Regression in SPSS (Frequentist) - Solutions

Developed by Naomi Schalken, Lion Behrens *and* Rens van de Schoot

last modified: 24-02-2017

Example Data

***Question**: Write down the null and alternative hypotheses that represent this question. Which hypothesis do you deem more likely?*

H0: Age^2 is not related to a delay in the PhD projects.

H1: Age^2 is related to a delay in the PhD projects.

Preparation - Importing and Exploring Data

***Question**: Have all your data been loaded in correctly? That is, do all data points substantively make sense? If you are unsure, go back to the .csv-file to inspect the raw data.*

The descriptive statistics make sense:

B3_difference_extra: Mean (9.97), SE (0.791)

E22_Age: Mean (31.68), SE (0.376)

E22_Age_Squared: Mean (1050.22), SE (35.970)

Regression Analysis

***Question**: Using a significance criterion of 0.05, is there a significant effect of age and age2?*

There is a significant effect of age and age^2, with b=2.657, p <.001 for age, and b=-0.026, p<.001 for age^2.

***Question**: What can you conclude about the hypothesis being tested using the correct interpretation of the p-value?*

Assuming that the null hypothesis is true in the population, the probability of obtaining a test statistic that is as extreme or more extreme as the one we observe is <0.1%. Because the effect of age^2 is significant, we reject the null hypothesis.

***Question**: How does your conclusion change if you follow this advice?*

Because the p-values for both regression coefficients were really small <.001, the conclusion doesn't change in this case.

Of course, we should never base our decisions on single criterions only. Luckily, there are several additional measures that we can take into account. A very popular measure is the confidence interval.

***Question**: What can you conclude about the hypothesis being tested using the correct interpretation of the confidence interval?*

Age: 95% CI [1.504, 3.810]

Age^2: 95% CI [-0.038, -0.014]

In both cases the 95% CI's don't contain 0, which means, the null hypotheses should be rejected. A 95% CI means, that if infinitely samples were taken from the population, then 95% of the samples contain the true population value. But we do not know whether our current sample is part of this collection, so we only have an aggregated assurance that in the long run if our analysis would be repeated our sample CI contains the true population parameter.

***Question:** What can you say about the relevance of your results? Focus on the explained variance and the standardized regression coefficients.*

R^2 = 0.063 in the regression model. This means that 6.3% of the variance in the PhD delays, can be explained by age and age^2. The standardized coefficients, age (1.262) and age^2 (-1.174), show that the effects of both regression coefficients are comparable, but the effect of age is somewhat higher.

***Question:** Drawing on all the measures we discussed above, formulate an answer to your research question.*

The variables age and age^2 are significantly related to PhD delays. However, the total explained variance by those two predictors is only 6.3%. Therefore, a large part of the variance is still unexplained.