HW 5: Correlations and linear models. Tests for categorial variables

1. Vowel reduction in Russian

Pavel Duryagin ran an experiment on perception of vowel reduction in Russian language. The dataset `shva` includes the following variables:
time1 - reaction time 1
duration - duration of the vowel in the stimuly (in milliseconds, ms)
time2 - reaction time 2
f1, *f2*, *f3* - the 1st, 2nd and 3rd formant of the vowel measured in Hz (for a short introduction into formants, see here)
vowel - vowel classified according the 3-fold classification (*A* - *a* under stress, *a* - *a/o* as in the first syllable before the stressed one, *y* (stands for shva) - *a/o* as in the second etc. syllable before the stressed one or after the stressed syllable, cf. *g*[*y*]*g*[*a*]*t*[*A*]*l*[*y*] *gogotala* 'guffawed').
In this part, we will ask you to analyse correlation between f1, f2, and duration. The dataset is available https://raw.githubusercontent.com/agricolamz/2018-MAG_R_course/master/data/duryagin_ReductionRussian.txt.

1.0

Read the data from file to the variable `shva`.

1.1

Scatterplot `f1` and `f2` using `ggplot()`. Design it to look like the following.

1.2

Plot the boxplots of `f1` and `f2` for each vowel using `ggplot()`. Design it to look like this and this.

```
# f1 boxplot

# f2 boxplot
```

1.3

Calculate Pearson's correlation of `f1` and `f2` (all data)

1.4 Calculate Pearson's correlation of `f1` and `f2` for each vowel

2 Linear regressions

2.1.1

Use the linear regression model to predict `f2` by `f1`.

2.1.2

Write down the equation for f2 using coefficients from the model (e.g. $y = b + kx$)

2.1.3

Provide the adjusted R^2

2.1.4

Add the regression line in the scatterplot 1.1.

2.1.5

Make a scatter plot for `f1` and `f2` grouped by vowels. Use `ggplot()` and `facet_wrap()`.

2.2.1

Use the linear regression model to predict `f2` by `f1` and `vowel`.

2.2.2

What is the intercept of the model?

2.2.3

Provide the adjusted R^2

2.2.4

Write down your general conclusions about the relationship between `f1`, `f2`, and `vowels`.

3. Dutch causative constructions

When the Dutch use two near-synonymous periphrastic causative verbs, *doen* and *laten*?

```
     De politie deed/liet de auto stoppen.
lit. the police did/let the car stop
     'The police stopped the car'
```

This is a data set on two rival constructions with *doen* and *laten* sampled from the newspaper corpora. The data frame includes 500 observations on the following 7 variables:

- `Aux` – verb: a factor with levels `doen` and `laten`
- `CrSem` – the semantic class of the Causer: a factor with levels `Anim` (animate) and `Inanim` (inanimate)
- `CeSem` – the semantic class of the Causee: a factor with levels `Anim` (animate) and `Inanim` (inanimate)
- `CdEvSem` – the semantic domain of the caused event expressed by the Effected Predicate: a factor with levels `Ment` (mental) and `NonMent` (e.g. physical or social)
- `CeSynt` – the syntactic status of the Causee: a factor with levels `Clause`, `Impl` (implicit, not expressed), `NP` (noun phrase), `PP` (prepositional phrase)
- `EPTrans` – transitivity or intransitivity of the effected predicate, a factor with two levels `Tr` and `Intr`
- `Country` – a factor with levels `BE` (Belgium) and `NL` (Netherlands)
- `Domain` – a factor with four levels for newspaper domains.

Data from Natalya Levshina's `RLing` package available (here)[https://raw.githubusercontent.com/agricolamz/2018-MAG_R_course/master/data/dutch_causatives.csv] Read more on the constructions in Levhina, Geerarts, Speelman 2014.

3.0

Read the data from file to the variable `d_caus`.

```
d_caus <- read.csv("https://raw.githubusercontent.com/agricolamz/2018-MAG_R_course/master/data/dutch_cau
summary(d_caus)
```

```
##     Aux        CrSem        CeSem        CdEvSem       CeSynt    EPTrans
##  doen : 85   Anim  :408   Anim  :317   Ment   :101   Clause: 22   Intr:239
##  laten:415   Inanim: 92   Inanim:183   NonMent:399   Impl  :134   Tr  :261
##                                                      NP    :268
##                                                      PP    : 76
##  Country  Domain
##  BE:220   E: 86
##  NL:280   F:116
##           M:150
##           P:148
```

3.1

We are going to test whether the association between `Aux` and other categorical variables (`Aux` ~ `CrSem`, `Aux` ~ `CeSem`, etc) is statistically significant. The assiciation with which variable should be analysed using Fisher's Exact Test and not using Pearson's Chi-squared Test? Is this association statistically significant?

3.2.

Test the hypothesis that `Aux` and `EPTrans` are not independent with the help of Pearson's Chi-squared Test.

3.3

Provide expected frequencies for Pearson's Chi-squared Test of `Aux` and `EPTrans` variables.

3.4.

Calculate the odds ratio for observed frequencies of `Aux` and `EPTrans` For 2×2 contigency table

$$\begin{matrix} a & b \\ c & d \end{matrix}$$

one can find *odds ratio* as $(a/c)/(b/d)$.

3.4.1

Find odds ratio for expected frequencies of `Aux` and `EPTrans`

What can you say about odds ratio of expected frequencies for arbitrary data?

3.5

Calculate effect size for this test using Cramer's V (phi).

3.6.

Report the results of independence test using the following template:

```
We have / have not found a significant association between variables ... and ... (p < 0.001).  The odds
```

3.7

Visualize the distribution using mosaic plot. Use `mosaic()` function from `vcd` library.

Below is an example of how to use mosaic() with three variables.

```
# mosaic(~ Aux + CrSem + Country, data=d_caus, shade=TRUE, legend=TRUE)
```

3.8

Why is it not recommended to run multiple Chisq tests of independence on different variables within your dataset whithout adjusting for the multiplicity? (i.e. just testing all the pairs of variables one by one)

3.9

Provide a short text (300 words) describing the hypothesis of this study and the results of your analysis.