Statistisk Modellering (ST523,ST813)
Exercise Session 6

All exercises should be prepared BEFORE the exercise session.

Application of F-tests

The exercises use the economic dataset on 50 countries, `savings`, from the `faraway` package. The data represents averages from 1960 to 1970 (to remove business cycle or other short-term fluctuations):

- `sr` is aggregate personal saving divided by disposable income;
- `pop15` is the percentage of population under 15;
- `pop75` is the percentage of population over 75;
- `dpi` is per capita disposable income in U.S. dollars;
- `ddpi` is the percentage rate of change in per capita disposable income.

Exercise 6.1

- Start out inspecting the dataset e.g. what is the sample size, how many variables and of which type. Produce a scatterplot matrix using the command `pairs`.
- Model `sr` as reponse variable depending on the remaining variables using a normal linear model. Fit a regression model and interpret the output.

Exercise 6.2
(Test of all the predictors)

- Perform an overall F-test. Does any of the predictors have significance in the model? In other words, can you reject $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$?
- Verify the results for the F-test from **R** 's regression summary by repeating the underlying calculations on your own.

Exercise 6.3
(Testing a single predictor)

Use the general F-testing approach to test the null hypothesis that `pop15` is *not* significant in the full model.

- The test performed should be relative to the other predictors in the model, namely, `pop75`, `dpi`, and `ddpi`. Stating the null hypothesis as $H_0 : \beta_1 = 0$ is somewhat imprecise - try to specify this formulation by stating null hypothesis together with the underlying model assumptions and an alternative hypothesis.
- Fit the model representing the null, i.e. `sr` $\sim$ `pop75 + dpi + ddpi`.
- Compute the residual sum of squares, the F-statistic, and the p-value.
- Relate this to the t-based test and p-value.
- Compare the two nested models using the built-in R function `anova`.

Exercise 6.4
(Testing a pair of predictors)

Test the hypothesis that both `pop75` and `ddpi` may be excluded from the model. Hint: Try to fit a model with and then without them and use the general F-test.

Exercise 6.5
(Testing a nested model)
We might hypothesize that the effect of young and old people on the savings rate was the same:

$$H_0 : \beta_{\text{pop15}} = \beta_{\text{pop75}}$$

In this case, the null model would take the form

$$y = \beta_0 + \beta_{\text{dep}}(\text{pop15} + \text{pop75}) + \beta_{\text{dpi}}\text{dpi} + \beta_{\text{ddpi}}\text{ddpi} + \varepsilon.$$

We can then compare this to the full model as follows:

```
> fit_1 <- lm(sr ~ ., data = savings)
> fit_0 <- lm(sr ~ I(pop15 + pop75) + dpi + ddpi, data = savings)
> anova(fit_0, fit_1)
```

Interpret the results and conclude. Is there evidence for that young and old people need to be treated separately in the context of this particular model?

Exercise 6.6
You want to test whether one of the coefficients can be set to a particular value. Try to test $H_0 : \beta_{\text{ddpi}} = 0.5$.
Hint: In this case, the null model would take the form

$$y = \beta_0 + \beta_{\text{pop15}}\text{pop15} + \beta_{\text{pop75}}\text{pop75} + \beta_{\text{dpi}}\text{dpi} + 0.5\text{ddpi} + \varepsilon.$$

Fit this model and compare it to the full model. Can you reject the null hypothesis?

```
> fit_1 <- lm(sr ~ ., data = savings)
> fit_0 <- lm(sr ~ pop15 + pop75 + dpi+ offset(0.5*ddpi), data = savings)
> anova(fit_0, fit_1)
```

Try to test the same point hypothesis using a t-statistic:

$$t = \frac{\widehat{\beta} - c}{\text{SE}(\widehat{\beta})}$$

where c is the point hypothesis.

Verify that the p-value is the same as before, and verify that the squared t-statistic is equal to the F-value.

Factor variables and parametrizations

Exercise 6.7

Parametrization in oneway layout

Consider a oneway layout together with the following parametrization:

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij},$$

where ε_{ij}, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$, are i.i.d. with zero mean and variance σ^2. In order to achieve identifiability one adds a linear constraint to $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$.

- What are the model matrices for the first-category-baseline-parametrization (i.e. assuming $\beta_1 = 0$) and for sum-to-zero-contrasts (i.e. assuming $\sum_{l=1}^{k} \beta_l = 0$).
- Explain how to interpret the coefficients β in both cases.