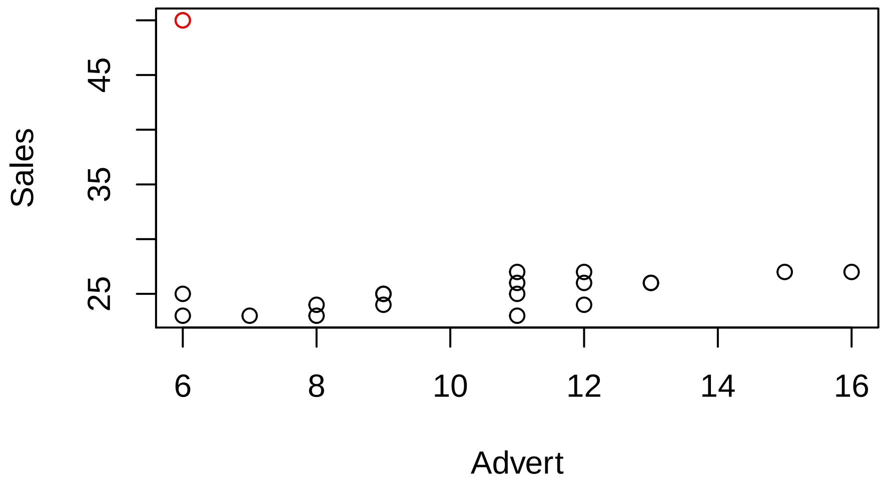
MOOC Econometrics Test Exercise 1

Tan Siow Boon

Sunday, November 8, 2015

Question 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?



It is obvious from the scatter plot that there is an outlier with extremely large sales value of 50 comparing to the rest of the entries. This outlier will skew the regression and result in bias estimates.

Question 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?

The results of the linear regression of sales vs advertising using R is as follows:

```
##
## Call:
## lm(formula = Sales ~ Advert, data = dat1)
##
## 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
```

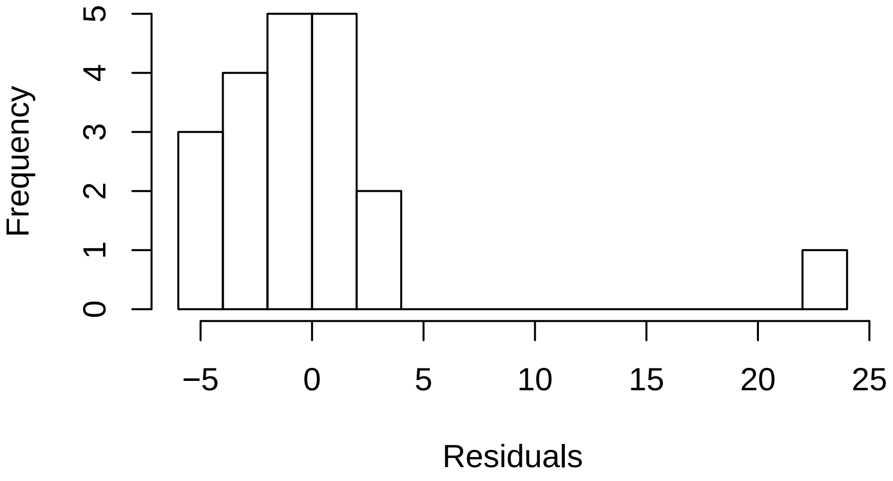
The coefficient a is 29.6269 and b is -0.3246. This is counter-intuitive as it implies that increasing adverstising will reduce sales.

The standard error and t-value of b are 0.4589 and -0.707 respectively. b is not signicantly different from 0 given the high p-value of 0.488. The 95% confidence interval of b also includes 0.

```
##                 2.5 %     97.5 %
## (Intercept) 19.371185 39.8826017
## Advert      -1.288711  0.6395612
```

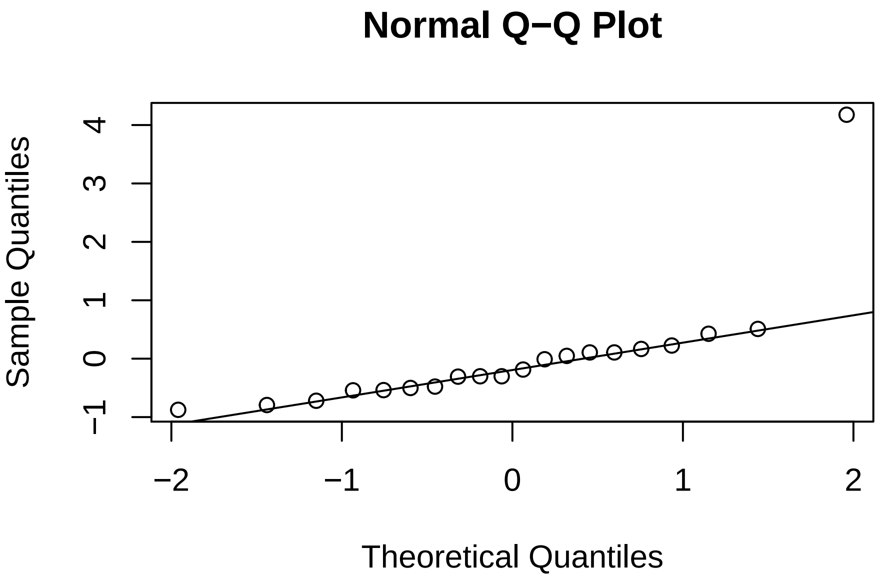
Question c)

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



The histogram shows that the distribution of the residuals is very far from Normal. While the mean of the residuals is very close to 0, the distribution is right skewed instead of symmetric around the mean as is expected in a Normal distribution.

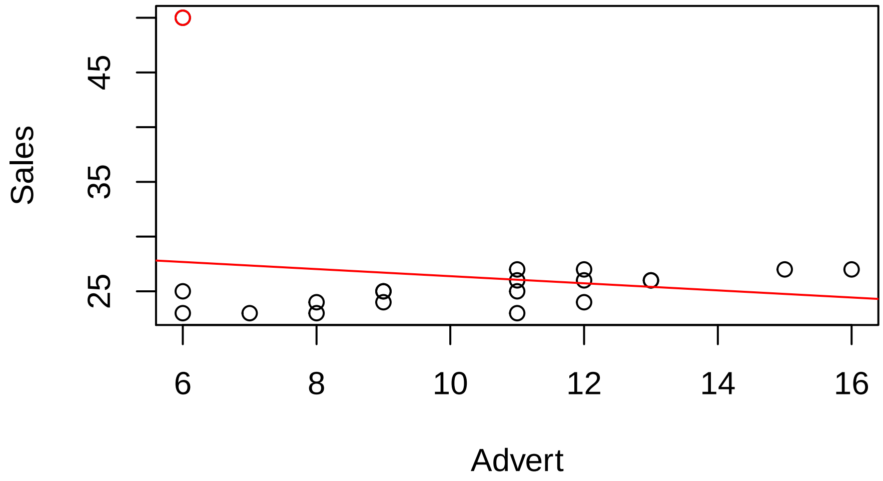
The following Quantile-Quantile plot also shows that the distribution of residuals is not Normal.



Question 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?

By appending the regression line on to the scatter plot, it can be seen quite clearly that the outlier skewed the whole regression. The regression model is not acceptable with very low R-squared of 0.02704 and coefficient b is not signicantly different from 0.



In order to get a more satisfactory regression model, the outlier has to be removed from the sample used to estimate the coefficients.

Question 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?

The results of the linear regression of sales vs advertising after removel of the outlier are as follows:

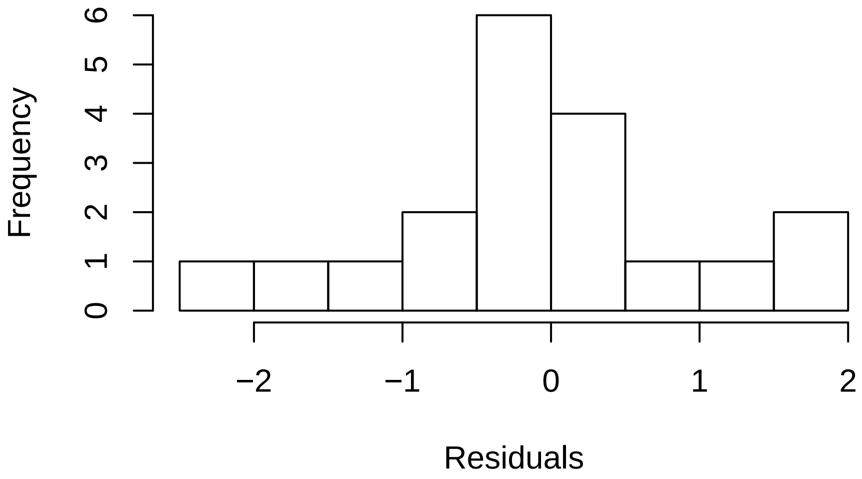
```
##
## Call:
## lm(formula = Sales ~ Advert, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2500 -0.4375  0.0000  0.5000  1.7500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.1250     0.9548  22.124 5.72e-14 ***
## Advert        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
```

The coefficient a is now 21.1250 and b is 0.3750. This is now more intuitive as it implies that increasing adverstising will increase sales.

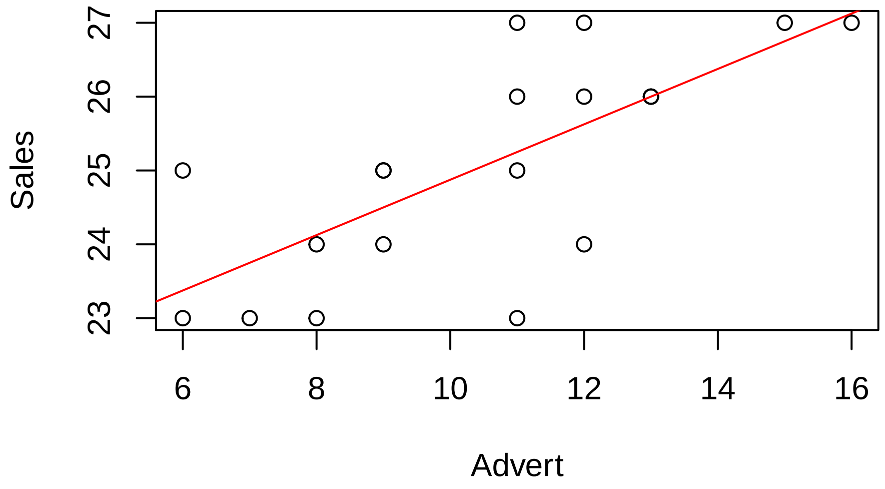
The standard error and t-value of b are 0.0882 and 4.252 respectively. b is signicantly different from 0 given the very low p-value of 0.000538. The 95% confidence interval of b also does not includes 0.

```
##                  2.5 %     97.5 %
## (Intercept) 19.1104466 23.1395534
## Advert       0.1889218  0.5610782
```

By removing the outlier, the distribution of the residuals is also more symmetric around the mean.



The resulting regression model is much better with R-squared of 0.5154 and coefficient b is signicantly different from 0.



Question f)

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

In part (b), the regression model generated is not acceptable:
- Coefficient b has counter-intuitive sign and it is not significently different from 0.
- R-squared of model is very low at 0.02704.

In part (e), the regression model generated with the removal of a single outlier shows much improvement:
- Coefficient b has intuitive sign and it is significently different from 0.
- R-squared of model is much higher at 0.5154.

From the above, we learned that linear regression only works well if its key assumptions are not violated. It is therefore important to perform exploratory data analysis and the necessary diagnostic tests to ensure that the data sample does not violate the key assumptions of linear regression. Data cleansing may also be required before performing the linear regression.