HW3

Yigao Li

February 9, 2018

3 - 3

(a)

i.

False.

$$Y(\text{Gender} = \text{male}|\text{IQ}, \text{GPA}) = ... + 0$$
$$Y(\text{Gender} = \text{female}|\text{IQ}, \text{GPA}) = ... + 35 - 10 \times \text{GPA}$$

There is not enough information to tell whether males earn more on average than female.

ii.

False.

Same reason with i.

iii.

True.

When GPA is high enough ($\text{GPA} > 3.5$), $Y(\text{Gender} = \text{male}|\text{IQ}, \text{GPA} > 3.5) > Y(\text{Gender} = \text{female}|\text{IQ}, \text{GPA} > 3.5)$. Thus, males earn more on average than females provided that the GPA is high enough.

iv.

False.

Same reason with iv. Females earn less on average than males provided that the GPA is high enough.

(b)

The regression equation is:

$$\hat{Y} = 50 + 20 \times \text{GPA} + 0.07 \times \text{IQ} + 35 \times \text{Gender} + 0.01 \times \text{GPA} \times \text{IQ} - 10 \times \text{GPA} \times \text{Gender}$$

When $\text{IQ} = 110$, $\text{GPA} = 4.0$ and $\text{Gender} = 1$(female), $\hat{Y} = 137.1$. The estimated starting salary of a female with IQ of 110 and a GPA of 4.0 after graduation is 137.1 thousand dollars.

(c)

False. Coefficient cannot indicate the effectiveness of a predictor.

3 - 9

(e)

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.3
```

```r
auto <- subset(Auto, select = -name)
model.1 <- lm(mpg ~ .^2, data = auto)
summary(model.1)
```

```
##
## Call:
## lm(formula = mpg ~ .^2, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.548e+01  5.314e+01   0.668  0.50475
## cylinders                  6.989e+00  8.248e+00   0.847  0.39738
## displacement              -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                 5.034e-01  3.470e-01   1.451  0.14769
## weight                     4.133e-03  1.759e-02   0.235  0.81442
## acceleration              -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                       6.974e-01  6.097e-01   1.144  0.25340
## origin                    -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight:year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

Among all interactive terms, “displacement × year”, “acceleration × year” and “acceleration × origin” are statistically significant.

(f)

```
model.2 <- lm(mpg ~ log(cylinders) + log(displacement) + log(horsepower) + log(weight) +
                log(acceleration) + log(year) + log(origin), data = auto)
summary(model.2)
```

```
##
## Call:
## lm(formula = mpg ~ log(cylinders) + log(displacement) + log(horsepower) +
##     log(weight) + log(acceleration) + log(year) + log(origin),
##     data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5987 -1.8172 -0.0181  1.5906 12.8132
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -66.5643    17.5053  -3.803 0.000167 ***
## log(cylinders)      1.4818     1.6589   0.893 0.372273
## log(displacement)  -1.0551     1.5385  -0.686 0.493230
## log(horsepower)    -6.9657     1.5569  -4.474 1.01e-05 ***
## log(weight)       -12.5728     2.2251  -5.650 3.12e-08 ***
## log(acceleration)  -4.9831     1.6078  -3.099 0.002082 **
## log(year)          54.9857     3.5555  15.465  < 2e-16 ***
## log(origin)         1.5822     0.5083   3.113 0.001991 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.069 on 384 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8454
## F-statistic: 306.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

There are 5 logarithm predictors that are statistically significant: “horsepower”, “weight”, “acceleration”, “year” and “origin”.

```
model.3 <- lm(mpg ~ sqrt(cylinders) + sqrt(displacement) + sqrt(horsepower) + sqrt(weight) +
                sqrt(acceleration) + sqrt(year) + sqrt(origin), data = auto)
summary(model.3)
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(cylinders) + sqrt(displacement) + sqrt(horsepower) +
##     sqrt(weight) + sqrt(acceleration) + sqrt(year) + sqrt(origin),
##     data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -9.5250 -1.9822 -0.1111  1.7347 13.0681
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -49.79814    9.17832  -5.426 1.02e-07 ***
## sqrt(cylinders)     -0.23699    1.53753  -0.154   0.8776
## sqrt(displacement)   0.22580    0.22940   0.984   0.3256
## sqrt(horsepower)    -0.77976    0.30788  -2.533   0.0117 *
## sqrt(weight)        -0.62172    0.07898  -7.872 3.59e-14 ***
## sqrt(acceleration)  -0.82529    0.83443  -0.989   0.3233
## sqrt(year)          12.79030    0.85891  14.891  < 2e-16 ***
## sqrt(origin)         3.26036    0.76767   4.247 2.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.21 on 384 degrees of freedom
## Multiple R-squared:  0.8338, Adjusted R-squared:  0.8308
## F-statistic: 275.3 on 7 and 384 DF,  p-value: < 2.2e-16
```

There are 4 square root predictors that are statistically significant: “horsepower”, “weight”, “year” and “origin”.

```r
model.4 <- lm(as.formula(paste("mpg ~ ", paste("poly(", colnames(auto[-1]), ",2)", collapse = '+'))),
              data = auto)
summary(model.4)
```

```
##
## Call:
## lm(formula = as.formula(paste("mpg ~ ", paste("poly(", colnames(auto[-1]),
##     ",2)", collapse = "+"))), data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6457 -1.5810  0.0953  1.3132 12.2519
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              23.4459     0.1392 168.426  < 2e-16 ***
## poly(cylinders, 2)1      13.8556    10.8088   1.282  0.20067
## poly(cylinders, 2)2      -1.5780     3.8730  -0.407  0.68392
## poly(displacement, 2)1 -17.4481    16.6992  -1.045  0.29676
## poly(displacement, 2)2   6.8516     7.1616   0.957  0.33933
## poly(horsepower, 2)1    -51.7980    10.2958  -5.031 7.57e-07 ***
## poly(horsepower, 2)2     12.3555     5.0675   2.438  0.01522 *
## poly(weight, 2)1        -58.4689    12.0798  -4.840 1.90e-06 ***
## poly(weight, 2)2         20.7826     4.9728   4.179 3.64e-05 ***
## poly(acceleration, 2)1 -12.9033     5.3901  -2.394  0.01716 *
## poly(acceleration, 2)2  10.3880     3.7693   2.756  0.00614 **
## poly(year, 2)1           56.6081     3.2370  17.488  < 2e-16 ***
## poly(year, 2)2           16.6217     2.9542   5.626 3.59e-08 ***
## poly(origin, 2)1         10.8816     4.2387   2.567  0.01064 *
## poly(origin, 2)2         -3.4645     3.2344  -1.071  0.28480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.756 on 377 degrees of freedom
## Multiple R-squared:  0.8798, Adjusted R-squared:  0.8753
## F-statistic:   197 on 14 and 377 DF,  p-value: < 2.2e-16
```

The regression model with quadratic terms has 9 statistically significant terms: "horsepower", "horsepower^2", "weight", "weight^2", "acceleration", "acceleration^2", "year", "year^2" and "origin".

Concluding from all previous results, quadratic transformation is the best regression because its multiple R-squared is the largest. Also, all forms of "cylinders" variable is not statistically significant in any regression.

3 - 10

(a)

```r
carseats <- Carseats
model.5 <- lm(Sales ~ Price + Urban + US, data = carseats)
summary(model.5)
```

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

(b)

The regression is:

$$\hat{\text{Sales}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Price} + \hat{\beta}_2 \times \text{Urban} + \hat{\beta}_3 \times \text{US},$$
$$\text{where Urban} = \begin{cases} 1 \text{ Yes} \\ 0 \text{ No} \end{cases} \quad \text{and US} = \begin{cases} 1 \text{ Yes} \\ 0 \text{ No} \end{cases}$$

β_0: If company charges nothing for car seats at a rural store not in the US, the estimated sales at this location is 13.043469 thousand dollars.
β_1: In the same store, if company charges 1 dollar more for each car seat, sales at this location decreases \$54.459 on average.

β_2: For a fixed number of dollars company charges for car seats, sales at a US urban location is on average \$21.916 less than a US rural location.
β_3: For a fixed number of dollars company charges for car seats, sales at a US location is on average \$1200.573 more than a similar non-US location.

(c)

Answered in part (b)

(d)

"Price" and "US"

(e)

```r
model.6 <- lm(Sales ~ Price + US, data = carseats)
summary(model.6)
```

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

(f)

All variables in (e) model are statistically significant. Generally speaking, model (e) does not improve much comparing to model (a). Coefficients slightly changed and multiple R-squared does not change.

(g)

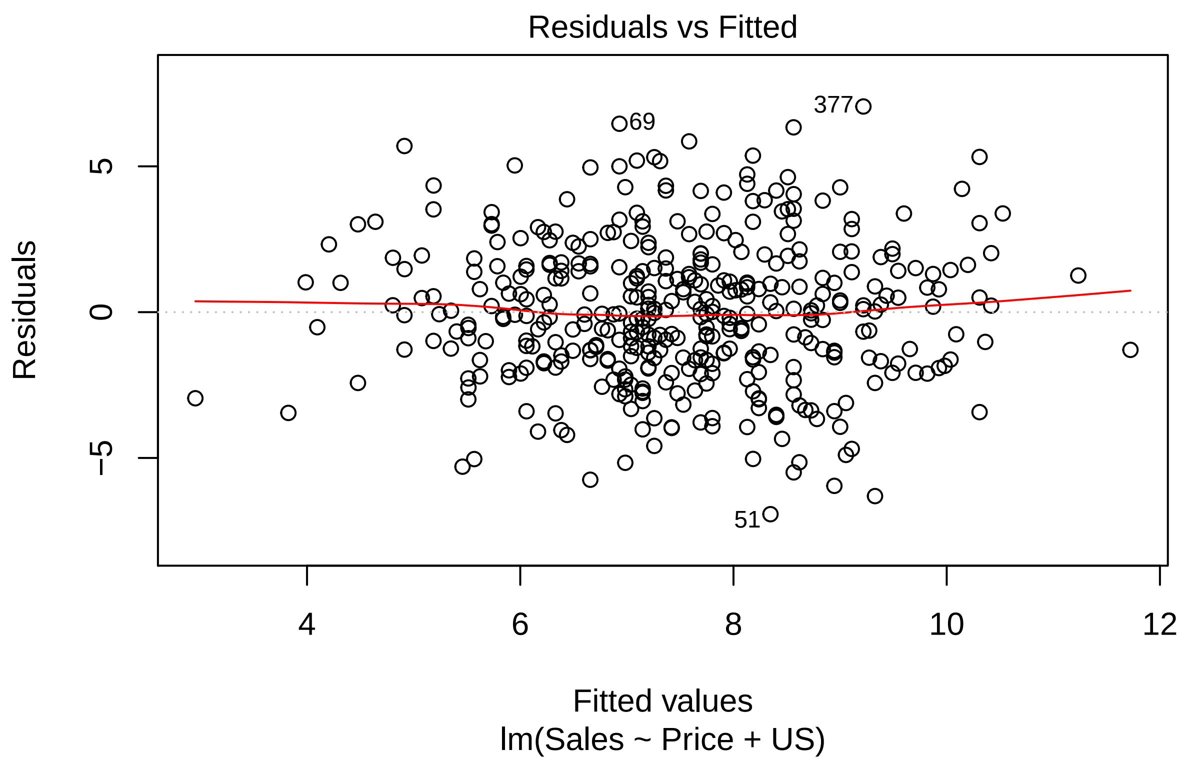
```r
confint(model.6)
```

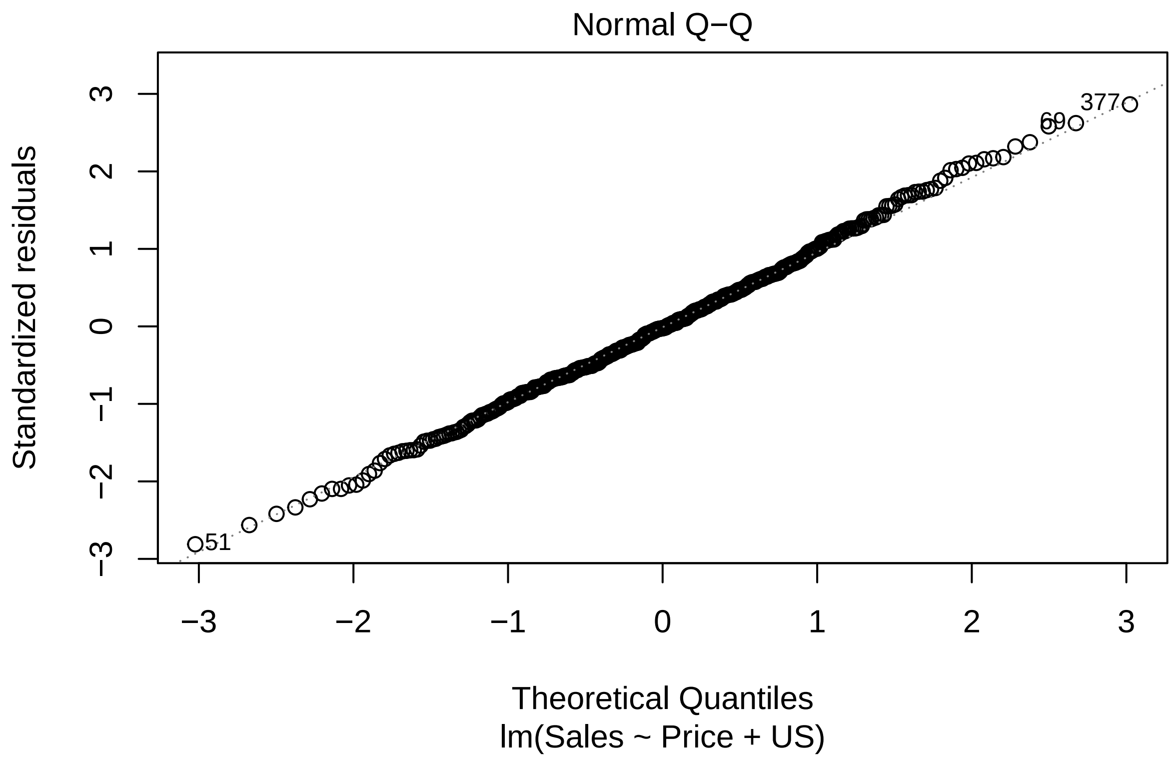
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
```

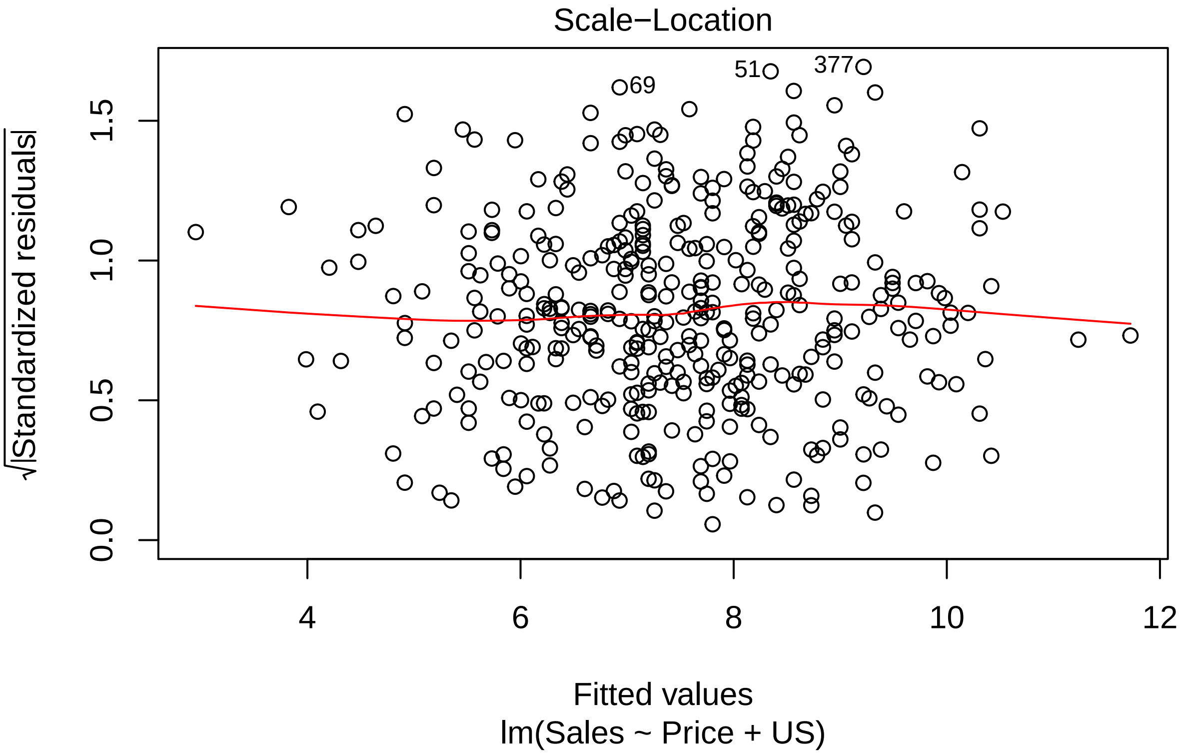
```
## Price        -0.06475984 -0.04419543
## USYes         0.69151957  1.70776632
```

(h)

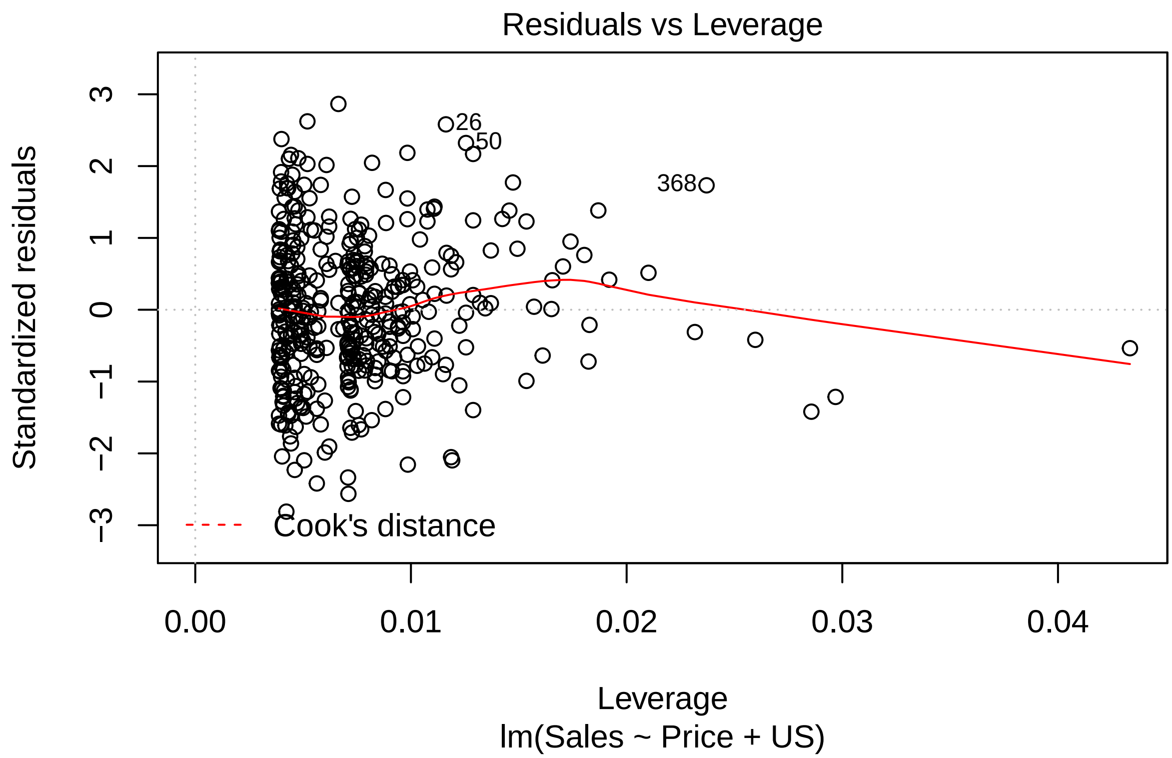
```r
plot(model.6)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Price + US)
377
69
51
−3
−2
−1
0
1
2
3



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
6
8
10
12
Fitted values
lm(Sales ~ Price + US)
69
51
377




```r
carseats[377,]
```

```
##     Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 377 16.27       141     60          19        319    92      Good  44
##     Education Urban  US
## 377        11   Yes Yes
```

Observation 377 is an outlier. It has the highest leverage in the model (e).

3 - 14

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

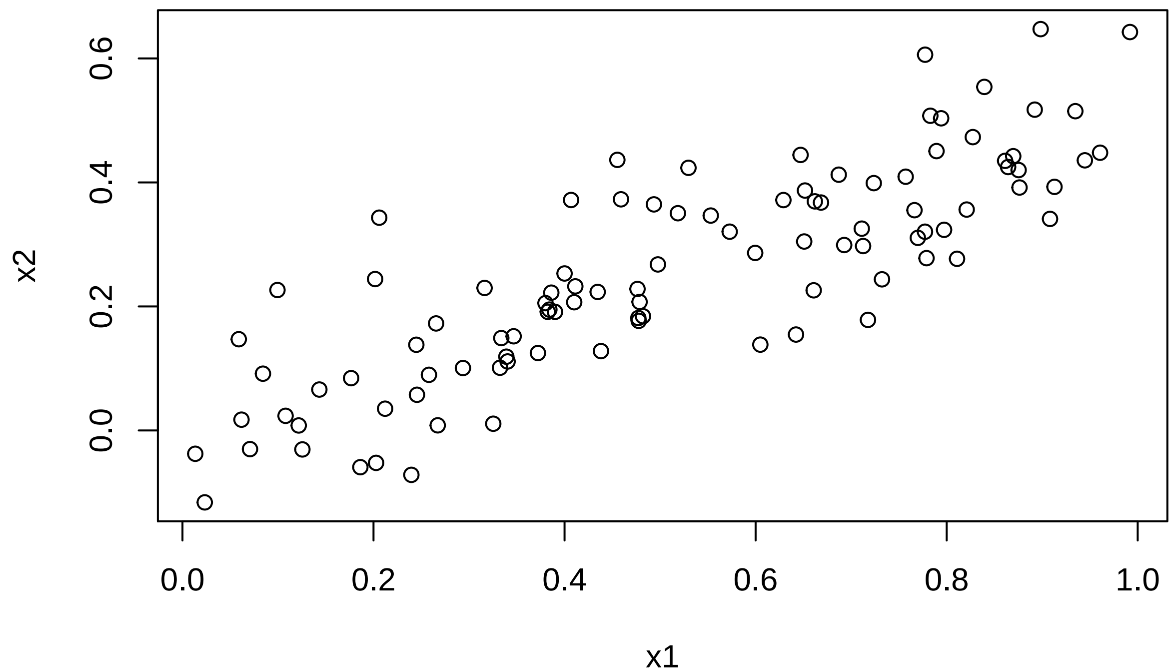
$\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$

(b)

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

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

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



(c)

```r
model.7 <- lm(y ~ x1 + x2)
summary(model.7)
```

```
##
## 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
```

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$\hat{\beta}_0 = 2.1305$, $\hat{\beta}_1 = 1.4396$, $\hat{\beta}_2 = 1.0097$
β_0 is approximately the same but the estimated β_1 is smaller than real β_1 and the estimated β_2 is larger than real β_2. From the model summary, we can reject null hypothesis for β_1 but not for β_2.