CHAPTER 7 – Tests of relationships

For this chapter, use the Salaries.csv file.

**Correlations**

Using the Salaries.csv file, we want to find the correlations between the years since the respondents completed their PhDs, their years in service and their salaries. Interpret the results. Which pairings are strongly correlated and which are not? Do the correlations appear to be significant? Use the *Spearman* test for this example.

file = read.csv("Salaries.csv")

colnames(file)



experience = with(file, data.frame(yrs.since.phd, yrs.service, salary))

experience = na.omit(experience)

options(scipen = 999) # to rid us of scientific notation

yearsPHD = experience$yrs.since.phd

yearsService = experience$yrs.service

salary = experience$salary

You could examine individual pairings, e.g.

cor(yearsPHD, yearsService, method="spearman")

plot(yearsPHD, yearsService, method="spearman")

However, you could use the code used in the section 'Multiple correlations – non-parametric – using Spearman/Kendall's tau-b'. The result could look like this:



Or this:





The dark blue formation shows a more well-defined relationship, between tenure and years since the PhD, as compared to the rather less well-correlated pairings with salary.

The results shows that years in service and years since PhD are strongly correlated. The correlation coefficient is 0.91 and it is significant since the *p* value is less than 0.05.

Salary and years since PhD are also positively correlated and this correlation is significant. However, it is not as strong as the first two since the coefficient is 0.48. The significant correlation shows that as years since PhD increases, the professors' salaries increase as well.

Lastly, years in service and salary have a significant positive correlation but could be seen as the weakest pairing, since the correlation coefficient is 0.425. The salary increases with years in service.

**Simple Linear Regression**

Fit a linear model that predicts a professor’s salary given his or her years in service.

model = lm(salary ~ yearsService, data = experience)

options(scipen = 999, digits=3) # Tidies output

summary(model)

Salary is the dependent variable, placed to the left of the tilde.



Model interpretation:

Based on the model fit measures, years in service can interpret 11.2% of the variability of a professor’s salary (use Multiple R-squared for a bivariate analysis). Moreover, the estimates show that for every year's increase in a professor’s tenure, his or her salary increases by $780.

**Multiple Regression**

Fit a multiple linear regression model that predicts a professor’s salary using his or her years since earning a PhD, years in service, and gender. Analyze to see if the assumptions were satisfied and perform the necessary corrections if not.

We add gender to the mix, using the 'sex' variable; salary remains the criterion (dependent variable).

experienceB = with(file, data.frame(yrs.since.phd, yrs.service, salary, sex)) # adds gender

experienceB = na.omit(experienceB)

yearsPHD = experienceB$yrs.since.phd

yearsService = experienceB$yrs.service

salary = experienceB$salary

gender = experienceB$sex

To test for multicollinearity, you want something like this:

library(mctest)

predictors = data.frame(yearsPHD, yearsService, gender) # new data frame with required variables

responseVariable = salary

options(digits = 3) # For clearer output

vifTol = imcdiag(predictors, responseVariable)

vifTol$idiags[1:3, 1:2] [1:3, 1:2] # the '3' refers to the number of variables



You want a low VIF and high tolerance. In this case, the opposite can be observed for the years since PhD and years in service variables. These show that they are more likely to be correlated with each other. Note that from the correlation results before this section, the two variables have an extremely high Pearson correlation coefficient, providing further confirmation that the assumption was not met.

Before we can use the Durbin-Watson test, we need to create our multiple regression model:

model1 = lm(salary ~ yearsPHD + yearsService + gender)

Then follow the instructions for running the **Durbin-Watson test**, as shown early in the section 'Assumptions for multiple regression' section. You should get an **autocorrelation result of 0.0537, D-W statistic 1.89 and a *p* value of something like 0.23 or 0.24**; the *p* value is non-deterministic. The Durbin Watson statistic should be between 1.5 and 2.5 to say that autocorrelation is not present. This was satisfied in the model since the DW Statistic is equal to 1.89. Moreover, the null hypothesis in this test statistic is that there is no autocorrelation between the residuals; since the *p* value is greater than 0.05, we do not reject the null hypothesis.

As in the same section, we can create a q-q plot:



For the test of normality, looking at the Q-Q plot, there is not much deviation from the normal line.

*Since there is multicollinearity between years of service and years since PhD, remove one of these variables and retain the other within the data set.* The basis for deciding on which to remove depends upon the researcher but for simplicity, retain yearsPHD since it has a higher correlation to salary based on the previous output.

Model2 = lm(salary ~ yearsPHD + gender)

Let us go through the assumption checks again:



There is high tolerance and low VIF for both yearsPHD and gender. Thus, the collinearity assumption has been met.

Autocorrelation: 0.0688 Durbin-Watson statistic: 1.86 p: 0.178 (p will vary a bit)



Looking at the assumptions in the same manner as we did before, you can verify that all were met this time. We can proceed to interpret the model results.



The model fit shows that 18,2% of the variability of salary can be explained by years since PhD and salary (as this is a bivariate analysis again, we use Multiple R-squared. For the estimates, for every 1-year increase since being a PhD graduate, annual salary increases by $958 on the average. Male professors are likely to earns $7924 more than their female counterparts.