Econometrics Week1 Assignment

Pradeepta Das

9 November 2020

Questions

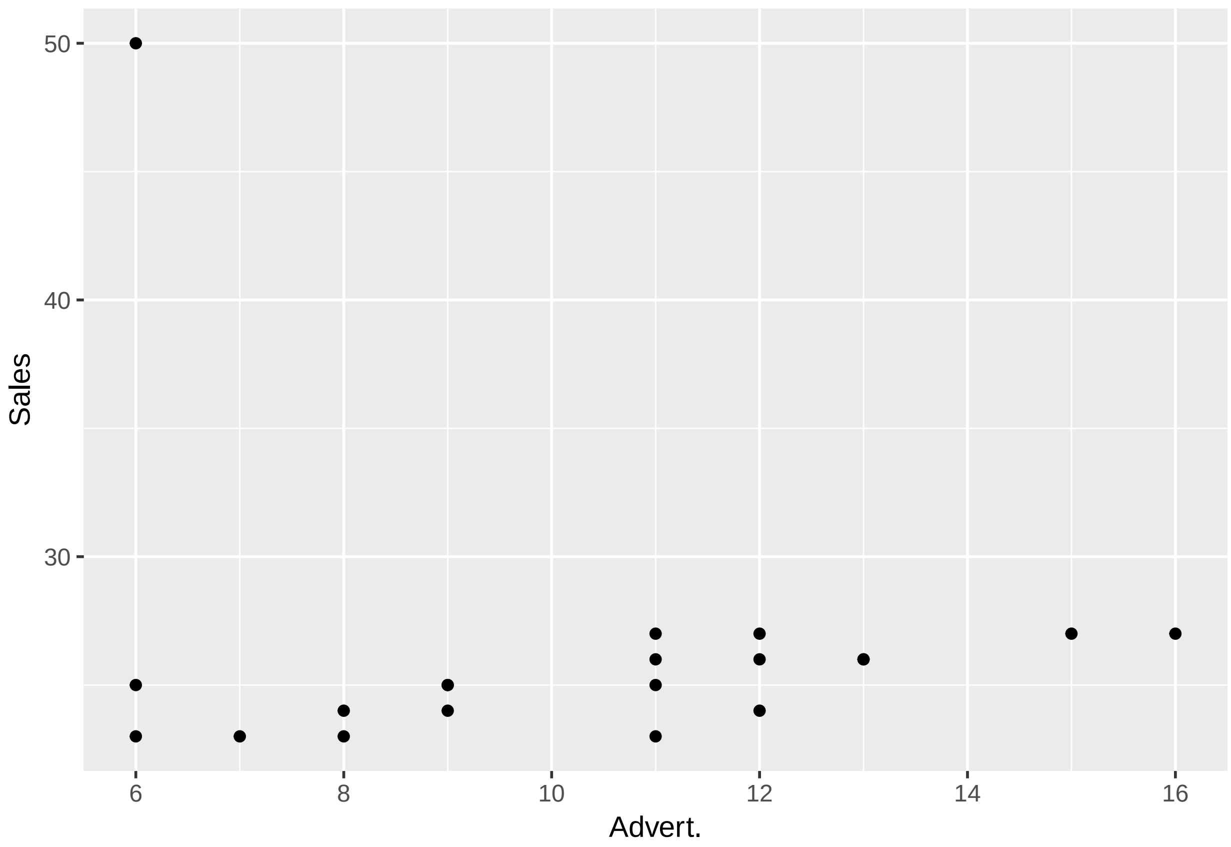
This exercise considers an example of data that do not satisfy all the standard assumptions of simple regression. In the considered case, one particular observation lies far off from the others, that is, it is an outlier. This violates assumptions A3 and A4, which state that all error terms ϵ_i are drawn from one and the same distribution withmean zero and fixed variance σ^2. The dataset contains twenty weekly observations on sales and advertising of adepartment 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.

(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?

```r
# read the data!
data_sales<-read.table(file, header = TRUE, sep = "", dec = ".")
data_sales
```

```
##    Observ. Advert. Sales
## 1        1      12    24
## 2        2      12    27
## 3        3       9    25
## 4        4      11    27
## 5        5       6    23
## 6        6       9    25
## 7        7      15    27
## 8        8       6    25
## 9        9      11    26
## 10      10      16    27
## 11      11      11    25
## 12      12       6    50
## 13      13      13    26
## 14      14      11    23
## 15      15      13    26
## 16      16       7    23
## 17      17       8    23
## 18      18       8    24
## 19      19      12    26
## 20      20       9    24
```

```r
# plot the advertisement vs sales points
data_sales %>% ggplot(aes(Advert.,Sales))+geom_point()
```



Looks like there is one outlier which doesn't fit in the pattern.

(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?

```
linear_model = lm(data_sales$Sales ~ data_sales$Advert.)
summary(linear_model)
```

```
##
## Call:
## lm(formula = data_sales$Sales ~ data_sales$Advert.)
##
## 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 ***
## data_sales$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
```

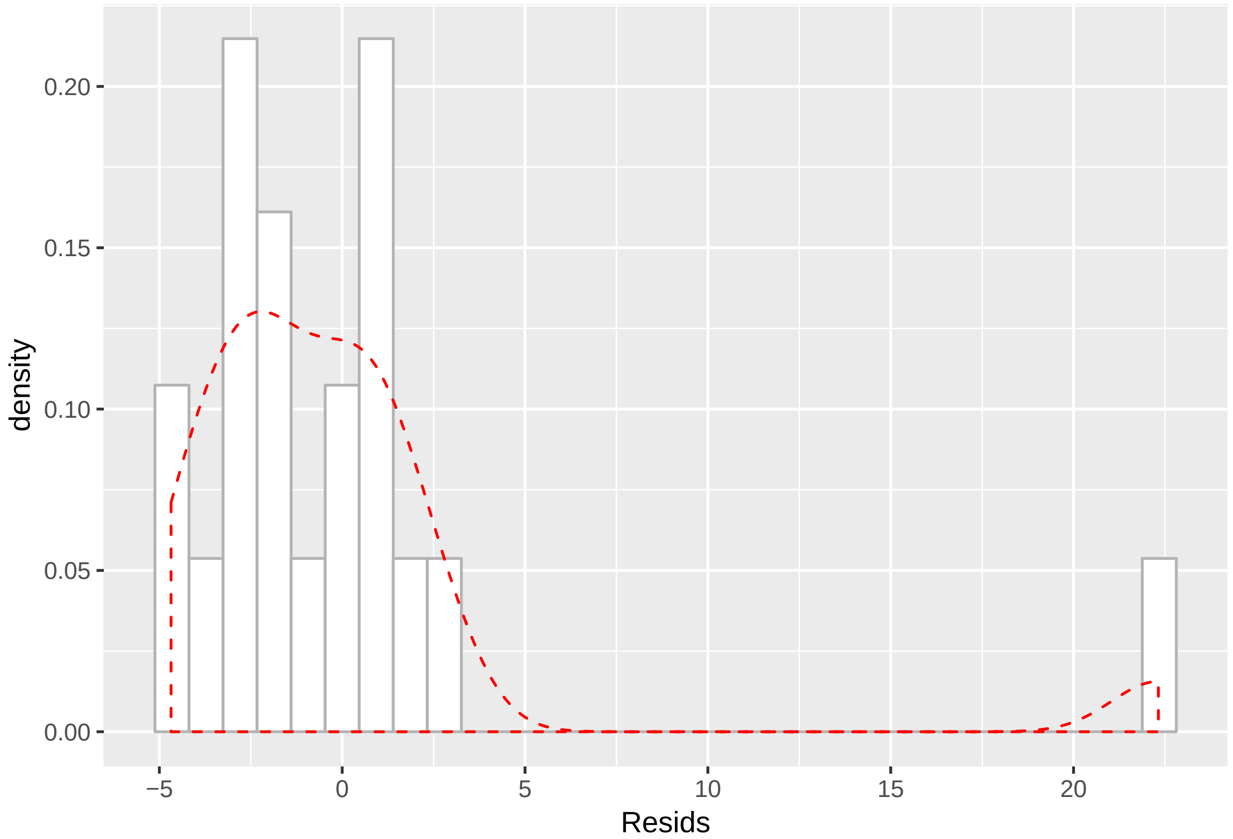
b (coefficient of Advertising) here doesn't seem to be significantly different to zero! The p-value is too high 0.488 > 0.05. This is mostly because of the large outlier that we observed. That one big large outlier holds the ability to make the linear regression model bad!!

The variability in the data is also not explained by the model. Because the R-squared value is 2%. Only 2% of the variability is explained. This indicates a low explanatory power of the model.

Also the F satistics is Not Significant indicating that the R squared (or the model) is not significant.

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

```
x<-data_sales$Advert.
predicted <- predict(linear_model, newdata=data.frame(data_sales$Advert.))
residuals <- data_sales$Sales - predicted
data_sales <- data_sales %>% mutate(Resids = residuals)
data_sales %>% ggplot(aes(Resids)) + geom_histogram(bins = 30, color = "grey70", fill = "white", aes(y
```



From the plot of histogram of residuals, we see a very highly right skewed distribution with majority of values lying in the range of +5 and -5 and one extreme value (outlier) which is making the distribution highly right skewed (non-normal).

Also this means that the residual terms do not have a mean = 0. which violates the A3 principle. Again the residual term's variances are not equal for a fixed n samples! This too violates the A4 principle. Because of these two violations, we conclude that the linear model doesn't work as expected in this case!

(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?

As we know, is necessary to drop out that observation to clean the data and get a better coefficient with the minimum error.

```r
#lets find the outlier
which.max(data_sales$Sales)
```

```
## [1] 12
```

```r
#The 12th obs is the outlier

#outlier value in terms of sales
data_sales$Sales[12]
```

```
## [1] 50
```

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

```r
data_sales <- data_sales[-which.max(data_sales$Resids),]
data_sales
```

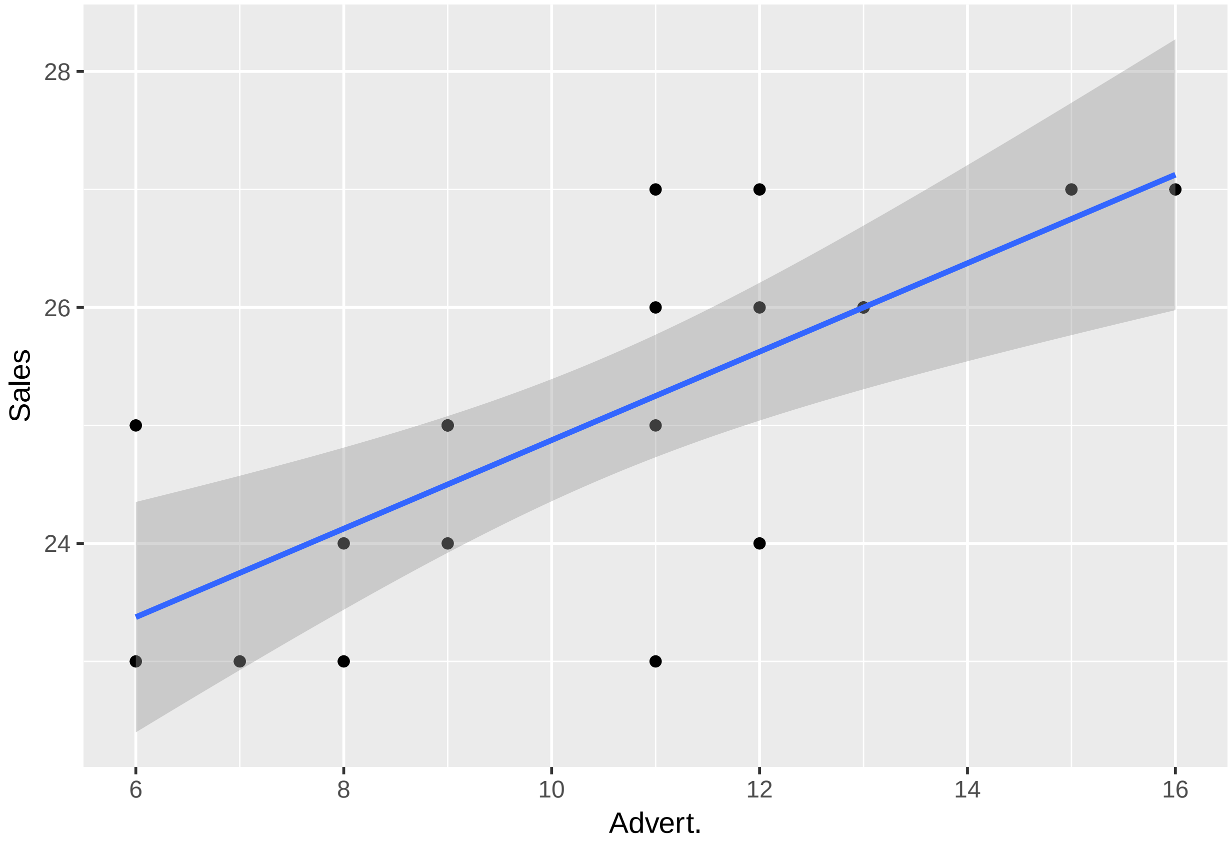
```
##    Observ. Advert. Sales      Resids
## 1        1      12    24 -1.73199382
## 2        2      12    27  1.26800618
## 3        3       9    25 -1.70571870
## 4        4      11    27  0.94343122
## 5        5       6    23 -4.67944359
## 6        6       9    25 -1.70571870
## 7        7      15    27  2.24173107
## 8        8       6    25 -2.67944359
## 9        9      11    26 -0.05656878
## 10      10      16    27  2.56630603
## 11      11      11    25 -1.05656878
## 13      13      13    26  0.59258114
## 14      14      11    23 -3.05656878
## 15      15      13    26  0.59258114
## 16      16       7    23 -4.35486862
## 17      17       8    23 -4.03029366
## 18      18       8    24 -3.03029366
## 19      19      12    26  0.26800618
## 20      20       9    24 -2.70571870
```

```r
linear_model = lm(data_sales$Sales ~ data_sales$Advert.)
summary(linear_model)
```

```
##
## Call:
## lm(formula = data_sales$Sales ~ data_sales$Advert.)
##
## 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 ***
## data_sales$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
```

Now we can reject the H_null: b=0, so beta is statically important. From the regression output, the slope coefficient(= 0.3750) of the model is highly significant (pvalue<0.001).Hence we reject the null hypothesis in favour of the alternate that the slope coefficient b is significantly different from 0.

```
data_sales %>% ggplot(aes(Advert.,Sales))+geom_point() + geom_smooth(method = "lm")
```



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

Comparing the summary regression output results from point b & e we see that the after removing the outlier, the slope coefficient has become significant.

Also from the scatter plot & the regression line, we see now a positive linear association with Multiple R-squared=0.5154 which implies that about 51% of the variation in Sales is being explained by variation in Advertsisng i.e. the explanatory power of the model has drastically improved from the original model.

Also from the F satistics, we see that it has become significant as compared to the original model impying the R squared (or the model) is significant.

After removal of the residual point A3 and A4 part of the restrictions also became compliant. Hence the model seem to be working better. We can test that by taking the mean of the residuals:

```
resid<-residuals(linear_model)
mean(resid)
```

```
## [1] 4.672483e-17
```

We can see that the mean is close to zero.