MOOC Econometrics: Test Exercise #1

Greg Faletto

Questions

This exercise considers an example of data that do not satisfy all the standard assumptions of simple regression. In the considered case, assumption A6 that the coefficients α and β are the same for all observations is violated. The dataset contains survey outcomes of a travel agency that wishes to improve recommendation strategies for its clients. The dataset contains 26 observations on age and average daily expenditures during holidays.

Problem (a)

Use all data to estimate the coefficients a and b in a simple regression model, where expenditures is the dependent variable and age is the explanatory factor. Also compute the standard error and the t-value of b.

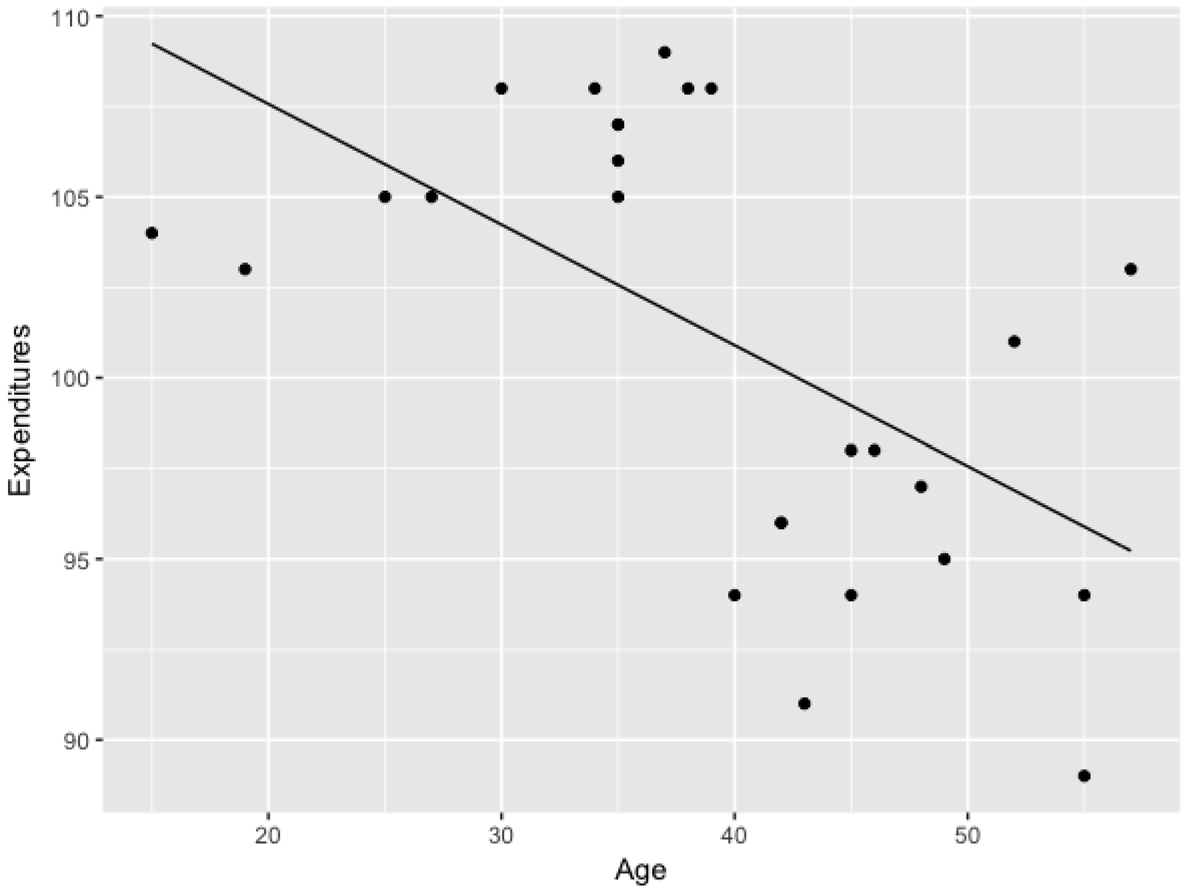
Solution

```
[1] "Regression line:"
[1] "b= -0.333596096606276"
[1] "a= 114.241107954932"
[1] "Standard error of b: 0.0953691827886391"
[1] "t-value of b: -3.49794437628353"
```

Problem (b)

Make the scatter diagram of expenditures against age and add the regression line $y = a + bx$ of part (a) in this diagram. What conclusion do you draw from this diagram?

Solution



The conclusion I draw from this diagram is that Age provides some predictive value in predicting Expenditures, but it seems as though people under 40 have significantly different spending habits as a function of age than people over 40. Specifically, for people under 40 expenditures seem to increase with age, and for people over 40 there isn't a clear trend. Splitting the data into two groups might yield better results.

Problem (c)

It seems there are two sets of observations in the scatter diagram, one for clients aged 40 or higher and another for clients aged below 40. Divide the sample into these two clusters, and for each cluster estimate the coefficients a and b and determine the standard error and t-value of b.

Solution

```
[1] "Regression line (40 and over):"
[1] "b= 0.146470828233374"
[1] "a= 88.8718890248878"
[1] "Standard error of b (40 and over): 0.197384418725913"
[1] "t-value of b (40 and over): 0.742058715570468"
[1] "95% Confidence interval for b: ( -0.248298009218451 ,  0.541239665685199 )."
[1] "Regression line (Under 40):"
[1] "b= 0.197971278778208"
[1] "a= 100.232277182585"
[1] "Standard error of b (under 40): 0.0443836675864513"
[1] "t-value of b (under 40): 4.46045335015626"
[1] "95% Confidence interval for b: ( 0.109203943605306 ,  0.286738613951111 )."
```

Problem (d)

Discuss and explain the main differences between the outcomes in parts (a) and (c). Describe in words what you have learned from these results.

Solution

In part (c), I found a significantly different way to model the data by splitting the data into two groups and modeling them separately. Based on our limited sample data set, this seems to provide more predictive value than simply grouping all the data together and making one model. We may need a larger sample before we can know for sure if this is the best model for the data. Sometimes it may be better to group the data into different groups and then model the data separately.

The 95% confidence interval for b includes 0 in the model for customers 40 and over. This means we do not have significant evidence to reject the null hypothesis that $b = 0$ for customers over 40–in other words, we do not have significant evidence suggesting a correlation between age and expenditures for customers over 40.

The 95% confidence interval for b does not 0 in the model for customers under 40. This means we have significant evidence to reject the null hypothesis that $b = 0$ for customers under 40. We have significant evidence suggesting a correlation between age and expenditures for customers over 40.