Chapter3

Tong Sun

1/23/2022

#3.1 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.

From Table 3.4, the null hypothesis for "TV" is that with the "radio" and "newspaper" existing, "TV" advertising has no influence on sales. The null hypothesis for "radio" is that with the "TV" and "newspaper" existing, "radio" advertising has no effect on sales. And the null hypothesis for "newspaper" is that with the "TV" and "radio" existing, "newspaper" advertising does not affect sales. The low p-values of "TV" and "radio" suggest that we should reject the null hypothesis for "TV" and "radio" and the high p-value of "newspaper" suggests that we should accept the null hypothesis for "newspaper".

#3.2 Carefully explain the differences between the KNN classifier and KNN regression methods.

The main difference is the fact that for the classifier approach, the algorithm assumes the outcome as the class of more presence, and on the regression approach the response is the average value of the nearest neighbors.

#3.5 Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form $\hat{y}_i = x_i\hat{\beta}$, where $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$.Show that we can write $\hat{y}_i = \sum a_j y_j$. What is a_j? Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

For the ith fitted value takes the form $\hat{y}_i = X_i\hat{\beta}$, where $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_j^2}$. And$\hat{y}_i = \sum a_j y_j$. We have,$\hat{y}_i = X_i\hat{\beta} = X_i\frac{\sum x_i y_i}{\sum x_j^2} = \sum \frac{x_j x_i}{\sum x_k^2} y_j$ Because variables' names do not matter inside of $\sum$, we have,$a_j = \frac{x_i x_j}{\sum x_k^2}$

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

We have equation1:$y = \beta_0 + \beta_1 x$, and from (3.4), we got that equation2:$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, also $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$. So from equation1,we have, $0 = \beta_0 + \beta_1 x - y$. If $(\bar{x}, \bar{y})$ is on the line, we have: $0 = \beta_0 + \beta_1\bar{x} - \bar{y}$. Also from equation2, we have$\bar{y} = \beta_0 + \beta_1\bar{x}$. So, we get $0 = \beta_0 + \beta_1\bar{x} - (\beta_0 + \beta_1\bar{x})$, $0 = 0$, which makes sense. In this case, we have a conclusion that the least squares line always passes through $(\bar{x}, \bar{y})$.

#3.11

###(a)

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The p-value of t-statistic is almost zero so we should reject the null hypothesis.

###(b)

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The p-value of t-statistic is near zero so we should reject the null hypothesis.

###(c)

Both outcomes from (a) and (b) are the same thing.

$$y = 2 * x + \epsilon$$

, which equals to,

$$x = 0.5 * (y - \epsilon)$$

.

###(d)

For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $t = \beta/SE(\beta)$, where $\hat{\beta}$ is given by (3.38), and where$SE(\hat{\beta}) = \sqrt{\frac{\sum (y_i - x_i\hat{\beta})^2}{(n-1)\sum x_i^2}}$. Firstly, show algebraically, the t-statistic can be written as $\frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum (y_i^2 - 2\beta x_i y_i + x_i^2\beta^2)}}$. We have $t = \frac{\sum x_i y_i}{\sum x_i^2}\sqrt{\frac{(n-1)\sum x_i^2}{\sum (y_i - x_i\beta)^2}}$, and$t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum (y_i^2 - 2\beta x_i y_i + x_i^2\beta^2)}}$, next we get$t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2 - \sum x_i y_i (2\sum x_i y_i - \sum x_i y_i)}}$.

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

###(e)

Here we change y into x and x into y, the formula from (d) will not change any more.

###(f)

```
##
## 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
```

```
##
## 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
```

The results of t-statistic are the same.

#3.12

###(a)

When the sum of the squares of the observed y are equal to the sum of the squares of the observed x, the coefficients will be the same.

###(b)

```
## Warning in summary.lm(fit5): essentially perfect fit: summary may be unreliable

##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -8.558e-16 -4.290e-17  4.600e-18  1.410e-16  1.173e-14
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## x 3.000e+00  1.334e-16 2.249e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.201e-15 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 5.059e+32 on 1 and 99 DF,  p-value: < 2.2e-16

## Warning in summary.lm(fit6): essentially perfect fit: summary may be unreliable

##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.624e-15 -4.547e-17 -3.270e-18  3.747e-17  2.846e-16
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## y 3.333e-01  1.022e-17 3.261e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.761e-16 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.064e+33 on 1 and 99 DF,  p-value: < 2.2e-16
```

###(c)

```
## [1] 81.05509

## [1] 81.05509

##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.1665 -0.4995  0.1140  0.6945  2.2833
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x -0.07768    0.10020  -0.775     0.44
##
## Residual standard error: 0.9021 on 99 degrees of freedom
## Multiple R-squared:  0.006034,   Adjusted R-squared:  -0.004006
## F-statistic: 0.601 on 1 and 99 DF,  p-value: 0.4401
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2182 -0.4969  0.1595  0.6782  2.4017
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y -0.07768    0.10020  -0.775     0.44
##
## Residual standard error: 0.9021 on 99 degrees of freedom
## Multiple R-squared:  0.006034,   Adjusted R-squared:  -0.004006
## F-statistic: 0.601 on 1 and 99 DF,  p-value: 0.4401
```

#3.13

###(a)

###(b)

###(c)

The length of y is 100.

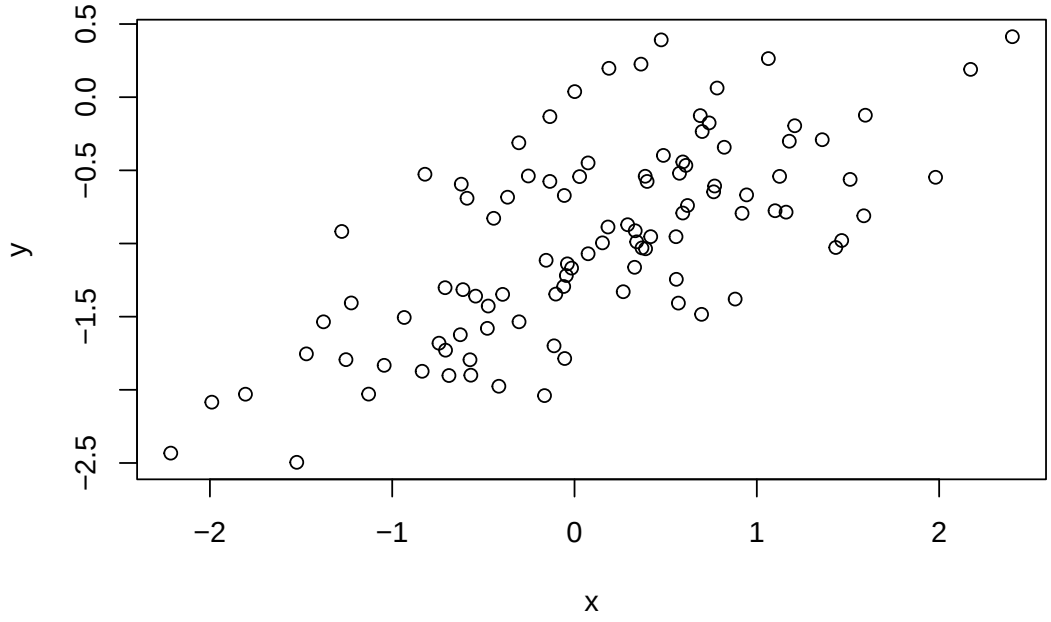
$$\beta_0 = -1$$

,and

$$\beta_1 = 0.5$$

.

###(d)



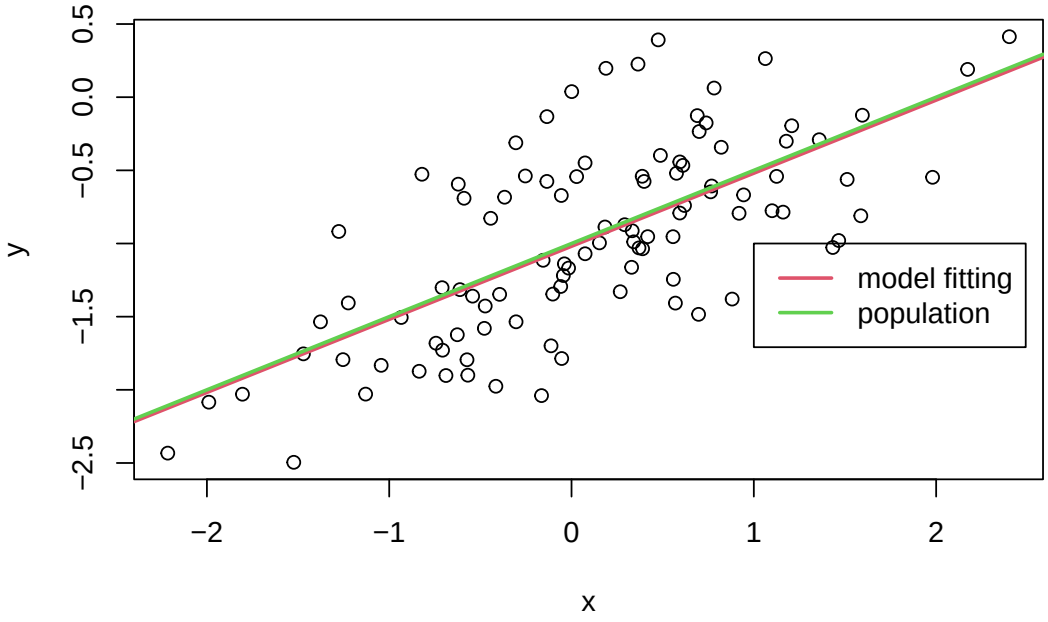
There is a linear relationship between y and x with a positive slope.

###(e)

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

This model has a coefficient that is close to that we generated in (c), and the p-values are near zero so we could reject the null hypothesis.

###(f)



###(g)

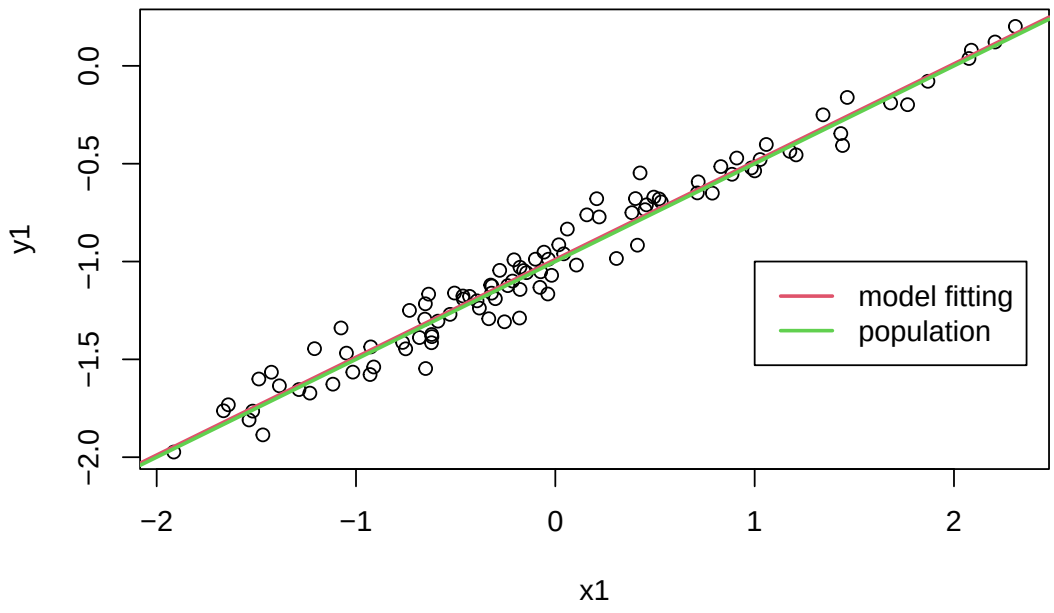
```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
## I(x^2)      -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

The model fitting is better after adding x^2 into the model because the R^2 increased slightly, but the RSE decreased a little. The p-value of the t-statistic is 0.164 which means there isn't any relationship between y and x^2.

###(h)

```
##
```

```
## Call:
## lm(formula = y1 ~ x1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.232416 -0.060361  0.000536  0.058305  0.229316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.989115   0.009035 -109.48   <2e-16 ***
## x1           0.499907   0.009472   52.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09028 on 98 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9657
## F-statistic:  2785 on 1 and 98 DF,  p-value: < 2.2e-16
```

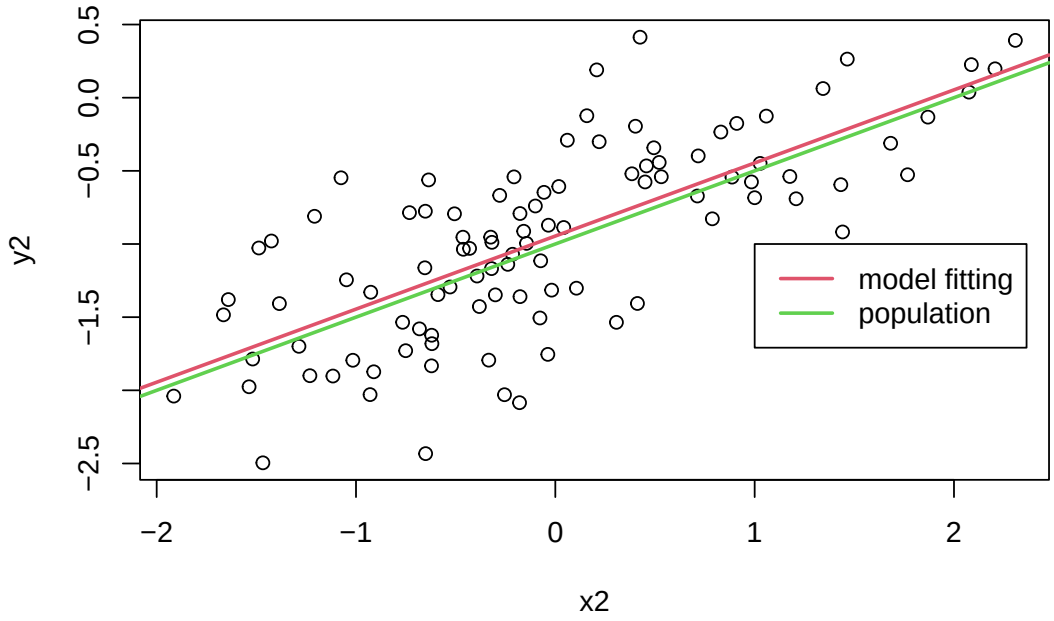


From the result, we can find that the RSE decreases considerably, changing into 0.09.

###(i)

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.16208 -0.30181  0.00268  0.29152  1.14658
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.94557    0.04517  -20.93   <2e-16 ***
## x2           0.49953    0.04736   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4514 on 98 degrees of freedom
## Multiple R-squared:  0.5317, Adjusted R-squared:  0.5269
## F-statistic: 111.2 on 1 and 98 DF,  p-value: < 2.2e-16
```



The error observed in R^2 and RSE increased considerably.

###(j)

```
##                  2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0070441 -0.9711855
## x1           0.4811096  0.5187039
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0352203 -0.8559276
## x2           0.4055479  0.5935197
```

The second and third fits' intervals are both narrower than the first fit's interval,although the third one is similar with the first one. All intervals seem to be centered on approximately 0.5.

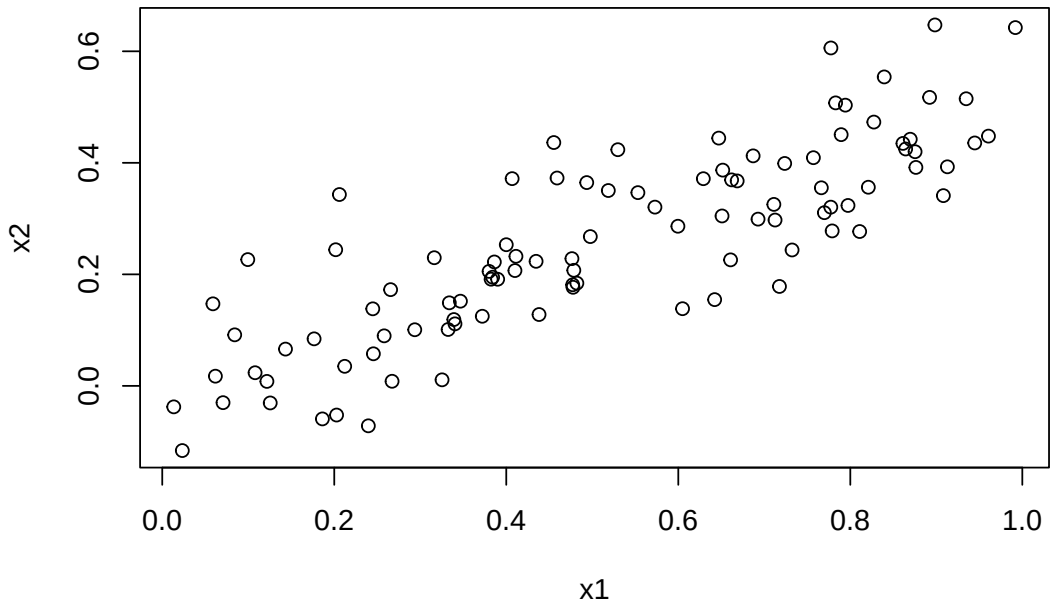
#3.14

###(a)

The regression coefficients are: $\beta_0 = 2, \beta_1 = 2, \beta_3 = 0.3$

###(b)

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



###(c)

```
##
## 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
```

$\beta_0 = 2.1305, \beta_1 = 1.4396, \beta_2 = 1.0097$

The regression coefficients are similar to the true values, but with some standard error. For the p-values of β_0 and β_1, they are both below 5%, so we should reject the null hypothesis for these two coefficients. However, for β_2, the p-value is 0.3754, much above 5% typical cutoff, we cannot reject the null hypothesis for this coefficient.

###(d)

```
##
## 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
```

Yes, the null hypothesis for this regression coefficient can be rejected because the p-value for its t-statistic is close to zero.

###(e)

```
##
## 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
```

Yes, we can reject the null hypothesis for this regression given the p-value for its t-statistic is near zero.

###(f)

No, it doesn't contradict. Because there is a linear relationship between x1 and x2, it's difficult to distinguish their influences on the model when using them to make regression at the same time. But when they are regressed separately, it is more clearly to see the relationship between y and each predictor.

###(g)

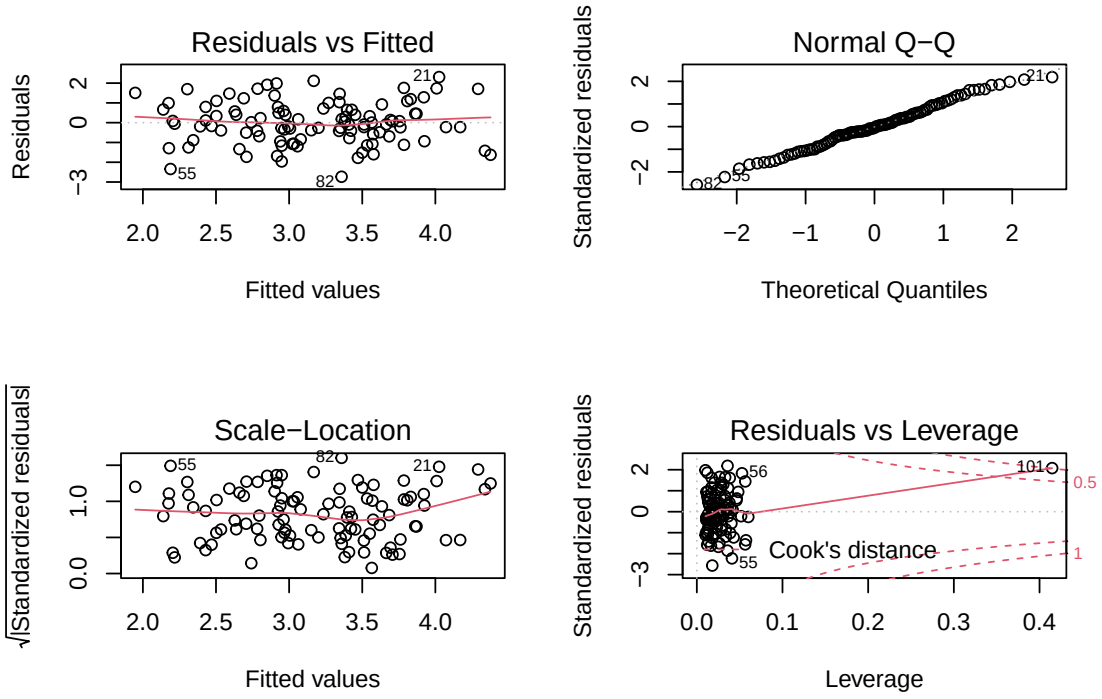
```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            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
```

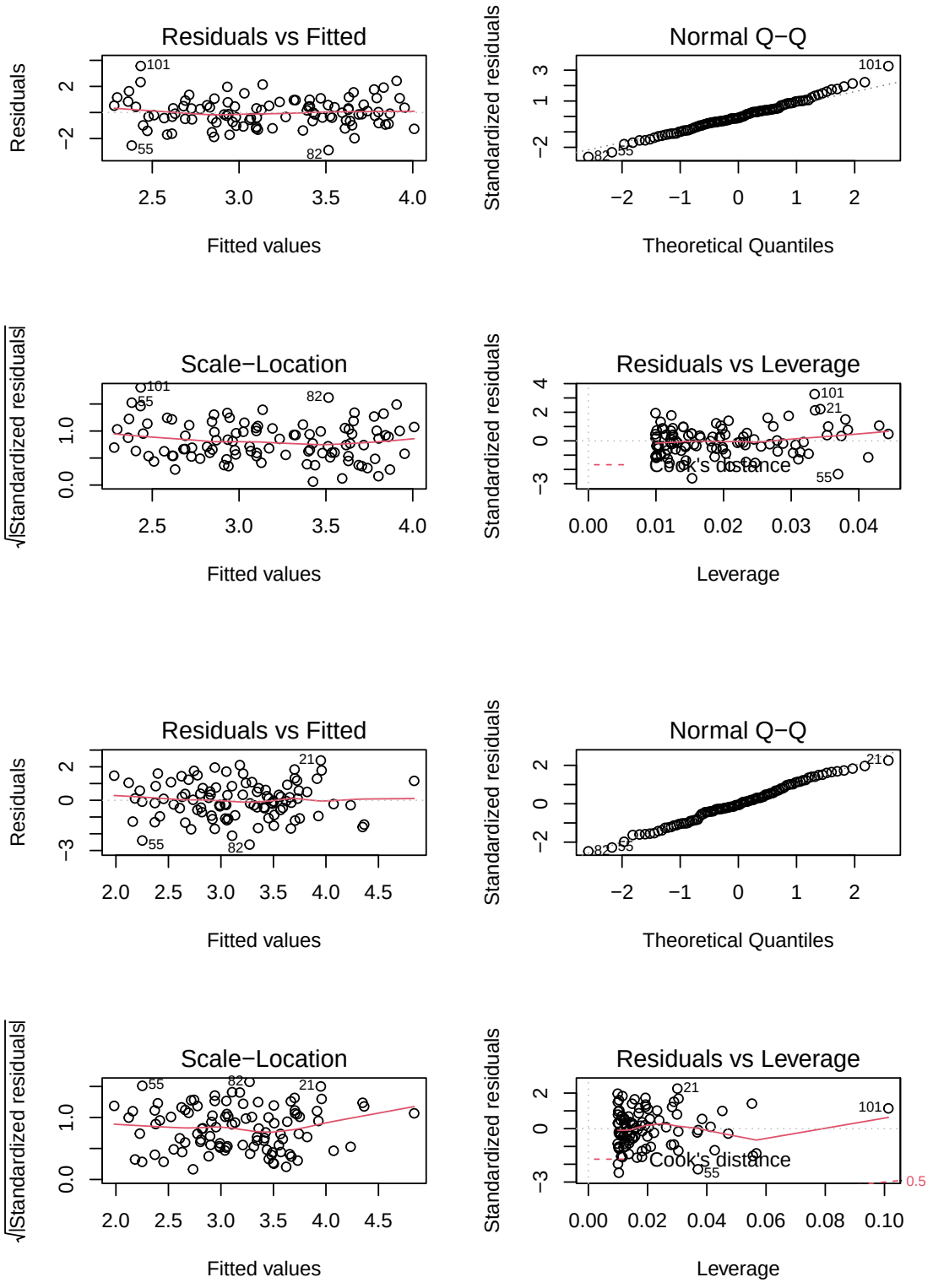
```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   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
```

```
##
## 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
```

For the first regressed model that including x1 and x2, the coefficient of x1 is not significant but that of x2 is statistically significant.





From the leverage plot of these three models, we can say the second one – the point does not become a high leverage point. And the first and third ones do.