Tutorial 3 Solutions

STAT 4040/7040

- Textbook key:
  - [ISL] *Introduction to Statistical Learning*
  - [ESL] *Elements of Statistical Learning*

1. **3.4 Solution:**

   a. We would expect the training RSS (residual sum of squares) to be smaller for the cubic regression since that functional form has more flexibility to fit the training data set with and thus it will be able to do a better job.

   b. Since the true model is linear, we would expect that on the test set the linear model would perform better than the cubic model.

   c. On the training set, the cubic model will always have a lower RSS measurement since the procedure used to fit the coefficients in the cubic model has more flexibility.

   d. On the test set, which model performs better depends on how non-linear the true underlying model is and how much data we were given to fit the models using. If the true underlying model is highly non-linear then we would expect the cubic model to perform better than the linear model. If however we were given only very few data points n then due to errors in fitting the cubic model it might perform worse on the test set than the linear model. This is because with only a few data points to fit the cubic model, it will have more variance than the linear model, thus it could result in a worse *out-of-sample* fit.

   3.10 Solution:

   a. Let's fit the linear regression model and get a summary of the results:

```
library(ISLR)
data(Carseats)

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

```
##
## 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 summary table indicates that the coefficient for *Price* is negative (and statistically significant), showing that as the price increases sales decrease. The coefficient for *UrbanYes* is negative (but not statistically significant). If it was significant, we could conclude that an urban environment has less sales than a rural environment. The coefficient for *USYes* is positive and statistically significant indicating that stores in the US have increased sales over ones that are not located in the US.

c. Let's write out the model:

$$\begin{aligned}\widehat{\text{Sales}} &= \hat{\beta}_0 + \hat{\beta}_1\text{Price} + \hat{\beta}_2\text{UrbanYes} + \hat{\beta}_3\text{USYes} \\ &= 3.043 - 0.054\text{Price} - 0.022\text{UrbanYes} + 1.201\text{USYes}\end{aligned}$$

```
* Where *Yes* indicates that the variable is true, thus we set that variable equal
```

d. We can reject the null hypothesis for *Sales* and *US*. For US let's explicitly write out the t-test and determine the p-value. We are testing:

$$H_0 : \beta_3 = 0, \quad H_a : \beta_3 \neq 0$$

Our observed test statistic is:

$$t_{obs} = \frac{\hat{\beta}_3 - \beta_3}{SE(\hat{\beta}_3)} = \frac{1.2005 - 0}{0.2590} = 4.63$$

Under the null hypothesis, this statistic has a T distribution with $n-(p+1) = 400-4 = 396$ degrees of freedom. Let's determine the p-value (remember we have a two sided alternative):

$$\begin{aligned}\text{p-value} &= P(T_{df=396} > 4.63) + P(T_{df=396} < -4.63) \\ &= 2P(T_{df=396} < -4.63)\end{aligned}$$

```
p.value <- 2*pt(-4.63, df=396)
p.value
```

```
## [1] 4.965445e-06
```

As this is less than our standard α level of 0.05, we reject the null hypothesis!

e. Let's refit the model after dropping the *Urban* covariate:

```
mod2 <- lm(Sales ~ Price + US, data=Carseats)
summary(mod2)
```

```
##
## 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. The model in (a) has an $R^2 = 0.2393$ (with $\hat{\sigma} = 2.472$) and the model in (e) has an $R^2 = 0.2393$ (with $\hat{\sigma} = 2.469$). These two models have the same value of R^2 but the second model has a smaller estimate of error and thus suggesting a better model for the data.

g. To get 95% confidence intervals for the regression coefficients for the model in (e) we can use *confint()*:

```
confint(mod2)
```

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

Let's do this "by hand" for β_3:

$$\hat{\beta}_3 \pm t^* SE(\hat{\beta}_3)$$

Where t^* is the value such that $P(T_{df=397} > t^*) = 0.05/2$.

```
1.19964 - qt(0.975, df=397)*0.25846
```

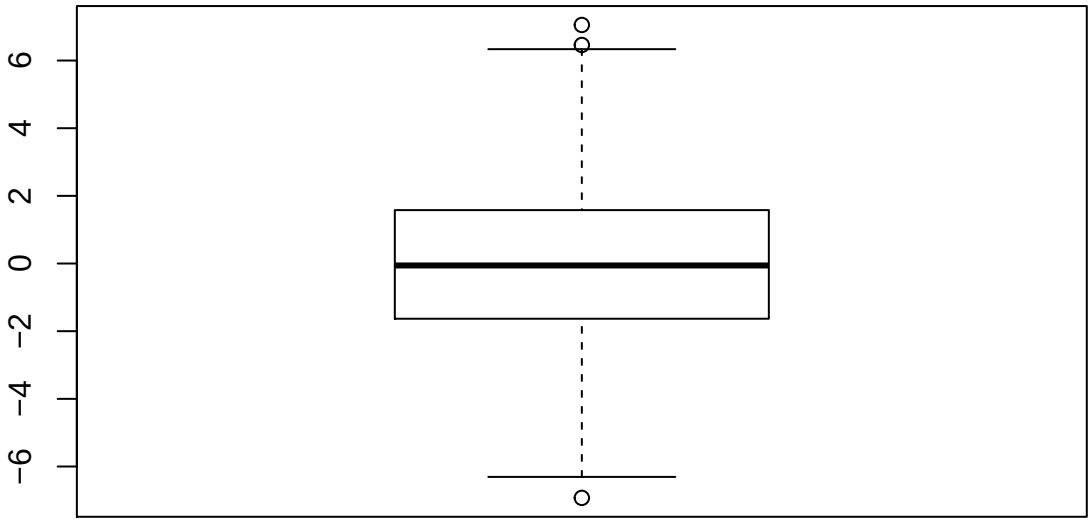
```
## [1] 0.6915186
```

```
1.19964 + qt(0.975, df=397)*0.25846
```

```
## [1] 1.707761
```

h. One way to check for outliers is to examine a boxplot of the residuals.

```
boxplot(mod2$res)
```



From the boxplot it appears we have three outliers. Let's get the values and id number:

```
box <- boxplot(mod2$res, plot=FALSE)
box$out
```

```
##        51        69       377
## -6.926851  6.459567  7.051506
```

Let's see which (if any) data points have high leverage. To examine leverage we get the "Hat" matrix (H). This matrix puts the "hat" onto y:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{H}\boldsymbol{y} \end{aligned}$$

```
X <- model.matrix(mod2)
H <- X%*%solve(t(X)%*%X)%*%t(X)
h.i <- diag(H)
h.i[1:5]
```

```
##           1           2           3           4           5
## 0.003921615 0.009003813 0.009954183 0.005636383 0.007927678
```

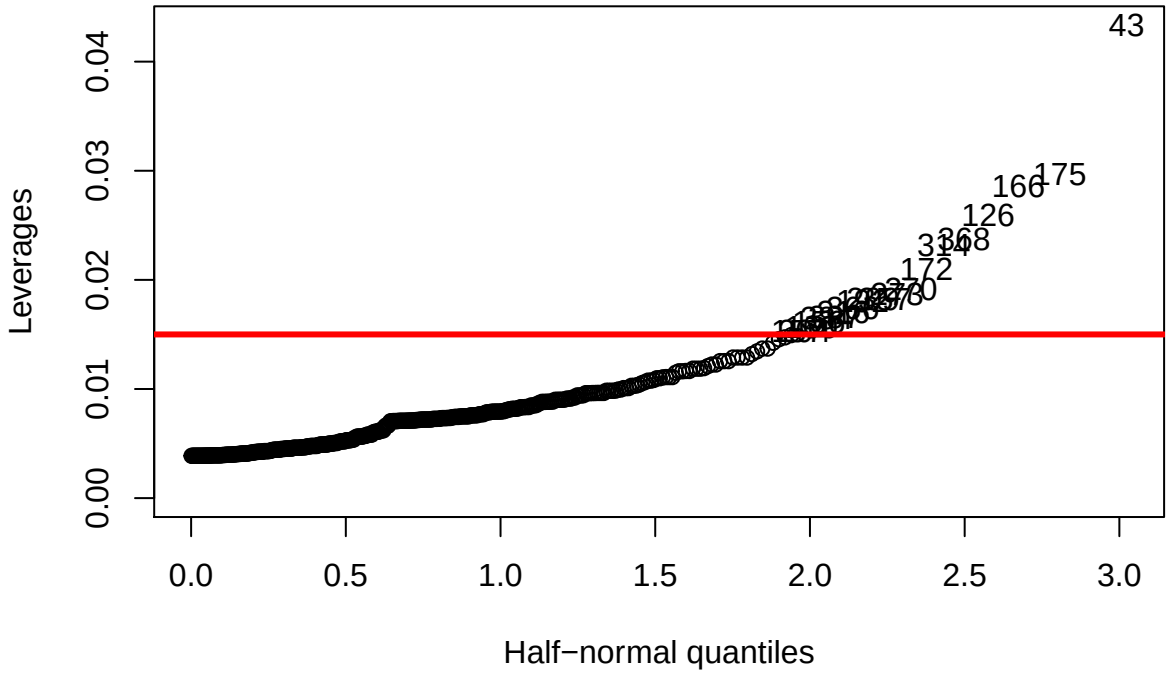
We can get these quickly using the *influence()* command:

```r
lev <- influence(mod2)$hat
lev[1:5]
```

```
##           1           2           3           4           5
## 0.003921615 0.009003813 0.009954183 0.005636383 0.007927678
```

Note that the leverages are positive. We will use a half normal plot (in the *faraway* library) to look for large leverages:

```r
library(faraway)
halfnorm(lev, ylab="Leverages", nlab=length(lev[lev>2*(3)/400]))
abline(h=2*(3)/400, lwd=3, col="red")
```



Based on the heuristic of looking for leverage values greater than $2\frac{p+1}{n}$, we can identify quite a few data points (a boxplot also shows a number of outliers). However, most data points, except for 43, don't seem to separate too much from the "pack".

3.14 Solution:

a. Let's run the commands:

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

The linear regression model is:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ &= 2 + 2x_1 + 0.3x_2 + \epsilon \\ \epsilon &\sim \text{normal}(0, \sigma^2 = 1) \end{aligned}$$

We can actually rewrite this in terms of just x_1:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_1 \\ &= \beta_0 + \beta_1 x_1 + \beta_2 (0.5x_1 + \epsilon_2) + \epsilon_1 \\ &= \beta_0 + \beta_1 x_1 + \beta_2 0.5x_1 + \epsilon_3 \\ &= 2 + 2x_1 + (0.3)0.5x_1 + \epsilon_3 \\ &= 2 + 2.15x_1 + \epsilon \\ \epsilon &\sim \text{normal}(0, \sigma^2 = 1) \end{aligned}$$

b. Let's determine first determine the variance of x_1 and x_2 and the covariance between x_1 and x_2:

$$\begin{aligned} X_1 &\sim \text{uniform}(0, 1) \\ V(X_1) &= 1/12 \end{aligned}$$

$$\begin{aligned} X_2 &= 0.5X_1 + Z \\ Z &\sim \text{normal}(0, \sigma^2 = 0.01) \\ V(X_2) &= V(0.5X_1 + Z) \\ &= 0.25(1/12) + 0.01 = 0.03083333 \end{aligned}$$

$$\begin{aligned} Cov(X_1, X_2) &= Cov(X_1, 0.5X_1 + Z) \\ &= Cov(X_1, 0.5X_1) + Cov(X_1, Z) \\ &= 0.5V(X_1) + 0 \\ &= 0.5(1/12) \end{aligned}$$

So the correlation is:

$$\begin{aligned} \rho &= \frac{0.5(1/12)}{\sqrt{(1/12)}\sqrt{(0.25(1/12) + 0.01)}} \\ &\approx 0.8219949 \end{aligned}$$

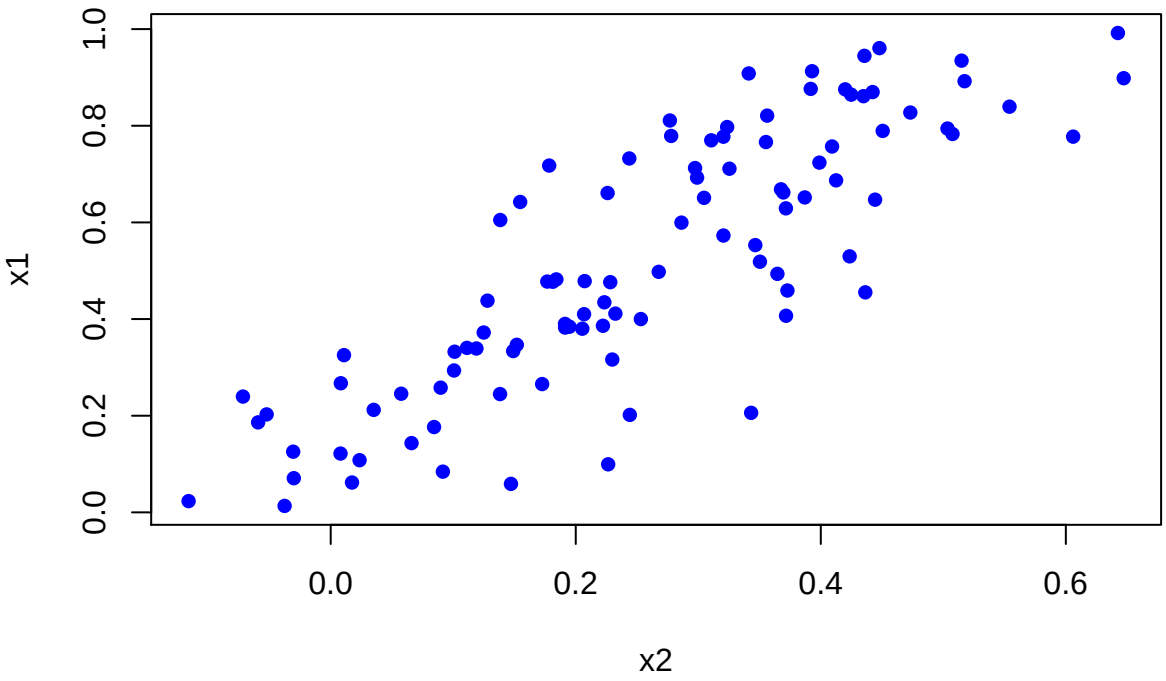
The estimated correlation is:

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

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

Let's examine a scatter plot of the two variables:

```r
plot(x2, x1, pch=16, col="blue")
```



The variables are quite highly correlated.

c. Let's fit the regression model and get a summary of the results:

```r
mod <- lm(y ~ x1 + x2)
summary(mod)
```

```
##
## 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
```

From the summary table, only x_1 is statistically significant (i.e. we can reject the null hypothesis $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ (barely). It appears we are not doing a great job of estimating the parameters. For example $\beta_1 = 2$, but we estimated $\hat{\beta}_1 = 1.440$.

d. Now let's fit the model with just x_1:

```
mod2 <- lm(y ~ x1)
summary(mod2)
```

```
##
## 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
```

We can now see that x_1 is an important covariate.

e. Now let's fit the model with just x_2:

```
mod3 <- lm(y ~ x2)
summary(mod3)
```

```
##
## 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
```

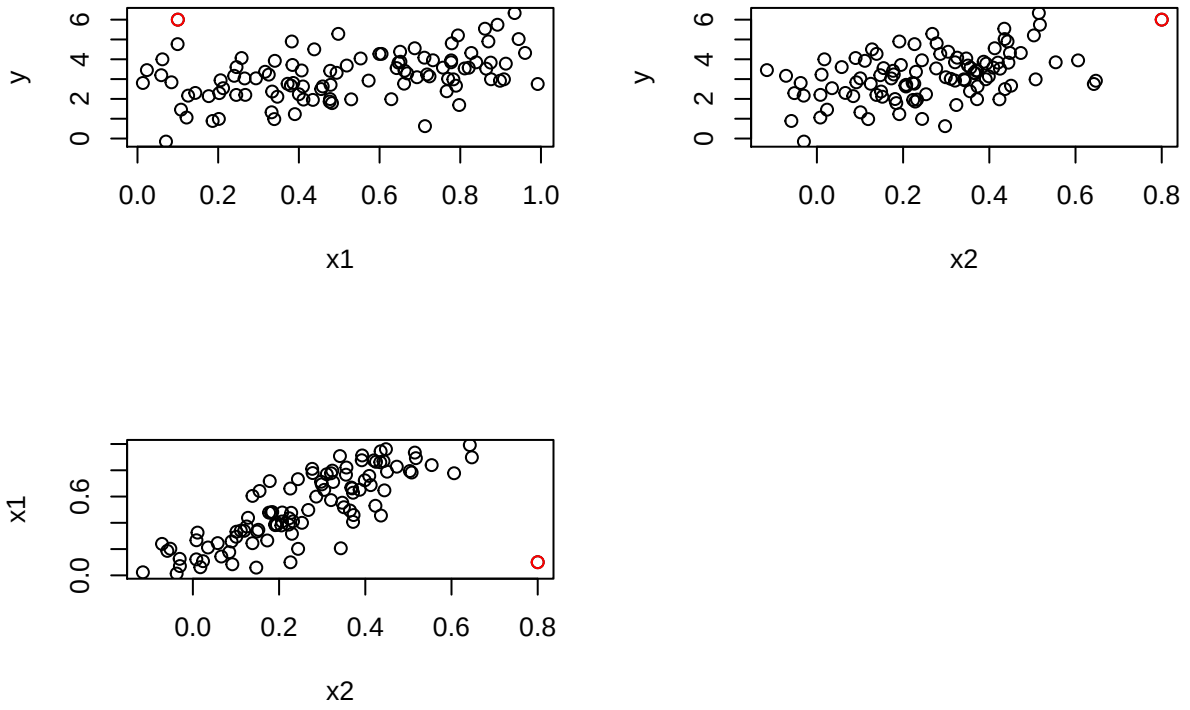
We can also see that x_2 is an important covariate.

f. These results don't contradict each other due to the strong correlation between the two variables. Essentially the two covariates represent similar pieces of information and including both in the model increases the standard error for each of them.

g. Let's add the additional data point and make some scatter plots. From the scatter plots of y against each x, the point doesn't appear to be too outlying. However, when looking at the plot of x_2 against x_1, we do see that the point is an outlier.

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

par(mfrow=c(2,2))
plot(x1, y); points(x1[101], y[101], pch=21, col="red")
plot(x2, y); points(x2[101], y[101], pch=21, col="red")
plot(x2, x1); points(x2[101], x1[101], pch=21, col="red")
```

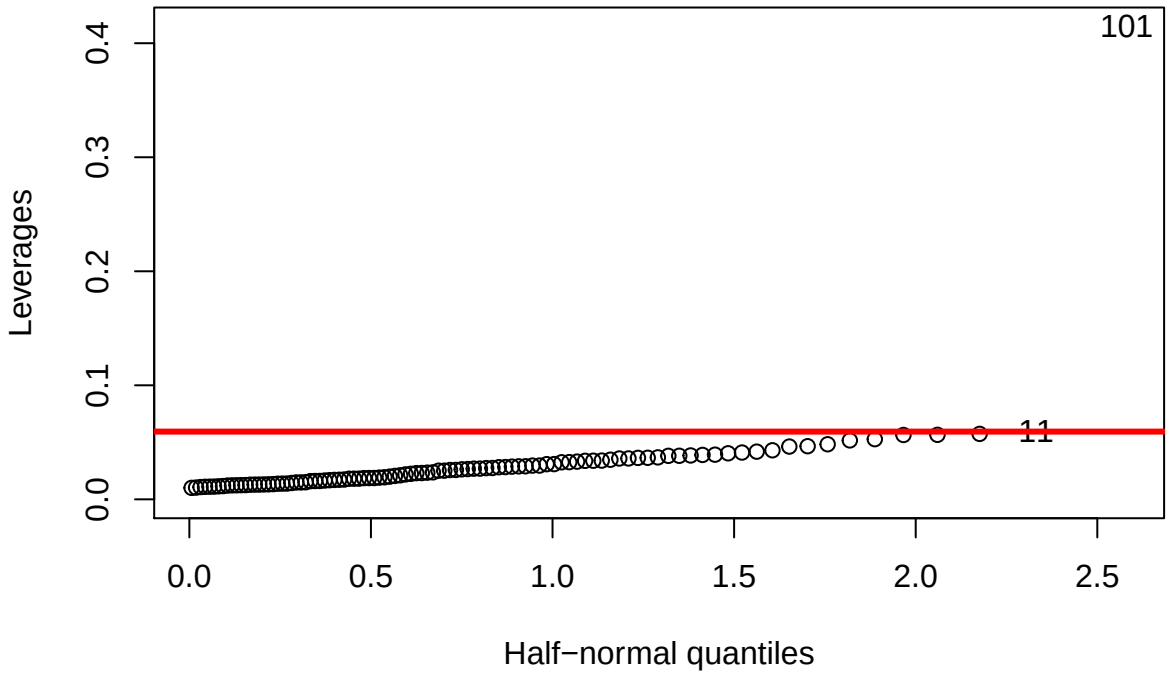


Let's see if this point is a high leverage point:

```r
mod4 <- lm(y ~ x1 + x2)
summary(mod4)
```

```
##
## 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
```

```r
lev <- influence(mod4)$hat
halfnorm(lev, ylab="Leverages", nlab=length(lev[lev>2*(3)/101]))
abline(h=2*(3)/101, lwd=3, col="red")
```



```
lev[101]
```

```
##       101
## 0.4147284
```

We do see that the new observation has quite high leverage compared to the rest of the data. We see that the point does change the estimated coefficients compared to the model in (c).