Chapter 9 exercises

The set of exercises given in the document are intended to give a more "hands on" approach to the material in Chapter 9. The execireses are given per lecture, and are of course intended to be solved after that lecture, only lectures that cover material in Chapter 9 is included in this document.

Exercise 1.1 Lecture 1

a) **Summary statistics:** For the data in Example 1.20, write $\bar{\boldsymbol{y}}$, $\boldsymbol{S}$, $\hat{\boldsymbol{\sigma}}$, and $\boldsymbol{R}$, when $\boldsymbol{y}_i = [x_i, y_i]$.

Exercise 1.2 Lecture 2

a) **Eigenvalue:** From the previous question, use Python to find the eigenvalue decomposition of $\boldsymbol{S}$, and use that to find a matrix $\boldsymbol{S}^{\frac{1}{2}}$, such that $\boldsymbol{S} = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{S}^{\frac{T}{2}}$.

Exercise 1.3 Lecture 3

a) **Non-linear error propagation:** When estimating multiple probabilities $p_1, ..., p_k$ (with $\sum p_i = 1$), the multivariate logit transformation, with

$$p_i(\boldsymbol{\theta}) = \frac{e^{\theta_i}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}}; \quad i \neq k,$$

and $p_k = 1 - \sum_{i=1}^{k-1} p_i$, is often used. The transformation imply that for any $\boldsymbol{\theta} \in \mathbb{R}^{k-1}$, then $p_i \in (0,1)$.
In estimation it turns out that, under general condistions, it is resonable to assume that estimates ($\hat{\boldsymbol{\theta}}$) of $\boldsymbol{\theta}$ follow a multivariate normal distribution. Now assume that $\boldsymbol{\theta} \sim N_2(\hat{\boldsymbol{\theta}}, \sigma^2 \boldsymbol{I})$. With $\hat{p}_i(\boldsymbol{\theta})$ as above, find approximate values of the mean and variance of

$$\boldsymbol{p}(\boldsymbol{\theta}) = \begin{bmatrix} p_1(\theta_1, \theta_2) \\ p_2(\theta_1, \theta_2) \\ p_3(\theta_1, \theta_2) \end{bmatrix}$$

using the approximative non-linear error propagation. It is enough to write up the Jacobian and the solution in terms of the Jacobian, as a function of p_i.

b) **Non-linear error propagation:** With $\hat{\boldsymbol{\theta}} = \boldsymbol{0}$ expplicitly write the approximative variance.

c) **Multivariate normal distribution:** Let $\boldsymbol{Z} \sim N_2(\boldsymbol{0}, \boldsymbol{I})$ be a standard normal random variable. Also let the matrix $\boldsymbol{A}$ be given by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

What is the distribution of $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z}$, and can you calculate the value of the pdf at $\boldsymbol{Y} = [1,1,1]$?

Exercise 1.4 Lecture 5

a) **Multivariate normal distribution:** Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \bar{\boldsymbol{y}}$, and $\boldsymbol{\Sigma} = \boldsymbol{S}$, with $\bar{\boldsymbol{y}}$ and $\boldsymbol{S}$ as in Exercise 1, and consider the random variable

$$Q = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}),$$

what is the expected value and the variance of Q?

b) **Orthogonal projections:** Consider the 3 orthogonal projection matrices

$$\boldsymbol{H}_1 = \frac{1}{6}\begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix}; \quad \boldsymbol{H}_2 = \frac{1}{6}\begin{bmatrix} 4 & -2 & -2 \\ -2 & 1 & 1 \\ -2 & 1 & 1 \end{bmatrix}; \quad \boldsymbol{H}_3 = \frac{1}{6}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & -3 \\ 0 & -3 & 3 \end{bmatrix}.$$

With $\boldsymbol{Z} \sim N_3(\boldsymbol{0}, \boldsymbol{I})$, what is the expectation of the random variables

$$Q_i = \boldsymbol{Z}^T \boldsymbol{H}_i \boldsymbol{Z}?$$

Hint: Check the sum: $\boldsymbol{H}_1 + \boldsymbol{H}_2 + \boldsymbol{H}_3$.

c) Say you have 3 item (item 1, 2, and 3). Now items are put on a scale in the following order (and the weight is denoted as y_i)

1. Item 1 is put on the scale.
2. Item 1 and 2 is put on the scale.
3. Item 1 and 3 is put on the scale.
4. Item 2 and 3 is put on the scale.

Now consider the following model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

here $\boldsymbol{\beta} \in \mathbb{R}^3$. Write the design matrix ($\boldsymbol{X}$) for the following two interpretations of $\boldsymbol{\beta}$.

1. The estimate of β_i is the weight of item i, i.e. $\boldsymbol{\beta} = [\mu_1, \quad , \mu_2, \quad \mu_3]$.
2. Set $\mu_i = \mu + \delta_i$, and, using the constraint $\sum \delta_i = 0$, let $\boldsymbol{\beta} = [\mu, \quad \delta_1, \quad \delta_2]$.

Using Python check that the parametrizations are equivalent (i.e. they have the same projection matrices).

Exercise 1.5 Lecture 6

a) For the sleep medicine data in Example 3.27:

- Find the design matrix ($\boldsymbol{X}$) for the model.

- Use the design matrix to find the projection matrix ($\boldsymbol{H}$)

Further use Python to calculate:

- $\boldsymbol{y}^T\boldsymbol{H}\boldsymbol{y}$
- $\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$
- The observed F-test statistics and the $p-value$ for the hypothesis that there is no effect of the sleep medicine
- compare p-value and F-test statistic with the p-value in Example 3.27.

b) What is the estimate of the residual variance for the model?

c) The data is not collected as a time series, and hence checking lag-1 autocorrelation does not make a lot of sense, however for the sake of illstration, test if the lag-1 autocorrelation can be assumed to be zero anyway.

Exercise 1.6 Lecture 7

a) Consider the data in the nutrient study in Example 3.55, and consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Set up the design matrix $\boldsymbol{X}_1$ when the interpretation of the parameters ($\boldsymbol{\beta}$) is the mean/average in each of the two groups.

b) Now consider two alternative parametrizations

$$\boldsymbol{X}_2 = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{0} \\ \boldsymbol{1} & \boldsymbol{1} \end{bmatrix}, \quad \boldsymbol{X}_3 = \begin{bmatrix} \boldsymbol{1} & \frac{1}{2}\boldsymbol{1} \\ \boldsymbol{1} & -\frac{1}{2}\boldsymbol{1} \end{bmatrix}.$$

Using Python show that the three parametrizations are equvivalent, and give an inpterpretaion of the paremeters (you may consider $(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T$). Hint: For setting up the design matrix you may use (e.g. for $\boldsymbol{X}_2$):

```
X2 = np.array([np.repeat([1,1],n),np.repeat([0,1],n)]).T
```

c) Use Theorem 9.29 and one of the above parametrizations to test if it can be assumed that the mean value is each group is the same (this should be done in Python). Does it matter which paramtrization you use for the test?

Exercise 1.7 Lecture 9

This exercise is about modelling of CO_2 concentrations in a room (and adapted from the 2023 June exam), the data for the exercise should be downloaded from the website, and can be read into Python by

```
dat = pd.read_csv("co2_data.csv", sep=';')
```

the data contain hourly measurement of CO_2 concentrations for a full week. The columns in the dataset are

- `Hour`: hours since the start of measurements
- `Day`: Days since the start of the measurements
- `CO2C`: The measured CO_2 concentration [ppm]

Now consider the following model

$$Y_i = \beta_0 + \sum_{j=1}^{q} \sin\left(2j\pi\frac{h_i}{24}\right)\beta_{1,q} + \cos\left(2j\pi\frac{h_i}{24}\right)\beta_{2,q} + \epsilon_i; \quad \epsilon_i \sim N(0,\sigma^2) \quad \text{and iid,} \tag{1-1}$$

where h_i is the hours since the begining of the data, and q is an order to be determined (i.e. how may basis functions should be included),

a) Plot the CO_2-concentration as a function of time (i.e. `Hour`), and comment on why the proposed model seems reasonable.

b) Starting from $q = 3$, and using significance level $\alpha = 0.05$, test how many sine/cosine pairs should be included in the model using Type I partioning.

c) If the parametrization is orthogonal, then the type I and type III test are equivalent (see Theorem 9.44). In the case studied here, is the Type I and Type III test equivalent?

d) Make a leverage plot (similar to Example 9.42) for the chosen model, and comment on the effect it will have on standardized and studentized residuals.
(Hint: you may extract diagonal elements of a matrix (`A`) by `A.diagonal(0)`).

Exercise 1.8 Lecture 10

This exercise use the data from Example 8.26, and you can read the data into Python by

```
dat = pd.read_csv("cars.csv", sep=';')
dat
```

```
    car  tire  fuel
1     1     1  22.5
2     1     2  21.5
3     1     3  22.2
4     2     1  24.3
5     2     2  21.3
6     2     3  21.9
7     3     1  24.9
8     3     2  23.9
9     3     3  21.7
10    4     1  22.4
11    4     2  18.4
12    4     3  17.9
```

In this exercise we only consider `car` as the explanatory variable.

a) Explicitly write down the design matrix when the interpretation of the parameters is the mean within each group. Also set up the design matrix when the interpretation is as in Chapter 8 (i.e. $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, and $\sum_i n_i \alpha_i = 0$).

b) Now assume that car 1 is considered a reference, and we hence want a parametrization where the first parameter is the average for car 1, and the other paramters is the difference between the consumption for car 1 and car i, $i \in \{2, 3, 4\}$. Again set up the design matrix.

c) Using Python find the parameter estimates for the last model and test the three hypotheses ($\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$), do this both with and without the Bonferroni correction.
Compare with the result of using

```
dat["car"] = pd.Categorical(dat["car"])
fit = smf.ols(formula = "fuel ~ car", data = dat).fit()
fit.summary(slim=True)
```

d) Make a residual analysis of the model.