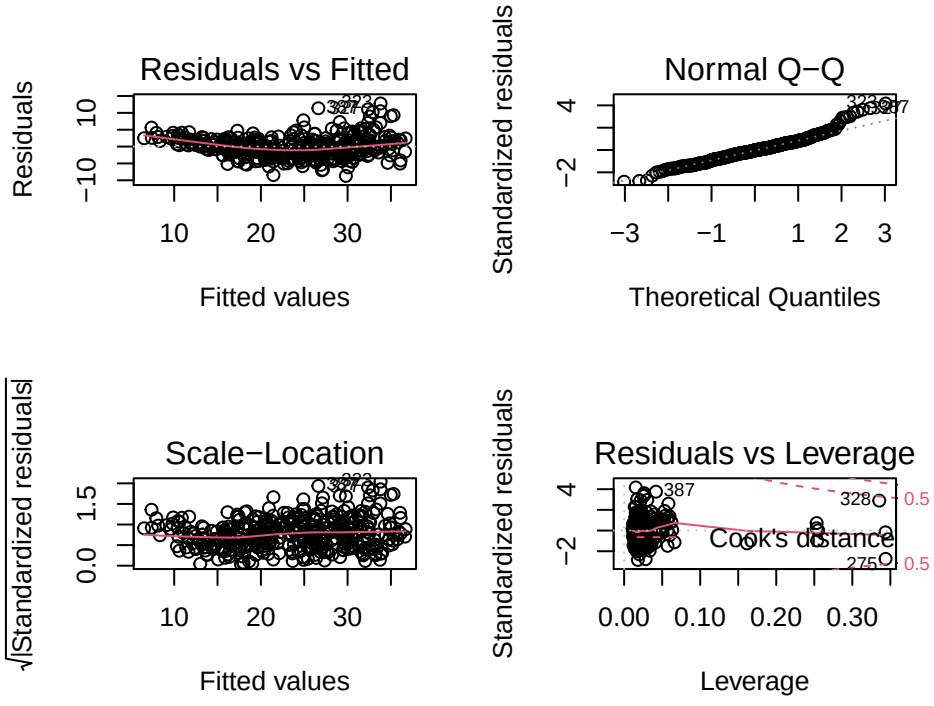
(d) Take out acceleration:

```
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) With interaction terms

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

(f) Trying some transformations

```
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*.

Problem 5

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 $\beta_2 = 0.3$.

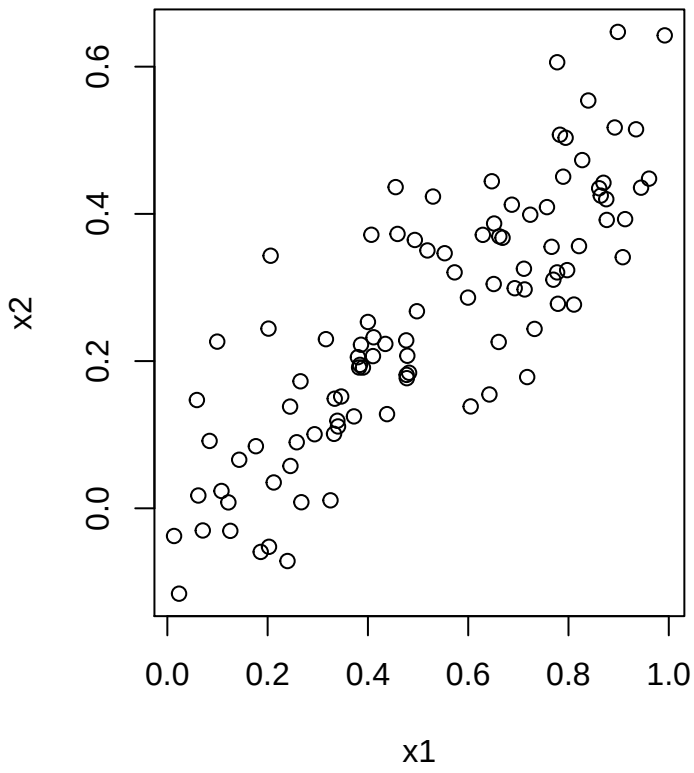
(b) What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

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

```
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

```
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
```

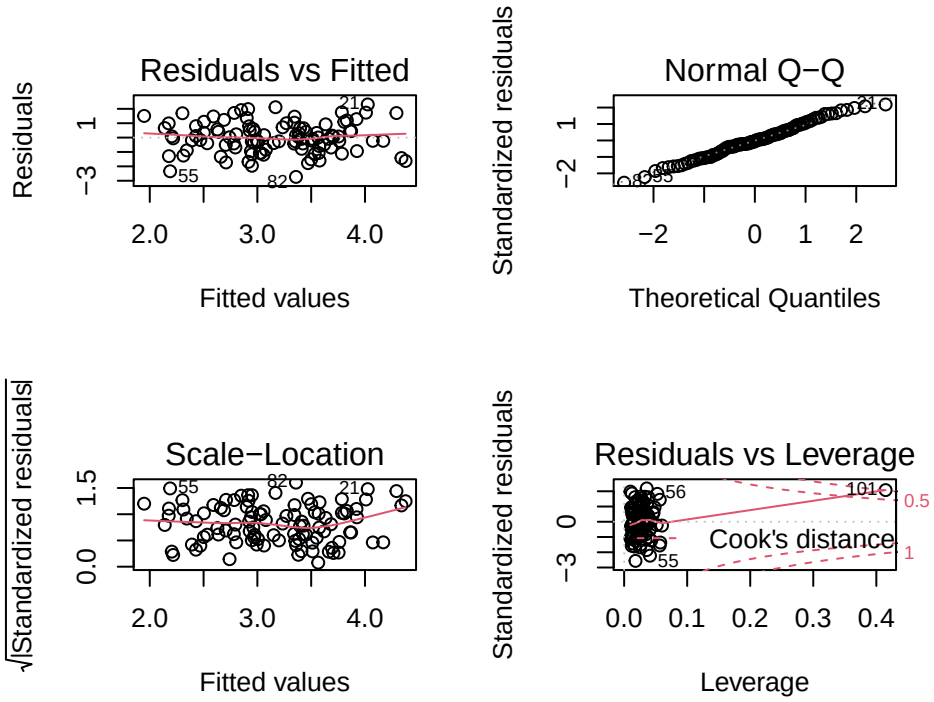
```
summary(lm(y ~ x2))
```

```
##
## 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
```

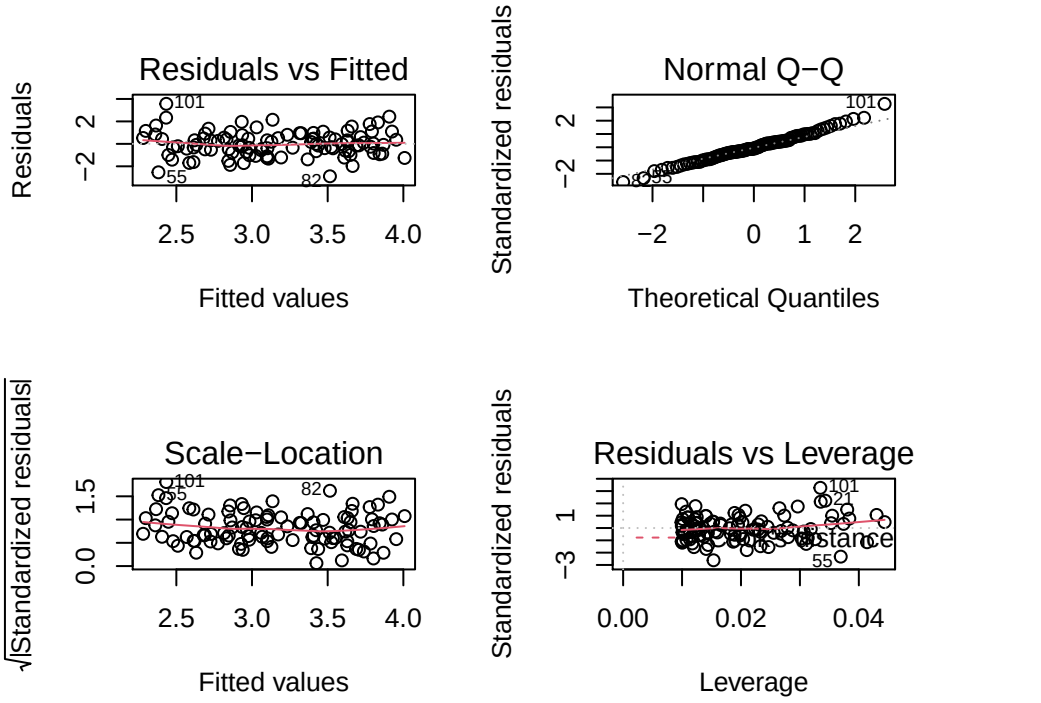
This does change the estimates of β_1 and β_2.

Plots

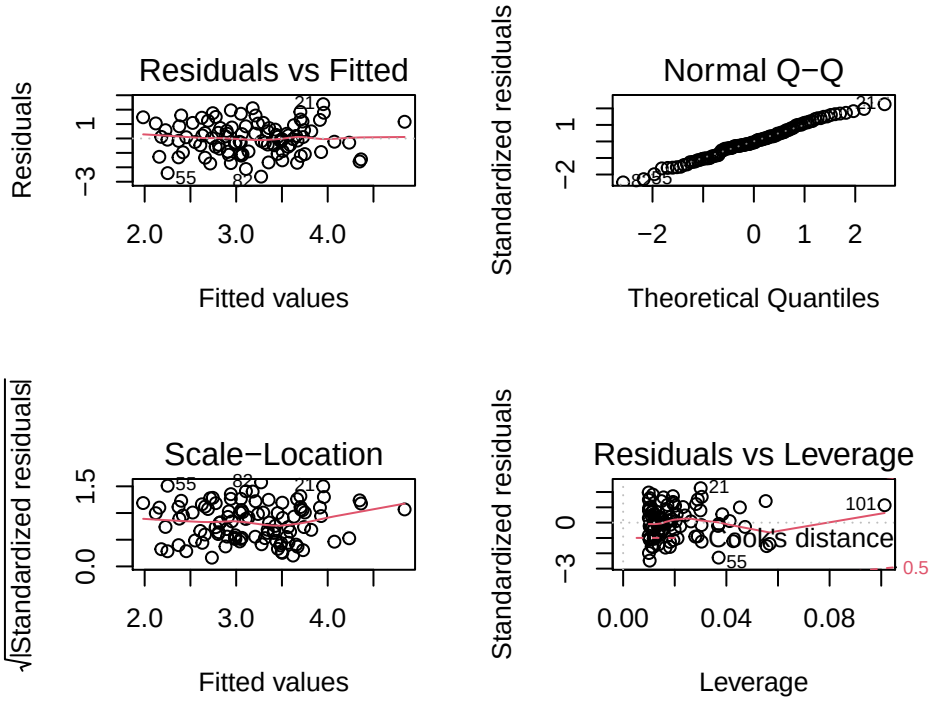
```
par(mfrow = c(2,2))
plot(lm(y ~ x1 + x2))
```




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




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



This extra point has high leverage.

Problem 6

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 $\beta_3 = 5$.

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

(c) Use regsubsets

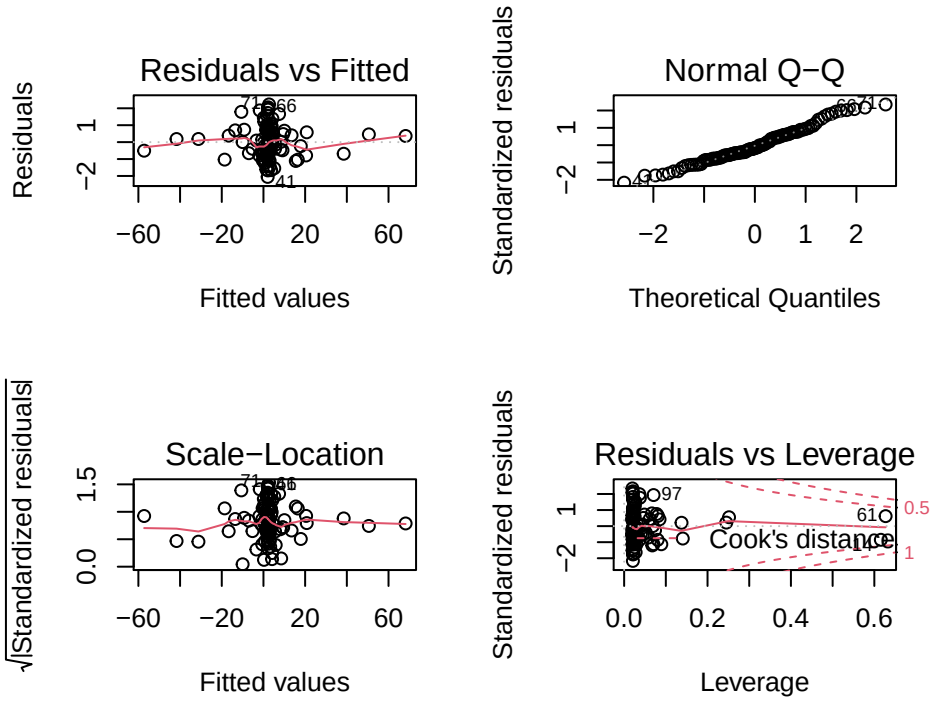
```
library(leaps)
new.data = data.frame(cbind(Y,X))
fit.y = regsubsets(Y ~ poly(X,10),data = new.data)
fit.res = summary(fit.y)
fit.stat = cbind(fit.res$adjr2,fit.res$cp,fit.res$bic)
colnames(fit.stat) = c("Adjr2","Cp","BIC")
print(fit.stat)
```

```
##          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
```

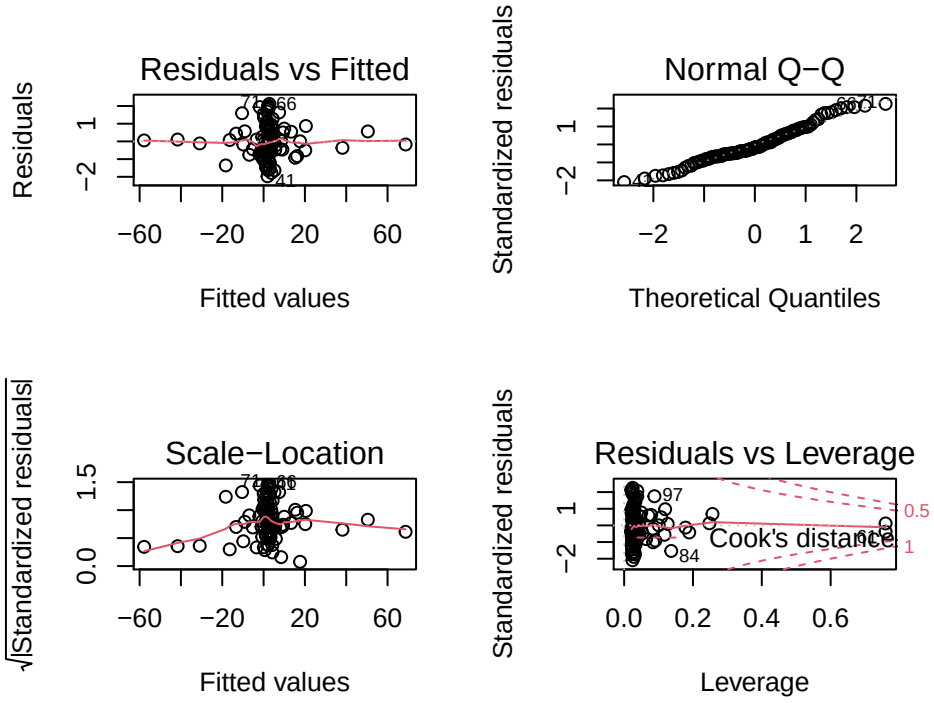
The model with the 4th degree appears to be the best subset.

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