LinearRegression-HW

Franky Zhang

1/25/2022

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.

***Answer*:** 1)TV: null hypothesis: TV is not related with sales. since p-value of TV is smaller than .0001, we can successfully reject null hypothesis and TV is related to sales under this multiple linear regression model.

2)radio: null hypothesis: radio is not related with sales. since p-value of radio is smaller than .0001, we can successfully reject null hypothesis and radio is related to sales under this multiple linear regression model.

3)newspaper: null hypothesis: newspaper is not related with sales. while p0-vlaue of newspaper is .86, it means there is a prob of 86% that the data exists under null hypothesis. hence, we fail to reject null hypothesis, and we are not sure whether newspaper is related to sales.

3.2

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

***Answer*:** KNN classifier attempts to predict the class to which the output variable belong by computing the local probability. KNN regression gives prediction of output variable by using local average.

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

$$\hat{\beta} = (\sum_{i=1}^{n} x_i y_i)/(\sum_{i'=0}^{n} x_{i'}^2)$$

Show that we can write

$$\hat{y_i} = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

what is $a_{i'}$.

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

***Answer*:** take $\hat{\beta} = (\sum_{i=1}^{n} x_i y_i)/(\sum_{i=1}^{n} x_{i'}^2)$ in to $\hat{y_i} = x_i \cdot \beta$:

$$\hat{y_I} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2} \cdot x_I$$

$$= \sum_{i=1}^{n} (\frac{x_i}{\sum_{i'=1}^{n} x_{i'}^2} y_i \cdot x_I)$$

thus,

$$\hat{y_i} = \sum_{i'=1}^{n} (\frac{x_i'}{(\sum_{i''=1}^{n} x_{i''}^2)} y_i' \cdot x_i)$$

hence, $a_{i'}$ here is

$$\frac{x_i'}{(\sum_{i''=1}^{n} x_{i''}^2)} \cdot x_i$$

3.6

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

***Answer*:** the number of observation points is n, $\hat{\beta_0}$ and $\hat{\beta_1}$ are estimated coefficients by LSE method. let the regression model be:

$$y_i = \hat{\beta_0} + \hat{\beta_1} \times x_i + \epsilon_i$$

sum them up:

$$\sum_{i=1}^{n} y_i = n \times \hat{\beta_0} + \hat{\beta_1} \times \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \epsilon_i$$

$$\sum_{i=1}^{n} y_i = n \times \hat{\beta_0} + \hat{\beta_1} \times \sum_{i=1}^{n} x_i$$

then take average:

$$\bar{y} = \hat{\beta_0} + \hat{\beta_1} \times \bar{x}$$

and this exactly prove $(\bar{x}, \bar{y}$ is on the least square line.

3.11

In this problem we will investigate the t-statistic for the null hypoth- esis H0 : $\beta = 0$ 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)

***Answer*:**

```
##
## 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
```

Comment: the estimated value $\hat{\beta}$ is 1.99 and *p_value* here is <2e-16, based on which we can reject Null hypothesis: $\beta = 0$.

(b)

***Answer*:**

```
##
## 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
```

Comment: the estimated value $\hat{\beta}$ is 0.39 and *p_value* here is <2e-16, based on which we can reject Null hypothesis: $\beta = 0$.

(c)

***Answer*:**

$$(\sum_{i=1}^{n} x_i^2) \cdot \hat{\beta}_x = (\sum_{i=1}^{n} y_i^2) \cdot \hat{\beta}_y$$

(d)

***Answer*:** Already know that

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

the estimation of β is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

take it in to $SE(\hat{(\beta)})$, and simplify it to:

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

then,

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

$$= \frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i y_i)^2}}$$

Algebraical proving is done. Then comes the numerically prove:

```
t_stat1 <- (sqrt(number - 1)*sum(x*y))/
  (sqrt(sum(x^2) * sum(y^2) - (sum(x*y))^2))
cat("t-statistic of lm.fit1 is", t_stat1)
```

```
## t-statistic of lm.fit1 is 18.72593
```

(e)

***Answer*:** the algebraical forms of *t-statistic* for y onto x and x onto y are both

$$\frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i y_i)^2}}$$

and the *t-statistic* of lm.fit1 and lm.fit2 are quite the same: 18.73

(f)

***Answaer*:**

```
##
## 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
```

the result shows their *t-statitic* are not exactly the same.

3.12

This problem involves simple linear regression without an intercept.

(a)

***Answer*:** the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X means:

$$\hat{\beta_x} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \hat{\beta_y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2}, and \beta_x = \beta_y$$

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2}$$

thus, $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2$, and their coefficients are the same as long as sum of x square equal to y square.

(b)

```
number <- 100
x <- rnorm(number)
y <- 2*x + rnorm(number)
fit1 <- lm(y~x+0)
summary(fit1)$coef
```

```
##   Estimate Std. Error  t value     Pr(>|t|)
## x 2.107575 0.09584885 21.98853 6.991473e-40
```

```
fit2 <- lm(x~y+0)
summary(fit2)$coef
```

```
##    Estimate Std. Error  t value     Pr(>|t|)
## y 0.3938373 0.01791104 21.98853 6.991473e-40
```

(c)

```
x <- rnorm(number) + 1
y <- c()
for (i in 1:number) {
  y[i] <- x[number + 1 -i]
}
fit3 <- lm(y~x+0)
summary(fit3)$coef
```

```
##    Estimate Std. Error  t value     Pr(>|t|)
## x 0.4232642 0.09105706 4.648341 1.034003e-05
```

```r
fit4 <- lm(x~y+0)
summary(fit4)$coef
```

```
##    Estimate Std. Error  t value     Pr(>|t|)
## y 0.4232642 0.09105706 4.648341 1.034003e-05
```

3.13

(a)

```r
set.seed(1)
number <- 100
x <- rnorm(number, mean = 0, sd = 1)
```

(b)

```r
set.seed(2)
eps <- rnorm(number, mean = 0, sd = .5) # sqrt(variance) = sd
```

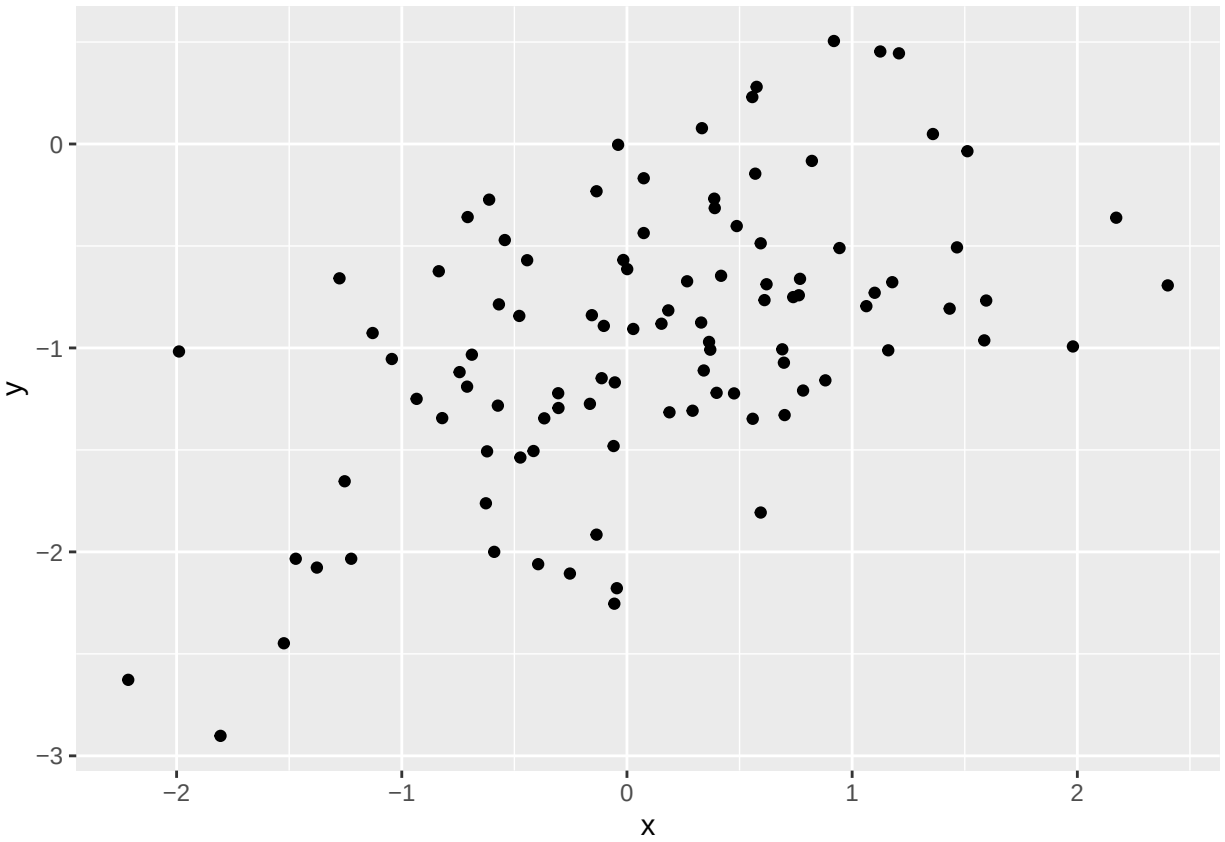
(c)

```r
y <- -1 + 0.5*x + eps
```

***Answer*:** the length of vector y is 100, the values of β_0 and β_1 are -1 and 0.5 in this model

(d)

```r
library(ggplot2)
ggplot(mapping = aes(x = x, y = y)) +
  geom_point()
```



Answer: the relationship between x and y are roughly linear, which mean as the value of x goes up the value of y increases.

(e)

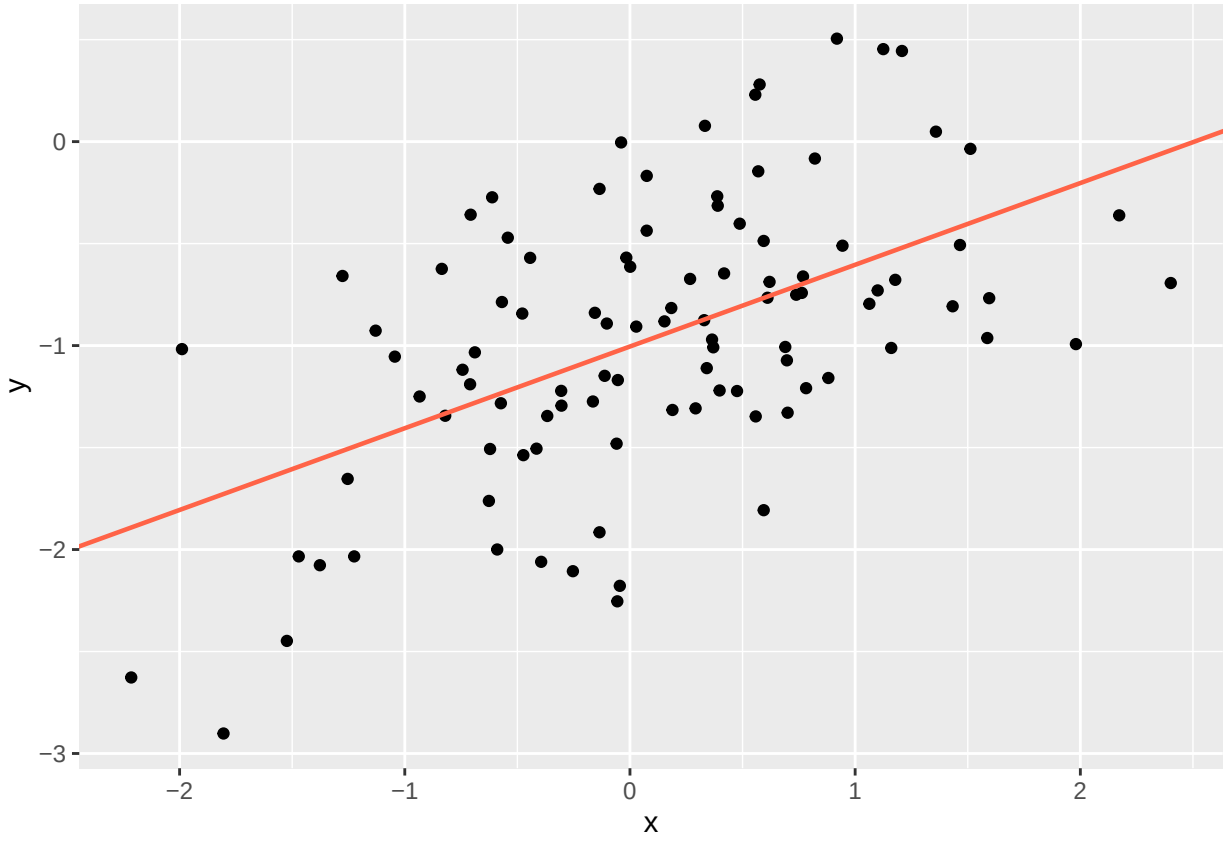
```
fit.lse <- lm(y~x)
summary(fit.lse)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22689 -0.40393 -0.04575  0.41574  1.14118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.00454    0.05804 -17.308  < 2e-16 ***
## x            0.40072    0.06446   6.216 1.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5761 on 98 degrees of freedom
## Multiple R-squared:  0.2828, Adjusted R-squared:  0.2755
## F-statistic: 38.64 on 1 and 98 DF,  p-value: 1.247e-08
```

$\hat{\beta}_0$ and $\hat{\beta}_1$ are -1.00 and 0.40, compare with $\beta_0 = 1$, $\beta_1 = .5$, the result is really close, but still with some irreducible error.

(f)

```
ggplot(mapping = aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = -1.00454, slope = 0.40072, color = "tomato1" , lwd = .8)
```



(g)

```
x2 <- x*x
fit.poly <- lm(y~x+x2)
summary(fit.poly)
```

```
##
## Call:
## lm(formula = y ~ x + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30604 -0.38957 -0.06695  0.40921  1.13539
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.92420    0.06967 -13.265  < 2e-16 ***
## x            0.41623    0.06394   6.509 3.33e-09 ***
## x2          -0.10121    0.05020  -2.016   0.0465 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5673 on 97 degrees of freedom
## Multiple R-squared:  0.3116, Adjusted R-squared:  0.2974
## F-statistic: 21.96 on 2 and 97 DF,  p-value: 1.362e-08
```

***Answer*:** the R-squared and adjust R-square are improved a little but not big enough to prove that the model fit is better.

(h)

```
set.seed(3)
eps <- rnorm(number, mean = 0, sd = 0.1)
y <- -1 + 0.5*x + eps
fit.less <- lm(y~x)
summary(fit.less)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.229491 -0.072903  0.001073  0.072741  0.170917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.998650   0.008666 -115.24   <2e-16 ***
## x            0.497737   0.009625   51.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08602 on 98 degrees of freedom
## Multiple R-squared:  0.9646, Adjusted R-squared:  0.9643
## F-statistic:  2674 on 1 and 98 DF,  p-value: < 2.2e-16
```

Here, the estimation of intercept and slope is nearly the exact value of β_0 and β_1. Additionally, the R-squared value increase from 0.28 to 0.96, which indicate that under this model, 96% of data variation can be explained.

(i)

```r
set.seed(4)
eps <- rnorm(number, mean = 0, sd = 1)
y <- -1 + 0.5*x + eps
fit.more <- lm(y~x)
summary(fit.more)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79189 -0.69979 -0.05084  0.61551  2.05659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.88849    0.09168  -9.691 5.67e-16 ***
## x            0.36239    0.10183   3.559 0.000577 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9101 on 98 degrees of freedom
## Multiple R-squared:  0.1144, Adjusted R-squared:  0.1054
## F-statistic: 12.66 on 1 and 98 DF,  p-value: 0.0005765
```

***Answer*:** the estimation of intercept and slope are acceptable, but the R-square is down to only 0.11, which means this model can only explain only 11% information of the data.

(j)

```r
cat("95% confidence interval of original data set \n")
```

```
## 95% confidence interval of original data set
```

```r
confint(fit.lse)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1197128 -0.8893648
## x            0.2727931  0.5286483
```

```r
cat("\n")
```

```r
cat("95% confidence interval of less noiser data set \n")
```

```
## 95% confidence interval of less noiser data set
```

```
confint(fit.less)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0158465 -0.9814536
## x            0.4786365  0.5168378
```

```
cat("\n")
```

```
cat("95% confidence interval of more noiser data set \n")
```

```
## 95% confidence interval of more noiser data set
```

```
confint(fit.more)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0704304 -0.7065526
## x            0.1603084  0.5644796
```

***Answer*:** As the extent of noise grows up, the confidence interval of both intercept and slope grows up, indicating the more uncertainty of the model. But the centers of intervals are still close to the true value.

3.14

This problem focuses on the *collinearity* problem.

(a)

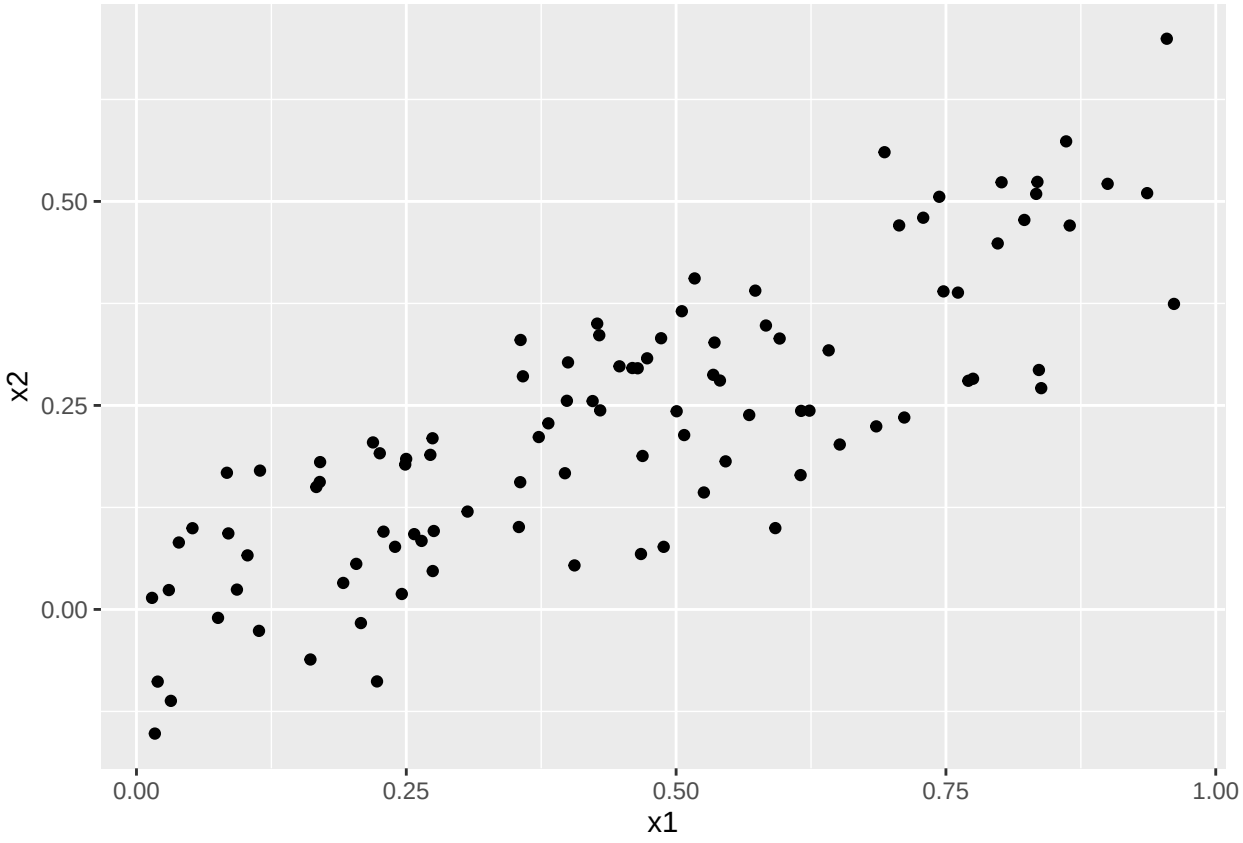
***Answer*:** the form of linear model is

$$y_i = 2 + 2 \times x_{1i} + 0.3 \times x_{2i} + \epsilon_i$$

the regression coefficients $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$.

(b)

```
ggplot(mapping = aes(x = x1, y = x2)) +
  geom_point()
```



***Answer*:** According to the plot, the relationship between x_1 and x_2 is linear.

(c)

```
fit.lse <- lm(y~x1+x2)
summary(fit.lse)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02507 -0.64102 -0.04811  0.64382  2.09265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7651     0.1876   9.409 2.54e-15 ***
## x1            3.1604     0.6444   4.904 3.77e-06 ***
## x2           -1.3863     0.9787  -1.417     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9377 on 97 degrees of freedom
## Multiple R-squared:  0.321,  Adjusted R-squared:  0.307
## F-statistic: 22.93 on 2 and 97 DF,  p-value: 7.002e-09
```

***Answer*:** According to the regression result, $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are 1.77, 3.16 and -1.83, which are really far away differernt from true value. By p-value of the model, we can successfully reject *Null Hypothesis* $H_0 : \beta_1 = 0$ but fail to reject $H_0 : \beta_2 = 0$.

(d)

```
fit.1 <- lm(y~x1)
summary(fit.1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14104 -0.68538 -0.07084  0.71946  2.09537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7858     0.1880   9.499 1.48e-15 ***
## x1            2.4065     0.3653   6.588 2.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9425 on 98 degrees of freedom
## Multiple R-squared:  0.307,  Adjusted R-squared:  0.2999
## F-statistic: 43.41 on 1 and 98 DF,  p-value: 2.235e-09
```

***Answer*:** *Null Hypothesis* $H_0 : \beta_1 = 0$ can be rejected.

(e)

```
fit.2 <- lm(y~x2)
summary(fit.2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5776 -0.7731  0.1079  0.7634  2.4292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2717     0.1741  13.051  < 2e-16 ***
## x2            2.5773     0.6134   4.202 5.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 98 degrees of freedom
## Multiple R-squared:  0.1527, Adjusted R-squared:  0.144
## F-statistic: 17.65 on 1 and 98 DF,  p-value: 5.847e-05
```

***Answer*:** *Null Hypothesis* $H_0 : \beta_2 = 0$ can be rejected.

(f)

***Answer*:** No contridiction! Since $x_{2i} = 0.5 \times x_{1i} + \epsilon_i$, the are highly correlated. Including both of them is likely to fall into collinearity trap and result in totally wrong result.

(g)

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
data.base <- data.frame(y = y, x1 = x1, x2 = x2)
rownames(data.base) <- c(1: 101)

fit.lse <- lm(y~x1+x2)
fit.1   <- lm(y~x1)
fit.2   <- lm(y~x2)
fit.lse
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##       1.903        1.637        1.126
```

```
fit.1
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Coefficients:
## (Intercept)           x1
##       1.914        2.206
```

```
fit.2
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Coefficients:
## (Intercept)           x2
##       2.220        2.872
```

***Answer*:** this new observation has nearly no effect on regression coefficients for all three models.

```
library(car)
outlierTest(fit.lse)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 101 4.164214         6.7744e-05    0.0068422
```

```
outlierTest(fit.1)
```

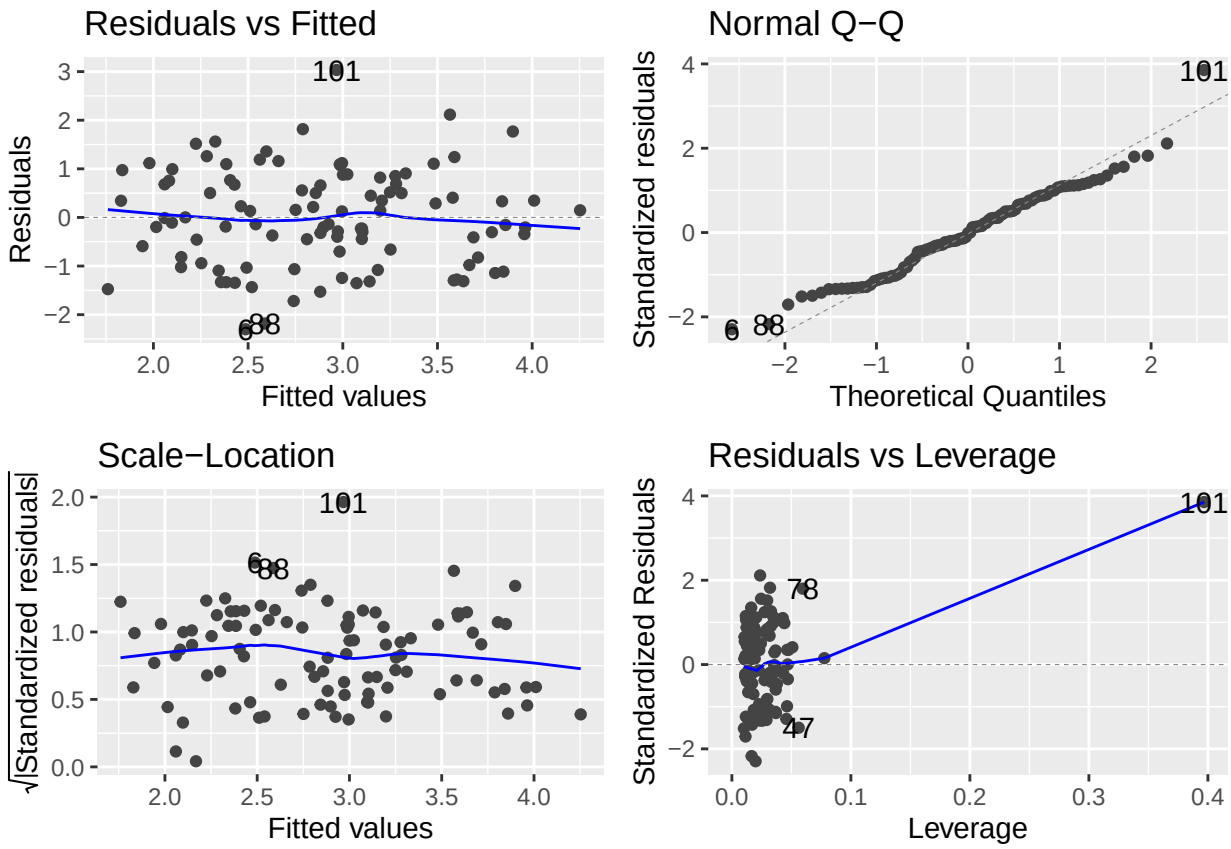
```
##     rstudent unadjusted p-value Bonferroni p
## 101 4.158221         6.8781e-05    0.0069468
```

```
outlierTest(fit.2)
```

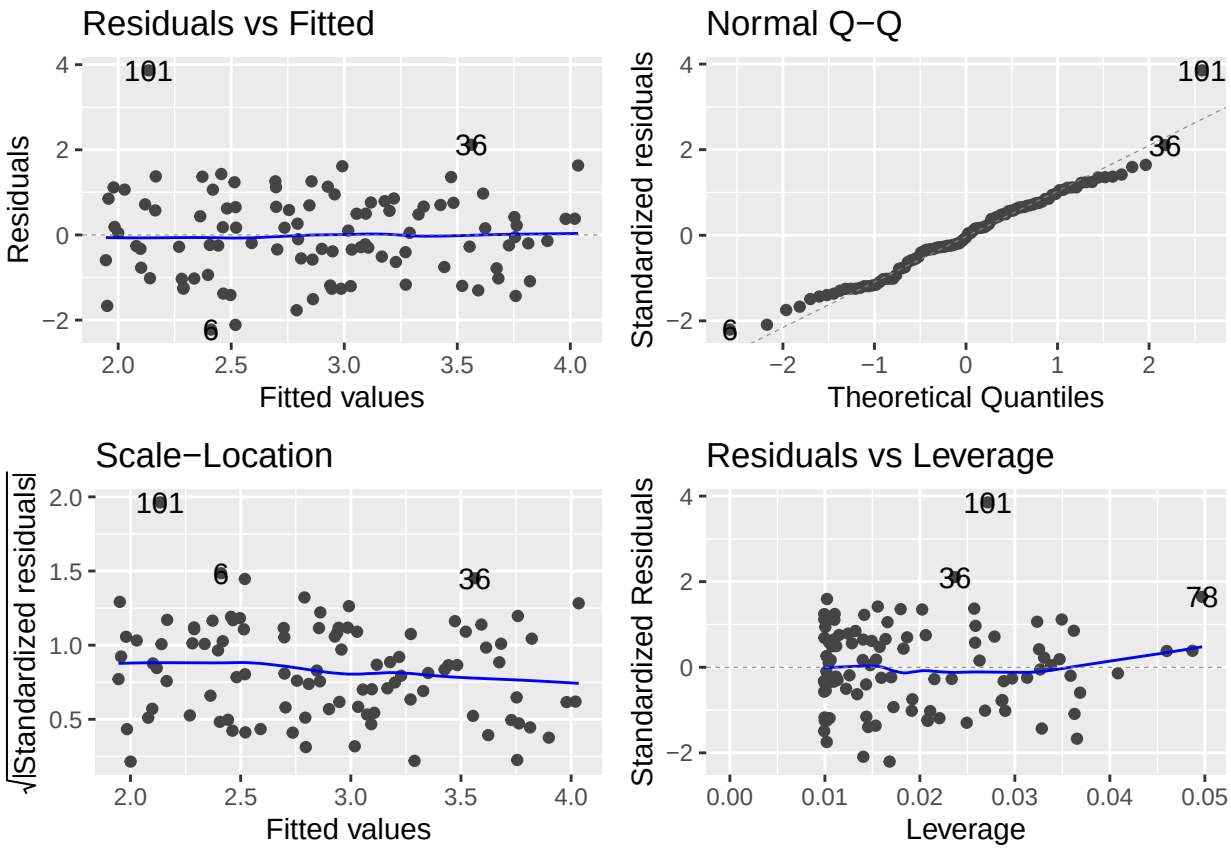
```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 6 -2.541823           0.012595           NA
```

the new observation is not an outlier for each model except *fit.lse*. but it is an high leverage point for each of these three models.

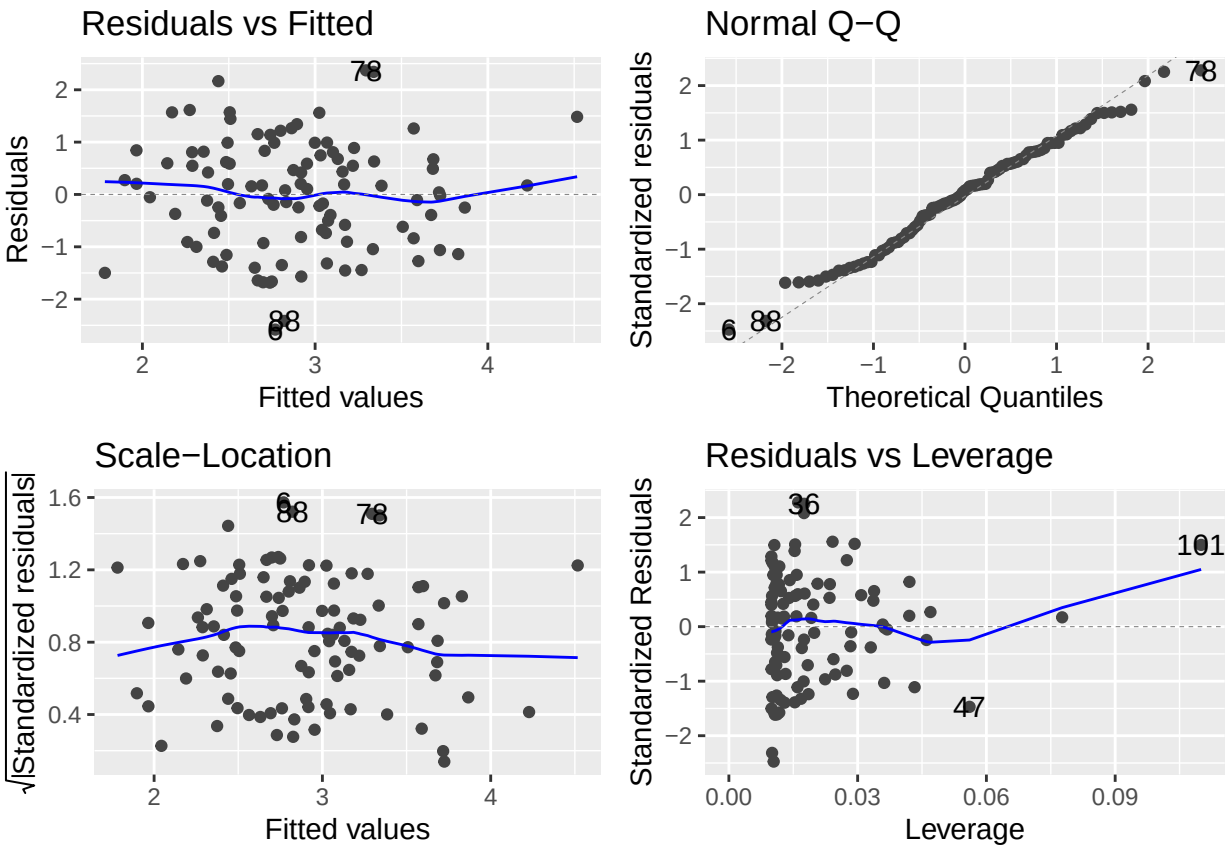
```
library(ggfortify)
autoplot(fit.lse)
```




```
autoplot(fit.1)
```




```
autoplot(fit.2)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized Residuals
Leverage