Week1

August 15, 2018

1 Week 1 Graded Assignment

1.1 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 ˆ2.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.

(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?

1.2 TestExer1-sales

Simulated data set on weekly sales and advertising of a department store. 1. Advertising: index of advertising efforts in current and previous week 2. Sales: sales volume in current week

1.2.1 Solutions

Libraries / Modules

In [35]:
```
# getting the necessary libraries to read the data
import pandas as pd

# import the libraries for any potential mathematical operation
import math

# getting libraries for the scatter diagram
import matplotlib.pyplot as plt
import seaborn as sns

# necessary function to display the chart here once it is generated

%matplotlib inline
```

1.3 Read data

In [82]:
```
testExer1=pd.read_excel("C:\Users\\vishe\Downloads\_689829b5c78b9b71bd384ed8fb8714c1_
testExer1
```

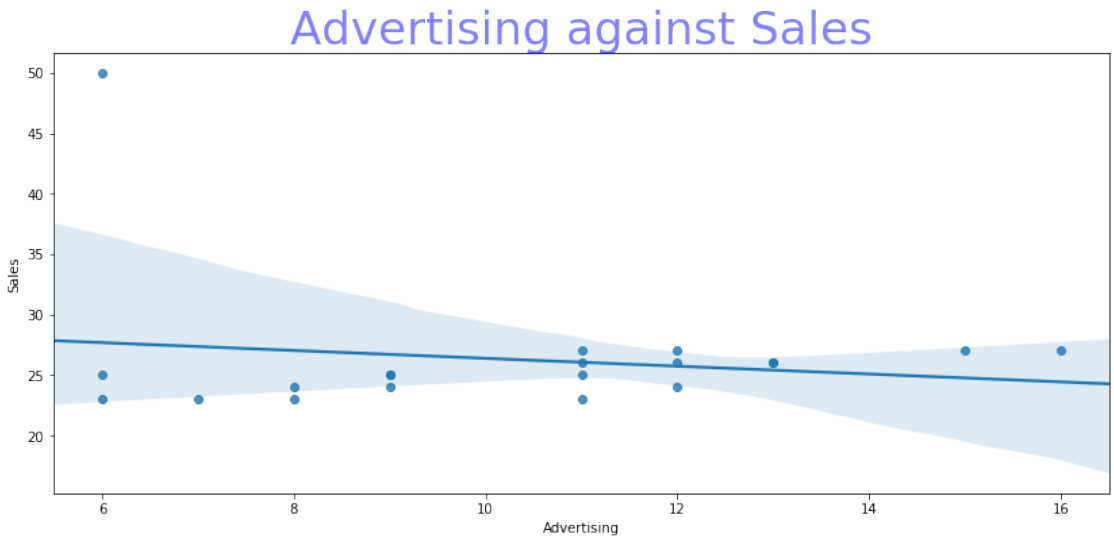
Out[82]:

	Observation	Advertising	Sales
0	1	12	24
1	2	12	27
2	3	9	25
3	4	11	27
4	5	6	23
5	6	9	25
6	7	15	27
7	8	6	25
8	9	11	26
9	10	16	27
10	11	11	25
11	12	6	50
12	13	13	26
13	14	11	23
14	15	13	26
15	16	7	23
16	17	8	23
17	18	8	24
18	19	12	26
19	20	9	24

1.3.1 Answer 1

In [37]:
```
plot = sns.regplot(data=testExer1, x= "Advertising", y= "Sales")
plot.figure.set_size_inches(14,6)
plot.axes.set_title('Advertising against Sales', fontsize=34,color="b",alpha=0.5)
```

```
Out[37]: Text(0.5,1,'Advertising against Sales')
```



intuitively and based on the scatter diagram, the regression line is expected to be a diagonal line starting from close to values-point [6, 30] aprox. dropping towards to values-point [16, 23] aprox. As the expected line appears to be diagonal (not perfectly horizontal) , we can intuitively conclude that Advertising affects Sales. It is also observed an outlier observation, at values-point [6, 50] aprox., in violation of the Assumpion A3 .

1.3.2 Answer 2 code and calculation

```
In [38]: ## calculating the value of b which is needed to derive a
         Y = testExer1.Sales              # the dependent variable
         X = testExer1.Advertising        # the independent variable
```

coefficient b

```
In [39]: b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
         print 'Value of b, the coefficient, rounded to 3 digits is', round(b,3)

Value of b, the coefficient, rounded to 3 digits is -0.325


In [40]: X_bar = testExer1.Advertising.mean()                # sample mean of age
         Y_bar = testExer1.Sales.mean()    # sample mean of expenditures

         print"Mean Advertising rounded to 3 digits is: ", round(X_bar, 3)
         print"Mean Sales rounded to 3 digits is: ", round(Y_bar, 3)

Mean Advertising rounded to 3 digits is:  10.25
Mean Sales rounded to 3 digits is:  26.3
```

intercept a

```
In [41]: a = Y_bar - b*X_bar
         print"Value of a, the intercept, rounded to 3 digits is:", round(a, 3)
```

```
Value of a, the intercept, rounded to 3 digits is: 29.627
```

Error

```
In [42]: testExer1["error"] = testExer1.Sales - a - b*testExer1.Advertising
         testExer1
```

Out[42]:

	Observation	Advertising	Sales	error
0	1	12	24	-1.731994
1	2	12	27	1.268006
2	3	9	25	-1.705719
3	4	11	27	0.943431
4	5	6	23	-4.679444
5	6	9	25	-1.705719
6	7	15	27	2.241731
7	8	6	25	-2.679444
8	9	11	26	-0.056569
9	10	16	27	2.566306
10	11	11	25	-1.056569
11	12	6	50	22.320556
12	13	13	26	0.592581
13	14	11	23	-3.056569
14	15	13	26	0.592581
15	16	7	23	-4.354869
16	17	8	23	-4.030294
17	18	8	24	-3.030294
18	19	12	26	0.268006
19	20	9	24	-2.705719

```
In [24]: sum_sq_error = (testExer1.error ** 2).sum() # calculating the sum of squares
         sum_sq_error
```

```
Out[24]: 613.1598145285934
```

Standard Error

```
In [43]: n = testExer1.shape[0]   # number of entries
         s = math.sqrt(1/(n-2.0) * sum_sq_error) # standard deviation
         print"The standard error rounded to 3 digits is: ", round(s, 3)
```

```
The standard error rounded to 3 digits is:  5.836
```

t-value From the lecture 1.4 and the correspoding slides

$$t_b = \frac{b - \beta}{s_b}$$

where

$$s_b{}^2 = \frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

and

$$\beta = b - \sum_{i=1}^{n} c_i e_i$$

where

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

The t-value of b (for =0)

```
In [44]: ## calculating sb^2 before arriving at t## calcu
         s_b_sq1 = s**2 / ((X - X_bar)**2).sum()
         s_b_sq=math.sqrt(s_b_sq1)
         s_b_sq

Out[44]: 0.45891097580291074

In [45]: t_b = (b )/s_b_sq
         print"The t value of b is: ", t_b # it is too low to round

The t value of b is:  -0.707272169275
```

Answer 2 Summary

```
In [46]: print"Quick Summary of Answer 1 results\n"
         print"Value of a, the intercept, rounded to 3 digits is: ", round(a, 3)
         print"Value of b, the coefficient, rounded to 3 digits is: ",round(b,3)
         print"The standard error rounded to 3 digits is: ", round(s, 3)
         print"The t value of b is: ", t_b

Quick Summary of Answer 1 results

Value of a, the intercept, rounded to 3 digits is:  29.627
Value of b, the coefficient, rounded to 3 digits is:  -0.325
The standard error rounded to 3 digits is:  5.836
The t value of b is:  -0.707272169275
```

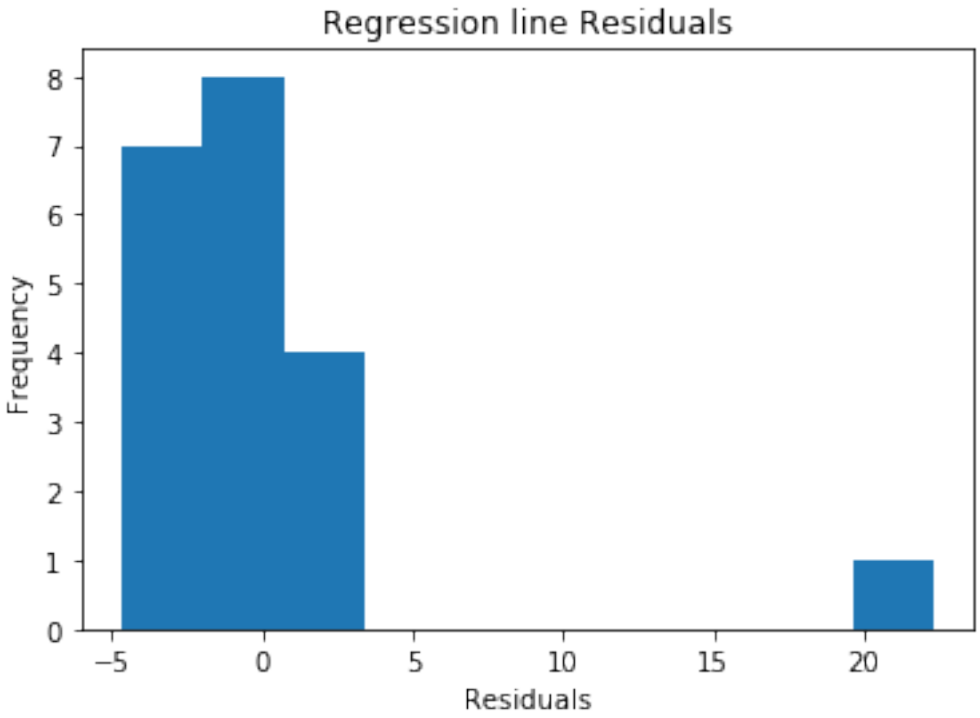
With a value of b=0.3246 and tb=-0.707272169275, it is concluded that b is negative and not significantly different from 0. Also, with a value of a=29.6269, it is concluded that a is significantly different from 0.

1.3.3 Answer 3

Residuals we have already estimated in earlier question

```
In [33]: plt.hist(testExer1.error)
         plt.title("Regression line Residuals")
         plt.xlabel("Residuals")
         plt.ylabel("Frequency")
         plt
```

```
Out[33]: <module 'matplotlib.pyplot' from 'C:\Users\vishe\Anaconda2\lib\site-packages\matplotli
```



Analysis: The histogram shows the most residuals concentrated in small values around zero, with the exception of a signle large residual (value over 22). Obviously, this is the observation #12 (Advertising = 6, Sales = 50) which describes the exceptional case of the special week when the store was also open in the evenings during this week. This violates Assumpion A3 and A4.

1.3.4 Answer 4

The most appropriate way for producing a better model for the store normal days, is to remove the one special/exceptional observation #12 from the data, and redefine the model.

1.3.5 Answer 5

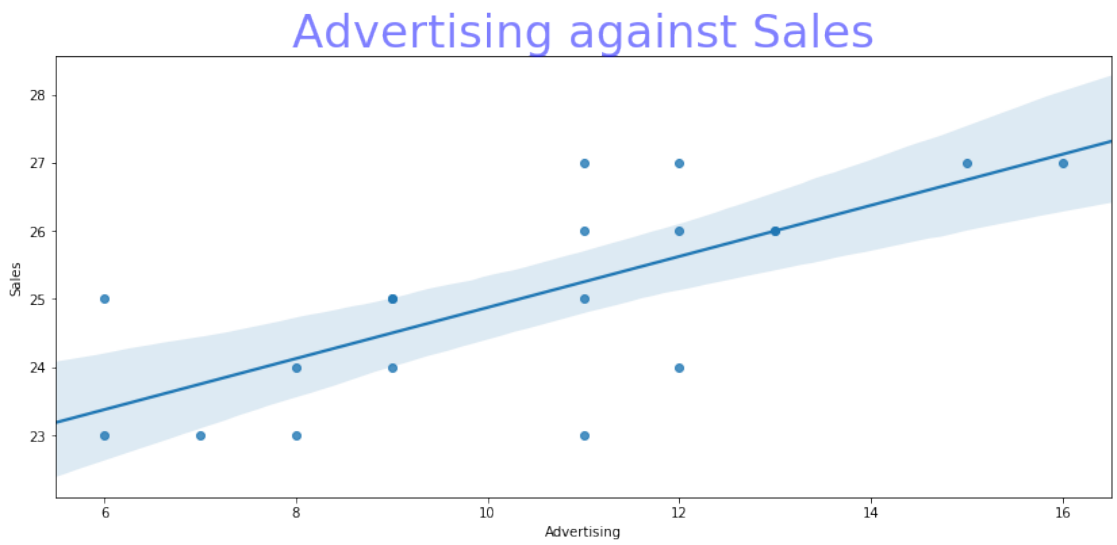
dropping the outlier

```
In [83]: testExer1=testExer1.drop(testExer1.index[11])
```

Plotting the graph

```
In [84]:  plot = sns.regplot(data=testExer1, x= "Advertising", y= "Sales")
          plot.figure.set_size_inches(14,6)
          plot.axes.set_title('Advertising against Sales', fontsize=34,color="b",alpha=0.5)
```

```
Out[84]: Text(0.5,1,'Advertising against Sales')
```



coefficient b

```
In [85]: ## calculating the value of b which is needed to derive a
         Y = testExer1.Sales                # the dependent variable
         X = testExer1.Advertising          # the independent variable
```

```
In [86]: b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
         print 'Value of b, the coefficient, rounded to 3 digits is', round(b,3)
```

```
Value of b, the coefficient, rounded to 3 digits is 0.375
```

Intercept a

```
In [87]: a = Y_bar - b*X_bar
         print"Value of a, the intercept, rounded to 3 digits is:", round(a, 3)
```

```
Value of a, the intercept, rounded to 3 digits is: 22.456
```

```
In [88]: n = testExer1.shape[0]
         n
```

```
Out[88]: 19
```

Standard Error

```
In [79]: testExer1["error"] = testExer1.Sales - a - b*testExer1.Advertising
         sum_sq_error = (testExer1.error ** 2).sum() # calculating the sum of squares
         n = testExer1.shape[0]   # number of entries
         s = math.sqrt(1/(n-2.0) * sum_sq_error) # standard deviation
         print"The standard error rounded to 3 digits is: ", round(s, 3)

The standard error rounded to 3 digits is:  1.781


In [90]: sum_sq_error

Out[90]: 50.77507812499995
```

t value

```
In [91]: # calculating sb^2 before arriving at t## calcultion
         s_b_sq1 = s**2 / ((X - X_bar)**2).sum()
         s_b_sq=math.sqrt(s_b_sq1)
         t_b = (b )/s_b_sq  ## t value calculation
         print"The t value of b is: ", t_b

The t value of b is:  2.52333814667
```

With a value of bnew=0.375 and tbnew=2.52333814667, it is concluded that now bnew is both positive and significantly different from 0. While, with a value of anew=22.456, it is concluded that anew still is significantly different from 0.

1.3.6 Answer 6

The regression line slope change caused by the special observation #12 removal, as indicated above by the significant b coefficient increase (and not only), was expected as the outlier Sales value (for a relatively low Advertising) that was removed was initially distorting the regression line very much. In other words, we now have a regression line slope increasing rather than decreasing, meaning a positive b instead of a negative one. This aligns with the intuition given by the two models scatter diagrams, provided previously in sections (a) and (e).

The a coefficient slight decrease was also expected, as the removed outlier had an extremely high y value. So, now the regression line starts slightly lower in the y-axis than before.

The standard error s significant decrease and the the t-value of b significant increase, both indicate that our regression model is now much better (more efficient), meaning that the expected variations are now much smaller than initially. This is also indicated by the significantly lower new sum of residuals square values (50.775 now, against 613.1598 initially).

1.3.7 – End of document –