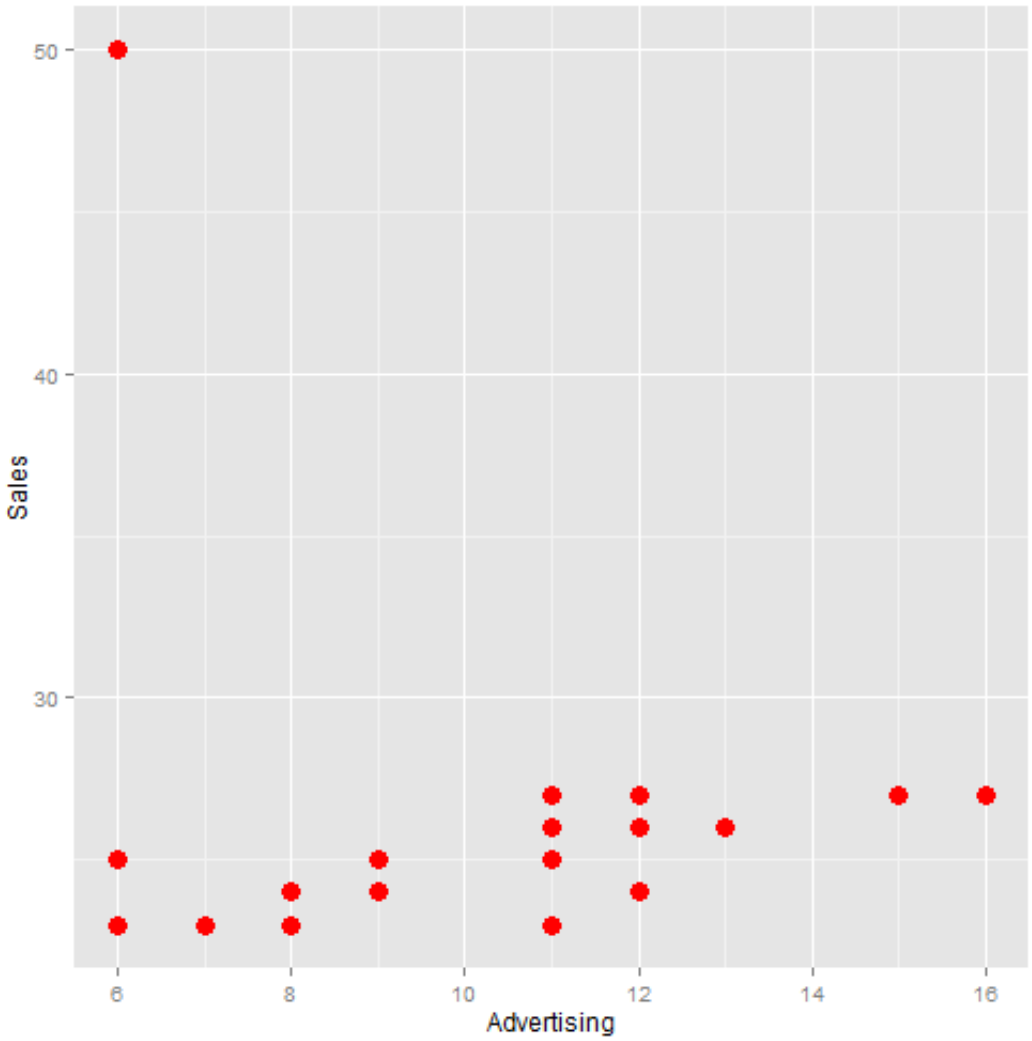
Test Exercise 1: Answers to the Questions

This exercise considers an example of data that do not satisfy all the standard assumptions of simple regression. In the considered case, one particular observation lies far off from the others, that is, it is an outlier. This violates assumptions **A3** and **A4**, which state that all error terms “i are drawn from one and the same distribution with mean zero and fixed variance 2. The dataset contains twenty weekly observations on sales and advertising of a department store. The question of interest lies in estimating the effect of advertising on sales. One of the weeks was special, as the store was also open in the evenings during this week, but this aspect will first be ignored in the analysis.

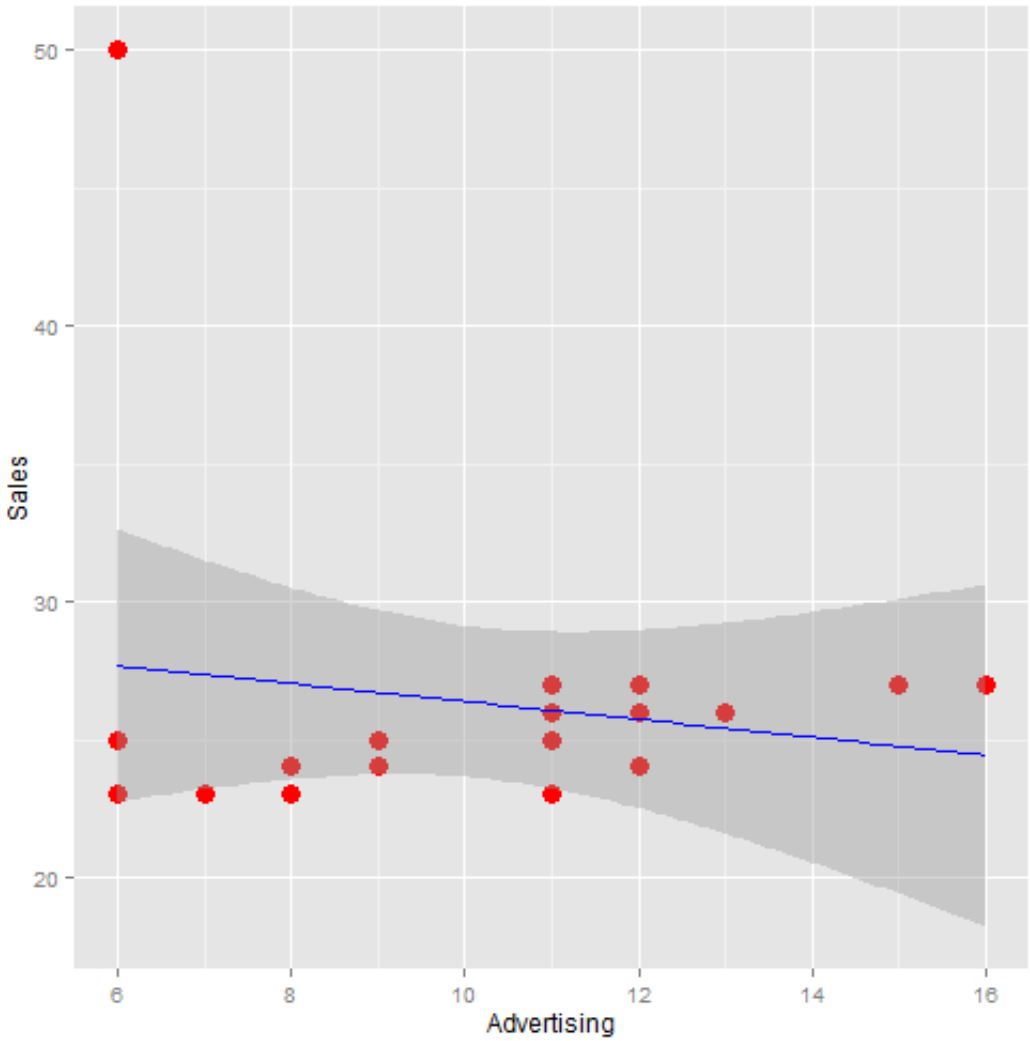
- Make the scatter diagram with sales on the vertical axis and advertising on the horizontal axis. What do you expect to find if you would fit a regression line to these data?



- Estimate the coefficients a and b in the simple regression model with sales as dependent variable and advertising as explanatory factor. Also compute the

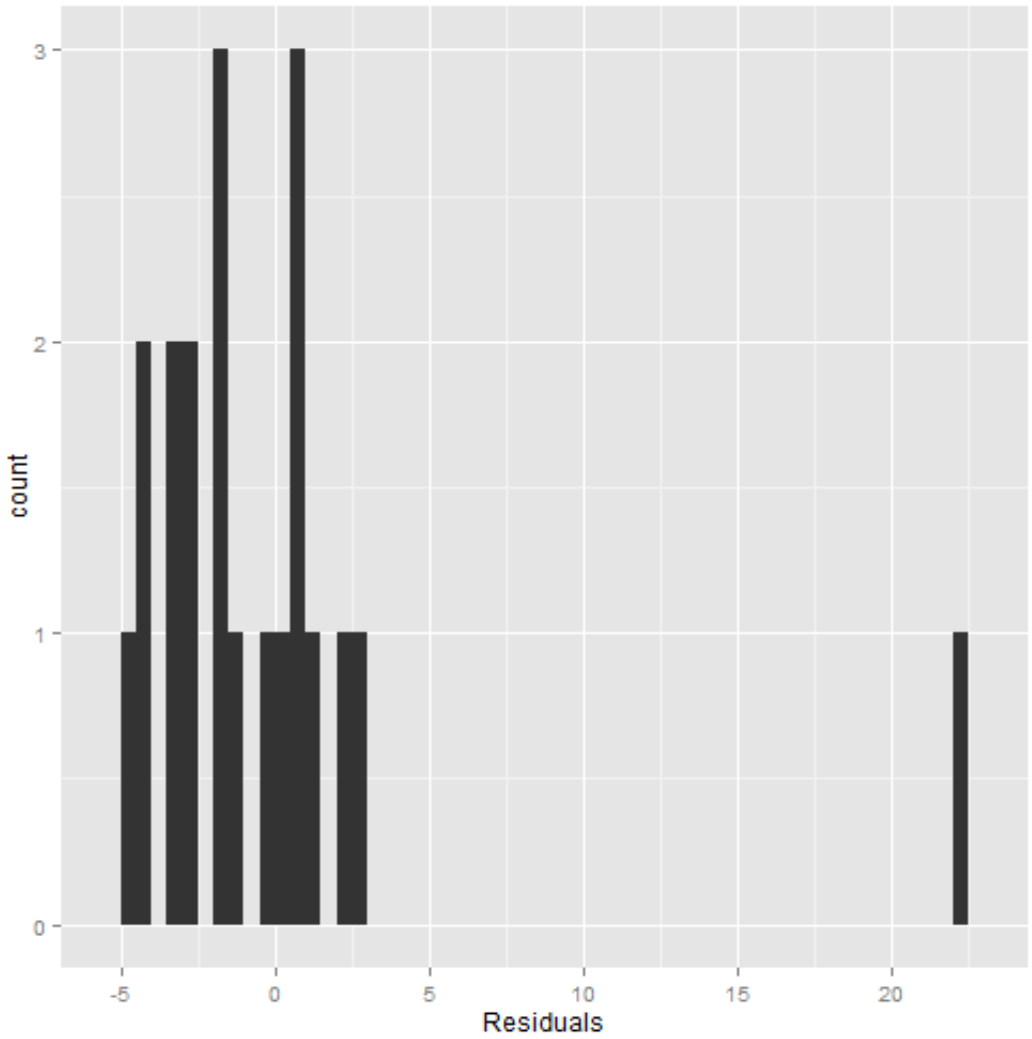
standard error and t-value of b. Is b significantly different from 0?

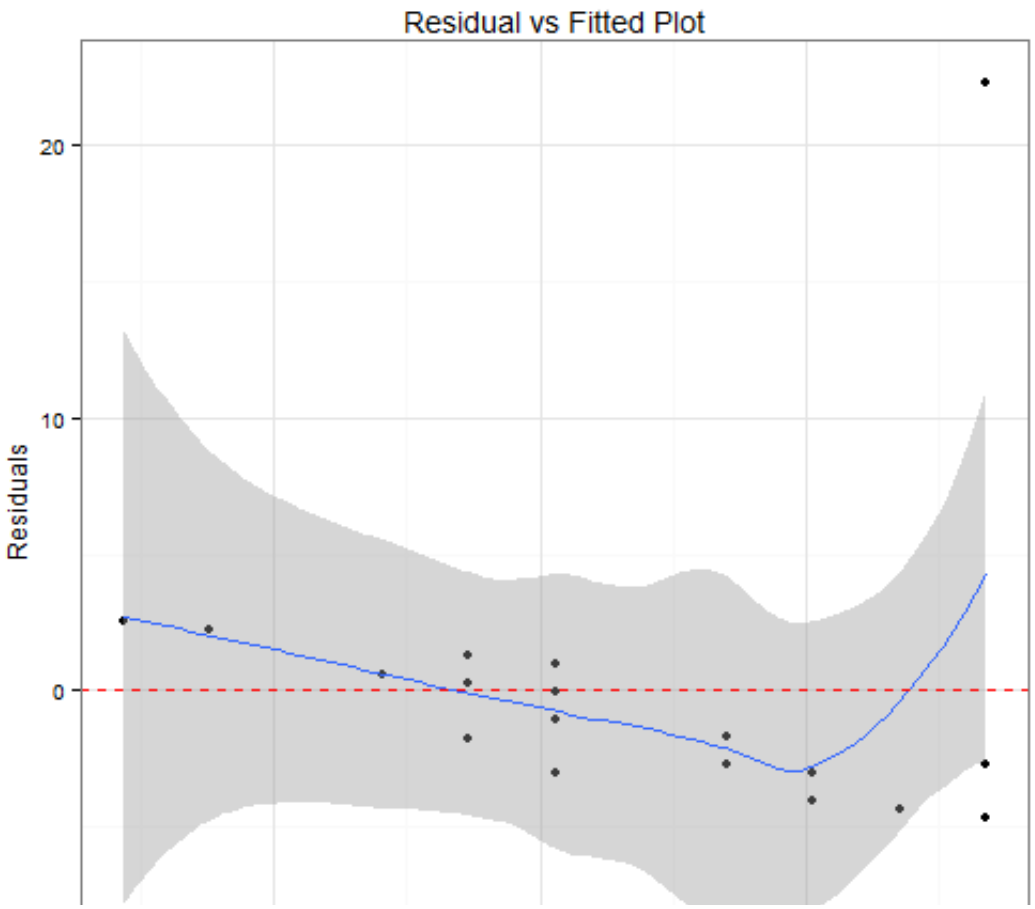
```
##
## Call:
## lm(formula = Sales ~ Advertising, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6794 -2.7869 -1.3811  0.6803 22.3206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.6269     4.8815   6.069 9.78e-06 ***
## Advertising  -0.3246     0.4589  -0.707    0.488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 5.836 on 18 degrees of freedom
## Multiple R-squared:  0.02704,    Adjusted R-squared:
-0.02701
## F-statistic: 0.5002 on 1 and 18 DF,  p-value: 0.4885
```

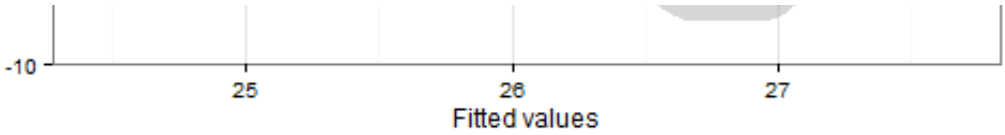


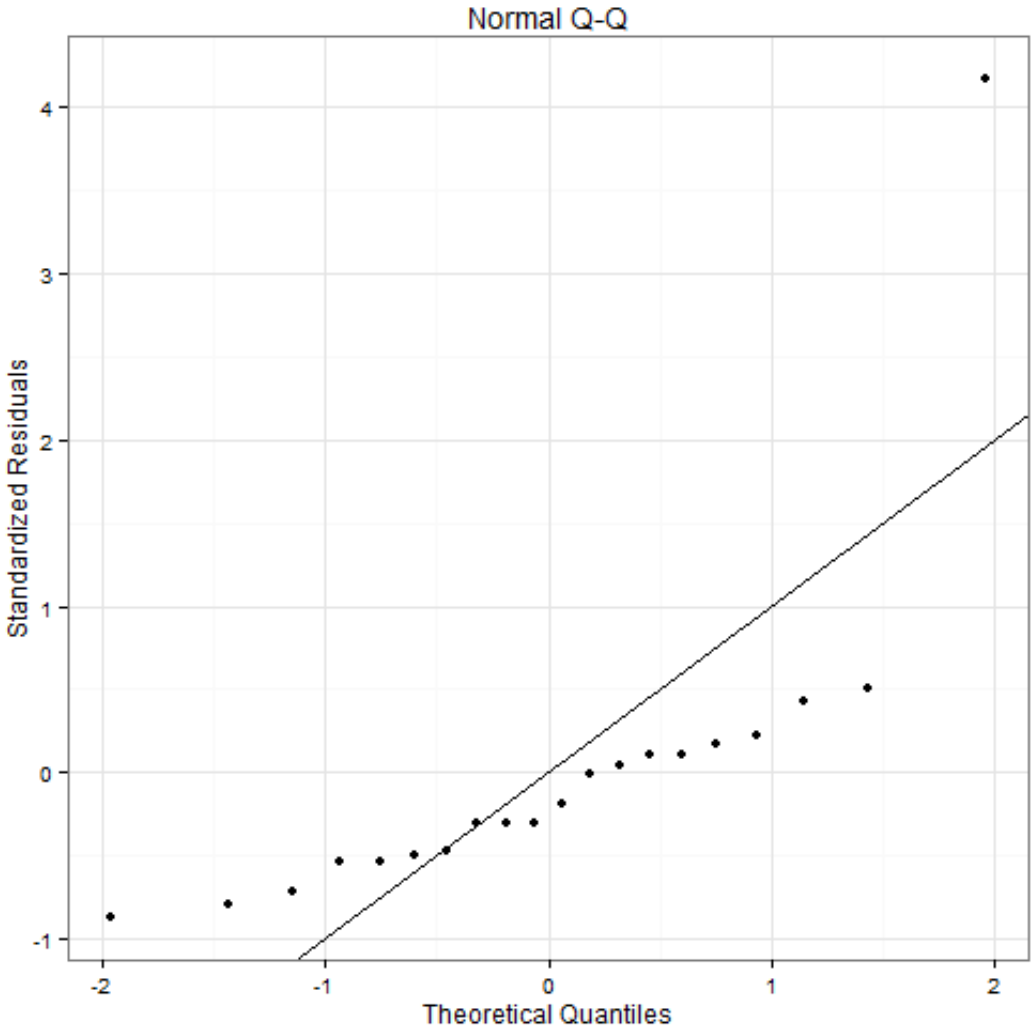
- Compute the residuals and draw a histogram of these residuals. What conclusion

do you draw from this histogram?


3
2
1
0
count
-5
0
5
10
15
20
Residuals


Residual vs Fitted Plot
20
10
0
Residuals



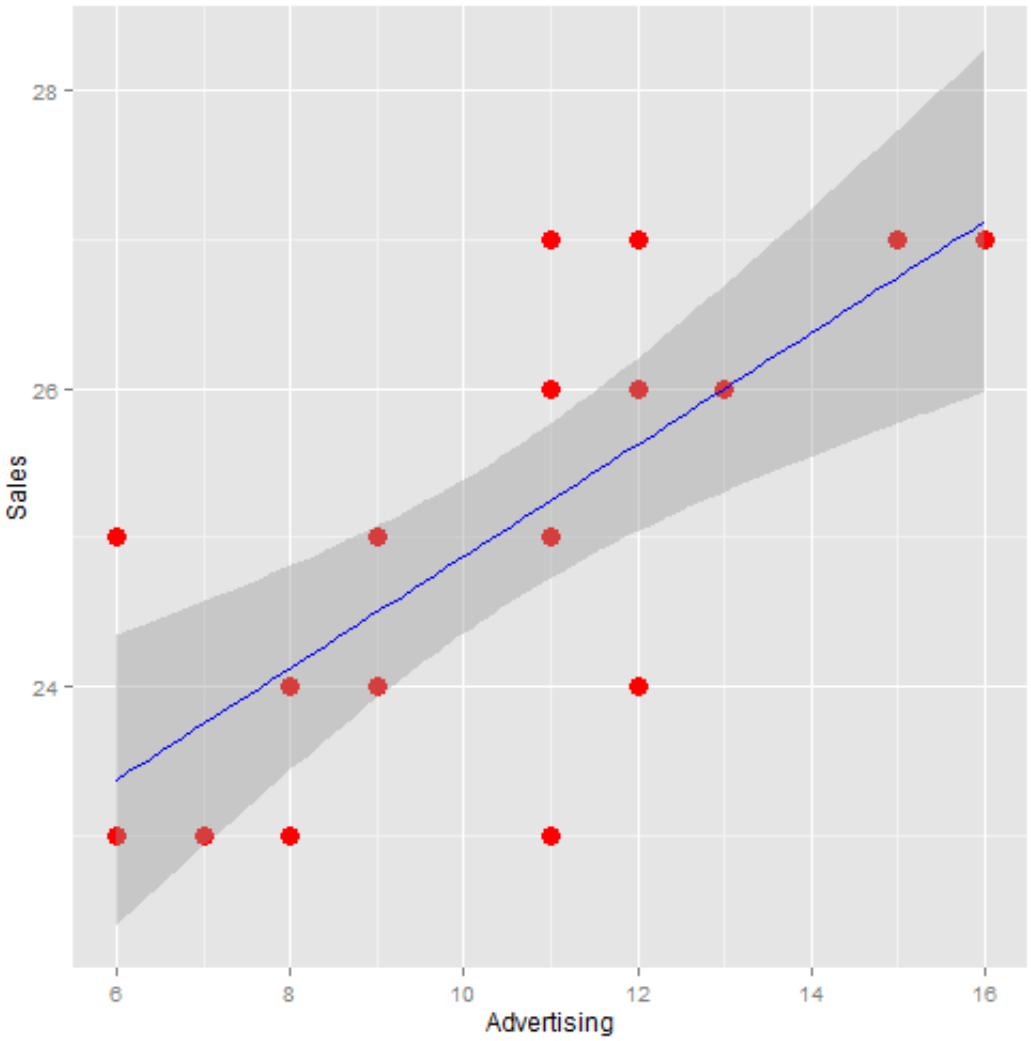


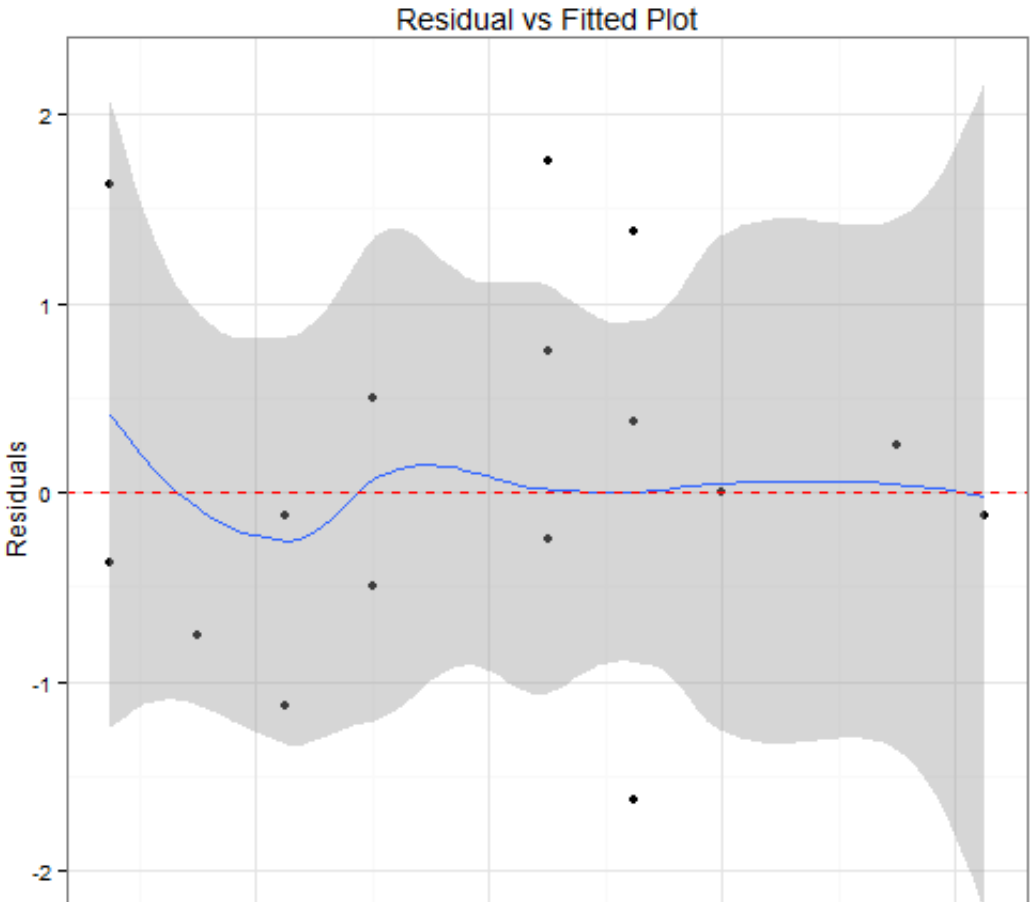
- Apparently, the regression result of part (b) is not satisfactory. Once you realize that the large residual corresponds to the week with opening hours during the evening, how would you proceed to get a more satisfactory regression model?

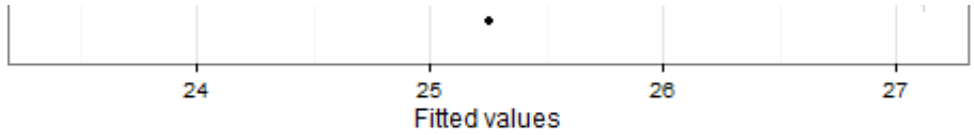
```
# delete this outlier point.
```

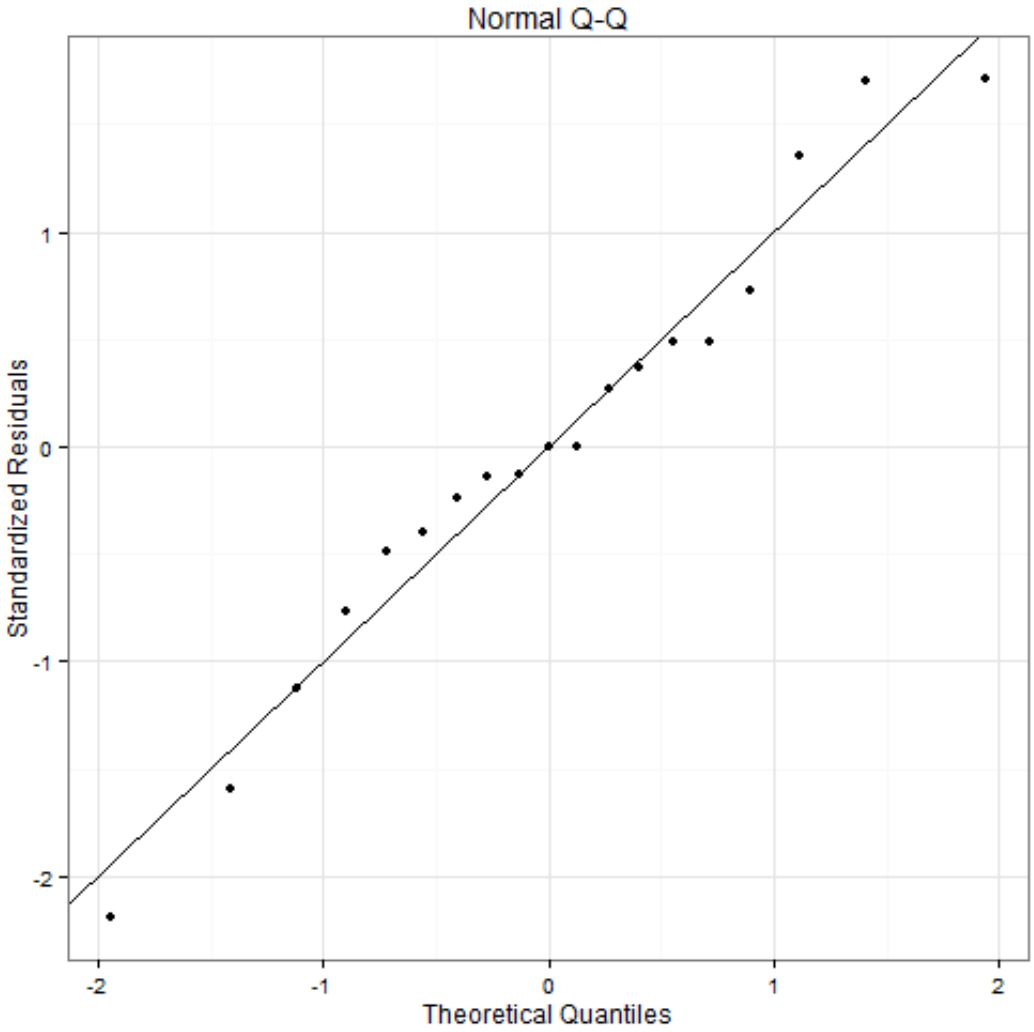
- Delete this special week from the sample and use the remaining 19 weeks to estimate the coefficients a and b in the simple regression model with sales as dependent variable and advertising as explanatory factor. Also compute the standard error and t-value of b. Is b significantly different from 0?

```
##
## Call:
## lm(formula = Sales ~ Advertising, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2500 -0.4375  0.0000  0.5000  1.7500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.1250     0.9548  22.124 5.72e-14 ***
## Advertising   0.3750     0.0882   4.252 0.000538 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 1.054 on 17 degrees of freedom
## Multiple R-squared:  0.5154, Adjusted R-squared:  0.4869
## F-statistic: 18.08 on 1 and 17 DF,  p-value: 0.0005379
```


28
26
24
Sales
6
8
10
12
14
16
Advertising


Residual vs Fitted Plot
2
1
0
-1
-2
Residuals





- Discuss the differences between your findings in parts (b) and (e). Describe in words what you have learned from these results.

```
# Outliers can have impact on the coefficients, hence they
should be remove first as pre-processing step before
analysis.
```