Notes regarding the research reports

Marcus Wurzer

This document contains notes regarding the intermediate and final reports for the regression project in MDS-STCO.

1 General guidelines

You have to hand in

- an R Markdown/Quarto file allowing me to go through the R code if needed to comprehend certain steps of your analysis. From this it follows that `echo = TRUE` has to be set for every code chunk without exception.
- a PDF generated from this R Markdown/Quarto file.

The document has to be written in the vein of the statistical reports for a colleague that some of you already had to write in the *Applied Statistics* bridging course (see chapter 15 of the "Data Analysis And Statistical Modeling" script for details), but with one important difference: A rigid scheme is not meaningful for more complex analyses and thus, no specific template is provided here (but those of you that didn't visit the course mentioned above may have a look at the two exemplary reports that have been uploaded to the "Regression Project" section of the Moodle course).

- Because of the many plots, tables and summaries that you will have to produce, it makes no sense to give a limit of pages. Concerning the graphs, please pay attention to their interpretablility and readability.
- Try to write coherent code and do also use functions like `apply()` etc. in order to reduce the lines of code.
- Every step of the analysis has to be done using R, i.e., do not perform data manipulation, visualizations etc. in other programs.
- The R packages used have to be embedded in the file (e.g., `library(car)`) in order to avoid errors because of functions not being available.
- It is permitted to use short comments within the code chunks, but the detailed descriptions and interpretations have to be included as running text.
- Do not use `paste()`, `cat()` and the like! As an example, if we performed the following χ^2-test

```
if (!require(carData)) install.packages("carData"); library(carData)
(tab <- table(TitanicSurvival[, 1:2]))
```

```
##          sex
## survived female male
##      no     127  682
##      yes    339  161
```

```
(x <- chisq.test(tab))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  tab
## X-squared = 363.62, df = 1, p-value < 2.2e-16
```

it would be considered bad style if you did the following:

```
if (x$p.value > 0.05) {
  cat("The test does not show a significant result.")
} else {
  cat("The test shows a significant result.")
}
```

```
## The test shows a significant result.
```

Either you just check the p-value and plainly write "The test shows a significant result" (it should be doable to compare the two numbers by visual inspection) or - if you really want to dynamically generate running text based on the results of the computations - you utilize inline R code:

```
if (x$p.value > 0.05) {
  res < "The test does not show a significant result."
} else {
  res <- "The test shows a significant result."
}
```

Embedding `` `r res` `` in the running text will then give you the following sentence which has been generated by evaluating the above inline expression:

The test shows a significant result.

2 Structure of the report

The following points are not to be seen as mere suggestions, but really have to be worked through. The number in square brackets indicate the maximum number of points you can get for the specific part.

2.1 Interim report (deadline: 2024-12-17)

The interim research report encompasses data management and descriptive statistics.

2.1.1 Introduction [5]

- Starting point; Objective of the analysis
- Methods of analysis used (*short* description in two sentences)

2.1.2 Data collection [5]

- Type of survey; facts concerning the execution of the survey (period etc.)
- Description of the data set/operationalization (type of sample, sample size, variables, scale levels, missing values etc.)
- Data preparation (missing value treatment, transformations, ...)

⇒ see *Fundamentals of empirical social research III* in the Meyer/Wurzer script for details about the above points

2.1.3 Descriptive analysis of the sample [35]

- Descriptive analysis of the analyzed variable(s)
  - Diagrams, numerical measures, tables, ...
  - All statistics have to be commented, in particular diagrams!

  - Are there any distinctive features? (e.g., group differences, trends, outliers, ...)
- In detail, the following plots have to be produced:
  - Univariate visualizations of all variables
  - Bivariate relationships between predictors and response to show the influence of the former on the latter
  - Joint influences of all possible pairs of predictors on the response to show possible interactions (exception: the interaction between the two metric variables doesn't have to be visualized)
- Summary of the descriptive analysis. Based on these descriptive findings, segue to the analysis of the questions about the population

2.2 Final report (deadline: 2024-01-21)

In the final report, the whole analysis process is documented, i.e., it consists of the data management and descriptive statistics parts that have already been produced, with added regression and conclusion sections.

2.2.1 Inferential statistics/Regression modeling [50]

- The model building is to be seen as an iterative process. Several steps will be needed until you have found your final model (and remember: "Variable selection is a means to an end and ot an end itself" (cf. J. J. Faraway: Linear Models with R).
- When building the model, all possible main effects and two-way interactions have to be considered.
  1. Produce a naive model first, then perform a diagnostic analysis to check for possible problems (e.g., normal distribution, outliers, linearity, ...). Try to resolve them with the methods we learned in the course (data transformations, polynomial regression, temporary removal of outliers, ...) before you write down any interpretations or perform model selection!
  2. Perform a model/variable selection. You can base inclusion or deletion of terms on statistical tests or the AIC, but you should try out several prodedures (backward, forward, stepwise). Are the results consistent or do you get completely different models that seem to be roughly equally as good? If there seem to be various possibilities, choose one of these models for the subsequent steps.
  3. Interpret the model:
     - pay special attention to the significance of the parameters, confidence intervals and explanatory power (R^2)
     - Write down the model equation including interpretation (!), i.e., how to the independent variables influence the dependent one ("effect size")
  - utilize effect plots to present your model visually. Do also add a posterior predictive check to look for systematic dicrepancies between real and simulated data.
  - relate the statistical findings to the research question

2.2.2 Conclusion and criticism [5]

This part should be about one page long and include the following points:

- Summary
  - What has been done?
  - Providing answers to the research question
- Possible problems
  - Data problems
  - Analysis problems
- Generalizability of the findings?
- Possible further questions