Week 1

Multifaceted

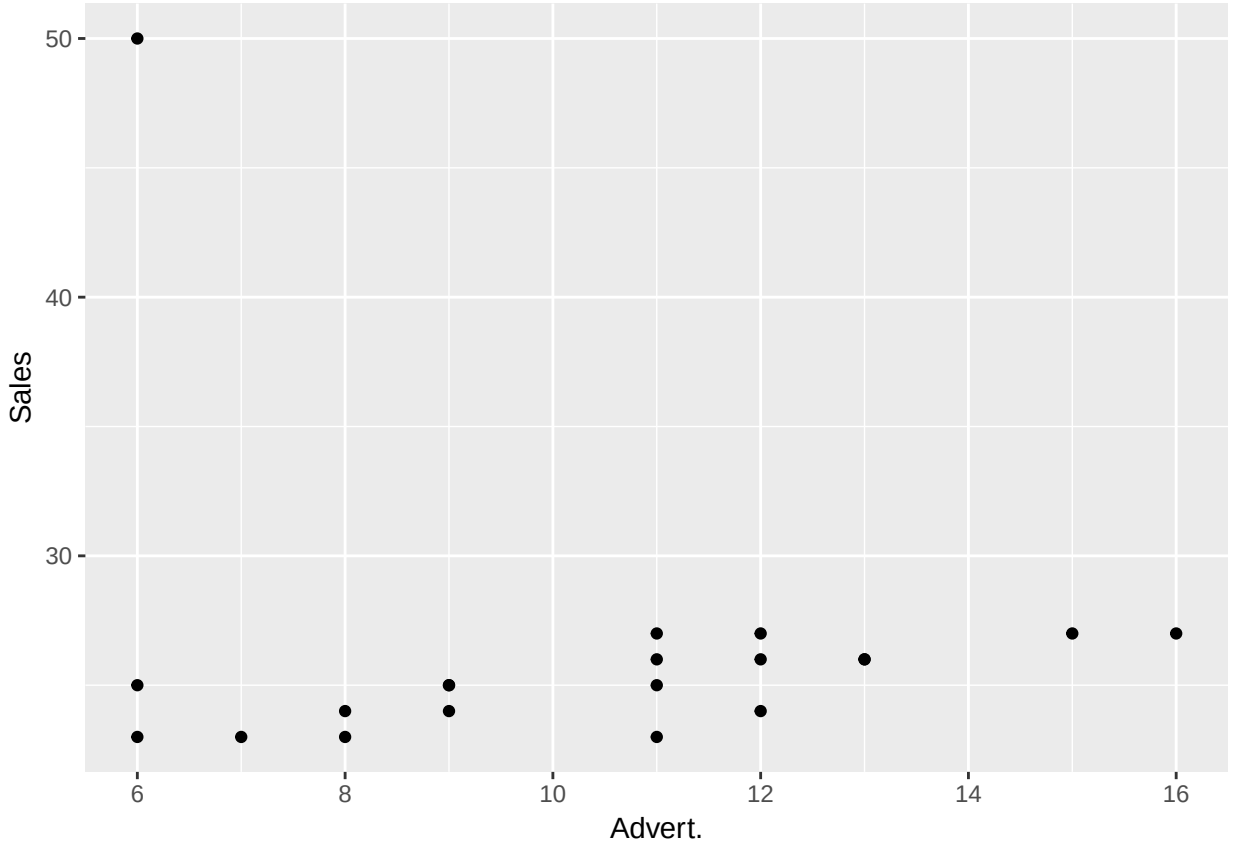
May 18, 2018

(a) 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?

```
data <- read_table2('TestExer-1-sales-round1.txt')
```

```
## Parsed with column specification:
## cols(
##   Observ. = col_integer(),
##   Advert. = col_integer(),
##   Sales = col_integer()
## )
```

```
ggplot(data, aes(x = Advert.,y = Sales)) + geom_point()
```



The model will be significantly affected by the outlier point.

(b) 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?

```
md <- lm(Sales~Advert., data = data)
summary(md)
```

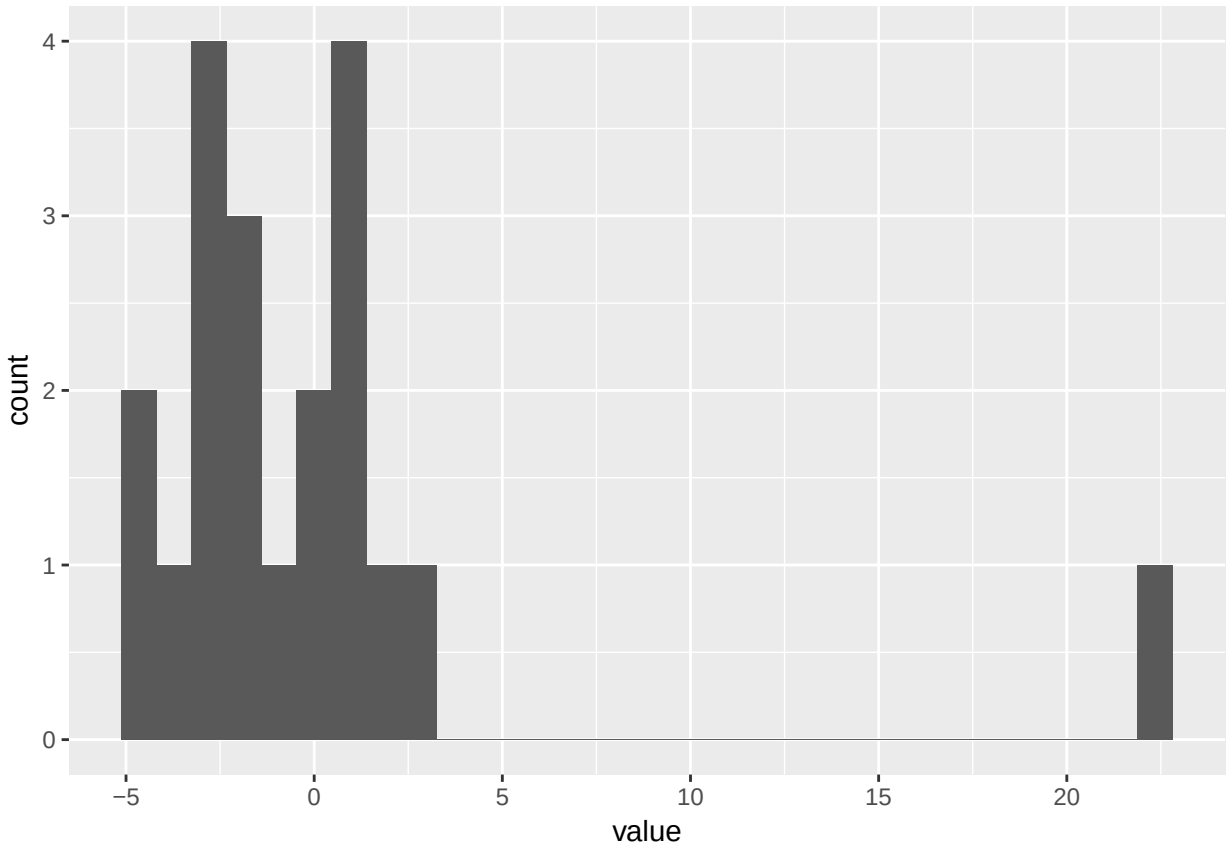
```
##
## Call:
## lm(formula = Sales ~ Advert., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6794 -2.7869 -1.3811  0.6803 22.3206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.6269     4.8815   6.069 9.78e-06 ***
## Advert.      -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
```

a = 29.6 b - -0.325 The standard error of b is 0.459. The t-value of b is -0.707, which is not significantly different from 0.

(c) Compute the residuals and draw a histogram of these residuals. What conclusion do you draw from this histogram?

```
res <- residuals(md)
res <- tibble(index = 1:length(res), value = res)
ggplot(res, aes(x = value)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There is an observation of which the estimation is wrong.

(d) 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 the outlier point.

(e) 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?

```
idx <- which(data$Sales > 30)
newdata <- data[-idx, ]
n_md <- lm(Sales~Advert., data = newdata)
summary(n_md)
```

```
##
## Call:
## lm(formula = Sales ~ Advert., data = newdata)
##
## 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 ***
## Advert.       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
```

a = 21.1 b = 0.375 The standard error of b is 0.0882. The t-value of b is 4.25, which is significantly different from 0.

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

e eliminate the effect of outlier. Therefore the result is more reliable. We have to plot the data and remove the outlier before fitting a linear model.