Florian Wickelmaier
Forschungsmethoden und Mathematische Psychologie
Fachbereich Psychologie
Universität Tübingen

Statistik III/Advanced Statistics
SS 2026

Exercises

Exercise 16. The data in file `orthography.txt` are taken from a diagnostic test of spelling difficulties in 5th and 6th graders from secondary schools in Baden-Württemberg. It includes information on the following variables:

X_1	`CFT`	Culture Fair Intelligence Test (Subtest 1)
X_2	`WM`	Phonological working memory performance (subtest of VLMT)
X_3	`sex`	Gender (dummy coded: 0 male, 1 female)
X_4	`school`	School type (dummy coded: 0 Hauptschule, 1 Realschule)
X_5	`class`	Grade level (dummy coded: 0 grade 5, 1 grade 6)
Y	`score`	Number of correct spellings in 20 words

Answer the following questions using linear regression analysis in R. For all statistical tests assume a Type I error rate of $\alpha = 0.05$.

a) Does each of the general cognitive abilities captured by X_1 and X_2 contribute to predicting spelling performance, and how large is the proportion of variance explained by the two variables in total?

b) Do the data meet the assumptions for conducting the regression analysis and the corresponding statistical tests? Perform appropriate graphical tests.

c) In order to decide whether there are gender-related differences, include variable X_3 as an additional predictor into the model.

   (a) Specify the general gender-specific regression equations for male and female students.

   (b) Does gender provide an additional contribution to the prediction of spelling performance? Convince yourself that the test of the regression coefficient β_3 as appearing in the output of `summary()` gives the same result as computing an incremental F-test (via function `anova()`) to the previously considered regression model with predictors X_1 and X_2.

   (c) How is the prediction of the regression model based on the predictors X_1, X_2, and X_3 to be interpreted geometrically? What is the interpretation of the parameter estimates?

d) How many correct spellings does the previous model predict across the observed range of CFT values for male and female students with an average working memory performance? Draw the predictions into a scatter plot of the data.

e) Now make a new model that includes all the predictors $X_1, \ldots, X_5$? Test the hypothesis $\beta_4 = \beta_5$. With the result from this test, comment on the statement: "School type has the greatest influence on the prediction of the spelling difficulties"?

f) Plan a replication study for the previous test. How many observations would be necessary to detect an effect of school type that is two additional correct spellings larger than the effect of class? To answer this question, run a simulation-based power analysis.

- Create a data frame for the five predictor variables:

`CFT`	uniformally sample values from $\{1, \ldots, 13\}$
`WM`	uniformally sample values from $\{0, \ldots, 12\}$
`sex`	uniformally sample from $\{0, 1\}$
`school`	assign 0 to one half and 1 to the other half of students
`class`	assign 0 to one half and 1 to the other half of students

- Set the parameters to

$$\boldsymbol{\beta} = (2, .17, .27, 1.4, 4, 2)' \quad \sigma = 3.6.$$

- Generate predicted response values using the regression model.
- Estimate the power of the test against H_0: $\beta_4 = \beta_5$.
- Adjust the number of observations until the test has 80% power.
- Comment on the plausibility of the simulation assumptions.