Stat 435 HW1

Zach Fechko (011711215)

2022-10-23

General guidelines

This HW covers contents related to simple linear regression and multiple linear regression.

Please show your work in order to get points. Providing correct answers without supporting details does not receive full credits. The homework assignment also trains your programming skills. So, please complete both the conceptual exercises and applied exercises. It is absolutely not OK to just submit your codes only. This will result in a considerable loss of points on your assignments or projects.

For an assignment or project, you DO NOT have to submit your answers or reports using typesetting software. However, your answers must be well organized and well legible for grading. Specifically, if you are not able to knit a .Rmd/.rmd file into an output file such as a .pdf, .doc, .docx or .html file that contains your codes, outputs from your codes, your interpretations on the outputs, and your answers in text (possibly with math expressions), you can organize your codes, their outputs and your answers in a document. Regardless of if you are using typesetting software or not, please organize your work in the format given below:

```
Problem or task or question ...
Codes ...
Outputs ...
Your interpretations ...
```

Please upload your answers in a document to the course space where assignment or project is provided.

Conceptual exercises

1. Consider a simple linear regression.

- (a) Provide the formulae for the *least squares estimate (LSE)* of the coefficients, based on observations $y_i, i = 1, \ldots, n$ for the response Y and observations $x_i, i = 1, \ldots, n$ for the predictor X.
- (b) Are these estimated coefficients unbiased?

- (c) Provide the formulae for the standard errors of these estimated coefficients, and state the conditions under which the formulae are valid.
- (d) Explain how the sample variance of the observations $x_i, i = 1, \ldots, n$ for X affects the standard errors of the estimated intercept and slope.
- (e) Give a test on if the slope in the model is zero, by stating the null hypothesis, the statistic with its degrees of freedom, and the decision rule. Under what conditions will this test work well?
- (f) In the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$, ε is the random error term with unknown standard deviation σ. Provide an estimator for σ, and describe how the sample size n of the observations affects the accuracy of this estimator.

2. Consider a simple linear regression model. State the definition of the R^2 statistic by providing also the definitions of RSS and TSS, and explain its connection with sample correlation and the LSE of the slope.
3. Consider a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, β_0 is the intercept, and $\beta_1, \cdots, \beta_p$ are the regression coefficients. Suppose we want to test that a particular subset of q of the p regression coefficients are zero.

- (a) Provide a test statistic for this by describing in detail each quantity that makes up the statistic.
- (b) Under what conditions will the test statistic work well?

4. Question 3(a) of Section 3.7 (the section for exercises) of the Textbook.

Conceptual exercise answers

Question 1

(a)

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$

(b)

$\hat{\beta}_1$ is unbiased because $E(\hat{\beta}_1) = \frac{S_{xy}}{S_{xx}} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(X,Y)}{S_{xx}/(n-1)} = \frac{Cov(X,Y)}{S_{xx}/n} = \frac{Cov(X,Y)}{S_{xx}/n} = \frac{Cov(X,Y)}{Var(X)} = \beta_1$

$\hat{\beta}_0$ is unbiased because $E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - E(\hat{\beta}_1 \bar{x}) = \bar{y} - \beta_1 \bar{x} = \beta_0$

(c)

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u_i}^2}{[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}}$$

2

The R^2 is defined as the fraction of the total variation in Y_i that is explained by X_i. Because we can write: $Y_i = \hat{Y}_i + \hat{u}_i$. Which implies that the R^2 is the ratio of the sample variance of $\hat{Y}_i$ to the sample variance of Y_i.
$R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(\hat{u_i})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
The value of R^2 is always between 0 and 1. The closer R^2 is to 1, the better the model fits the data.

3

(a)

In order to find a test statistic for multivariate linear regression, we need to use the F-statistic, which is:

4

Using the equation

$$\text{Salary} = \beta_0 + \beta_1 \text{GPA} + \beta_2 \text{IQ} + \beta_3 \text{Level} + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

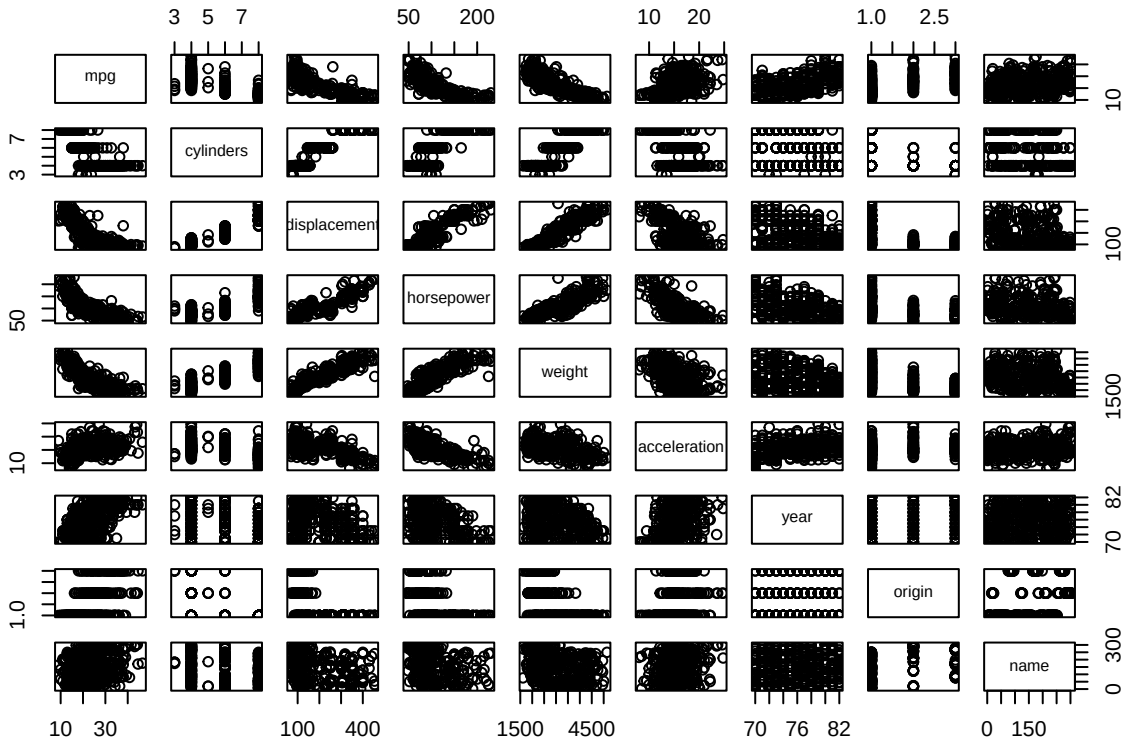
For fixed IQ and GPA
$\text{Salary}_{\text{High School}} = 50 + 20(1) + 0.07(2) + 35(0) + 0.01(x_1 x_2) - 10(x_1 x_3) + \varepsilon$ $\text{Salary}_{\text{College}} = 50 + 20(1) + 0.07(2) + 35(1) + 0.01(x_1 x_2) - 10(x_1) + \varepsilon$
Which gives us $\text{Salary}_{\text{College}} = \text{Salary}_{\text{High School}} + 35 - 10(x_1)$
From here we can use $\text{Salary}_{\text{College}} - \text{Salary}_{\text{High School}} = 35 - 10(x_1)$ Assuming the salary difference is $>= 0$, we get
$35 - 10(x_1) \leq 0 x_1 \geq 3.5$
Assuming the salary difference is $<= 0$, we get
$35 - 10(x_1) \geq 0 x_1 \leq 3.5$
Therefore, for a fixed value of IQ and GPA, **option iii is correct**

Applied exercises

1. Exercise 9 of Section 3.7 of the Textbook.

a.

```
# produce a scatterplot of all the variables in the Auto
# data set
pairs(Auto)
```



b.

```
# compute the matrix of correlations between the variables
# in the Auto data set, omitting the name variable
cor(Auto[, -9])
```

```
##                     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.

```
lm.auto <- lm(mpg ~ . - name, data = Auto)
summary(lm.auto)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

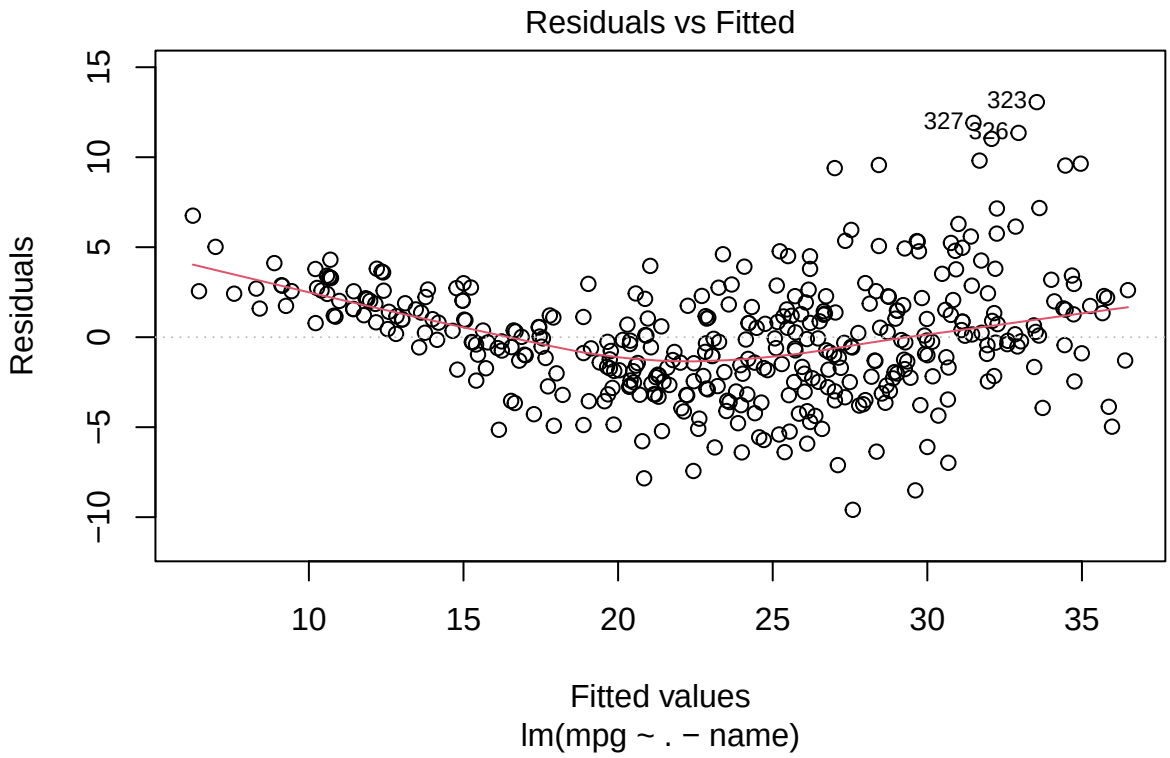
There are multiple predictors that have a relationship with the response `mpg` because their p-values are significant. Those predictors are `displacement`, `year`, and `weight`

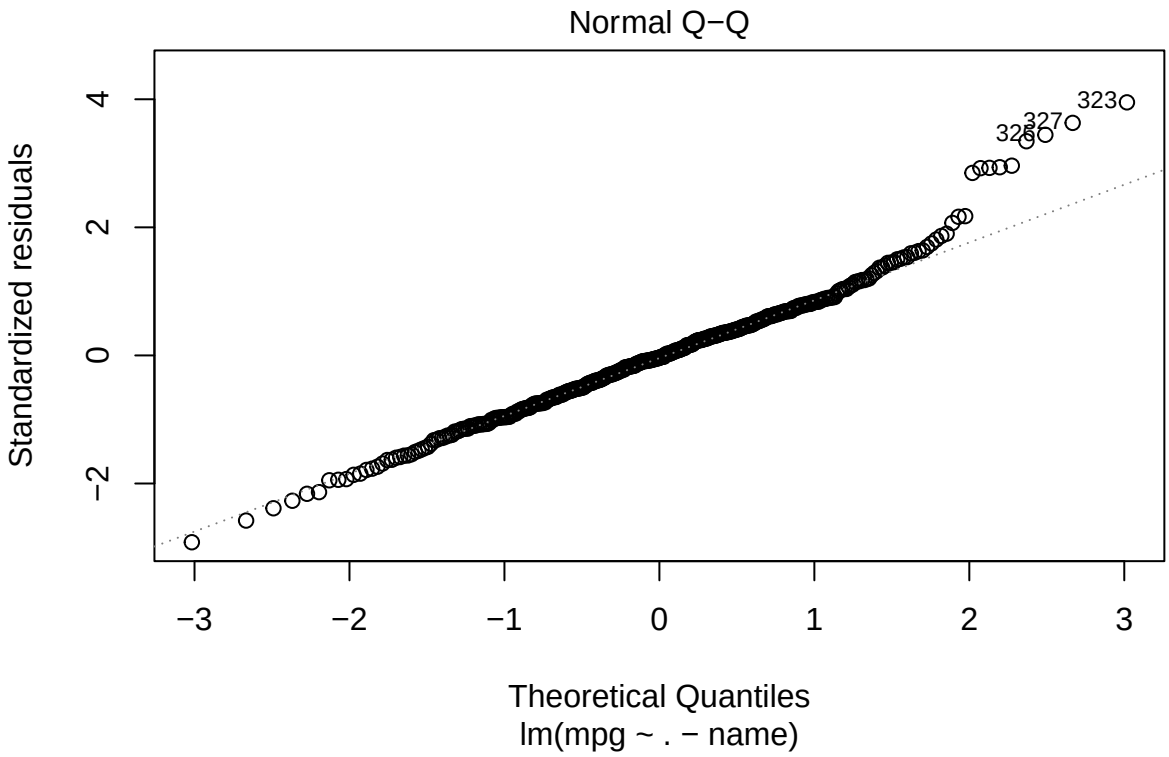
The predictors that have a statistically significant relationship with the response `mpg` are `displacement`, `year`, `origin`, and `weight`

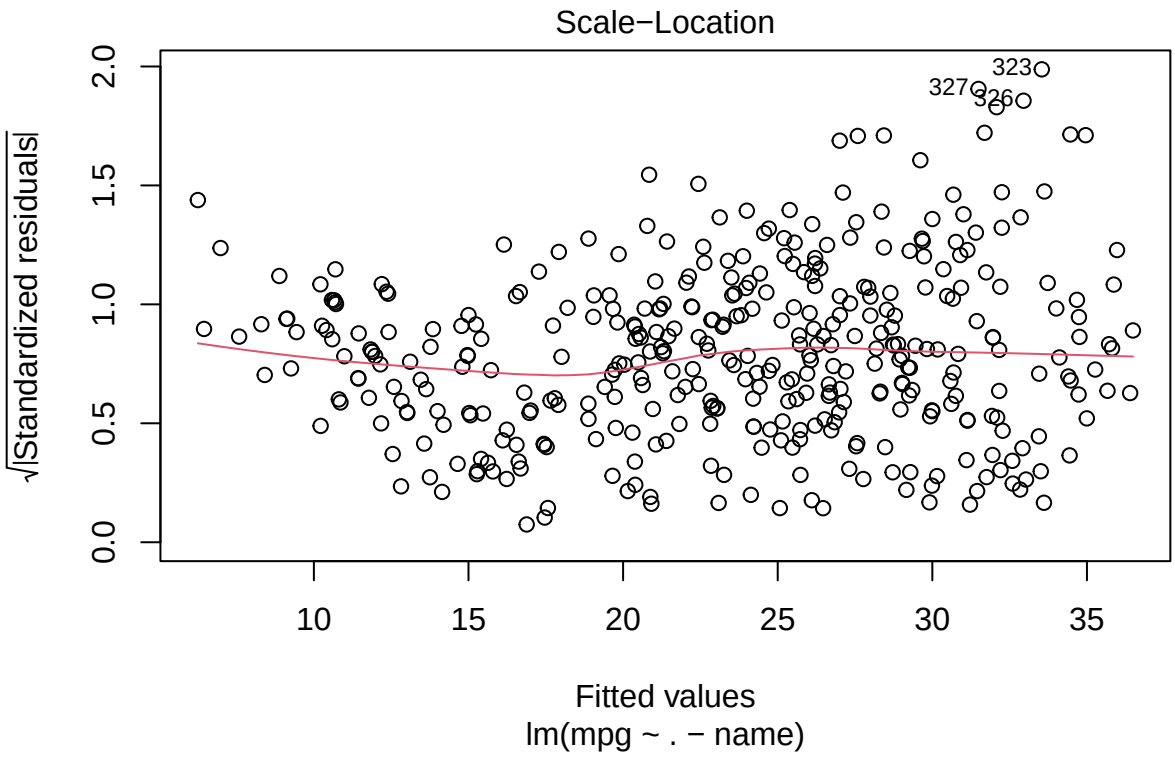
The coefficient of `year` is 0.7507, which means that for every year that passes, the mpg increases by 0.7507.

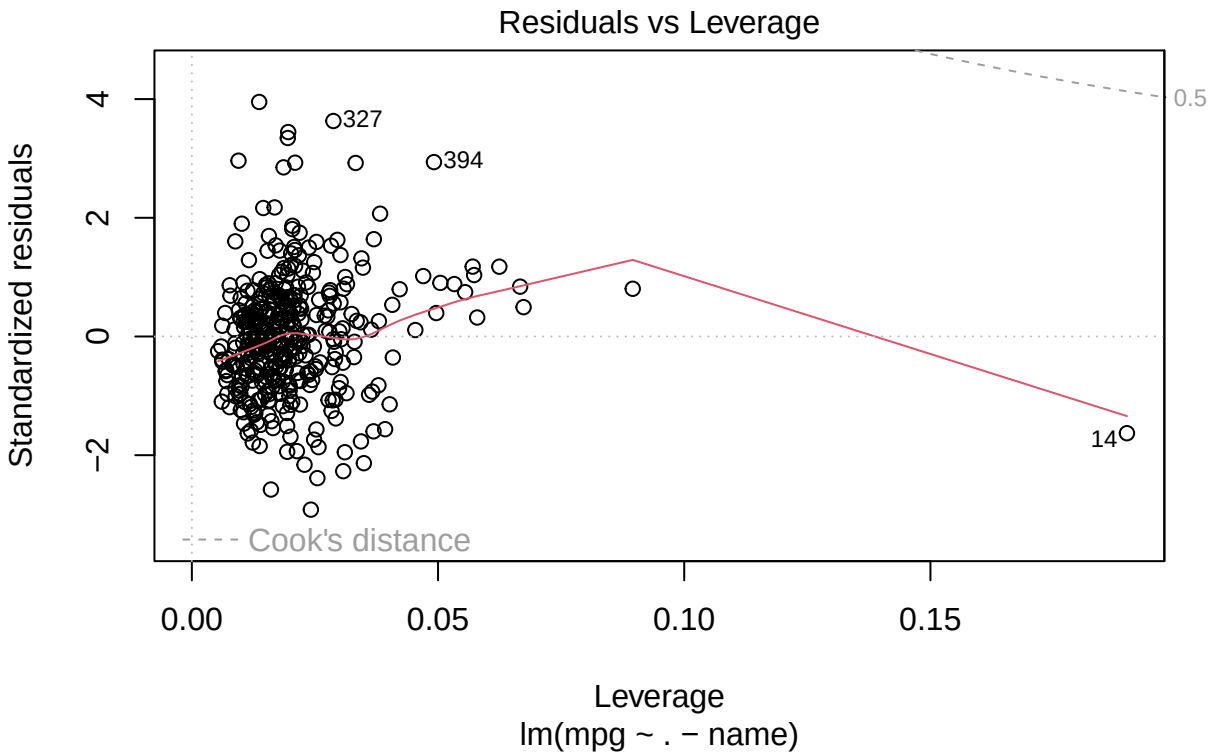
d.

```
plot(lm.auto)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ . − name)
323
327
326
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
323
327
326
10
15
20
25
30
35
Fitted values
lm(mpg ~ . − name)



The residuals vs fitted graph shows a u shape, which suggests that the data is non-linear. It also shows that the variance isn't constant with the funnel shape near the end. Based on the normal Q-Q plot, we can see that the most of the residuals are normally distributed escept for the ones from x= 2 to 3. Based on the Scale-Location plot, it looks like there aren't any outliers because the data fits between [0, 2]. Based on the Residuals vs Leverage plot, it doesn't look like there are any observations that provide high leverage.

e.

The two included interactions, `displacement * horsepower` and `horsepower * origin` are statistically significant

```
dp.interact <- lm(mpg ~ . - name + displacement * horsepower,
    data = Auto)
hporigin.interact <- lm(mpg ~ . - name + horsepower * origin,
    data = Auto)
```

f.

Messing around with the data I found that `log(acceleration)` while still significant, is less significant than `acceleration`. Squaring `horsepower` and squaring `weight` doesn't change their respective significances. And squaring `cylinders` makes `cylinders` and `horsepower` significant.

2. Exercise 10 of Section 3.7 of the Textbook.

```
# using the carseats data (a) Fit a multiple regression
# model to predict Sales using Price, Urban, and US.
car <- Carseats
head(car)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```
lm1 <- lm(Sales ~ Price + Urban + US, data = car)
summary(lm1)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = car)
##
## 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.

`Price`: For every increase of 1 in Price, Sales decreases by ≈ 54 units

`Urban`: If the store is in an urban area, Sales decreases by ≈ 22 units

`US`: If the store is in the US, Sales increases by 1200 units

c.

$$\text{Sales} = 13.043469 - 0.054459(\text{Price}) - 0.021916(\text{Urban}) + 1.200573(\text{US})$$

d.

We can reject the null hypothesis $H_0 : \beta_j = 0$ using the `Price` and `US` predictors.

e.

```
lm2 <- lm(Sales ~ Price + US, data = car)
summary(lm2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = car)
##
## 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.

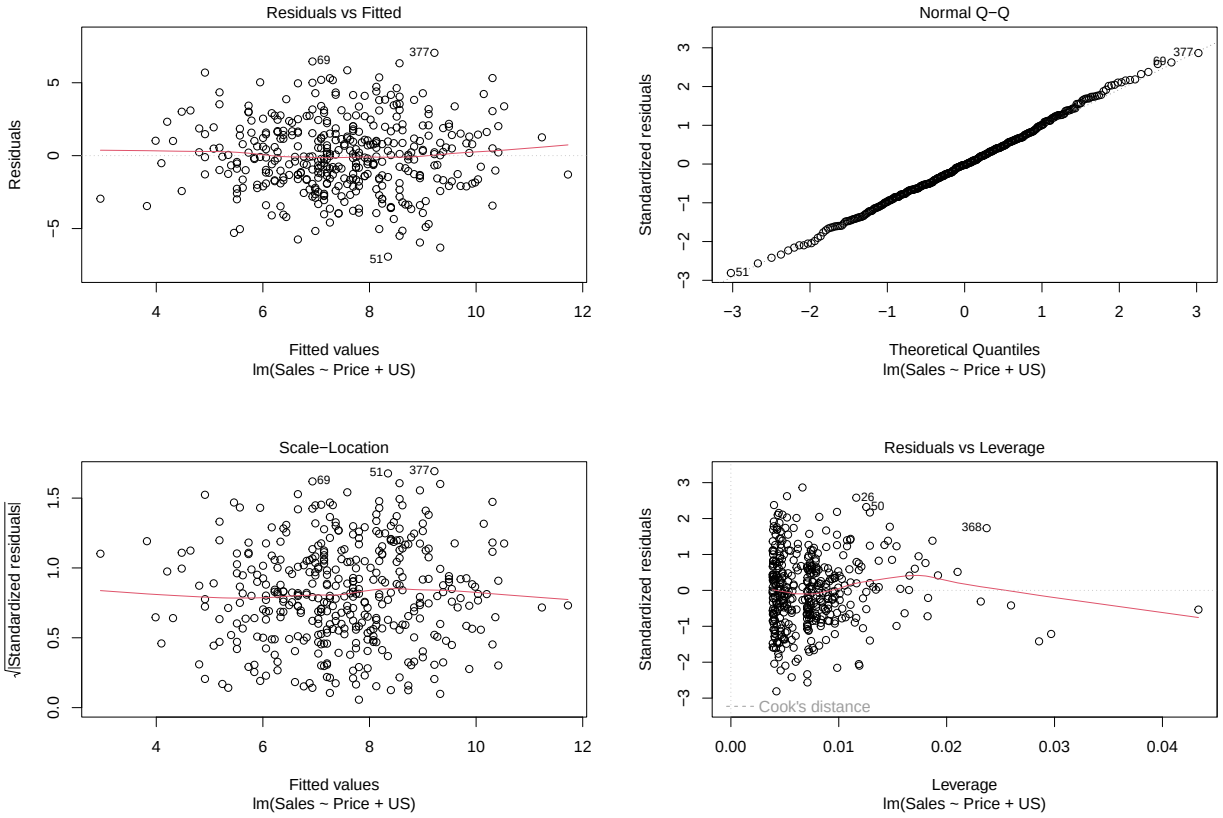
Both the models are able to explain the variation in `Sales`because the R^2 and adjusted R^2 are really close to each other. The residiual standard error of the models are small which means the model fits the data well. ### g.

```
confint(lm2)
```

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

h.

```
plot(lm2)
```



Looking at the plots, we can see that there is no evidence of outliers or high leverage points.
3. Exercise 14 of Section 3.7 of the Textbook.
a.

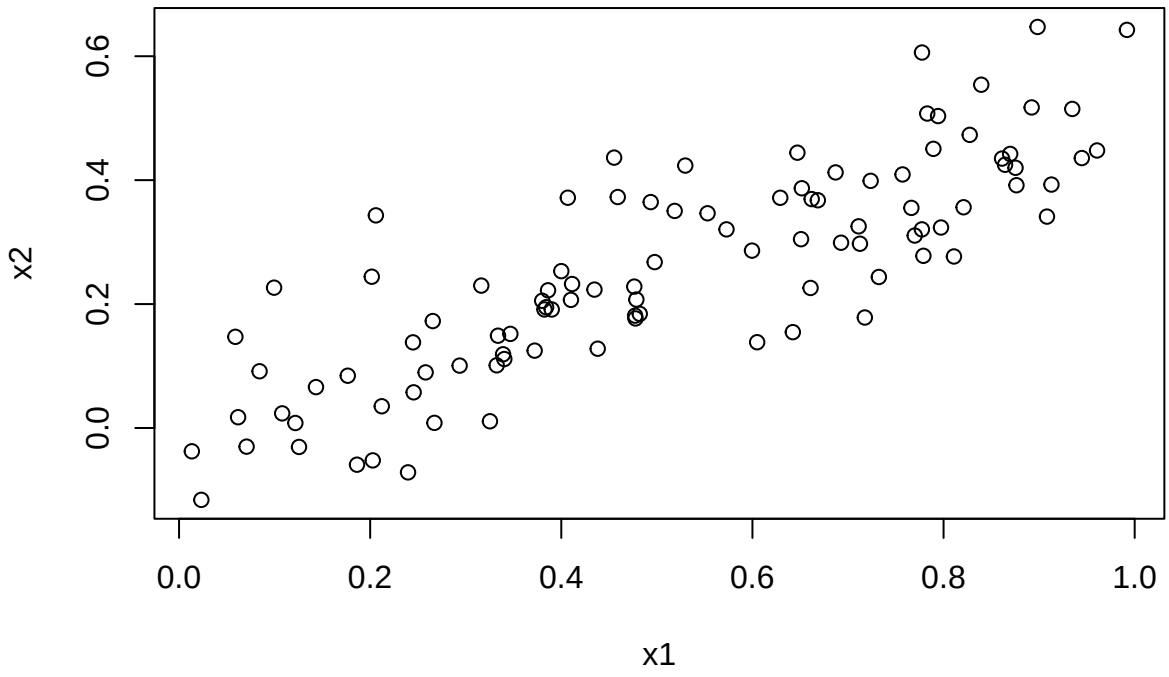
```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100)/10
y <- 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

The linear model is $y = 2 + 2x_1 + 0.3x_2 + \varepsilon$
b.

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

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

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



c.

```
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
```

$\hat{\beta}_0 = 2.1305$

$\hat{\beta}_1 = 1.4396$

$\hat{\beta}_2 = 1.0097$

The estimated value for the intercept is 2.1305, which is close to the true value of 2, this is the only estimator that is close to the true value. At the 5% significance level, we can reject the null hypothesis for β_1 but not β_2.

d.

```
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
```

In this model, the estimated value for `x1` is significant and we can more confidently reject the null hypothesis here than in the previous model

e.

```
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
```

Compared to the full model in part c, the estimate of β_1 for `x2` is significant and we can reject the null hypothesis.

f.

The results in the last three problems do contradict each other. When we fitted models for `x1` and `x2` individually, their estimates were significant. Compared to the full model in part c, where x1 could be considered significant and x2 was not significant at all. A reason for this would be some form of collinearity between the two predictors which could reduce the accuracy of the estimates. This is evident with the values of the estimators' standard errors, which are higher in the combined model and smaller in the individual models.

g.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)

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
```

With the introduction of the new observation, `x1` is no longer significant in the full model, but significant individually. `x2` is now significant in every model. In the full model `y ~ x1 + x2` and the individual model `y ~ x2`, the new observation is a high leverage point. Whereas in the individual model `y ~ x1`, the new observation is an outlier.