Test Exercise 1

Coursera/Erasumus U., Econometric Methods and Applications

Anthony Nguyen

2018 10 26

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 with mean zero and fixed variance σ^2. The data set 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.

(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?

(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?

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

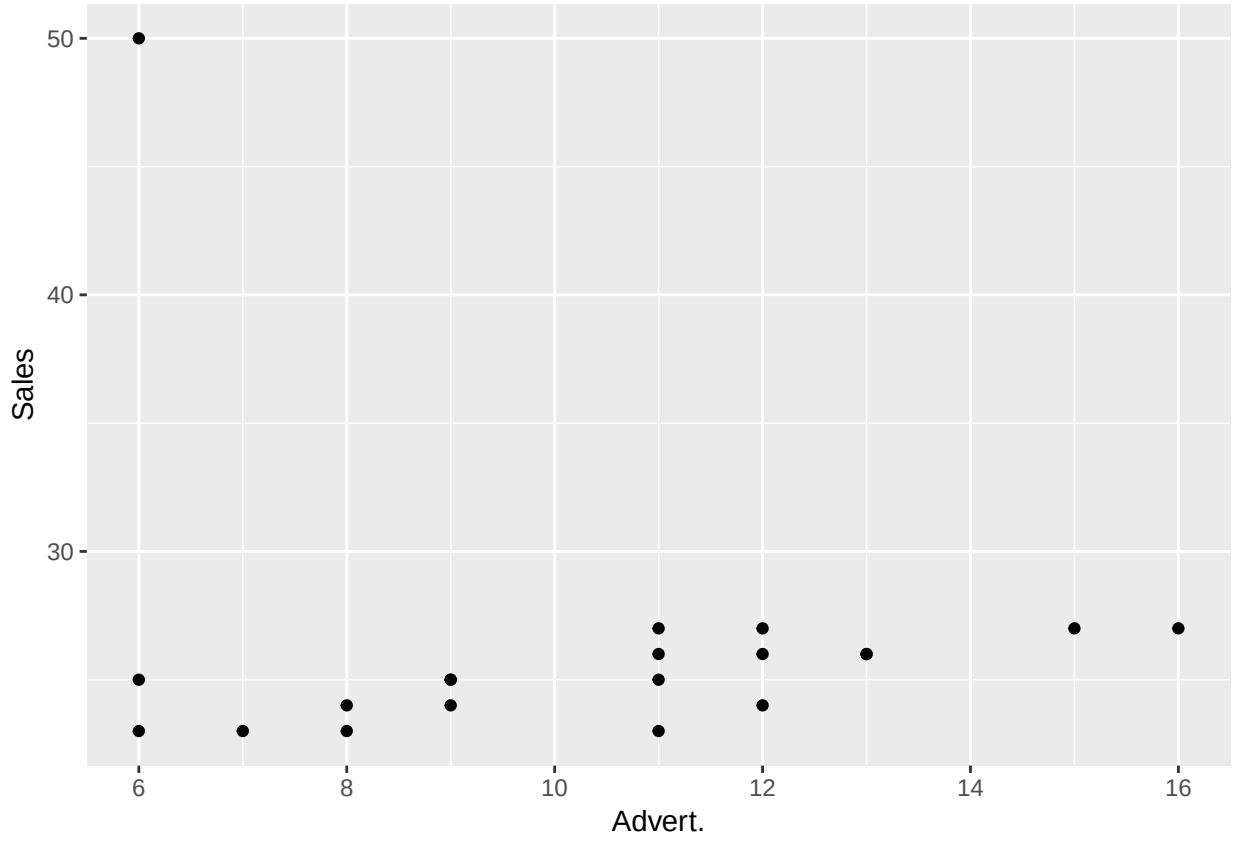
(d) Apparently, the regression result of part (b) is not satisfactory. Once you realize that the large residual corresponds to the week with opening hours during the evening, how would you proceed to get a more satisfactory regression model?

(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?

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

Answers

(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?



Looking at the scatter plot for Sales vs. Advertising, it seems like there may be a slightly positive linear relationship, if the outlier point around x = 6 is removed.

If the outlier point is left and a regression line is plotted, I would expect the line to actually take a slightly negative value for the slope, just to account for the outlier which is so large.

(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?

We can compute the values for b and a using the standard formulas, where:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
## [1] "b = -0.324574961360124"
```

$$a = \bar{y} - b\bar{x}$$

```
## [1] "a = 29.6268933539413"
```

To compute the standard error and t-value of b, first we calculate a vector e with all of the residual error terms using:

$$e_i = y_i - a - bx_i$$

Once we have e, we can calculate the standard error of b (SE_b) using:

$$SE_b = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

```
## [1] "where SE_b= 0.458910975802911"
```

For the t-statistic, we can use:

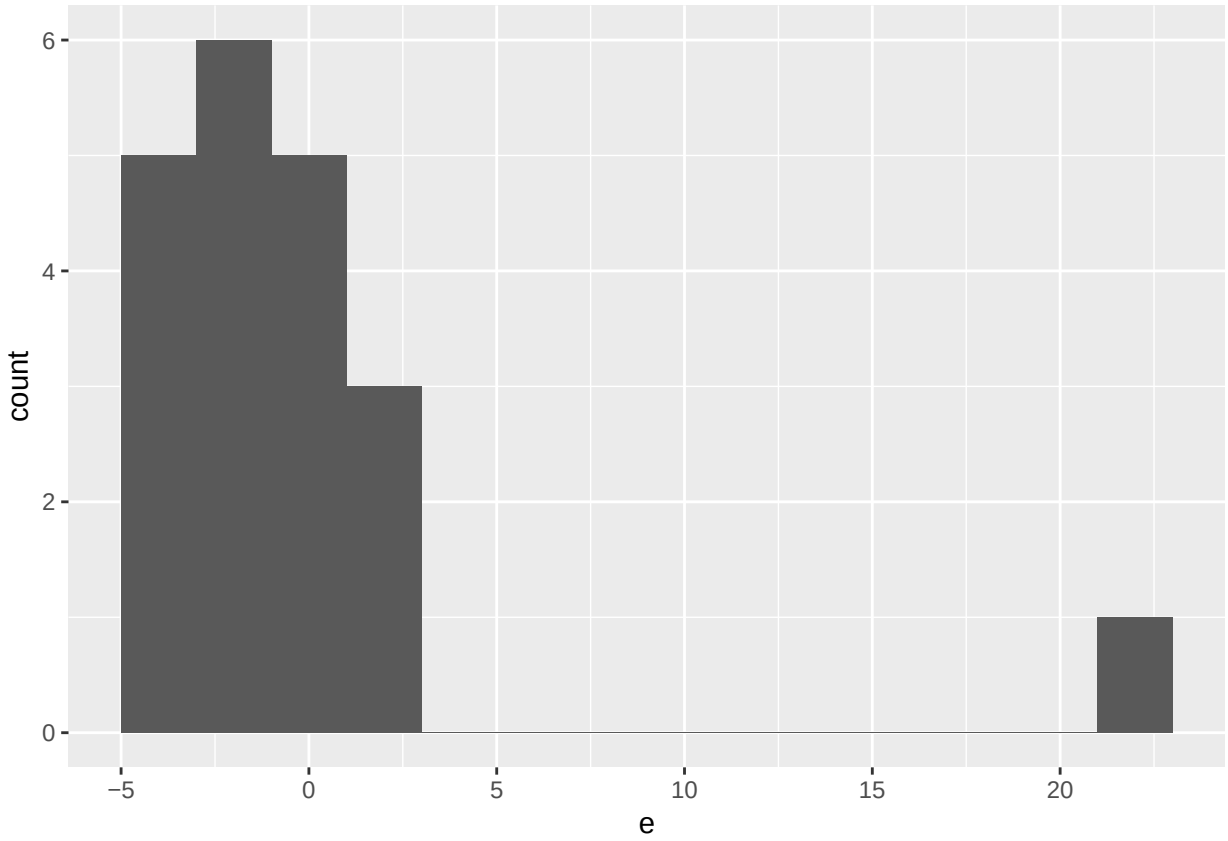
$$t = \frac{b - \beta}{SE_b}$$

Given that β is 0, then:

$$t = \frac{-0.3246 - 0}{0.4589} = -0.707$$

To answer the last part of the question, the t-value in this case is quite close to zero, which implies that there would not be a significant difference between b and zero (meaning a large p-value, and not-rejecting the null hypothesis).

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



Looking at the histogram, it's clear that there is an outlier in the data. This confirms what we saw in part (a) looking at the scatter plot.

(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?

Realizing the large residual corresponds to the week with opening hours during the evening, we could remove that particular row of data and re-run the regression.

(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?

After removing the outlier row (Observ. = 12) and re-running the regression, we find that the new values for a, b, the standard and t-value of b are:

$a = 21.1250$
$b = 0.3750$
$SE_b = 0.0882$
$t_b = 4.252$

We can see that after removing the outlier, the t-statistic value has increased fourfold, and is now, significantly different from zero.

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

Looking at the different coefficients for parts (b) and (e), I can see that some of the predictions from part (a) were indeed correct–meaning, that the initial regression line had a slightly negative slope to account for the one outlier observation, but when this observation was removed, the slope of the regression line took on a positive value (b value goes from -0.3246 to 0.3750).

In a similar fashion, the intercept (a) terms goes down from 29.627 to 21.125 not having to account for the outlier value.

Based on these results, there seems to be an inverse relationship between the standard error and the t-statistic–where the standard error approaches zero, the t-value becomes greater, and hence, more likely to be statistically significant.
We can also see that the residual error will be much greater the further the data points are away from the regression line and the associated mean values.
