Econometrics: Test Exercise 1

Jason

11/10/2017

## Introduction

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.

### Loading in the dataset

travelData <- read.table("TestExer1-holiday expenditures-round2.txt", header=TRUE, sep="\t")  
  
head(travelData)

## Observ. Age Expenditures  
## 1 1 49 95  
## 2 2 15 104  
## 3 3 43 91  
## 4 4 45 98  
## 5 5 40 94  
## 6 6 35 107

### Question 1:

**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.**

Using the following equation: where:

* = Expenditures
* = Age

x\_bar = mean(travelData$Age)  
y\_bar = mean(travelData$Expenditures)  
x\_bar

## [1] 39.34615

y\_bar

## [1] 101.1154

Therefore using these values of and we can now calculate the values of a and b.

x = travelData$Age  
y = travelData$Expenditures  
  
b = sum((y-y\_bar)\*(x-x\_bar))/(sum((x-x\_bar)^2))  
a = y\_bar-(b\*x\_bar)  
  
b

## [1] -0.3335961

a

## [1] 114.2411

Therefore

To calculate the standard error of b, we first need to calculate the standard error of the regression

Where e is the residuals

model <- lm(travelData$Expenditures ~ travelData$Age)  
  
residual <- residuals(model)  
  
s <- sqrt((1/(26-2))\*sum(residual^2))  
  
s

## [1] 5.073322

After calculating the standard error, we will then calculate the standard error of b using the following formula

s\_b = sqrt((s^2)/sum((x-x\_bar)^2))  
  
s\_b

## [1] 0.09536918

To calculate the t value of b, we use the following formula:

t\_b = b/s\_b  
t\_b

## [1] -3.497944

Our results can be confirmed with the following summary function of the linear model:

summary(model)

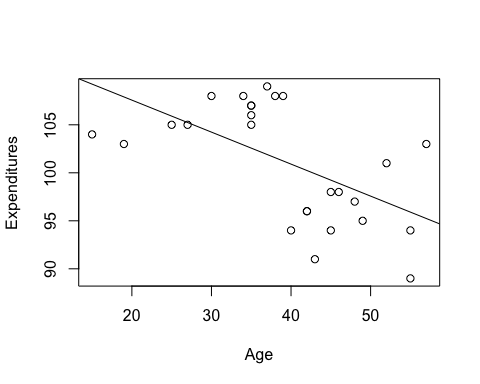
##   
## Call:  
## lm(formula = travelData$Expenditures ~ travelData$Age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8965 -4.2301 -0.8984 4.3525 7.7739   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 114.24111 3.88208 29.428 < 2e-16 \*\*\*  
## travelData$Age -0.33360 0.09537 -3.498 0.00185 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.073 on 24 degrees of freedom  
## Multiple R-squared: 0.3377, Adjusted R-squared: 0.3101   
## F-statistic: 12.24 on 1 and 24 DF, p-value: 0.001852

Summary:

### Question 2:

**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?**

plot(x, y, xlab="Age", ylab="Expenditures")  
abline(model)



From the scatter diagram above, it is very noticable that for lower ages, expenditures are much higher compared to values of age that are higher. The regression line also shows this relantionship.

### Question 3:

**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.**

We will first split the data into the two groups

travelDataYoung <- subset(travelData, Age < 40)  
nrow(travelDataYoung)

## [1] 13

travelDataOld <- subset(travelData, Age >= 40)

Starting with the data of people age less the 40 and using the same process as question 1, we calculate will first calculate the mean of age and expenditures in the data

x\_bar = mean(travelDataYoung$Age)  
y\_bar = mean(travelDataYoung$Expenditures)  
x\_bar

## [1] 31.07692

y\_bar

## [1] 106.3846

Using these values of and we can now calculate the values of a and b.

x = travelDataYoung$Age  
y = travelDataYoung$Expenditures  
  
b = sum((y-y\_bar)\*(x-x\_bar))/(sum((x-x\_bar)^2))  
a = y\_bar-(b\*x\_bar)  
  
b

## [1] 0.1979713

a

## [1] 100.2323

Therefore

Calculating the standard error of the regression

model <- lm(travelDataYoung$Expenditures ~ travelDataYoung$Age)  
  
residual <- residuals(model)  
  
s <- sqrt((1/(13-2))\*sum(residual^2))  
  
s

## [1] 1.153056

After calculating the standard error, we will then calculate the standard error of b

s\_b = sqrt((s^2)/sum((x-x\_bar)^2))  
  
s\_b

## [1] 0.04438367

Now we calculate the t value of b

t\_b = b/s\_b  
t\_b

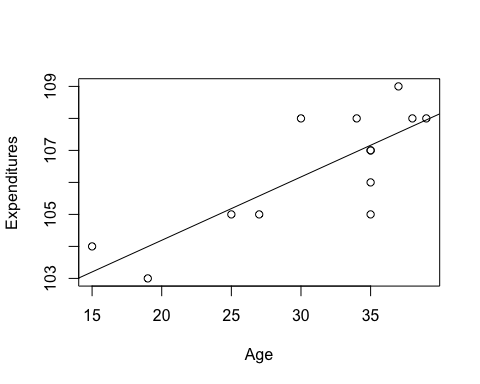
## [1] 4.460453

Our results can be confirmed with the following summary function of the linear model:

summary(model)

##   
## Call:  
## lm(formula = travelDataYoung$Expenditures ~ travelDataYoung$Age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1613 -0.5775 -0.1613 0.7982 1.8286   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 100.23228 1.41590 70.79 5.55e-16 \*\*\*  
## travelDataYoung$Age 0.19797 0.04438 4.46 0.000962 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.153 on 11 degrees of freedom  
## Multiple R-squared: 0.644, Adjusted R-squared: 0.6116   
## F-statistic: 19.9 on 1 and 11 DF, p-value: 0.0009619

plot(x, y, xlab="Age", ylab="Expenditures")  
abline(model)



Moving to the data where age is greater than 40, we calculate will first calculate the mean of age and expenditures in the data

x\_bar = mean(travelDataOld$Age)  
y\_bar = mean(travelDataOld$Expenditures)  
x\_bar

## [1] 47.61538

y\_bar

## [1] 95.84615

Using these values of and we can now calculate the values of a and b.

x = travelDataOld$Age  
y = travelDataOld$Expenditures  
  
b = sum((y-y\_bar)\*(x-x\_bar))/(sum((x-x\_bar)^2))  
a = y\_bar-(b\*x\_bar)  
  
b

## [1] 0.1464708

a

## [1] 88.87189

Therefore

Calculating the standard error of the regression

model <- lm(travelDataOld$Expenditures ~ travelDataOld$Age)  
  
residual <- residuals(model)  
  
s <- sqrt((1/(13-2))\*sum(residual^2))  
  
s

## [1] 3.832903

After calculating the standard error, we will then calculate the standard error of b

s\_b = sqrt((s^2)/sum((x-x\_bar)^2))  
  
s\_b

## [1] 0.1973844

Now we calculate the t value of b

t\_b = b/s\_b  
t\_b

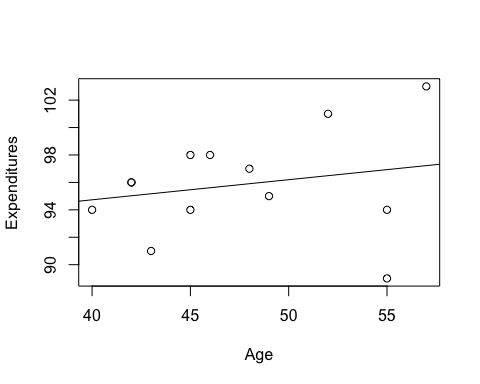
## [1] 0.7420587

Our results can be confirmed with the following summary function of the linear model:

summary(model)

##   
## Call:  
## lm(formula = travelDataOld$Expenditures ~ travelDataOld$Age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.9278 -1.4631 0.9763 2.3905 5.7793   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 88.8719 9.4585 9.396 1.37e-06 \*\*\*  
## travelDataOld$Age 0.1465 0.1974 0.742 0.474   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.833 on 11 degrees of freedom  
## Multiple R-squared: 0.04767, Adjusted R-squared: -0.0389   
## F-statistic: 0.5507 on 1 and 11 DF, p-value: 0.4736

plot(x, y, xlab="Age", ylab="Expenditures")  
abline(model)



Summary:

With the young age data:

With the old age data:

### Question 4:

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

The regression slope is now positive for part (c) unlike in part (a) where it was negative. This aligns with the data in the scatter plot from part (b) if the data was seperated. The coefficients in part (c) has also decreased as well.

The standard error of b decreased and the t-value of b increased for the young age data in part (c) compared to part(a) which indicates that the regression model for the seperate datasets was more accurate than the regression model when the dataset was combined.