RSM8512 Assignment - Linear Regression

Yanbing Chen 1009958752 2023/10/02

Question 1 [25 marks]

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

Answer: The null hypothesis for 'TV' is that in the presence of radio ads and newspaper ads, TV has no effect on sales. Similarly, the null hypothesis for radio is that in the presence of TV and newspaper, radio has no effect on sales. The similar null hypothesis for newspaper too.

The p-value of TV and radio is <0.0001, which means that we should reject null hypothesis for these two factors. However, the p-value of newspaper is bigger that 0.05, which means we should accept null hypothesis and think that newspaper ads has no effect on sales.

Question 3 [25 marks]

Suppose we have a data set with five predictors, X1 =GPA, X2 = IQ, X3 = Gender (1 for Female and 0 forMale), X4 = Interaction between GPA and IQ, and X5 = 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

$$\hat{\beta}_0 = 50, \quad \hat{\beta}_1 = 20, \quad \hat{\beta}_2 = 0.07, \quad \hat{\beta}_3 = 35, \quad \hat{\beta}_4 = 0.01, \quad \hat{\beta}_5 = -10$$

(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: We can generate a model based on the info above:

$$Y = 50 + 20 * X1 + 0.07 * X2 + 35 * X3 + 0.01 * X1 * X2 - 10 * X1 * X3$$

$$salary(x_3 = 0) = 50 + 20 * X1 + 0.07 * X2 + 0.01 * X1 * X2 - 10 * X1 * X3$$
$$salary(x_3 = 1) = 50 + 20 * X1 + 0.07 * X2 + 35 + 0.01 * X1 * X2 - 10 * X1 * X3$$
$$salary = salary(x_3 = 0) - salary(x_3 = 1) = 35$$

Therefore, iii is correct, when GPA is high enough, males earn more on average than females.

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

Answer:

$$salary = 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * 4.0 * 110 - 10 * 4.0 * 1 = 137.1$$

The salary of a female with IQ of 110 and a GPA of 4.0 is 137.1.

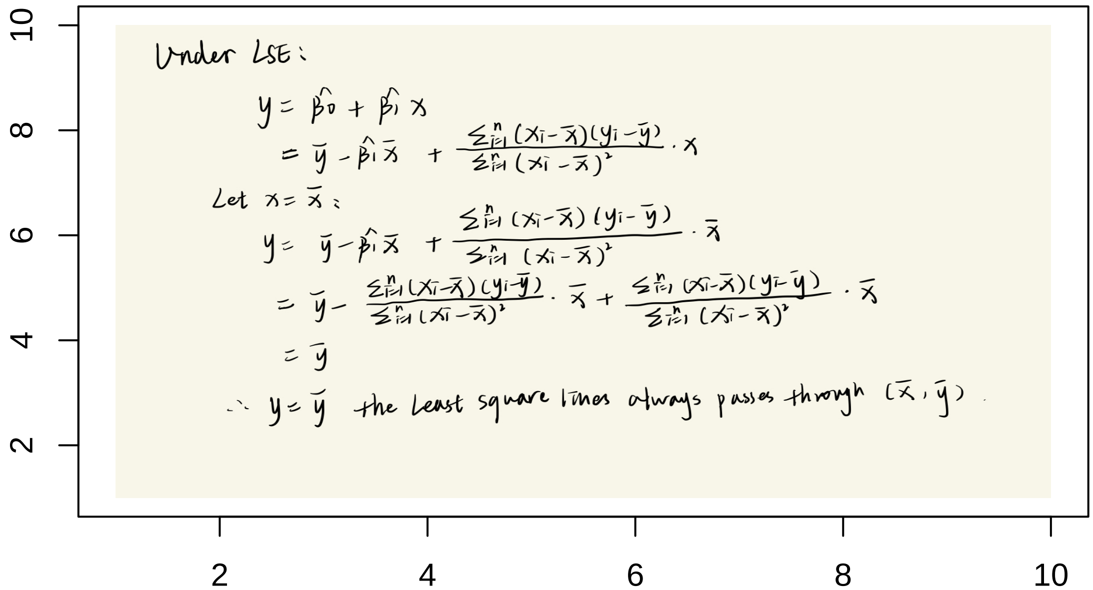
(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Answer: False. We must examine the p-value of the regression coefficient to determine if the interaction term is statistically significant or not.

Question 6 [25 marks]

Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

```
img <- image_read("Linear Regression Q3.png")
plot(1:10, type="n", ann=FALSE)
rasterImage(as.raster(img), 1, 1, 10, 10)
```



Question 8 [25 marks]

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` as the predictor. Use the summary() function to print the results. Comment on the output. For example:

```
model1 = lm(mpg ~ horsepower, data = Auto)
summary(model1)
```

```
##
## 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
```

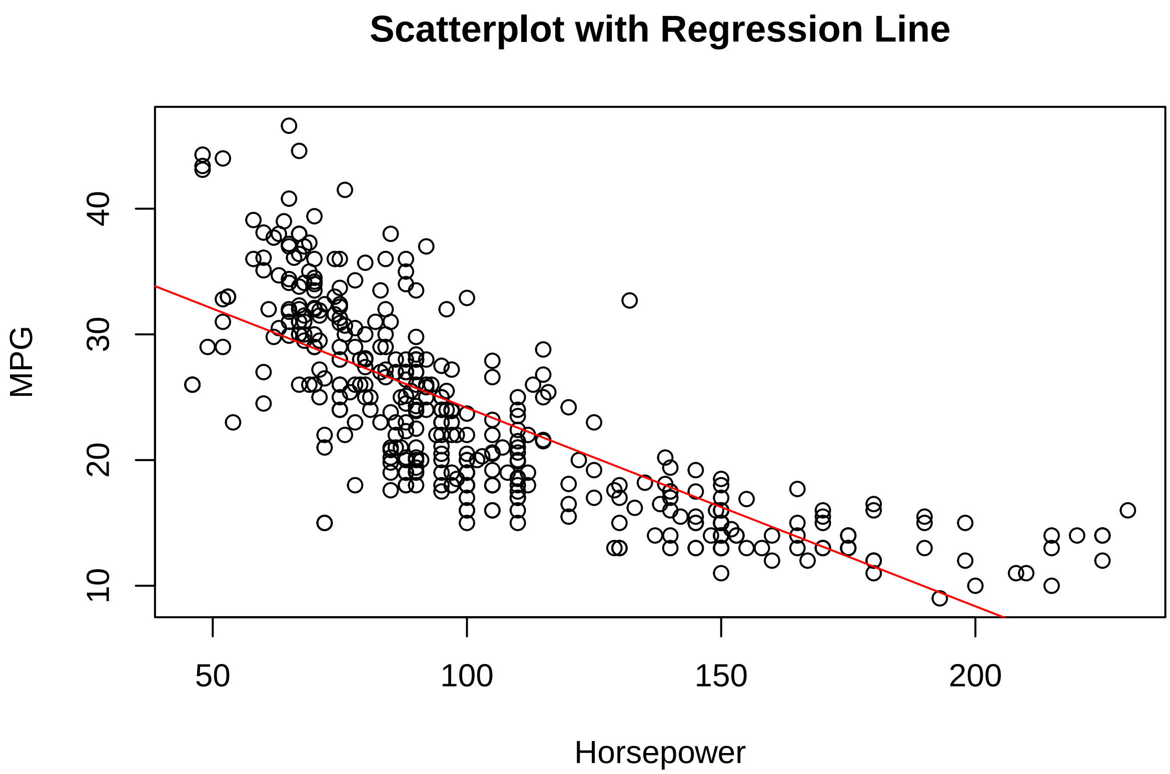
```r
#cor(Auto$mpg, Auto$horsepower)
```

i. Is there a relationship between the predictor and the response? **Answer:** Yes, there is relationship between the predictor and the response.
ii. How strong is the relationship between the predictor and the response? **Answer:** Not too strong for the R^2 is 0.6059, and just under two-thirds of the variability in `mpg` is explained by a linear regression on `horsepower`.
iii. Is the relationship between the predictor and the response positive or negative? **Answer:** Negative
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals? **Answer:** According to model1, we have

$$mpg = 39.935861 - 0.157845 * 98 = 24.46705$$

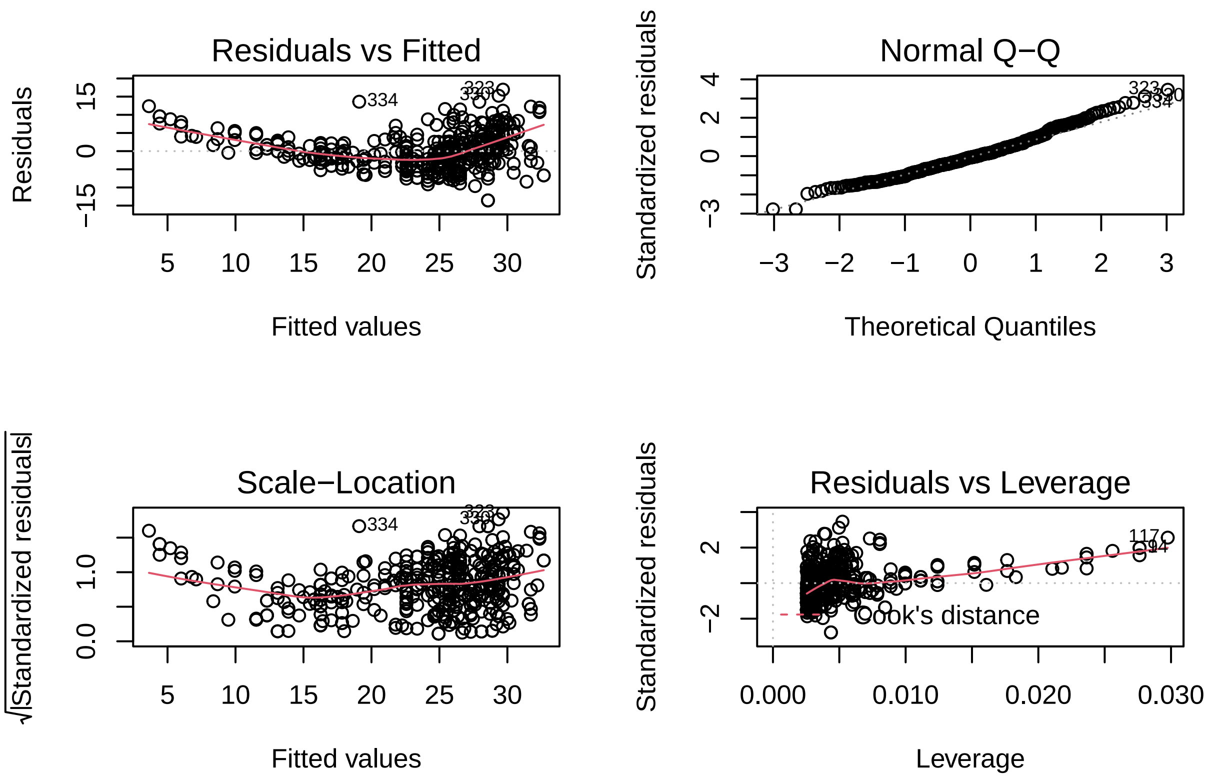
(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```r
plot(x = Auto$horsepower,y = Auto$mpg, main="Scatterplot with Regression Line", xlab="Horsepowe
abline(model1, col = 'red')
```



(c) Use the plot() function to produce diagnostic plots of the least squares regression.

```
par(mfrow=c(2,2)) # To display 4 plots in a 2x2 grid
plot(model1)
```



Question 11 [25 marks]

In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
set.seed (1)
x=rnorm (100)
y=2*x+rnorm (100)
```

(a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command lm(y□x+0).

```
lm1 = lm(y~x+0)
#summary_fit = summary(lm1)
#coef(lm1)
#summary_fit$coefficients["x", "Std. Error"]
```

```
#summary_fit$coefficients["x", "t value"]
#summary_fit$coefficients["x", "Pr(>|t|)"]
```

Answer: The coefficient estimate $\hat{\beta}$ is 1.9939, the standard error of this coefficient estimate is 0.1065, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$ are 18.73 and 2.642197e-34 separately. The p-value of the t-statistic is near to zero so we need to reject the null hypothesis.

(b) Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

```
lm2 = lm(x~y+0)
summary_fit2 = summary(lm2)
#summary_fit$coefficients["x", "Pr(>|t|)"]
```

Answer: The coefficient estimate is 0.3911, the standard error is 0.02089, and the corresponding t-statistic and p-values are 18.73 and 2.642197e-34 separately. The p-value of the t-statistic is near to zero so we need to reject the null hypothesis.

(c) What is the relationship between the results obtained in (a) and (b)? **Answer:** Both results in (a) and (b) reflect the same line created in 11a. In other words, y = 2x + ϵ could also be written x=0.5(y−ϵ).

(d) For the regression of Y onto X without an intercept, the t statistic for $H_0 : \beta = 0$ takes the form $\frac{\hat{\beta}}{SE(\hat{\beta})}$, where $\hat{\beta}$ is given by (3.38), and where

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^{n} x_{i'}^2}}$$

(These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$t = \frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n})(\sum_{i'=1}^{n} y_{i'}^2) - (\sum_{i'=1}^{n} x_{i'} y_{i'})^2}}$$

Answer: Based on the two formula above, we have

$$t = \frac{\beta}{\text{SE}(\beta)} = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2 - (\sum x_i y_i)^2}}$$

```r
(sqrt(length(x)-1) * sum(x*y)) / (sqrt(sum(x*x) * sum(y*y) - (sum(x*y))^2))
```

```
## [1] 18.72593
```

The t-statistic is 18.72593.

(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

```r
(sqrt(length(y)-1)*sum(y*x))/(sqrt(sum(y*y)*sum(x*x)-(sum(y*x)^2)))
```

```
## [1] 18.72593
```

The t-statistics for x onto y is 18.7253.

(f) In R, show that when regression is performed with an intercept, the t-statistic for $H_1 : \beta = 0$ is the same for the regression of y onto x as it is for the regression of x onto y.

```r
lm3 = lm(y~x)
lm4 = lm(x~y)
summary(lm3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389    0.698
## x            1.99894    0.10773  18.556   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
summary(lm4)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91    0.365
## y            0.38942    0.02099   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Answer: Based on the result above, we can see that the t-statistic value is the same for the two linear regression.

Question 14 [20 marks]

This problem focuses on the collinearity problem.

(a) Perform the following commands in R:

```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 x1 and x2. Write out the form of the linear model. What are the regression coefficients?

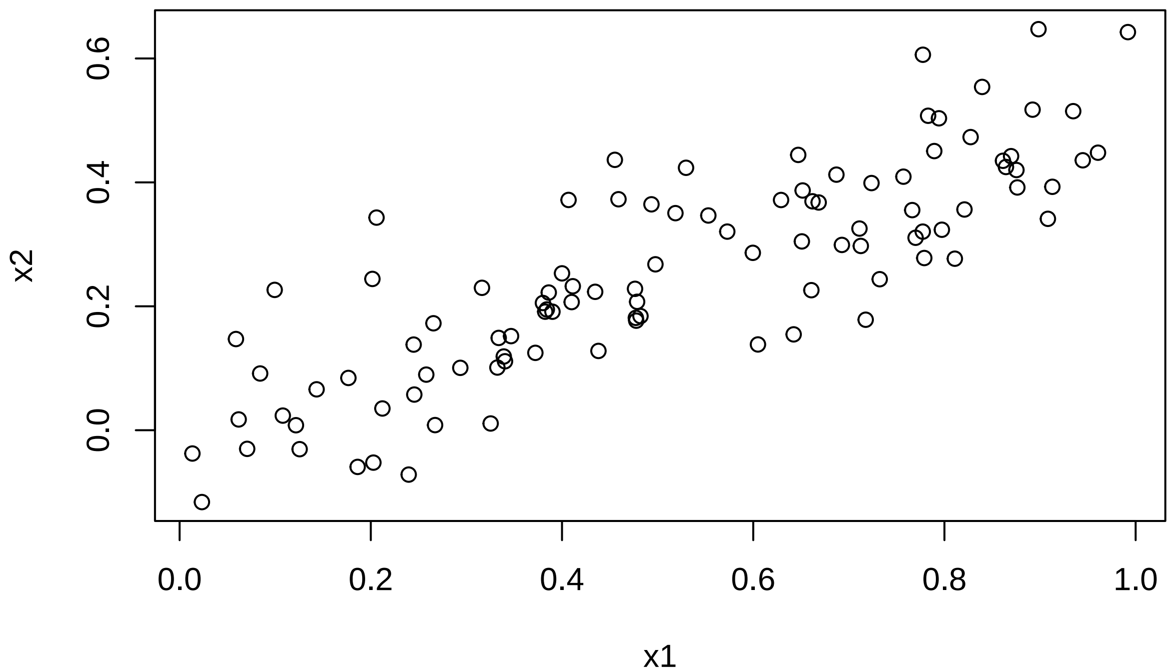
Answer: The form of the modle is multiple linear regression model. The regression coefficients are: $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

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

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

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



(c) Using this data, fit a least squares regression to predict y using x1 and x2. 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 $H_0 : \beta_2 = 0$?

```
lm.fit = lm(y~x1+x2)
summary(lm.fit)
```

```
##
## 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
```

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
lm.fit1 = lm(y~x1)
summary(lm.fit1)
```

```
##
## 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
```

Answer: the p-value of t-statistic is 2.661e-06, which is less than 0.05, so we can reject the null hypothesis.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```r
lm.fit2 = lm(y~x2)
summary(lm.fit2)
```

```
##
## 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
```

Answer: the p-value of t-statistic is 1.366e-05, which is less than 0.05, so we can reject the null hypothesis.

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

Answer: not contradict with each other because x1 and x2 have collinearity, it is hard to distinguish their effects when regressed upon together. When they are regressed upon separately, the linear relationship between y and each predictor is indicated more clearly.

(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)
```

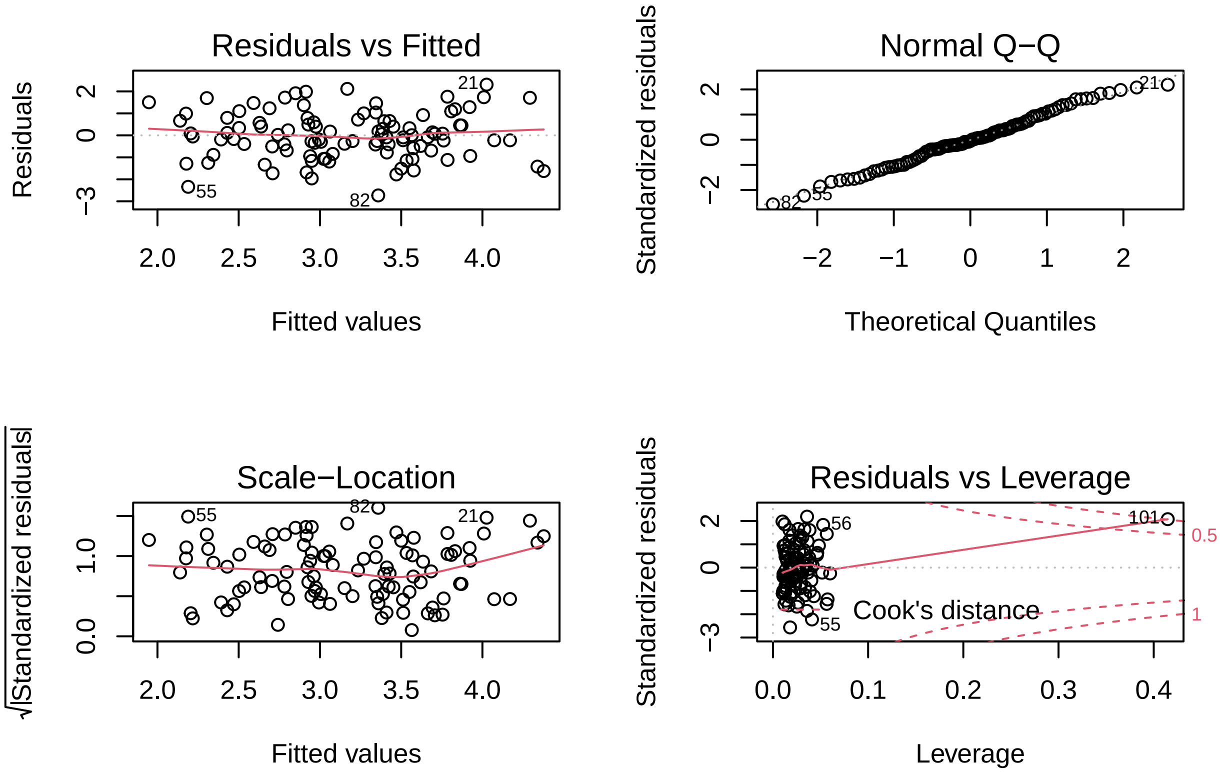
Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

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

```
##
## 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
```

In model1, x1 is not significant but x2 is significant.

```
par(mfrow=c(2,2))
plot(model_c)
```

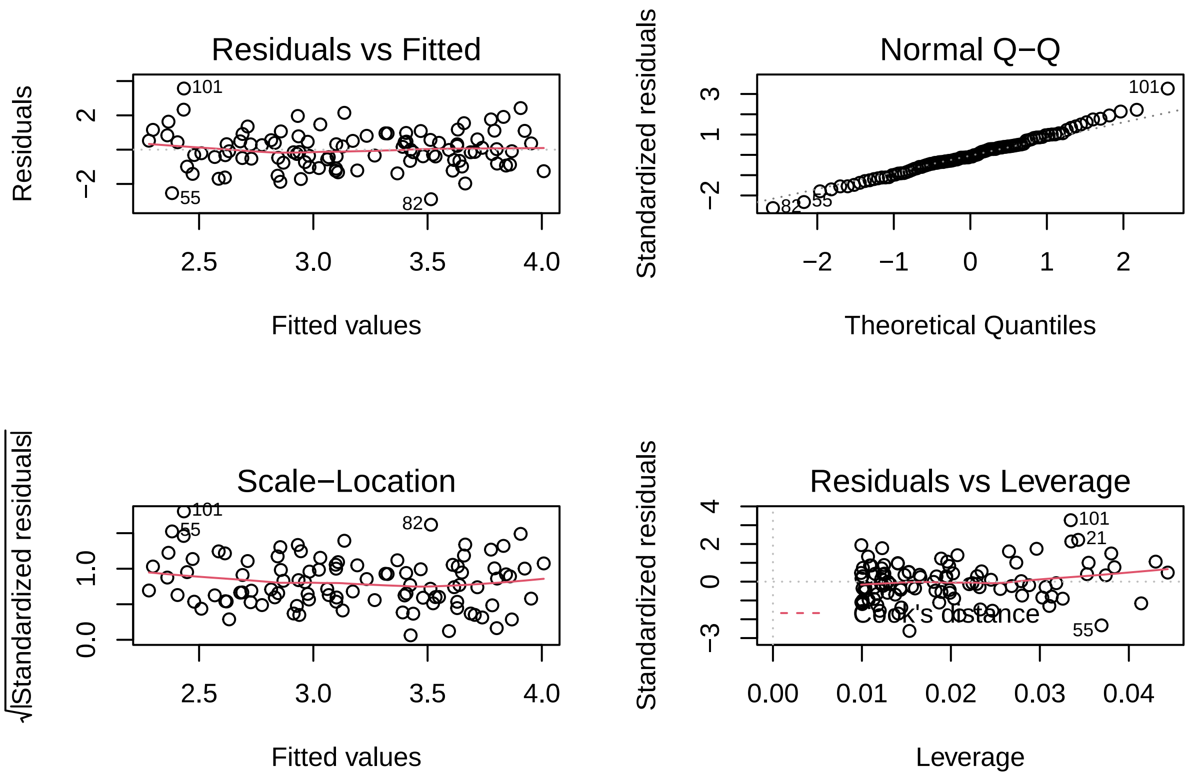



```r
model_d = lm(y~x1)
summary(model_d)
```

```
##
## 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
```

```r
par(mfrow=c(2,2))
plot(model_d)
```

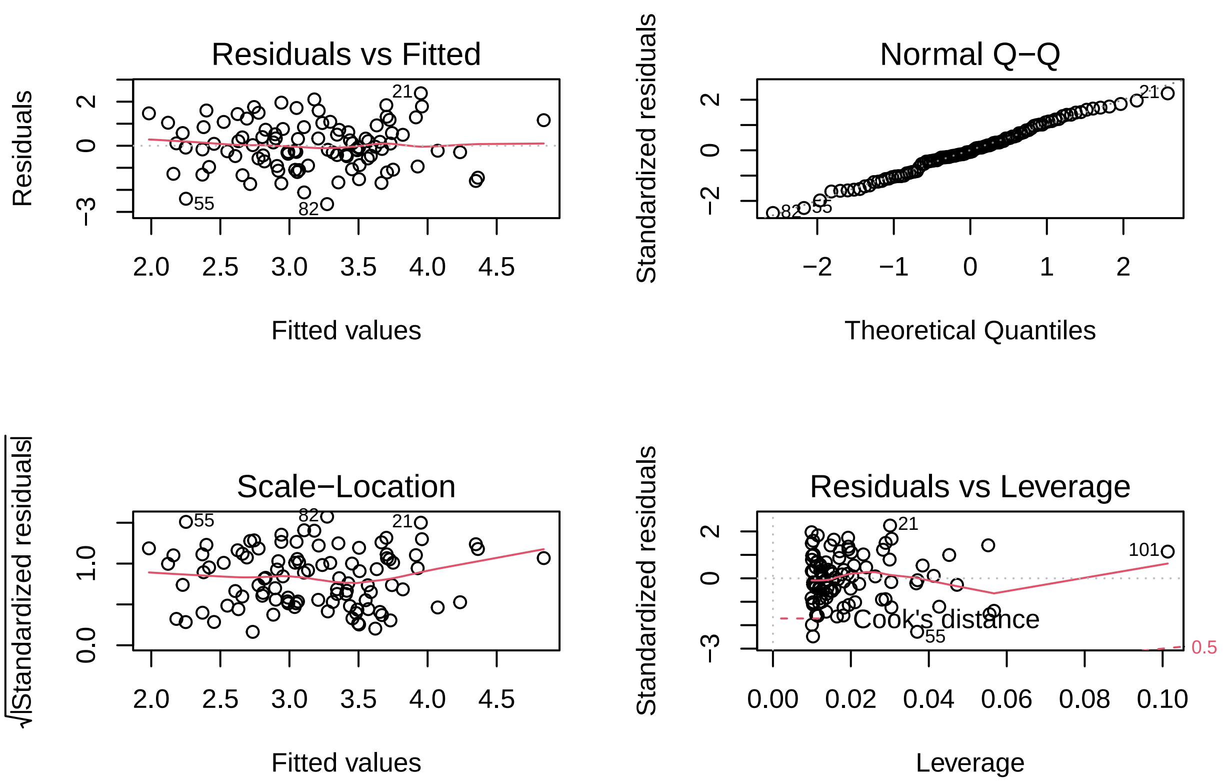



```r
model_e = lm(y~x2)
summary(model_e)
```

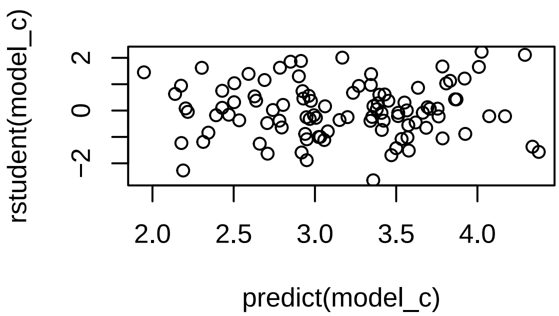
```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
```

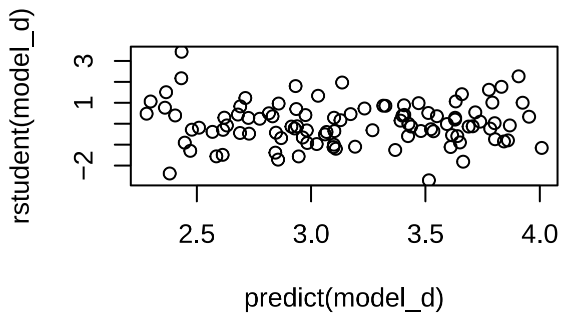
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 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
```

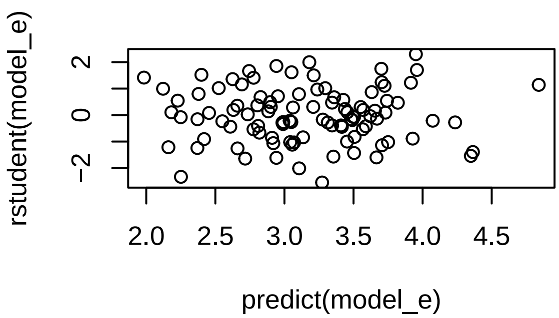
```
par(mfrow=c(2,2))
plot(model_e)
```




```
par(mfrow=c(2,2))
plot(predict(model_c), rstudent(model_c))
plot(predict(model_d), rstudent(model_d))
plot(predict(model_e), rstudent(model_e))
```







By checking the residual plots above, we find only model_c(y~x1+x2) has outlier observation.

BONUS [10 marks]: If you were to conduct the simulation in a) many times (with different draws for the random variables) so that you had a large number of datasets, what you would expect the average values (over all simulations) for the coefficients from the regression in c) to be? Explain your answer.