MA679 Linear Regression Homework

Jiachen Feng

2021/2/3

3.1

The null hypothesis is **There is no relationship between Sales and the these kinds of advertising budgets**. Based on the p-values, we can conclude that there is some relationship between sales and TV advertising budgets, and also some relationship between sales and radio advertising budgets, for the reason the p-values are extremely small. Additionally, it seems that there is no relationship between sales and newspaper advertising budgets.

3.2

The KNN regression methods is closely related to the KNN classifier. KNN classifier decides the prediction point to a certain class, however, KNN regression methods is a method calculating the average of all responses. KNN classifier puts a categorical variable as outcomes, and KNN regression methods is able to process quantitative variables.

3.5

Through mathematical transformation, $a_{i^1} = \frac{x_i x_{i1}}{\sum_{j=1}^{n} {x_j}^2}$

3.6

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

Therefore, $\bar{y} = \hat{y} - \hat{\beta}_1 x + \hat{\beta}_1 \bar{x}, \ldots$

3.11

(a)

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

fit_1 <- lm(y~x+0)
summary(fit_1)
```

```
##
## 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
```

Based on the p-value, there is a relationship between y and x. The coefficient 1.9939 is close to 2, which means the result of this regression is convincing.

(b)

```
fit_2 <- lm(x~y+0)
summary(fit_2)
```

```
##
## 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 suggests that there is a relationship between x and y. However, the coefficient 0.3911 is not very close to 0.5. Considering the number of the sample point is 100(not large), the result is reasonable.

(c) The sample points are the same. The t-statistics and p-value associated with the null hypothesis are the same.

(d)

```
X <- data.frame(x)
Y <- data.frame(y)
dat <- cbind(X,Y)
dat$xy <- dat$x*dat$y
dat$x2 <- dat$x^2
dat$y2 <- dat$y^2

sebeta <- (sqrt(100-1)*sum(dat$xy))/(sqrt(sum(dat$x2)*sum(dat$y2)-(sum(dat$xy)^2)))
sebeta
```

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

The *sebeta* value calculated by R is 18.72593, and the t-value is 18.73. The two value is close.

(e) The sample points are the same. For the formula from (d), the results are the same.

(f)

```
fit_3 <- lm(y~x)
fit_4 <- lm(x~y)
summary(fit_3)
```

```
##
## 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
```

```
summary(fit_4)
```

```
##
## 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
```

3.12

(a) Y=X.

(b)

```
summary(fit_1)
```

```
##
```

```
## 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
```

```
summary(fit_2)
```

```
##
## 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
```

(c)

```
set.seed(1)
x <- rnorm(100)
y <- x

fit_1 <- lm(y~x+0)
summary(fit_1)
```

```
## Warning in summary.lm(fit_1): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.888e-16 -1.689e-17  1.339e-18  3.057e-17  2.552e-16
##
```

```
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## x 1.000e+00  6.479e-18 1.543e+17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.833e-17 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.382e+34 on 1 and 99 DF,  p-value: < 2.2e-16
```

```
fit_2 <- lm(x~y+0)
summary(fit_2)
```

```
## Warning in summary.lm(fit_2): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.888e-16 -1.689e-17  1.339e-18  3.057e-17  2.552e-16
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## y 1.000e+00  6.479e-18 1.543e+17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.833e-17 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.382e+34 on 1 and 99 DF,  p-value: < 2.2e-16
```

3.13

(a)

```
set.seed(1)
x <- rnorm(100)
```

(b)

```
eps <- rnorm(100,sd=.5)
```

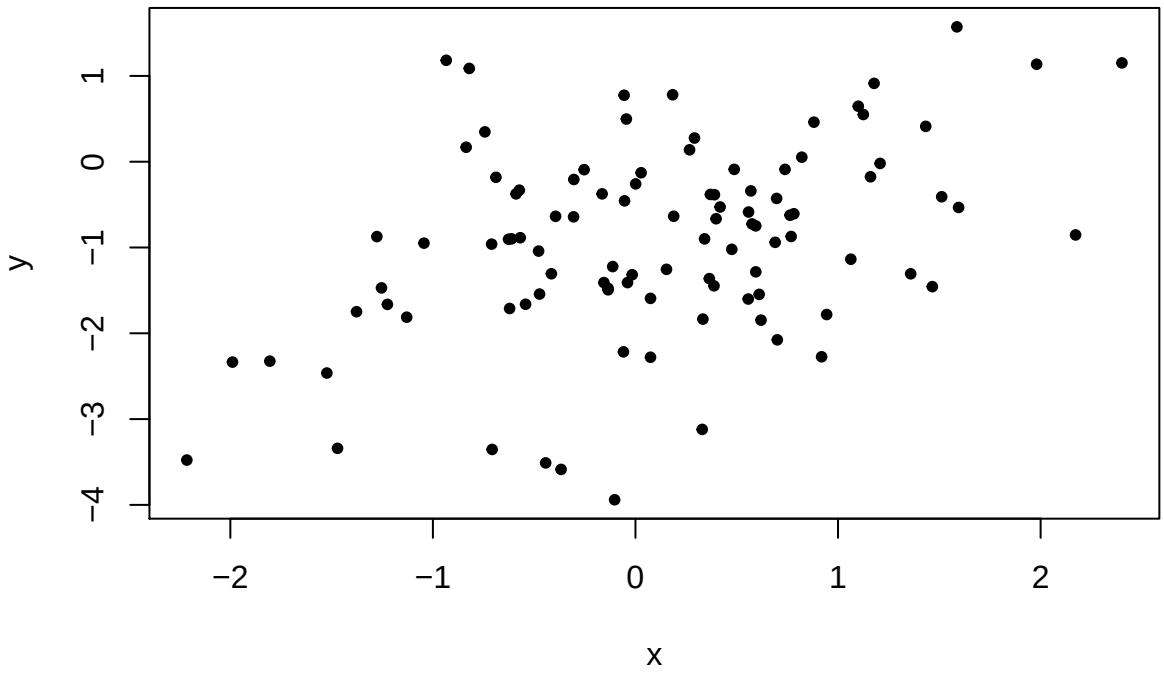
(c)

```
y <- -1+.5*x+rnorm(100)
```

The length is 100, $\beta_0 = -1$,$\beta_1 = 0.5$.

(d)

```r
plot(x,y,pch=20)
```



(e)

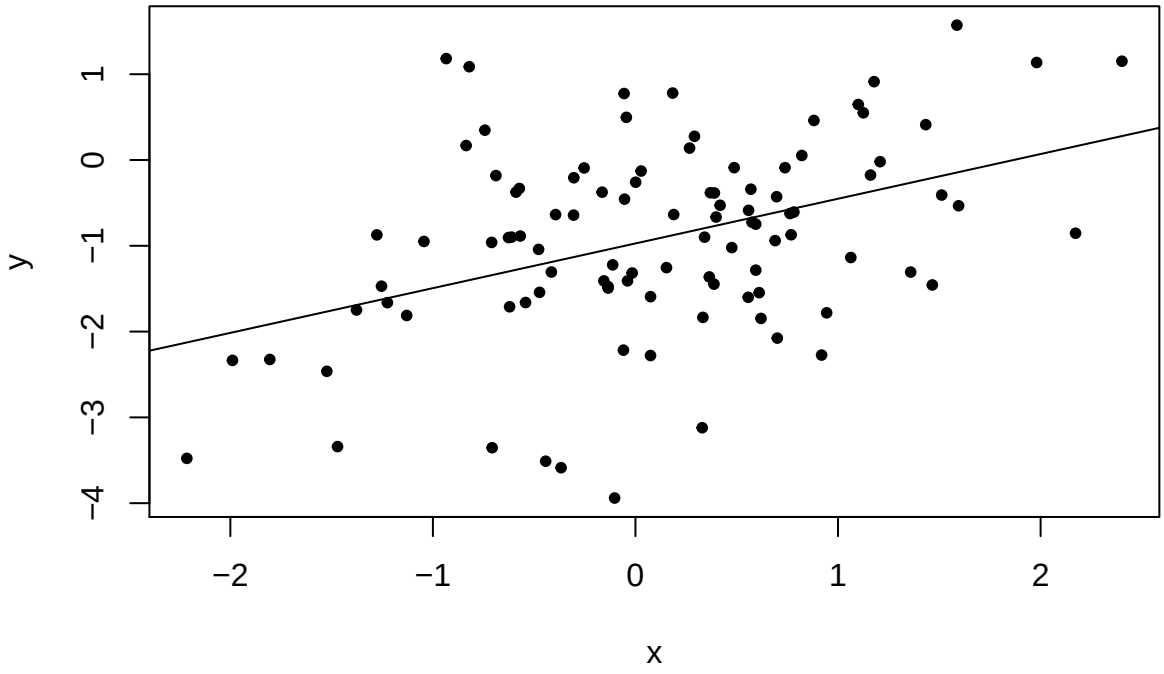
```r
fit_5 <- lm(y~x)
summary(fit_5)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91411 -0.48230 -0.04533  0.64924  2.64157
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9726     0.1047  -9.289 4.22e-15 ***
## x             0.5212     0.1163   4.481 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1616
## F-statistic: 20.08 on 1 and 98 DF,  p-value: 2.013e-05
```

$\hat{\beta}_0$ is close to β_0, and $\hat{\beta}_1$ is close to β_1.

(f)

```
plot(x,y,pch=20)
abline(coef(fit_5)[1],coef(fit_5)[2])
```



(g)

```
fit_6 <- lm(y~x+x^2)
summary(fit_6)
```

```
##
## Call:
## lm(formula = y ~ x + x^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91411 -0.48230 -0.04533  0.64924  2.64157
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9726     0.1047  -9.289 4.22e-15 ***
## x             0.5212     0.1163   4.481 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1616
## F-statistic: 20.08 on 1 and 98 DF,  p-value: 2.013e-05
```

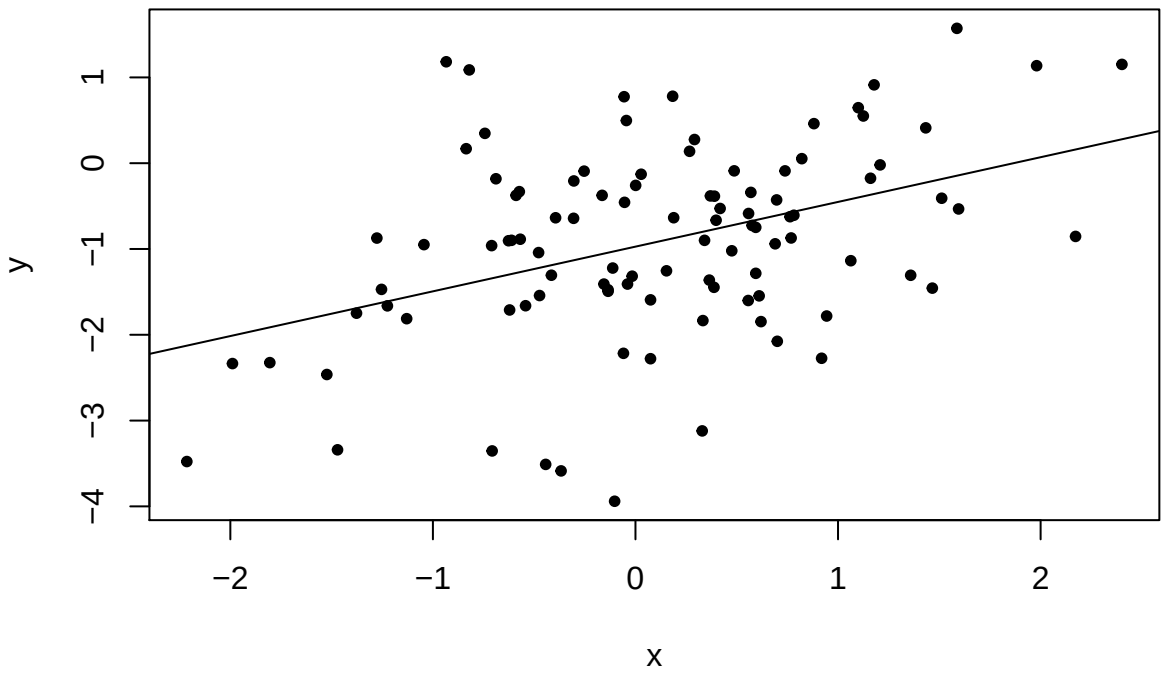
No. The p-value and t-value do not change.

(h)

```
set.seed(1)
x <- rnorm(100)
eps <- rnorm(100,sd=.01)
y <- -1+.5*x+rnorm(100)
fit_7 <- lm(y~x)
summary(fit_7)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91411 -0.48230 -0.04533  0.64924  2.64157
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9726     0.1047  -9.289 4.22e-15 ***
## x             0.5212     0.1163   4.481 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1616
## F-statistic: 20.08 on 1 and 98 DF,  p-value: 2.013e-05
```

```
plot(x,y,pch=20)
abline(coef(fit_7)[1],coef(fit_7)[2])
```




```
fit_8 <- lm(y~x+x^2)
summary(fit_8)
```

```
##
## Call:
## lm(formula = y ~ x + x^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91411 -0.48230 -0.04533  0.64924  2.64157
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9726     0.1047  -9.289 4.22e-15 ***
## x             0.5212     0.1163   4.481 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1616
## F-statistic: 20.08 on 1 and 98 DF,  p-value: 2.013e-05
```

The result does not seem to change.

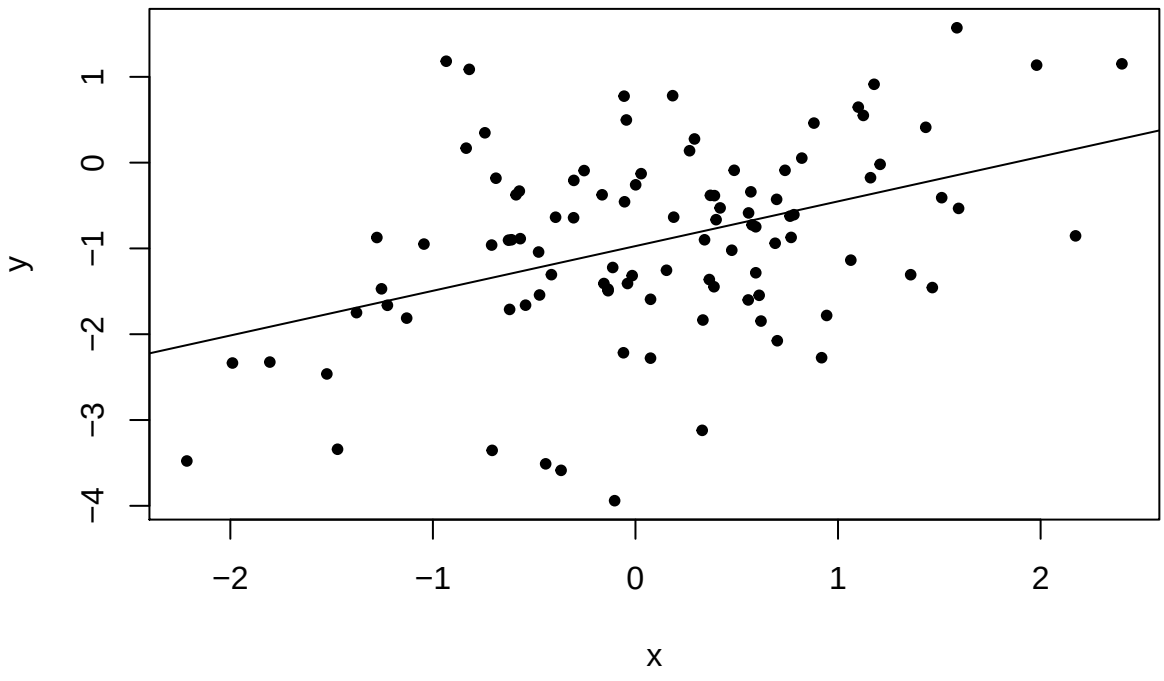
(i)

```
set.seed(1)
x <- rnorm(100)
```

```r
eps <- rnorm(100,sd=100)
y <- -1+.5*x+rnorm(100)
fit_9 <- lm(y~x)
summary(fit_9)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91411 -0.48230 -0.04533  0.64924  2.64157
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9726     0.1047  -9.289 4.22e-15 ***
## x             0.5212     0.1163   4.481 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1616
## F-statistic: 20.08 on 1 and 98 DF,  p-value: 2.013e-05
```

```r
plot(x,y,pch=20)
abline(coef(fit_9)[1],coef(fit_9)[2])
```



```r
fit_10 <- lm(y~x+x^2)
summary(fit_10)
```

```
##
## Call:
## lm(formula = y ~ x + x^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91411 -0.48230 -0.04533  0.64924  2.64157
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9726     0.1047  -9.289 4.22e-15 ***
## x             0.5212     0.1163   4.481 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1616
## F-statistic: 20.08 on 1 and 98 DF,  p-value: 2.013e-05
```

The result does not seem to change.

(j)

```r
confint(fit_5)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.1804128 -0.7648497
## x            0.2903769  0.7519568
```

```r
confint(fit_7)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.1804128 -0.7648497
## x            0.2903769  0.7519568
```

```r
confint(fit_9)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.1804128 -0.7648497
## x            0.2903769  0.7519568
```

They are the same.

3.14

(a)

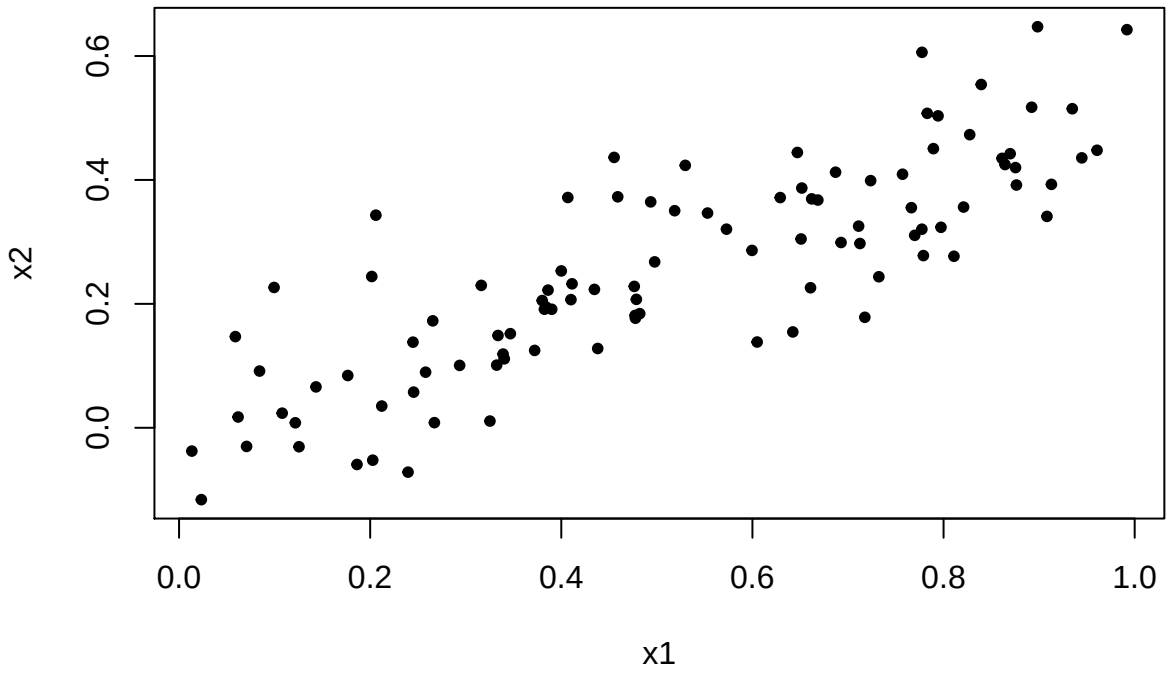
```r
set.seed(1)
x1 <- runif(100)
x2 <- 0.5*x1+rnorm(100)/10
y <- 2+2*x1+0.3*x2+rnorm(100)
```

$$y = 2 + 2 * x_1 + 0.3 * x_2 + \epsilon$$

The coefficients are 2,2,0.3, relatively.

(b)

```
plot(x1,x2,pch=20)
```



$$y = 0.5 * x_1 + 0.1 * \epsilon$$

(c)

```
fit_11 <- lm(y~x1+x2)
summary(fit_11)
```

```
##
## 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
```

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ correspond to the first column of the output result. The null hypothesis $H_0 : \beta_1 = 0$ is rejected, and the null hypothesis $H_1 : \beta_2 = 0$ is retained.

(d)

```
fit_12 <- lm(y~x1)
summary(fit_12)
```

```
##
## 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
```

The coefficients are close to $\hat{\beta}_0, \hat{\beta}_1$. The null hypothesis can be rejected, because the p-value is less than 0.05.

(e)

```
fit_13 <- lm(y~x2)
summary(fit_13)
```

```
##
## 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
```

The null hypothesis can be rejected, because the p-value is less than 0.05.

(f) No. The multiple linear regression is used to fit a model with two predictors, and the two predictors have interaction between each other, for the given formula $x_2 = 0.5 * x_1 + 0.1 * \epsilon$.

(g)

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

fit_14 <- lm(y~x1+x2)
summary(fit_14)
```

```
##
## 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
```

```
fit_15 <- lm(y~x1)
summary(fit_15)
```

```
##
## 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
```

```
fit_16 <- lm(y~x2)
summary(fit_16)
```

```
##
## 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
```

The new observation is an outlier, because the new observation is far from the previously generated fit line.