Econometrics: Methods and Applications

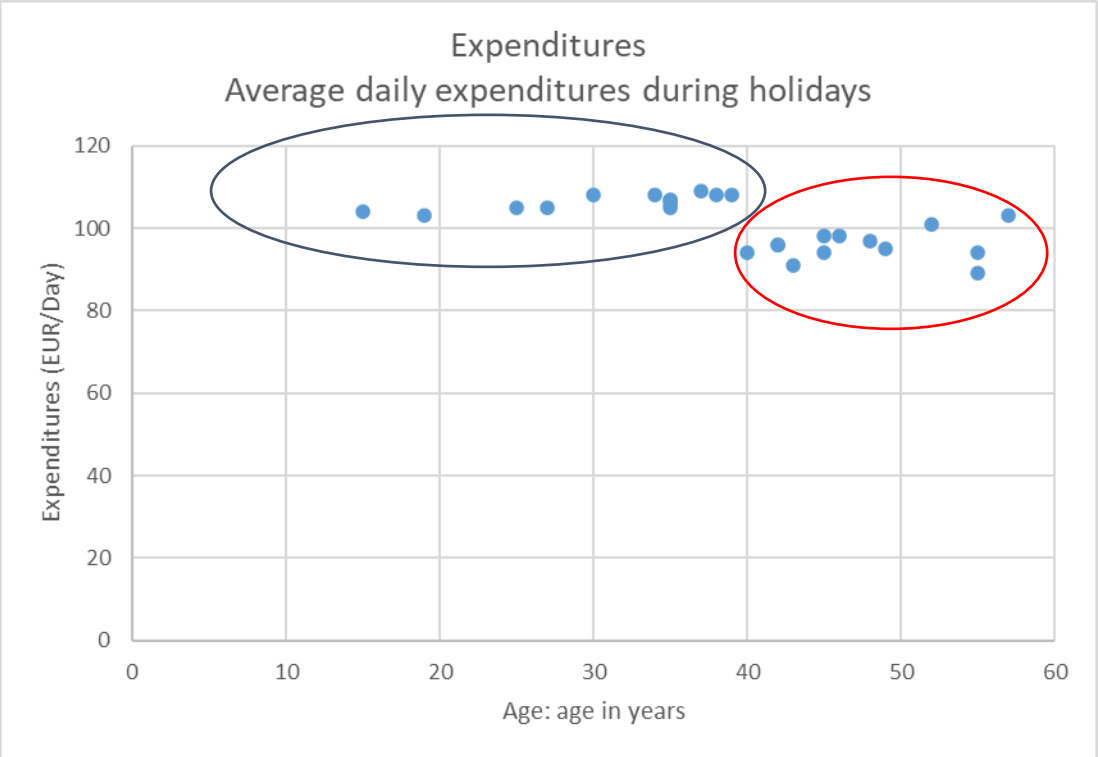
Erasmus University Rotterdam

Peer-graded Assignment: Test Exercise 1

In this assignment you will

- Get hands-on experience with performing simple regressions.
- Get feeling for consequences of violations of regression assumptions.
- Obtain some experience with how to diagnose that an assumption is violated

Step 1) Data Visualization



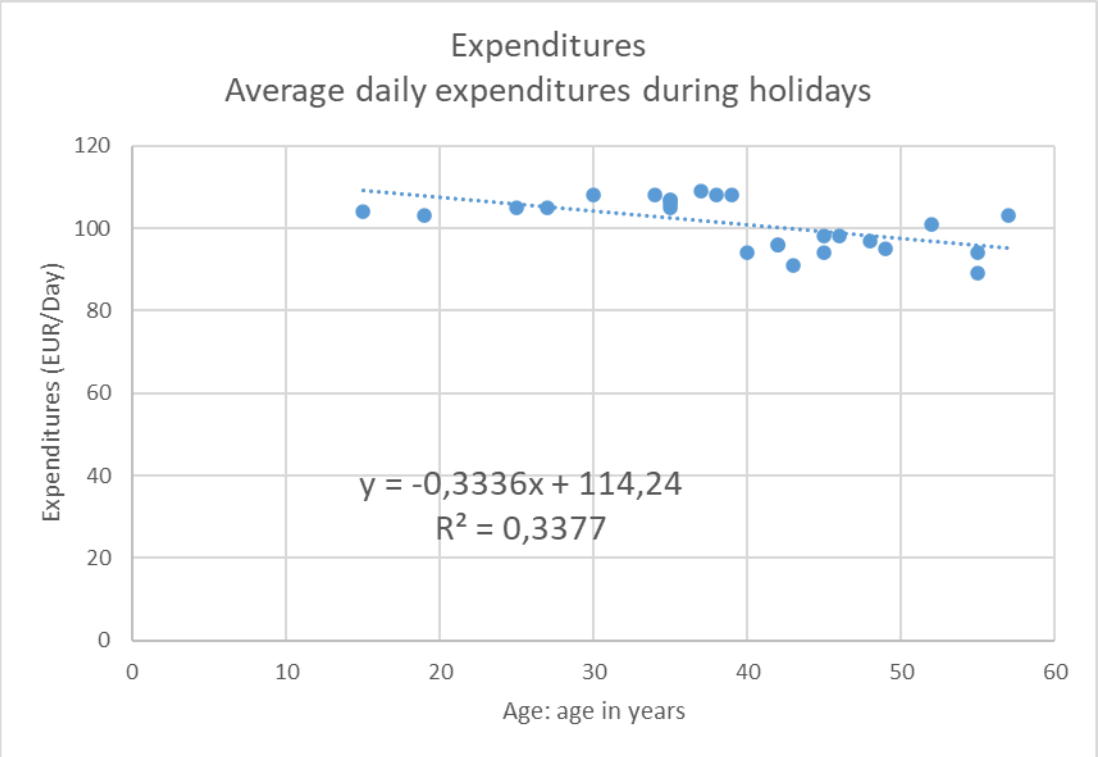
Plotting the data shows that there is a break between two groups of holiday makers, with people below age 40 spending more on average than people of higher age.

Step 2.) Calculate mean and standard deviation for the entire dataset

We calculate the mean and standard deviation for the 26 clients (i.e., observations):

	Age	Expenditures
Mean	39.3	101.12
STD	10.6	6.11
Standard deviation for a sample		

Step 3.) Run a linear regression on the entire dataset



Parameters of the linear regression

alpha	=	+114.24	(y-intercept)
beta	=	-0.3336	(slope)

The linear regression has a negative slope. Therefore, older people tend to spend less during their holidays than younger people.

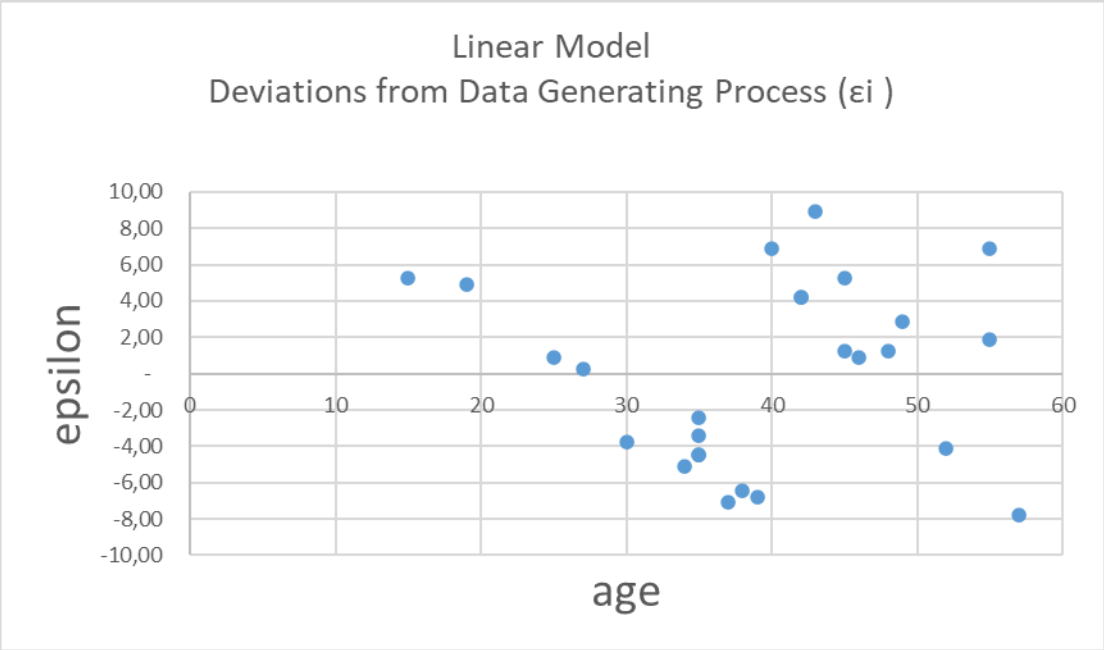
The R^2 of 0.3377 shows that only 33.8% of the variation in daily expenditure is explained by the age. However, there are violations of the assumptions of linear regressions:

No.	Assumption	Violation
A1	Data Generating Process is $y_i = \alpha + \beta * x_i + \varepsilon_i$	More than one x var. Choose among x-var.
A2	The n observations of x_i are fixed numbers.	Random x_i
A3/A6	The n error terms ε_i are random, with $E(\varepsilon_i) = 0$. $E(\varepsilon_i) = 0$, alpha, beta fixed	Parameter breaks
A4	The variance of the n errors is fixed	Heteroskedastic errors
A5	The errors are uncorrelated	Correlated errors
A6	α and β are unknown but fixed for all n observations. ε normal	Often not needed
A7	$\varepsilon_1 \ldots \varepsilon_n$ are jointly normally distributed. with A3, A4, A5: $\varepsilon_i \sim NID(0; \delta^2)$.	

Step 4.) Search for Parameter Breaks

	Age	Expenditures	epsilon
Mean	39,3	101,12	0,00
STD	10,6	6,11	4,97

The n error terms ε_i have an expected value of Null (i.e. $E(\varepsilon_i) = 0$. but they are not random.



The linear model predicts higher expenditure for younger people. However, people below 30 tend to spend more on vacations than the model predicts, and people between 30 and 40 tend to spend less than the model predicts. Similarly, people between 40 and 50 tend to spend more than the model predicts, and for people above 50 there are strong positive and negative deviations from the model prediction.

Conclusion

The assumption A3 that the n error terms ε_i are random, is violated. There are two groups of holiday goers with different spending habits. Therefore, we should use two Data Generating Processes, one for people below the age of 40, and one model for people aged 40 and higher.