Chapter 9

The general linear model

Contents

9.1 Matrix formulation of summary statistics

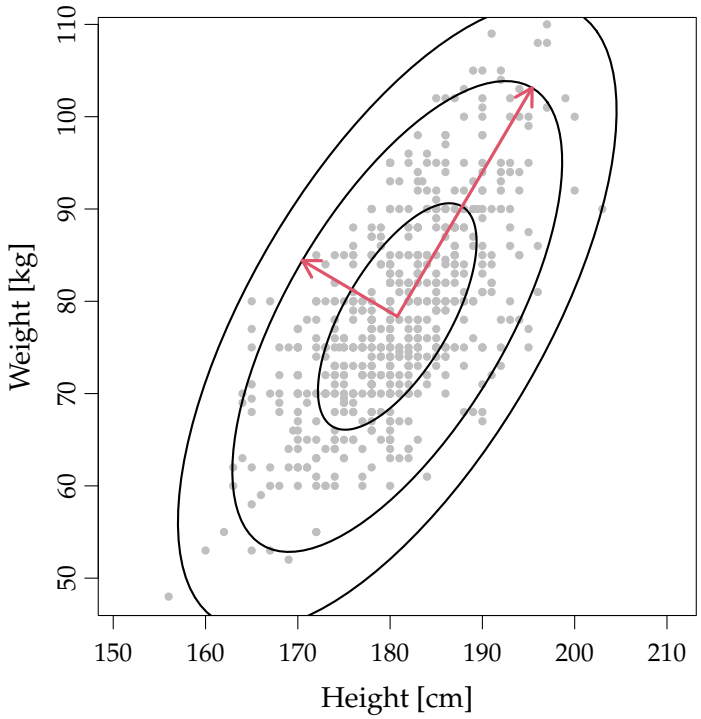
In this chapter we will focus on second order moment representations, i.e. average/mean, variance/sample variance, and covariances/sample covariances. The choice of second order moment representation is closely related to the multivariate normal (Gaussian) distribution, which is characterized by the second order moment representation. We start by a small example.

|||| Example 9.1 Height and weight

The scatter-plot below show height and weight (gray dots) of around 600 males in the age 25-50 years. From the plot it is clear that there is some correlation between the two variables, and hence that a good description of data include the correlation between the two.

The contour lines are related to a multivariate normal distribution, that is estimated to describe the data as good a possible, and define prediction regions. The red arrows are eigen-vectors of the variance-covariance matrix.

In this case observations are two dimensional and one observation consist of the observed height and the observed weight.



Assume that we have associated observations of different variables (e.g. height

and weight of a number of persons). In this section we will be interested in average, observed variance, observed covariance and observed correlation. The k-dimensional observations will be denoted by

$$\boldsymbol{y}_i = \begin{bmatrix} y_{1,i} \\ \vdots \\ y_{k,i} \end{bmatrix}, \tag{9-1}$$

if the are N observation then the average vector is given by

$$\bar{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{bmatrix} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_i, \tag{9-2}$$

recall that the observed covariance between two vector of observations, $\boldsymbol{y}_{l,\cdot}$ and $\boldsymbol{y}_{m,\cdot}$, is given by

$$s_{lm} = \frac{1}{N-1} \sum_{i=1}^{N} (y_{li} - \bar{y}_l)(y_{mi} - \bar{y}_m), \tag{9-3}$$

which can be collected in an observed variance-covariance matrix by

$$\boldsymbol{S} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T. \tag{9-4}$$

The matrix $\boldsymbol{S}$ is often, in particular when reported, decomposed into standard deviation and correlations

$$\boldsymbol{S} = \hat{\boldsymbol{\sigma}} \boldsymbol{R} \hat{\boldsymbol{\sigma}}, \tag{9-5}$$

where $\hat{\boldsymbol{\sigma}}$ is a diagonal matrix with the observed standard deviation in the diagonal (i.e $\hat{\sigma}_{ii} = \sqrt{S_{ii}}$ and $\hat{\sigma}_{ij} = 0$ for $i \neq j$), and $\boldsymbol{R}$ is the collection of all pairwise correlations. As a direct consequence we can write the correlation matrix as

$$\boldsymbol{R} = \hat{\boldsymbol{\sigma}}^{-1} \boldsymbol{S} \hat{\boldsymbol{\sigma}}^{-1}. \tag{9-6}$$

The main advantage of (9-6) is that the correlation coefficients are easy to interpret, while covariances are not.

|||| Example 9.2 Height and weight cont.

For the data presented in Example 9.1 the second order moment representation can be calculated as

```
S = dat.cov()
mu = dat.mean()
print(mu)


height    180.774671
weight     78.351891
dtype: float64


print(S)


           height     weight
height  53.304992   50.00331
weight  50.003310  108.62449
```

hence average height is about 180 cm and the average weight is about 78 kg. Further the variances and covariances is also calculated and the shape of the ellipsoids in Example 9.1 is described by those. As noted in the text it is usual practice to report standard deviations and correlation, as presented below, rather than the variance-covariance matrix.

```
R = dat.corr()
sig_hat = np.sqrt(dat.var())
print(sig_hat)


height     7.301027
weight    10.422307
dtype: float64


print(R)


          height    weight
height  1.000000  0.657129
weight  0.657129  1.000000
```

hence the standard deviations are 7 cm and 10 kg, respectively, and the correlation is about 0.66.

9.2 Preliminaries from linear algebra

This chapter rely on a many results from linear algebra, and we state a some results that are important for the further development. Some of these are stated without proof.

Lemma 9.3 Eigenvalue decomposition of symmmetric matrices

For a quadratic matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the iegenvalue decomposition can be written as

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}, \tag{9-7}$$

where $\boldsymbol{V}$ is the eigen-vectors and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues along the diagonal
If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix then the eigenvalue decomposition can be written as

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{T}, \tag{9-8}$$

i.e. $\boldsymbol{V}^{-1} = \boldsymbol{V}^{T}$. Further the rank of $\boldsymbol{A}$ is equal to the number a non-zero eigenvalues.

Example 9.4

In Example 9.1 we plotted the observed data along with some ellipsoids (we will get back to those). In the same plot there are two red arrows, these represent the eigen-vectors with length proportional to the eigen values of the variance-covariance matrix of the observed data. In Python the eigenvalues and eigen vector can be calculated by

```
Eigen = eig(S)
Eigvals, Eigvectors = eig(S)
print(Eigvals)


[ 23.821 138.108]


print(Eigvectors)


[[-0.861 -0.508]
 [ 0.508 -0.861]]
```

Hence the arrows both start in the observed average, $\bar{\boldsymbol{y}}$, and extend to

$$\bar{\boldsymbol{y}} + k\sqrt{23.8} \cdot \begin{bmatrix} -0.86 \\ 0.51 \end{bmatrix}; \quad \text{and} \quad \bar{\boldsymbol{y}} + k\sqrt{138.1} \cdot \begin{bmatrix} 0.51 \\ 0.86 \end{bmatrix}, \tag{9-9}$$

as stated above we will get back to the exact choice of k, but it is related to a prediction interval/region for the observations.

We state the following permutation result for permutation in traces

|||| Lemma 9.5 Permutation in traces

For matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ such that that the products $\boldsymbol{ABC}$, $\boldsymbol{BCA}$ and $\boldsymbol{CAB}$ can be formed then

$$Trace(\boldsymbol{ABC}) = Trace(\boldsymbol{BCA}) = Trace(\boldsymbol{CAB}). \tag{9-10}$$

We will sometimes need to update the matrix inverses, and the following lemma and corollary is useful for that.

Lemma 9.6 Rank-1 update of matrix inverse

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a matrix such that $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is well defined (i.e. $\boldsymbol{X}^T\boldsymbol{X}$ have full rank) and further let $\tilde{\boldsymbol{X}} = \begin{bmatrix}\boldsymbol{X} & \boldsymbol{v}\end{bmatrix}$, with $\boldsymbol{v} \in \mathbb{R}^n$ a vector, then

$$(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1} = \begin{bmatrix} \boldsymbol{A} + \frac{\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} & \frac{-\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} \\ \frac{-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} & \frac{1}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} \end{bmatrix} \tag{9-11}$$

We will use matrices of the form $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ (which, as we will show, is an othogonal projection matrix) often, and the following corollary to Lemma 9.5 apply

Corollary 9.7 Rank-1 update of projection matrix

Let $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ be as in Lemma 9.6, define $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$, then

$$\begin{aligned}\tilde{\boldsymbol{H}} =& \boldsymbol{H} + \frac{1}{k}\left(\boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H} - \boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H} - \boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T + \boldsymbol{v}\boldsymbol{v}^T\right) \\ =& \boldsymbol{H} + \frac{1}{k}\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{v}\boldsymbol{v}^T\left(\boldsymbol{I}-\boldsymbol{H}\right)\end{aligned} \tag{9-12}$$

with $k = \boldsymbol{v}^T\boldsymbol{v} - \boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v} = \boldsymbol{v}^T\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{v}$.

Proof

From Lemma 9.6, we have

$$(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1} = \begin{bmatrix} \boldsymbol{A} + \frac{\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} & \frac{-\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} \\ \frac{-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} & \frac{1}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}} \end{bmatrix} \tag{9-13}$$

and hence

$$\begin{aligned}\tilde{\boldsymbol{H}} &= \begin{bmatrix}\boldsymbol{X} & \boldsymbol{v}\end{bmatrix}\begin{bmatrix} \boldsymbol{A} + \frac{\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v}} & \frac{-\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v}} \\ \frac{-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v}} & \frac{1}{\boldsymbol{v}^T\boldsymbol{v}-\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v}} \end{bmatrix}\begin{bmatrix}\boldsymbol{X}^T \\ \boldsymbol{v}^T\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{X}\boldsymbol{A} + \frac{\boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}} - \frac{\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}} & \frac{-\boldsymbol{H}\boldsymbol{v}}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}} + \frac{\boldsymbol{v}}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}^T \\ \boldsymbol{v}^T\end{bmatrix} \\ &= \boldsymbol{H} + \frac{\boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H}}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}} - \frac{\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H}}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}} - \frac{\boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}} + \frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{v}}\end{aligned} \tag{9-14}$$

which is the stated result.

9.3 Multivariate distributions

We will focus the multivariate normal distribution, but start by some general definitions and results, related to multivariate distributions.

|||| Definition 9.8 Multivariate probability density functions

A multivariate probability density function for the random variable $\boldsymbol{Y} \in \mathbb{R}^n$, is a function from $\mathbb{R}^n$ into $\mathbb{R}_0$,

$$f(\boldsymbol{y}) = f(y_1, y_2, ..., y_n) \geq 0, \tag{9-15}$$

such that

$$\int f(\boldsymbol{y}) d\boldsymbol{y} = \int \int \cdots \int f(y_1, y_2, ..., y_n) dy_1 dy_2 \cdots dy_n = 1, \tag{9-16}$$

further the marginal distribution for Y_i is given by

$$f_{Y_i}(y_i) = \int \int \cdots \int f(y_1, y_2, ..., y_n) dy_1 \cdots dy_{i-1} dy_{i+1} \cdots dy_n. \tag{9-17}$$

If a random variable $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T]^T$ have the joint density $f_{\boldsymbol{Y}}(\boldsymbol{y})$, then the marginal density of $\boldsymbol{Y}_1$ is

$$f_{\boldsymbol{Y}_1}(\boldsymbol{y}_1) = \int f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y}_2. \tag{9-18}$$

The density function is the fundamental property of a random variable that describe everything about the random variable, here we are mostly interested in the second order moment representation (mean, variance and covariance).

Definition 9.9 Second order moment representation

If a random vector $\boldsymbol{Y} \in \mathbb{R}^n$ have the probability density function $f_{\boldsymbol{Y}}$ then the mean and variance of Y_i is

$$\begin{aligned} E[Y_i] =& \mu_i = \int y_i f_{Y_i}(y_i) dy_i \\ V[Y_i] =& \sigma_{ii} = \int (y_i - \mu_i)^2 f_{Y_i}(y_i) dy_i, \end{aligned} \tag{9-19}$$

and the covariances between Y_i and Y_j is

$$Cov[Y_i, Y_j] = \sigma_{ij} = \int (y_i - \mu_i)(y_j - \mu_j) f_{Y_i, Y_j}(y_i, y_j) dy_i dy_j. \tag{9-20}$$

Further the mean value vector of a random vector $\boldsymbol{Y} = [Y_1, ..., Y_n]^T$ is defined by

$$\boldsymbol{\mu} = E[\boldsymbol{Y}] = \begin{bmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{bmatrix}, \tag{9-21}$$

and the variance-covariance matrix is

$$\boldsymbol{\Sigma} = V[\boldsymbol{Y}], \tag{9-22}$$

where the elements of $\boldsymbol{\Sigma}$ are $\boldsymbol{\Sigma}_{ij} = Cov[Y_i, Y_j]$. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is referred to as the second order moment representation.

The covariance matrix between two random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ (not necessarily of the same dimension) is

$$\boldsymbol{\Sigma}^{12} = Cov[\boldsymbol{Y}_1, \boldsymbol{Y}_2], \tag{9-23}$$

meaning the $\boldsymbol{\Sigma}^{12}_{ij} = Cov[Y_{1,i}, Y_{2,j}]$. Now we can write the variance-covariance matrix of the random vector $[\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T]^T$ as

$$V \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{bmatrix}, \tag{9-24}$$

where (of course) $\boldsymbol{\Sigma}^{12} = \left(\boldsymbol{\Sigma}^{21}\right)^T$. We are now ready for the calculation rules for random vectors.

Theorem 9.10 Covariance calculation rules

Let the variance-covariance matrix of $[\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T]^T$ be as in (9-24) and let $\boldsymbol{b}$ be a vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices of appropriate dimensions, then

$$E[\boldsymbol{A}\boldsymbol{Y}_1 + \boldsymbol{b}] = \boldsymbol{A}E[\boldsymbol{Y}_1] + \boldsymbol{b} \tag{9-25}$$

$$Cov[\boldsymbol{A}\boldsymbol{Y}_1, \boldsymbol{B}\boldsymbol{Y}_2] = \boldsymbol{A}Cov[\boldsymbol{Y}_1, \boldsymbol{Y}_2]\boldsymbol{B}^T = \boldsymbol{A}\boldsymbol{\Sigma}^{12}\boldsymbol{B}^T \tag{9-26}$$

and as a special case

$$V[\boldsymbol{A}\boldsymbol{Y}_1] = \boldsymbol{A}\boldsymbol{\Sigma}^{11}\boldsymbol{A}^T. \tag{9-27}$$

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be such that $\boldsymbol{A}\boldsymbol{Y}_1 + \boldsymbol{B}\boldsymbol{Y}_2$ can be formed, then

$$V[\boldsymbol{A}\boldsymbol{Y}_1 + \boldsymbol{B}\boldsymbol{Y}_2] = \boldsymbol{A}\boldsymbol{\Sigma}^{11}\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{\Sigma}^{22}\boldsymbol{B}^T + \boldsymbol{A}\boldsymbol{\Sigma}^{12}\boldsymbol{B}^T + \boldsymbol{B}\boldsymbol{\Sigma}^{21}\boldsymbol{A}^T. \tag{9-28}$$

In addition to the second order moment representation, independence is a very important concept in statistics, the formal definition is

Definition 9.11 Independence of random vectors

Let $f_{\boldsymbol{Y}}$ be the joint distribution of the random vector $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T]^T$, then $\boldsymbol{Y}_1$, and $\boldsymbol{Y}_2$ are independent if

$$f_Y(\boldsymbol{y}) = f_{Y_1}(\boldsymbol{y}_1) f_{Y_2}(\boldsymbol{y}_2). \tag{9-29}$$

The definition imply that if $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent then $Cov[\boldsymbol{Y}_1, \boldsymbol{Y}_2] = \boldsymbol{0}$. In general the opposite is not true (i.e. no correlation does not imply independence).

Section 9.3.1 below consider the matrix formulation of error propagation. It is not used in the further development but included for completeness of matrix formulations.

9.3.1 Error propagation

We consider a random vector $\boldsymbol{Y} \in \mathbb{R}^n$ with

$$\begin{aligned} E[\boldsymbol{Y}] &= \boldsymbol{\mu} \\ V[\boldsymbol{Y}] &= \boldsymbol{\Sigma}, \end{aligned} \tag{9-30}$$

now consider an (possibly nonlinear) function $\boldsymbol{f}(\boldsymbol{Y}) \in \mathbb{R}^m$, the function $\boldsymbol{f}$ can be approximated around any point $\boldsymbol{y}_0$ by the Taylor approximation

$$\boldsymbol{f}(\boldsymbol{Y}) = \boldsymbol{f}(\boldsymbol{y}_0) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{y}_0)(\boldsymbol{Y} - \boldsymbol{y}_0) + "HOT", \tag{9-31}$$

where $"HOT"$ is short for Higher Order Terms. Now if we choose $\boldsymbol{y}_0 = \boldsymbol{\mu}$, we can write

$$\boldsymbol{f}(\boldsymbol{Y}) \approx \boldsymbol{f}(\boldsymbol{\mu}) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu}), \tag{9-32}$$

notice here that $\boldsymbol{\mu}$ and $\boldsymbol{f}(\boldsymbol{\mu})$ are non random vectors, and the Jacobian, $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{\mu})$, is a non-random matrix, and therefore we can directly write

$$\begin{aligned} E[\boldsymbol{f}(\boldsymbol{Y})] \approx& \boldsymbol{f}(\boldsymbol{\mu}) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{\mu}) E[(\boldsymbol{Y} - \boldsymbol{\mu})] \\ =& \boldsymbol{f}(\boldsymbol{\mu}), \end{aligned} \tag{9-33}$$

and

$$\begin{aligned} V[\boldsymbol{f}(\boldsymbol{Y})] \approx& \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{\mu}) V[(\boldsymbol{Y} - \boldsymbol{\mu}] \boldsymbol{J}_{\boldsymbol{f}}^T(\boldsymbol{\mu}) \\ =& \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{\mu}) \boldsymbol{\Sigma} \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{\mu})^T. \end{aligned} \tag{9-34}$$

|||| Example 9.12 Body Mass Index

Body mass index (BMI) is often used as an indicator of the health of a person, BMI is defined as

$$BMI = \frac{w}{h_m^2}, \tag{9-35}$$

where w is the weight [kg] and h_m [m] is the height, in our case we measure height in cm and therefore we get

$$BMI = \frac{w}{h_{cm}^2} 10^4, \tag{9-36}$$

and the Jacobian is

$$\boldsymbol{J}_{BMI}(h, v) = \begin{bmatrix} -2\frac{v}{h^3} & \frac{1}{h^2} \end{bmatrix}^T \cdot 10^4, \tag{9-37}$$

based on the data from Example 9.1 we can approximate the variance of BMI (for the considered population) by

```
mu
```

```
height    180.774671
weight     78.351891
dtype: float64
```

```
h = mu["height"]
w = mu["weight"]
J = np.array([-2 * w / h**3 * 10000, 1 / h**2 * 10000])
J @ dat.cov() @ J.T
```

```
np.float64(5.804464687904143)
```

hence the variance is approximated by 5.8 kg^2/m^4 or a standard deviation of 2.4 kg/m^2.

9.3.2 The multivariate Gaussian distribution

In this section we cover some important results for the multivariate normal distribution and the relation to the χ^2-distribution. These are important for the development of statistical tests related to the general linear model (LM[1]), which is the main topic of the chapter.

A common definition of the multivariate normal distribution is that the pdf of the random variable $\boldsymbol{Y} \in \mathbb{R}^n$ is

$$f_Y(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}, \tag{9-38}$$

and the parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) are the second order moment representation, i.e.

$$\begin{aligned} E[\boldsymbol{Y}] &= \boldsymbol{\mu} \\ V[\boldsymbol{Y}] &= \boldsymbol{\Sigma}, \end{aligned} \tag{9-39}$$

and we write

$$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{9-40}$$

[1] We use the abbreviation LM rather than GLM as GLM is usually used for the more general generalized linear model.

We will sometimes omit the subscript n if it is clear from context (or if it is not important).

Example 9.13

The ellipsoids in the figure in Example 9.1 are level curves in a 2-dimensional normal with mean value equal the observed average and variance-covariance equal the observed variance-covariance matrix (see Example 9.2).

Theorem 9.14 Independence of normal random variables

If $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T]^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and

$$Cov[\boldsymbol{Y}_1, \boldsymbol{Y}_2] = \boldsymbol{0}, \tag{9-41}$$

then $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent.

Proof

See Exercise 1.

■

Note that the assumption of the joint distribution is important in Theorem 9.14, i.e. it is not enough that the marginal distribution of the random variables in the vector are normal. The next example illustrate the point.

Example 9.15

Let $Y_1 \sim N(0,1)$ and let $P(X = -1) = P(X = 1) = \frac{1}{2}$ independent of Y_1, and define $Y_2 = XY_1$, then the marginal distribution of Y_2 is the standard normal and

$$\begin{aligned} Cov[Y_1, Y_2] =& Cov[Y_1, XY_1] = E[Y_1 XY_1] = E[X]E[Y_1^2] \\ =& E[X]V[Y_1] = E[X] = 0, \end{aligned} \tag{9-42}$$

hence no correlation, but clearly the variables are not independent, as knowledge of Y_1 limit the number of possible outcomes of Y_2 to two possible values (Y_1 or $-Y_1$). For a graphical simulation based analysis see Exercise 2.

Theorem 9.16 Normalization of normal random vectors

If $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the pdf of $\boldsymbol{Y}$ as defined in (9-38) (implying that $\boldsymbol{\Sigma}$ is positive definite), then

$$\boldsymbol{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim N(\boldsymbol{0}, \boldsymbol{I}) \tag{9-43}$$

with $\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}$ (implying that $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{T}{2}} = \boldsymbol{\Sigma}$), where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}$ in the diagonal and $\boldsymbol{V}$ is the corresponding eigenvectors.

Proof

$\boldsymbol{\Sigma}$ is a real symmetric matrix and hence it can be written as (see Lemma 9.3)

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T, \tag{9-44}$$

and $\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^{-1}$, also $\boldsymbol{V}$ is an orthogonal basis (hence $\boldsymbol{V}^{-1} = \boldsymbol{V}^T$), and hence

$$\begin{aligned} V[\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y}] =& \boldsymbol{\Sigma}^{-\frac{1}{2}} V[\boldsymbol{Y}] \boldsymbol{\Sigma}^{-\frac{T}{2}} \\ =& \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\boldsymbol{V}^{-T}\boldsymbol{\Lambda}^{-\frac{1}{2}} \\ =& \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}} \\ =& \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-\frac{1}{2}} \\ =& \boldsymbol{I}, \end{aligned} \tag{9-45}$$

and since $E[\boldsymbol{Y}] = \boldsymbol{\mu}$ the proof is completed.

■

The definition (9-38) clearly require $\boldsymbol{\Sigma}$ to be inevitable, and a more general definition, which we will need in the following, is

Definition 9.17 Multivariate normal distribution

Let Z_i, $i = 1, ..., n$, be iid. standard normal random variables, s.t. $(\boldsymbol{Z} = [Z_1, ..., Z_n]^T)$

$$\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}). \tag{9-46}$$

Then the random vector $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b}$, with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, follow an m-dimensional multivariate normal distribution with

$$\begin{aligned} E[\boldsymbol{Y}] &= \boldsymbol{b} \\ V[\boldsymbol{Y}] &= \boldsymbol{A}\boldsymbol{A}^T, \end{aligned} \tag{9-47}$$

this holds also when $\boldsymbol{A}\boldsymbol{A}^T$ is not positive definite.

The definition imply that any linear combination of a multivariate normal random vector is also a multivariate normal random vector and further if the covariance between two elements of a multivariate normal vector is zero the they are independent.

As an example suppose we have n iid. standard normal random variables (Z_i) and form the average of those $(\bar{Z})$ and consider the difference between the averages and the individual random variables (we denote these as residuals, $\boldsymbol{r}$)

$$\boldsymbol{r} = \begin{bmatrix} Z_1 - \bar{Z} \\ \vdots \\ Z_n - \bar{Z} \end{bmatrix} = \boldsymbol{A}\boldsymbol{Z}; \quad \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}), \tag{9-48}$$

with

$$\boldsymbol{A} = \begin{bmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & 1 - \frac{1}{n} \end{bmatrix} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{E}. \tag{9-49}$$

The matrix $\boldsymbol{A}$ is in $\mathbb{R}^{n \times n}$, but any column (or row) can be written as the (negative) sum of the remaining columns and therefore the rank of $\boldsymbol{A}$ is equal $n - 1$ (not n, see Exercise 3). Further in this special case, we have

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^2 = \boldsymbol{A}. \tag{9-50}$$

For a proof of the claims in (9-50) see Exercise 3. We will come back to the particular properties (9-50) of the matrix (9-49), but for now we can simply write

$$\boldsymbol{r} \sim N(\boldsymbol{0}, \boldsymbol{A}), \tag{9-51}$$

this imply that the pdf of $\boldsymbol{r}$ cannot be written explicitly (the inverse of $\boldsymbol{A}$ does not exist), and further that the covariance between r_i and r_j is not 0 (implying that they are not independent).

We can also show that $\boldsymbol{r}$ and $\bar{Z}$ are independent, to that end consider

$$\begin{bmatrix} \boldsymbol{r} \\ \bar{Z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \\ \frac{1}{n}\mathbf{1}^T \end{bmatrix} \boldsymbol{Z}, \tag{9-52}$$

hence the vector $[\boldsymbol{r}^T, \bar{Z}]^T$ follow a multivariate normal distribution and if the covariance between the two is zero then $\boldsymbol{r}$ and $\bar{Z}$ are independent,

$$\begin{aligned} Cov[\boldsymbol{r}, \bar{Z}] =& Cov\left[\boldsymbol{A}\boldsymbol{Z}, \frac{1}{n}\mathbf{1}^T\boldsymbol{Z}\right] \\ =& \frac{1}{n}\boldsymbol{A}Cov\left[\boldsymbol{Z}, \boldsymbol{Z}\right]\mathbf{1} \end{aligned} \tag{9-53}$$

since $V[\boldsymbol{Z}] = \boldsymbol{I}$ it reduce to

$$Cov[\boldsymbol{r}, \bar{Z}] = \frac{1}{n}\boldsymbol{A}\mathbf{1}, \tag{9-54}$$

and since the row-sums of $\boldsymbol{A}$ is zero (see Exercise 3) we get

$$Cov[\boldsymbol{r}, \bar{Z}] = \mathbf{0}, \tag{9-55}$$

hence $\boldsymbol{r}$ and $\bar{Z}$ are independent. For the development of statistical test we need to derive the relation between the multivariate normal and the χ^2-distribution, this is the subject of the next section.

9.4 The multivariate normal and the χ^2-distribution

From the definition of the χ^2-distribution (see Theorem 2.78) we know that, if $\boldsymbol{Z} \sim N_n(\mathbf{0}, \boldsymbol{I})$ then

$$\boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2(n). \tag{9-56}$$

A simple consequence of Theorem 9.16 is

Corollary 9.18

With $\boldsymbol{Y} \in \mathbb{R}^n$ as in Theorem 9.16 then

$$(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2(n). \tag{9-57}$$

Proof

See Exercise 4.

■

Corollary 9.18 imply that if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$P((\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) \leq \chi^2_{1-\alpha}) = 1-\alpha, \tag{9-58}$$

and hence level curves of the pdf describe probability regions that can be determined from the χ^2-distribution.

Example 9.19

In Example 9.1 we saw level curves of the Gaussian pdf, these are described by curves where

$$(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) = \chi^2_{1-\alpha} \tag{9-59}$$

with $\boldsymbol{\mu}$ equal the observed average of height and weight, and $\boldsymbol{\Sigma}$ equal the observed variance-covariance (see Example 9.4). Also the values of α is set at 0.5, 0.05 and 0.005 respectively for the three curves. Hence the length of the red arrow in the plot of Example 9.1 is

$$\chi^2_{0.95} \cdot 23.8; \quad \text{and} \quad \chi^2_{0.95} \cdot 138.1 \tag{9-60}$$

with $\chi^2_{0.95}$ a quantile of the χ^2-distribution with 2 degrees of freedom, i.e. (referring to Example 9.4) $k = \sqrt{\chi^2_{0.95}}$.

Using the from given in (9-48), we can write the quadratic form as (using $\boldsymbol{r} = \boldsymbol{Z} - \mathbf{1}\bar{Z}$)

$$\begin{aligned}\boldsymbol{Z}^T\boldsymbol{Z} =& (\boldsymbol{r}+\mathbf{1}\bar{Z})^T(\boldsymbol{r}+\mathbf{1}\bar{Z}) \\ =& \boldsymbol{r}^T\boldsymbol{r} + (\boldsymbol{Z}-\mathbf{1}\bar{Z})^T\mathbf{1}\bar{Z} + \mathbf{1}^T\bar{Z}(\boldsymbol{Z}-\mathbf{1}\bar{Z}) + \bar{Z}\mathbf{1}^T\mathbf{1}\bar{Z} \\ =& \boldsymbol{r}^T\boldsymbol{r} + (n\bar{Z}-n\bar{Z})\bar{Z} + \bar{Z}(n\bar{Z}-n\bar{Z}) + n\bar{Z}^2 \\ =& \boldsymbol{r}^T\boldsymbol{r} + n\bar{Z}^2\end{aligned} \tag{9-61}$$

since $n\bar{Z}^2 \sim \chi^2(1)$, and $\bar{Z}$ and $\boldsymbol{r}$ are independent then we must have

$$\boldsymbol{r}^T\boldsymbol{r} \sim \chi^2(n-1). \tag{9-62}$$

Example 9.20

Assume that $\boldsymbol{Y} \sim N_n(\mathbf{1}\mu, \sigma^2\boldsymbol{I})$, this is equivalent to

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I}), \tag{9-63}$$

and the "residuals" can be written as

$$\begin{aligned} \boldsymbol{r} =& \boldsymbol{Y} - \mathbf{1}\bar{Y} \\ =& \mathbf{1}\mu + \boldsymbol{\epsilon} - \mathbf{1}\mu - \mathbf{1}\bar{\epsilon} \\ =& \boldsymbol{\epsilon} - \mathbf{1}\bar{\epsilon}, \end{aligned} \tag{9-64}$$

hence not depending on μ, and in light of the discussion above we also have that $\frac{1}{\sigma^2}\boldsymbol{r}^T\boldsymbol{r} \sim \chi^2(n-1)$, and further if $\mu = 0$ (the null-hypothesis) then $\bar{Y}^2/(\sigma^2/n) \sim \chi^2(1)$, and hence if $\mu = 0$ then

$$F_{obs} = \frac{\frac{\bar{Y}^2}{\sigma^2/n}\frac{1}{n}}{\frac{1}{\sigma^2}\boldsymbol{r}^T\boldsymbol{r}/(n-1)} = \frac{n\bar{Y}^2}{\boldsymbol{r}^T\boldsymbol{r}/(n-1)} \sim F(1, n-1), \tag{9-65}$$

F_{obs} is a test statistic and conclusions about μ can be based on critical values or p-values.

The derivations above is a special case of Cochran's theorem, which we will state below, but first we need the concept of orthogonal projection matrices, as stated in the next definition.

Definition 9.21 Orthogonal projections

A matrix $\boldsymbol{P}$ is an orthogonal projection matrix if and only if

- $\boldsymbol{P}$ is symmetric, i.e. $\boldsymbol{P} = \boldsymbol{P}^T$
- $\boldsymbol{P}$ is idempotent, i.e. $\boldsymbol{P}^2 = \boldsymbol{P}$.

If $\boldsymbol{P}$ is a projection matrix then so is $\boldsymbol{I} - \boldsymbol{P}$, this is easily shown by

$$\begin{aligned} (\boldsymbol{I} - \boldsymbol{P})^T =& \boldsymbol{I}^T - \boldsymbol{P}^T = \boldsymbol{I} - \boldsymbol{P} \\ (\boldsymbol{I} - \boldsymbol{P})^2 =& \boldsymbol{I} + \boldsymbol{P}^2 - \boldsymbol{P} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{P}. \end{aligned} \tag{9-66}$$

Using the results above it is easy to show that the matrix $\boldsymbol{A}$ in (9-49) is an orthogonal projection matrix (see Exercise 5).

Lemma 9.22 Properties of orthogonal projection matrices

If $\boldsymbol{P}$ is an orthogonal projection matrix, then

1. The eigenvalues λ_i of $\boldsymbol{P}$ are either 0 or 1, and $Rank(\boldsymbol{P}) = \sum_i \lambda_i$.
2. $Rank(\boldsymbol{P}) = Trace(\boldsymbol{P})$.

Proof

Let $\boldsymbol{\Lambda}$ and $\boldsymbol{V}$ a diagonal matrix with the eigen-values along the diagonal, and the collection of eigen-vectors. Then 1) $\boldsymbol{P}^2 = \boldsymbol{P}$ and hence $\boldsymbol{V\Lambda V}^T = \boldsymbol{V\Lambda V}^T\boldsymbol{V\Lambda V}^T = \boldsymbol{V\Lambda}^2\boldsymbol{V}^T$ or $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^2$ implying that $\lambda_i = \lambda_i^2$ which can only happen if $\lambda_i = 0$ or $\lambda_i = 1$ and hence the number of non-zero eigenvalues (which is the rank) is $\sum_i \lambda_i$. 2) see Exercise 6.

■

We again turn to the simple example (9-48). We have already seen that $\boldsymbol{A}$ is a projection matrix and that $Rank(\boldsymbol{A})$ is $n-1$, using the results in Lemma 9.22 we also get

$$Trace(\boldsymbol{A}) = \sum_{i=1}^{n}\left(1 - \frac{1}{n}\right) = n - 1. \tag{9-67}$$

The main result for construction test statistics is Cochran's theorem as given below.

Theorem 9.23 Cochran's theorem

Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and let $\boldsymbol{H}_i$ be orthogonal projection matrices such that

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} = \frac{1}{\sigma^2}\sum_{i=1}^{K}\boldsymbol{Y}^T\boldsymbol{H}_i\boldsymbol{Y} \tag{9-68}$$

i.e. $\sum_{i=1}^{K}\boldsymbol{H}_i = \boldsymbol{I}_n$, with $Rank(\boldsymbol{H}_i) = p_i$, and $\sum_i p_i = n$ then

1. $\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_i\boldsymbol{Y} \sim \chi^2(p_i)$
2. $\boldsymbol{Y}^T\boldsymbol{H}_i\boldsymbol{Y}$ and $\boldsymbol{Y}^T\boldsymbol{H}_j\boldsymbol{Y}$ are independent for $i \neq j$.

As we will see in later sections Cochran's theorem is useful for constructing test statistics and determine their distributions. We prove the theorem in Section 9.4.1 below.

The independence condition in Theorem 9.23 is equivalent to

$$Cov[\boldsymbol{Y}^T\boldsymbol{H}_i, \boldsymbol{H}_j\boldsymbol{Y}] = \boldsymbol{0}. \tag{9-69}$$

In the simple example in eps. (9-48) we have

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{Z} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{Z} = \boldsymbol{H}_1\boldsymbol{Z} + \boldsymbol{H}_2\boldsymbol{Z} \tag{9-70}$$

and it is easy to show that $Cov[\boldsymbol{H}_1\boldsymbol{Z}, \boldsymbol{H}_2\boldsymbol{Z}] = \boldsymbol{0}$ (see Exercise 9).

We can also use Cochran's theorem to find the distribution of $\boldsymbol{r}^T\boldsymbol{r}$, the following is obviously true

$$\begin{aligned} \boldsymbol{Z} =& \boldsymbol{A}\boldsymbol{Z} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{Z} \\ =& \boldsymbol{r} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{Z}, \end{aligned} \tag{9-71}$$

now $Rank(\boldsymbol{I} - \boldsymbol{A}) = n - 1$, and hence by Cochran's theorem

$$\begin{aligned} \boldsymbol{r}^T\boldsymbol{r} =& \boldsymbol{Z}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{Z} \\ =& \boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2(n-1). \end{aligned} \tag{9-72}$$

This conclude the fundamental tools we need for the development of test statistics in the general linear model. The next section present the proof of Cochran's Theorem.

9.4.1 Proof of Cochran's Theorem*

Note that $\frac{1}{\sigma}\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and hence Y_i and Y_j are independent for all $i \neq j$. Therefore

$$\frac{1}{\sigma^2}\boldsymbol{Y}\boldsymbol{Y}^T = \frac{1}{\sigma^2}\sum_{i=1}^{n} Y_i^2 \sim \chi^2(n), \tag{9-73}$$

and further for any sub-sum

$$\frac{1}{\sigma^2}\sum_{i=1}^{p} Y_i^2 \sim \chi^2(p). \tag{9-74}$$

Now consider the case $K = 2$,

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} = \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_2\boldsymbol{Y}, \tag{9-75}$$

and let $\boldsymbol{V}_i$ be the eigen-vectors corresponding to $\boldsymbol{H}_i$, and $\boldsymbol{\Lambda}_i$ diagonal matrices with the corresponding eigenvalues, and consider the linear transformation $\boldsymbol{Z} = \boldsymbol{V}_1\boldsymbol{Y}$, then

$$\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{Y}^T\boldsymbol{V}_1^T\boldsymbol{V}_1\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y}, \tag{9-76}$$

and insert in (9-75)

$$\begin{aligned}\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} =& \frac{1}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{Z} \\ =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{V}_1^T\boldsymbol{H}_1\boldsymbol{V}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{V}_1^T\boldsymbol{H}_2\boldsymbol{V}_1\boldsymbol{Y} \\ =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{\Lambda}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{V}_1^T\boldsymbol{H}_2\boldsymbol{V}_1\boldsymbol{Y},\end{aligned} \tag{9-77}$$

without loss of generality we can assume that the first p_1 diagonal elements of $\boldsymbol{\Lambda}$ is 1 and the remaining are zero and hence

$$\boldsymbol{Y}^T\boldsymbol{\Lambda}_1\boldsymbol{Y} = \sum_{i=1}^{p_1} Y_i^2 \tag{9-78}$$

and therefore

$$\boldsymbol{Y}^T\boldsymbol{V}_1^T\boldsymbol{H}_2\boldsymbol{V}_1\boldsymbol{Y} = \sum_{i=p_1+1}^{n} Y_i^2. \tag{9-79}$$

The two terms are independent since they depend on different Y's, and it follows that

$$\begin{aligned}\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{\Lambda}_1\boldsymbol{Y} \sim& \chi^2(p_1) \\ \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{V}_1^T\boldsymbol{H}_2\boldsymbol{V}_1\boldsymbol{Y} \sim& \chi^2(n-p_1).\end{aligned} \tag{9-80}$$

This conclude the proof of the case $K = 2$. For $K > 2$ we first consider $K = 3$,

$$\begin{aligned}\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_2\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_3\boldsymbol{Y} \\ =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_R\boldsymbol{Y} \\ =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{\Lambda}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{\Lambda}_1)\boldsymbol{Y}\end{aligned} \tag{9-81}$$

with $\boldsymbol{H}_R = \boldsymbol{H}_2 + \boldsymbol{H}_3$, now consider the splitting $\boldsymbol{Y} = [\boldsymbol{Y}_1^T \quad \boldsymbol{Y}_R^T]^T$. Note that $\boldsymbol{Y}_R \sim N_{n-p_1}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{Y}_R^T\boldsymbol{Y}_R = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{Y}$ following the arguments for the case $K = 2$ we have

$$\begin{aligned}\boldsymbol{Y}_R^T\boldsymbol{Y}_R =& \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{Y} \\ =& \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{H}_2(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{H}_3(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{Y} \\ =& \boldsymbol{Y}_R^T\tilde{\boldsymbol{H}}_2\boldsymbol{Y}_R + \boldsymbol{Y}_R^T\tilde{\boldsymbol{H}}_3\boldsymbol{Y}_R,\end{aligned} \tag{9-82}$$

and

$$\begin{aligned}\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{Y} =& \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)\boldsymbol{\Lambda}_2\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1)(\boldsymbol{I} - \boldsymbol{\Lambda}_2)\boldsymbol{Y} \\ =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{\Lambda}_2\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2)\boldsymbol{Y}\end{aligned} \tag{9-83}$$

where the first term on the rhs follow a $\chi^2(p_2)$-distribution and the second term follow a $\chi^2(n - p_1 - p_2)$-distribution, and hence the quadratic form can be written as

$$\begin{aligned}\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_2\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_3\boldsymbol{Y} \\ =& \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{\Lambda}_1\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{\Lambda}_2\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2)\boldsymbol{Y}.\end{aligned} \tag{9-84}$$

Cases where $K > 3$ follow by induction.

9.5 The general linear model

The models covered in Chapter 3, 5, 6, and 8 can all be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}). \tag{9-85}$$

Any model that can be written in the form (9-85) is called a general linear model. $\boldsymbol{Y}$ is the outcome of interest, the known matrix $\boldsymbol{X}$ is called the design matrix, $\boldsymbol{\beta}$ is the mean value parameters that we should estimate based on the design matrix and the outcomes, $\boldsymbol{\epsilon}$ is the residual errors, with variance σ^2, and further we assume that all residuals are iid.

In this section we will cover the general linear model in very general terms, and in later sections we will present different examples (including the model covered in Chapters 3, 5, 6, and 8). As we do not know the mean parameter we will have to rely on estimates/estimators of them, i.e. we observe

$$\begin{aligned}\boldsymbol{Y} =& \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}, \quad \boldsymbol{r} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \\ =& \hat{\boldsymbol{Y}} + \boldsymbol{r}, \quad \boldsymbol{r} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}),\end{aligned} \tag{9-86}$$

where $\boldsymbol{r}$ is the observed residuals (i.e. the realized version of $\boldsymbol{\epsilon}$), $\boldsymbol{\Sigma}$ depend on design matrix ($\boldsymbol{X}$) and σ^2.

Now define the residual sum of squares as

$$RSS(\boldsymbol{\beta}) = \boldsymbol{r}^T\boldsymbol{r} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{9-87}$$

from the perspective of RSS the best estimator is

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta} RSS(\boldsymbol{\beta}), \tag{9-88}$$

the result of this minimization problem is given in the next theorem:

|||| Theorem 9.24 Least square estimator

Assuming that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible and that $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, then the least square estimator ($\hat{\boldsymbol{\beta}}$) of the mean value parameters ($\boldsymbol{\beta}$) in the general linear model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{9-89}$$

further $\hat{\boldsymbol{\beta}}$ is a central estimator ($E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$) and the variance-covariance matrix of the estimator is

$$\mathrm{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{9-90}$$

Throughout this document we will assume that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, and if this is not the case then we will discuss the action needed to me make $\boldsymbol{X}^T\boldsymbol{X}$ invertible (basically removing columns in the design matrix). Cases where one for some reason insist (which may be relevant) on a design matrix where $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible will not be discussed here.

We give the proof of Theorem 9.24 below

|||| Proof

When we want to find the minimum of RSS, we need to differentiate RSS with respect to the parameters ($\boldsymbol{\beta}$). To that end we write RSS as a quadratic form

$$RSS(\boldsymbol{\beta}) = \boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta}, \tag{9-91}$$

since $\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta}$ is a scalar we have $\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta})^T = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$ and hence

$$RSS(\boldsymbol{\beta}) = \boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}, \tag{9-92}$$

and find the derivative wrt. $\boldsymbol{\beta}$ can be found as

$$\begin{aligned}\nabla RSS(\boldsymbol{\beta}) = \frac{\partial RSS}{\partial \boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X} + (\boldsymbol{X}^T\boldsymbol{X})^T)\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{Y} \\ &= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{Y},\end{aligned} \tag{9-93}$$

setting $\nabla RSS(\boldsymbol{\beta}) = \boldsymbol{0}$ and solving for $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{9-94}$$

taking the expectation of $\hat{\boldsymbol{\beta}}$ we get

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\beta}}] &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned} \tag{9-95}$$

Hence $\hat{\boldsymbol{\beta}}$ is a central estimator for $\boldsymbol{\beta}$. The variance of the parameter estimator is given by

$$\begin{aligned}\mathrm{V}[\hat{\boldsymbol{\beta}}] &= \mathrm{V}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{V}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\mathrm{V}[\boldsymbol{X}\boldsymbol{\beta}] + \mathrm{V}[\boldsymbol{\varepsilon}])\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.\end{aligned} \tag{9-96}$$

■

For any reasonable design matrices this imply that $\mathrm{V}[\hat{\boldsymbol{\beta}}] \to 0$ as the number of observation go to infinity, implying that the estimator is consistent.

|||| Definition 9.25 Orthogonal parametrization

A parametrization is called orthogonal if $(\boldsymbol{X}^T\boldsymbol{X})_{ij} = 0$ for $i \neq j$.

An orthogonal parametrization imply that the covariance between parameters is zero. We will see later on in this chapter that the same model can be parameterized in different, but equivalent ways, implying that different design matrices may be associated with the same model. Orthogonal design is (given everything equal) preferable as changes in one parameter does not changes other

parameters. Also one way of dealing with multicollinarity is orthogonalization of the desing matrix.

9.5.1 Estimators or estimates

In the derivations above we have considered the observation as a random variables (and hence used $\boldsymbol{Y}$), and in that setting $\hat{\boldsymbol{\beta}}$ is also a random variable. When we have actual observation of the system we denote the observation by $\boldsymbol{y}$ (this not a random vector) and then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is also a vector of actual numbers (not a random vector) that is refereed to as an estimate.

In the following we will need both interpretations of $\hat{\boldsymbol{\beta}}$, but it should be clear from the context which we are referring to. In general we can say that what we actually observe are estimates, but when constructing appropriate test statistic we consider the estimator. For example the distribution used in the partial t-test is derived using the estimator, $\hat{\boldsymbol{\beta}}$, while when we calculate the test statistic in a specific problem (which is used for calculating a p-value or compared to a critical value), we use the estimate $\hat{\boldsymbol{\beta}}$.

9.5.2 Geometric interpretation of the general linear model (LM)

The estimator/estimate $\hat{\boldsymbol{\beta}}$ define an orthogonal projection of the observations into the space of fitted values, which is defined by the design matrix $\boldsymbol{X}$. Using the parameter estimate $\hat{\boldsymbol{\beta}}$ we can write the fitted values as

$$\begin{aligned}\hat{\boldsymbol{y}} =& \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{y},\end{aligned} \tag{9-97}$$

where the matrix $\boldsymbol{H}$ is defined by the design matrix [2]. The observed residuals can be written as

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}, \tag{9-98}$$

in which case the residuals are observed numbers or we can write

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \tag{9-99}$$

with $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, in which case $\boldsymbol{r}$ is a random vector, both $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ follow a multivariate normal distribution (we will get back to the mean and variance-covariance of those). Many results apply regardless of the interpretation of $\boldsymbol{r}$,

[2] $\boldsymbol{H}$ is often referred to as the "hat"-matrix, as it puts a hat on $\boldsymbol{Y}$

the exception is of course results related to resulting distributions, which only apply for the random variable interpretation.

The matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is an orthogonal projection matrix (see Definition 9.21) as

$$\begin{aligned}\boldsymbol{H}^T =& (\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H} \\ \boldsymbol{H}^2 =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}.\end{aligned} \tag{9-100}$$

The observed residuals of the model can be written as

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \tag{9-101}$$

the matrix $\boldsymbol{I} - \boldsymbol{H}$ is also an orthogonal projection matrix, and further the residuals and the fitted values are orthogonal

$$\begin{aligned}\boldsymbol{r}^T\hat{\boldsymbol{Y}} =& \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}\boldsymbol{Y} \\ =& \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{Y} = 0.\end{aligned} \tag{9-102}$$

The dimension of the linear subspace defined by the column space of $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is of course p and further the trace of $\boldsymbol{H}$ is equal p, as we can write (using Theorem 9.5)

$$\begin{aligned}Trace(\boldsymbol{H}) =& Trace(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ =& Trace((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) \\ =& Trace(\boldsymbol{I}_p) = p.\end{aligned} \tag{9-103}$$

Hence the dimension of the linear subspace defined by the design matrix $\boldsymbol{X}$ is p and further

$$Trace(\boldsymbol{I} - \boldsymbol{H}) = Trace(\boldsymbol{I}) - Trace(\boldsymbol{H}) = n - p. \tag{9-104}$$

Two models (defined by their design matrices) are equivalent if the resulting orthogonal projection matrices are equal, i.e. if

$$\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T = \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T = \boldsymbol{H}_2. \tag{9-105}$$

Hence a model depend in the projection matrix not on the particular parametrization. We will see examples of this in the next section, where we formulate the first statistical models/methods as LMs.

In statistical models the projections are usually from high dimensional space (n is usually way larger than 3), and hence difficult to illustrate graphically, the following simple example can hopefully illustrate the projection principle in an simple example.

|||| Example 9.26 Items on a balance

Two items A and B are weighted on a balance, first separately then together, giving the observations y_1, y_2, y_3, and the model

$$\begin{aligned} Y_1 &= \beta_A + \epsilon_1 \\ Y_2 &= \beta_B + \epsilon_2 \\ Y_3 &= \beta_A + \beta_B + \epsilon_3 \end{aligned} \tag{9-106}$$

with $\epsilon_i \sim N(0, \sigma^2)$. β_A is the weight of item A and β_B is the weight of item B.

Or in matrix notation

$$\boldsymbol{Y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} + \epsilon = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{9-107}$$

with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Hence

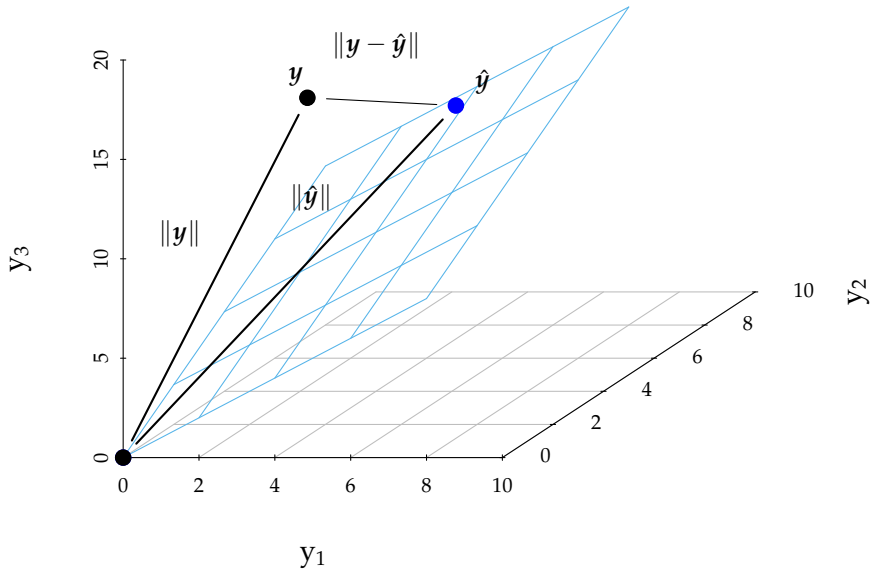
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{3}\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{bmatrix}\boldsymbol{y} \tag{9-108}$$

and

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{3}\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{9-109}$$

The projection $\boldsymbol{H}$ defines a 2-dimensional surface in $\mathbb{R}^3$. In the plot below the "blue" surface define the 2 dimensional surface into which any point is projected, the exact location on the surface is determined by the actual observation, as illustrated in the plot. Further the plot illustrate a norm interpretation of the projection.

To highlight the geometric interpretation the usual norm of the vectors are also indicated in the plot.



The example highlight the geometric interpretation of the projections, in the example we have

- Norm of the observations

$$||\boldsymbol{y}|| = \sqrt{\sum_{i=1}^{n} y_i^2} = \sqrt{\boldsymbol{y}^T\boldsymbol{y}} \tag{9-110}$$

- Norm of fitted values

$$||\hat{\boldsymbol{y}}|| = \sqrt{\sum_{i=1}^{n} \hat{y}_i^2} = \sqrt{\boldsymbol{y}^T\boldsymbol{H}\boldsymbol{y}} \tag{9-111}$$

- Norm of residuals

$$||\boldsymbol{y} - \hat{\boldsymbol{y}}|| = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = \sqrt{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}} \tag{9-112}$$

and further as $\hat{\boldsymbol{y}}$ and $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ are orthogonal it follows (Pythagoras) that

$$||\boldsymbol{y}||^2 = ||\hat{\boldsymbol{y}}||^2 + ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2 \tag{9-113}$$

intuitively we would argue that the data is well explained by the model if $||\hat{\boldsymbol{y}}||^2$ is large compared to $||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2$. When we develop tests in the following it is based on ratios between squared norms of orthogonal projections.

9.6 One-sample t-test as a LM

The one-sample t-test can be written as a general linear model with $\boldsymbol{X} = \mathbf{1}$, i.e. a vector of ones, the orthogonal projection matrix is in this case given by

$$\boldsymbol{H} = \frac{1}{n}\boldsymbol{E}, \tag{9-114}$$

where $\boldsymbol{E}_{ij} = 1$ for all (i, j) and $Trace(\boldsymbol{H}) = 1$ hence the dimension of the model is 1. The model can be written in the form $\boldsymbol{Y} \sim N(\mathbf{1}\mu, \sigma^2\boldsymbol{I})$, and then following corollary to Cochran's theorem apply

|||| Corollary 9.27 One-sample t-test as a projection

If $\boldsymbol{Y} \sim N(\mathbf{1}\mu, \sigma^2\boldsymbol{I})$ then the partitioning of variation can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \tag{9-115}$$

and, regardless of the value of μ, then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \chi^2(n-1). \tag{9-116}$$

further if $\mu = 0$ then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} \sim \chi^2(1). \tag{9-117}$$

Implying that if $\mu = 0$ then

$$F = \frac{\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y}/1}{\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/(n-1)} = \frac{\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/(n-1)} \sim F(1, n-1). \tag{9-118}$$

which can be used to test the null-hypothesis $\mu = 0$.

Proof

First note that $\frac{1}{\sigma^2}(\boldsymbol{Y} - \mathbf{1}\mu) \sim N(\mathbf{0}, \boldsymbol{I})$ (no matter the value of μ), and hence

$$\frac{1}{\sigma^2}(\boldsymbol{Y} - \mathbf{1}\mu)^T(\boldsymbol{Y} - \mathbf{1}\mu) = \frac{1}{\sigma^2}(\boldsymbol{Y} - \mathbf{1}\mu)^T\boldsymbol{H}(\boldsymbol{Y} - \mathbf{1}\mu) + \frac{1}{\sigma^2}(\boldsymbol{Y} - \mathbf{1}\mu)^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \mathbf{1}\mu),$$

and in light of Cochran's Theorem we have that

$$\begin{aligned} \frac{1}{\sigma^2}(\boldsymbol{Y}^T - \mathbf{1}\mu)\boldsymbol{H}(\boldsymbol{Y} - \mathbf{1}\mu) \sim& \chi^2(1) \\ \frac{1}{\sigma^2}(\boldsymbol{Y} - \mathbf{1}\mu)^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \mathbf{1}\mu) \sim& \chi^2(n-1). \end{aligned} \tag{9-119}$$

now consider the second term, the claim is that

$$(\boldsymbol{Y} - \mathbf{1}\mu)^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \mathbf{1}\mu) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{9-120}$$

for any choice of $\mu \in \mathbb{R}$,

$$\begin{aligned} (\boldsymbol{Y} - \mathbf{1}\mu)^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \mathbf{1}\mu) =& \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} - \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\mathbf{1}\mu \\ & - \mu\mathbf{1}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} + \mu\mathbf{1}^T(\boldsymbol{I} - \boldsymbol{H})\mathbf{1}\mu \end{aligned} \tag{9-121}$$

now with $\boldsymbol{H} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ we have $\mathbf{1}^T\boldsymbol{H} = \mathbf{1}^T\mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T = \mathbf{1}^T$, and of course also $\boldsymbol{H}\mathbf{1} = \mathbf{1}$, and hence

$$\frac{1}{\sigma^2}(\boldsymbol{Y} - \mathbf{1}\mu)^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \mathbf{1}\mu) = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}. \tag{9-122}$$

Finally, it is clear that if $\mu = 0$ then $(\boldsymbol{Y}^T - \mathbf{1}\mu)\boldsymbol{H}(\boldsymbol{Y} - \mathbf{1}\mu) = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y}$. And the proof is completed by comparing to definition of the F-distribution (see Theorem 2.96)

■

In Chapter 3 we saw that the test statistics should be compared to a t-distribution with $n-1$ degrees of freedom. If $t \sim t(n-1)$ then $t^2 \sim F(1, n-1)$ and hence the results are equivalent.

In the construction above $\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} \sim \chi^2(1)$ is valid as long as the null-hypothesis and the model assumption is correct, while $\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \chi^2(n-1)$ holds as long as the model assumption are correct and a central estimator for σ^2 can be found by considering the expectation

$$E\left[\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\right] = (n-1) \tag{9-123}$$

or

$$\frac{1}{(n-1)}E\left[\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\right] = \sigma^2 \tag{9-124}$$

and a central estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-1}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{9-125}$$

hence the usual and well known variance estimator.

9.6.1 Assumptions and how to check them

The assumption in the general linear model is that $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, i.e.

1. ϵ_i is normal
2. $V[\epsilon_i]$ is constant (i.e. does not depend on i)
3. $Cor[\epsilon_i, \epsilon_j] = 0$ for all (i, j), implying independence

we do not actually observe ϵ_i but rather we observe

$$\boldsymbol{r} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{H})) \tag{9-126}$$

and conclusions on the residuals will be based on r_i rather than ϵ_i. For the simple case we consider in this section the first two assumptions apply also to r_i as $r_i \sim N(0, \sigma^2(1-h_{ii}))$ and $h_{ii} = \frac{1}{n}$, is independent of i (note this does not apply to the general case). Also it is clear that strictly speaking the third assumption is not fulfilled for the observed residuals as $Cor[r_i, r_j] = -\frac{1}{n-1}$ (see Exercise 7), however the independence assumption is in general hard (or impossible) to check, we treat will an exception below.

9.6.2 Checking lag-1 autocorrelation

A notable case where independence can be checked is when the observations are taken with a clear ordering (typically in time), in this case the correlation between residuals should be checked. There is a extended theory on models that model correlation structures in time (time series analysis), which we will not treat here. We will however stress that the independence assumption should be checked for time series data, a simple check is to calculate the lag 1 autocorrelation (to stress the time dependence we have replaced i by t)

$$\rho_t(1) = \frac{Cov[\epsilon_t, \epsilon_{t+1}]}{\sqrt{V[\epsilon_t]V[\epsilon_{t+1}]}}, \tag{9-127}$$

assuming the correlation and variance in constant (independent of t), we can write

$$\rho(1) = \frac{Cov[\epsilon_t, \epsilon_{t+1}]}{V[\epsilon_t]}, \tag{9-128}$$

and again since we only observe r_t we will have to base the inference on r_t, i.e. the estimator (note that $\bar{r} = 0$)

$$\hat{\rho}(1) = \frac{\sum_{t=1}^{n-1} r_t r_{t+1}}{\sum_{t=1}^{n} r_t^2}, . \tag{9-129}$$

We will not go in details of this estimator, just mention that under the hypothesis that $Cov[\epsilon_t, \epsilon_{t+1}] = 0$ then asymptotically (i.e n large), $\hat{\rho}(1) \sim N(0, 1/n)$ (see Exercise 8). And hence the lag 1 auto-correlation can be compared to that distribution, in practice this imply that we test the hypothesis

$$H_0 : \quad \rho(1) = 0 \tag{9-130}$$

by comparing the estimated lag 1 auto correlation ($\hat{\rho}(1)$) to a quantile (usually the 0.975 quantile) of normal distribution with mean 0 and standard deviation $1/\sqrt{n}$.

9.7 Encoding

A LM is invariant to linear transformations of the design matrix, more specifically if

$$\boldsymbol{X}_2 = \boldsymbol{X}_1 \boldsymbol{T} \tag{9-131}$$

such that $\boldsymbol{T}^{-1}$ exist then

$$\begin{aligned} \boldsymbol{H}_2 =& \boldsymbol{X}_2 (\boldsymbol{X}_2^T \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T \\ =& \boldsymbol{X}_1 \boldsymbol{T} (\boldsymbol{T}^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \boldsymbol{T})^{-1} \boldsymbol{T}^T \boldsymbol{X}_1^T \\ =& \boldsymbol{X}_1 \boldsymbol{T} \boldsymbol{T}^{-1} (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{T}^{-T} \boldsymbol{T}^T \boldsymbol{X}_1^T \\ =& \boldsymbol{X}_1 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T = \boldsymbol{H}_1. \end{aligned} \tag{9-132}$$

Hence the two model defined by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are equivalent and we refer to different parametrizations (defined by $\boldsymbol{T}$) as encoding.

Example 9.28

Say we want to estimate the average height of males (25-50 years) based the data-set presented in Example 9.1. We can do that by considering the model

$$\boldsymbol{Y} = \boldsymbol{X}\mu + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \tag{9-133}$$

with $\boldsymbol{X} = \boldsymbol{1}$, the unit of μ will be the same as data (here cm). The projection matrix is given by

$$\boldsymbol{H} = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T. \tag{9-134}$$

Now let's say that we insist on having the parameter (μ) given in meters ($\mu_m = \mu_{cm}/100$) we can write the model as

$$\begin{aligned} \boldsymbol{Y} =& \boldsymbol{X}\mu_m 100 + \boldsymbol{\epsilon} \\ =& \boldsymbol{X}_m \mu_m + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \end{aligned} \tag{9-135}$$

with $\boldsymbol{X}_m = 100 \cdot \boldsymbol{1}$ and in this case we get

$$\boldsymbol{H}_m = \boldsymbol{X}_m(\boldsymbol{X}_m^T\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m^T = \frac{100^2}{n100^2}\boldsymbol{1}\boldsymbol{1}^T = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T = \boldsymbol{H}_{cm}. \tag{9-136}$$

9.8 Two sample t-test as a LM

The two sample t-test (assuming equal variance in the two groups) can be defined by the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{0}_{n_1} \\ \boldsymbol{0}_{n_2} & \boldsymbol{1}_{n_2} \end{bmatrix}, \tag{9-137}$$

in which case $\boldsymbol{\beta} = [\mu_1, \quad \mu_2]^T$. A more common parametrization of the design matrix is

$$\boldsymbol{X}_2 = \begin{bmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{0}_{n_1} \\ \boldsymbol{1}_{n_2} & \boldsymbol{1}_{n_2} \end{bmatrix}, \tag{9-138}$$

in which case $\boldsymbol{\beta} = [\mu_1, \quad \mu_2 - \mu_1]^T$. The two models are equivalent since

$$\boldsymbol{X}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \boldsymbol{X}\boldsymbol{T} = \boldsymbol{X}_2. \tag{9-139}$$

The usual null hypothesis ($\mu_1 = \mu_2 = \mu$) have the design matrix

$$\boldsymbol{X}_0 = \mathbf{1}. \tag{9-140}$$

The main result of this section is collected in the next corollary.

Corollary 9.29 Two-sample t-test as a projection

If $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, with $\boldsymbol{X}$ as in (9-137) (or any other equivalent parametrization e.g. (9-138)) then the orthogonal partitioning of variation can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}, \tag{9-141}$$

where $\boldsymbol{H}_1$ is based on (9-137) and $\boldsymbol{H}_0$ is based on (9-140). Regardless of the value of $\boldsymbol{\beta}$, then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y} \sim \chi^2(n-2) \tag{9-142}$$

further if $\mu_1 = \mu_2$ (corresponding to $\beta_2 = 0$ in (9-138)) then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y} \sim \chi^2(1). \tag{9-143}$$

and if $\mu_1 = \mu_2 = 0$ (corresponding to $\beta_1 = \beta_2 = 0$) then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} \sim \chi^2(1). \tag{9-144}$$

Implying that if $\mu_1 = \mu_2$ then

$$F_1 = \frac{\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}/1}{\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}/(n-2)} = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}/(n-2)} \sim F(1, n-2), \tag{9-145}$$

and if further $\mu_1 = \mu_2 = 0$ then

$$F_0 = \frac{\frac{1}{\sigma^2}\boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y}/1}{\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}/(n-2)} = \frac{\boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}/(n-2)} \sim F(1, n-2). \tag{9-146}$$

Proof

The proof follow the same steps as the proof of Corollary 9.27, i.e. use that $\boldsymbol{Z} = \frac{1}{\sigma}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \sim N(\boldsymbol{0}, \boldsymbol{I})$, and write the partitioning of variation in terms of $\boldsymbol{Z}$. The details are left to the reader as Exercise 10.

■

It follows from Cochran's Theorem (Theorem 9.23) that $\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}$ and $\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}$ are independent (see Exercise 11), and as a consequence that ($\hat{\boldsymbol{\mu}}_0 = \boldsymbol{H}_0\boldsymbol{Y}$, $\hat{\boldsymbol{\mu}}_1 = \boldsymbol{H}_1\boldsymbol{Y}$)

- $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ are independent
- $\hat{\boldsymbol{\mu}}_0$ and $\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_1$ are independent
- $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ and $\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_1$ are independent

where $\hat{\boldsymbol{\mu}}_i = \boldsymbol{H}_i\boldsymbol{Y}$ are the fitted values based on the projection $\boldsymbol{H}_i$.

The result is in line with the results in Chapter 3, where we found the test-statistics t to be $t(n-2)$-distributed under the null hypothesis, and in that case $t^2 \sim F(1, n-2)$. Further a central estimator for σ^2 is

Corollary 9.30 Variance estimator

With $\boldsymbol{Y}$ and the projections as in Corollary 9.29, then a central estimator for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}}{n-2}, \tag{9-147}$$

the estimator is equal the pooled variance estimator presented in Example 2.85, furthermore the variance of the estimator is

$$V[\hat{\sigma}^2] = \frac{2\sigma^2}{n-2}. \tag{9-148}$$

Proof

The proof follow directly from Corollary 9.29, but see Exercise 12 for more details.

■

In light of the two corollaries (9.27 and 9.29) it is clear that the test statistics are constructed by comparing variance estimators, that are valid under different assumptions (hypothesis). The estimator in Corollary 9.30 is valid no matter what the mean value in each group is (but assuming equal variance in the two groups), while the estimation that could be constructed considering (9-143) or (9-144) are only valid under specific hypothesis ($\mu_1 = \mu_2$ or $\mu_1 = \mu_2 = 0$).

9.8.1 Interpretation of parameters

The two encoding (9-137) and (9-138) result in different interpretation of the estimated parameters. In the case (9-137) the parameters is the group means and confidence intervals for the parameters are confidence intervals for the mean in each group under the assumption of equal variance in the two groups. In the encoding (9-138) the second parameter is the difference in group means and a confidence interval for the second parameter is a confidence interval for the difference in group means, again under the assumption of equal variance in the two groups. See Exercise 15 for an other example of a parametrization.

9.9 Successive testing and partitioning of variation

The discussion of projections and Cochran's theorem suggest that we can formulate a series of nested hypothesis. Nested imply that simpler models are included in the more complicated models by fixing some parameters to specific values (usually zero). The partitioning of variation can be done in different ways, usually referred to as Type I, II and III, for the setup we consider here the relevant once are I, and III.

9.9.1 Type I partitioning of variation

Formally if hypothesis H_i belong to a linear subspace of $\mathbb{R}^n$ of dimension p_i, we can write

$$H_0 \subset H_1 \subset \cdots \subset H_M \subset \mathbb{R}^n \tag{9-149}$$

in practice this is usually realized by adding columns to the design matrix, and an example is

$$\begin{aligned} \boldsymbol{X}_0 =& \mathbf{1} \\ \boldsymbol{X}_1 =& [\mathbf{1} \quad \tilde{\boldsymbol{X}}_1] \\ & \vdots \\ \boldsymbol{X}_i =& [\boldsymbol{X}_{i-1} \quad \tilde{\boldsymbol{X}}_i] \\ & \vdots \\ \boldsymbol{X}_M =& [\boldsymbol{X}_{M-1} \quad \tilde{\boldsymbol{X}}_M], \end{aligned} \tag{9-150}$$

each design matrix, $\boldsymbol{X}_i$, result in the projection matrix $\boldsymbol{H}_i$,

$$\boldsymbol{H}_i = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T, \tag{9-151}$$

and the residual variation is estimated by the projection matrix $\boldsymbol{I} - \boldsymbol{H}_M$. We note here that the results we present here are about projection matrices not the specific parametrization of the design matrix, the construction (9-150) is however a useful way of making projections concrete.

Now define

$$\begin{aligned} SS_0 =& \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y}; \\ SS_i =& \boldsymbol{Y}^T(\boldsymbol{H}_i - \boldsymbol{H}_{i-1})\boldsymbol{Y}; \quad i = \{1, ..., M\} \\ SSE =& \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}; \end{aligned} \tag{9-152}$$

the dimension for each level is

$$\begin{aligned} df_0 =& Trace(\boldsymbol{H}_0) \\ df_i =& Trace(\boldsymbol{H}_i) - Trace(\boldsymbol{H}_{i-1}); \quad i = \{1, ..., M\} \\ df_{SSE} =& n - Trace(\boldsymbol{H}_M). \end{aligned} \tag{9-153}$$

If $\boldsymbol{X}_i \in \mathbb{R}^{n\times p_i}$, then $df_i = p_i - p_{i-1}$ (and $\tilde{\boldsymbol{X}}_i \in \mathbb{R}^{n\times df_i}$). From a statistical test perspective we have

$$F_i = \frac{SS_i/df_i}{SSE/df_{SSE}} \sim F(df_i, df_{SSE}), \tag{9-154}$$

and statistical test can be based on the partitioning presented here. The partitioning is called Type I partitioning of the variation and the test is conditioning on the higher sources being zero. So for example F_i is conditioning on $\tilde{\boldsymbol{X}}_j$ not being included in the model for $j > i$. Formally we collect the results in the following theorem

|||| Theorem 9.31 Type I partioning and tests

If $\boldsymbol{Y} \sim N(\boldsymbol{X}_M\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, with $\boldsymbol{X}_M$ as in (9-150), $\boldsymbol{H}_i$ as in (9-151), and $\boldsymbol{\beta} = [\beta_0, \tilde{\boldsymbol{\beta}}_1^T, ..., \tilde{\boldsymbol{\beta}}_M^T]^T$, with $\tilde{\boldsymbol{\beta}}_i$ parameters corresponding to $\tilde{\boldsymbol{X}}_i$. Then the orthogonal partitioning of variation can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \sum_{i=1}^{M} \boldsymbol{Y}^T(\boldsymbol{H}_i - \boldsymbol{H}_{i-1})\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}, \tag{9-155}$$

and, regardless of the value of $\boldsymbol{\beta}$, then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y} \sim \chi^2(n - p_M), \tag{9-156}$$

further if $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{0}$, for $j > i$ then

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})\boldsymbol{Y} \sim \chi^2(df_j); \quad \text{for all } j > i. \tag{9-157}$$

Implying that

$$F_j = \frac{\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})\boldsymbol{Y}/df_j}{\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}/df_{SSE}} = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})\boldsymbol{Y}/df_j}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}/df_{SSE}} \sim F(1, df_j), \tag{9-158}$$

for $j > i$.

Proof

We start by (9-156); first note that

$$\frac{1}{\sigma}(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}) \sim N(\boldsymbol{0}, \boldsymbol{I}), \tag{9-159}$$

and hence $\frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}) \sim \chi^2(n)$. In light of Cochran's Theorem we have

$$\begin{aligned}\frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}) =& \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta})^T\boldsymbol{H}_M(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}) + \\ & \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta})^T(\boldsymbol{I} - \boldsymbol{H}_M)(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}),\end{aligned} \tag{9-160}$$

the two terms on the rhs follow independent χ^2-distributions with p_m and $n - p_m$ degrees of freedom, respectively. Hence the first claim in the theorem is that

$$(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta})^T(\boldsymbol{I} - \boldsymbol{H}_M)(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}, \tag{9-161}$$

which is true as $\boldsymbol{X}_M^T\boldsymbol{H}_M = \boldsymbol{X}_M^T\boldsymbol{X}_M(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T = \boldsymbol{X}_M^T$. For the claims in (9-157), it correspond to

$$(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta})^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})(\boldsymbol{Y} - \boldsymbol{X}_M\boldsymbol{\beta}) = \boldsymbol{Y}^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})\boldsymbol{Y} \tag{9-162}$$

when $\tilde{\boldsymbol{\beta}}_j = 0$ for $j > i$, and using the notation $\boldsymbol{\beta}_i = [\beta_0, \tilde{\boldsymbol{\beta}}_1^T, ..., \tilde{\boldsymbol{\beta}}_i^T]^T$ we can write (9-162) as

$$(\boldsymbol{Y} - \boldsymbol{X}_i\boldsymbol{\beta}_i)^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})(\boldsymbol{Y} - \boldsymbol{X}_i\boldsymbol{\beta}_i) = \boldsymbol{Y}^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})\boldsymbol{Y} \tag{9-163}$$

and since $\boldsymbol{X}_i^T\boldsymbol{H}_j = \boldsymbol{X}_i^T$ for $j > i$ (see Exercise 14) the proof is done.

■

The results are often collected in an analysis of variance (ANOVA) table as in Table 9.1, usually the hypothesis H_0 is that all observation have the same mean value ($Y_i \sim N(\mu, \sigma^2)$ and iid.), and also in the test setup it is assumed that the model H_M is sufficient, in the sense that the residual under that model are iid. normally distributed with zero mean. The mean sum of squares are all central estimators of the variance under the hypothesis of no effect (see Exercise 13).

In Type I partitioning of variation the total variation can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=0}^{M} SS_i + SSE. \tag{9-164}$$

In the Type I partitioning of variation the order in which variable enter the model in general matters, as the test statistics are conditioning on the previ-

Source of variation	df	Sum of Squares	Mean SS	F-statistics
H_0	df_0	SS_0	$\frac{SS_0}{df_0}$	$\frac{SS_0/df_0}{SSE/df_{SSE}}$
H_1	df_1	SS_1	$\frac{SS_1}{df_1}$	$\frac{SS_1/df_1}{SSE/df_{SSE}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H_M	df_M	SS_M	$\frac{SS_M}{df_M}$	$\frac{SS_M/df_M}{SSE/df_{SSE}}$
Residual	df_{SSE}	SSE	$\frac{SSE}{df_{SSE}}$	

Table 9.1: Partitioning of variation and resulting test statistics.

ous null-hypothesis already being accepted. For exploratory data analysis and testing the Type III partitioning of variation is therefore often preferred.

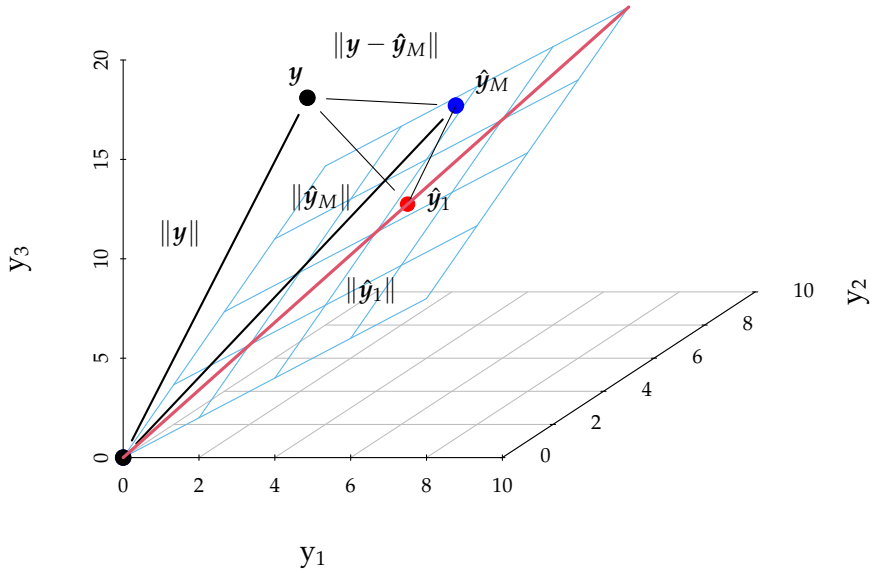
|||| Example 9.32 Items on a scale

We continue the example with items on a scale, again two items are put on scale and weighted first separately then together. In this example we assume that the recorded values are differences to a nominal value, hence the null hypothesis is that the expected difference is zero for both item. There is in this case a fairly obvious hierarchy of hypothesis: $H_0: \quad \mu_1 = \mu_2 = 0$, $H_1: \quad \mu_1 = \mu_2 = \mu$ and the full model H_M that allow different expected values for the two items (which is also assumed to be sufficient). In this case the design matrices could be

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}; \quad \boldsymbol{X}_M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \tag{9-165}$$

and for the null hypothesis there would be no design matrix as the mean value of both items is zero.

The projections are illustrated in the plot below.



From a geometric point of view the norms can be divided into (again using orthogonality)

$$||y||^2 = ||y_M||^2 + ||y - y_M||^2, \tag{9-166}$$

and further the norm of $\hat{y}_M$ can be described as

$$||\hat{y}_M||^2 = ||\hat{y}_1||^2 + ||\hat{y}_M - \hat{y}_1||^2, \tag{9-167}$$

and combining we get

$$||y||^2 = ||\hat{y}_1||^2 + ||\hat{y}_M - \hat{y}_1||^2 + ||y - y_M||^2. \tag{9-168}$$

When testing the described hypothesis's we compare these norms. It seems reasonable that if the expected value of the two item are different then $||\hat{y}_1 - \hat{y}_M||^2$ is large and also if the expected value of the items is not zero then $||\hat{y}_1||^2$ is large. The magnitude is evaluated relative to the variation of residuals ($||y - y_M||^2$), with the precise statements given by the described F-tests. In the presented case the magnitude of $||\hat{y}_1 - \hat{y}_M||$ seems small while $||\hat{y}_1||$ is large compared to $||y - y_M||$, but the precise statement should be based on statistical tests.

The example above illustrate the geometric interpretation of the developed test-statistics.

9.9.2 Type III partitioning of variation

In the Type III partitioning of variation every effect is tested in the setting of the Type I, but formulated as if the effects entered last in the model, i.e. start with the design matrix

$$\boldsymbol{X}_M = [\mathbf{1} \quad \tilde{\boldsymbol{X}}_1 \quad \cdots \quad \tilde{\boldsymbol{X}}_M], \tag{9-169}$$

and the design matrix for testing level i is

$$\boldsymbol{X}_{-i} = [\mathbf{1} \quad \tilde{\boldsymbol{X}}_1 \quad \cdots \quad \tilde{\boldsymbol{X}}_{i-1} \quad \tilde{\boldsymbol{X}}_{i+1} \quad \cdots \quad \tilde{\boldsymbol{X}}_M], \tag{9-170}$$

and the projection is written in a similar way as the Type I partitioning, i.e.

$$\begin{aligned} \boldsymbol{H}_M &= \boldsymbol{X}_M (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \\ \boldsymbol{H}_{-i} &= \boldsymbol{X}_{-i} (\boldsymbol{X}_{-i}^T \boldsymbol{X}_{-i})^{-1} \boldsymbol{X}_{-i}^T, \end{aligned} \tag{9-171}$$

and the partitioning of variation is

$$\boldsymbol{Y}^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_{-i} \boldsymbol{Y} + \boldsymbol{Y}^T (\boldsymbol{H}_M - \boldsymbol{H}_{-i}) \boldsymbol{Y} + \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H}_M) \boldsymbol{Y}, \tag{9-172}$$

there will be M (or $M+1$ if the intercept is included) of those. The result is collected in the theorem below

Theorem 9.33 Type III partioning and test

If $\boldsymbol{Y} \sim N(\boldsymbol{X}_M\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, with $\boldsymbol{X}_M$ as in (9-169), $\boldsymbol{H}_{-i}$ as in (9-171), and $\boldsymbol{\beta} = [\beta_0, \tilde{\boldsymbol{\beta}}_1^T, ..., \tilde{\boldsymbol{\beta}}_M^T]^T$, with $\tilde{\boldsymbol{\beta}}_i$ parameters corresponding to $\tilde{\boldsymbol{X}}_i$. Then the orthogonal partitioning of variation can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_{-i}\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}_M - \boldsymbol{H}_{-i})\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}, \tag{9-173}$$

and, regardless of the value of $\boldsymbol{\beta}$, then (with $p = Rank(\boldsymbol{H}_M)$)

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y} \sim \chi^2(n-p) \tag{9-174}$$

further if $\tilde{\boldsymbol{\beta}}_i = \boldsymbol{0}$, then (with $p_i = Rank(\boldsymbol{H}_{-i})$)

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H}_M - \boldsymbol{H}_{-i})\boldsymbol{Y} \sim \chi^2(p-p_i). \tag{9-175}$$

Implying that if $\tilde{\boldsymbol{\beta}}_i = \boldsymbol{0}$, then

$$F_i = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_M - \boldsymbol{H}_{-i})\boldsymbol{Y}/(p-p_i)}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}/(n-p)} \sim F(p-p_i, n-p). \tag{9-176}$$

Proof

Follow the steps in the proof of Theorem 9.31.

■

The Type III partitioning is often presented in a table similar to Table 9.1, with

$$SS_i = \boldsymbol{Y}^T(\boldsymbol{H}_M - \boldsymbol{H}_{-i})\boldsymbol{Y} \tag{9-177}$$

and the mean sum of squares in a similar way. However due to the construction of the sum of squares, the individual sum of squares does not sum up to the total sum of squares.

9.9.3 Variance estimator

Having estimated the mean value parameters, we also need an estimator for the variance, given the discussion above, the answer is quite straight forward, and

given in the theorem below

Corollary 9.34 Variance estimator

Provided that the model under H_M is sufficient then

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}}{df_{SSE}} \tag{9-178}$$

with $df_{SSE} = n - Trace(\boldsymbol{H}_M)$, is a central estimator for σ^2, and further

$$\frac{df_{SSE}\hat{\sigma}^2}{\sigma^2} \sim \chi^2(df_{SSE}). \tag{9-179}$$

Proof

Follow directly from Theorem 9.31 and 9.33

■

One might re-calibrate the variance estimator, using a reduced model, meaning that we replace $\boldsymbol{H}_M$ by $\boldsymbol{H}_i$ for some i as identified by the model reduction.

9.9.4 Type I or Type III?

An obvious question might of course be if there is a Type II partitioning of variation, and there is. The Type II partitioning, is however related to models that include interactions (or polynomials), and we will skip that for now, but give some comments to how to perform model reduction.

Using the notation of Equation (9-150), and $SS(\boldsymbol{X}_1|\boldsymbol{X}_2)$ meaning the sum of square contribution related to $\boldsymbol{X}_1$ when we have already controlled for $\boldsymbol{X}_2$, then the Type I partitioning correspond to a sequential test, testing for the signifi-

cance of

$$\begin{aligned} &SS(\tilde{\boldsymbol{X}}_1|\boldsymbol{X}_0) \\ &SS(\tilde{\boldsymbol{X}}_2|\boldsymbol{X}_1) \\ &\vdots \\ &SS(\tilde{\boldsymbol{X}}_M|\boldsymbol{X}_{M-1}), \end{aligned} \tag{9-180}$$

hence in each test we condition on (or control for) all preceding levels (effects). The Type III partitioning correspond to controlling for all other levels (effects)

$$\begin{aligned} &SS(\tilde{\boldsymbol{X}}_1|\boldsymbol{X}_0, \tilde{\boldsymbol{X}}_2, ..., \tilde{\boldsymbol{X}}_M) \\ &SS(\tilde{\boldsymbol{X}}_2|\boldsymbol{X}_1, \tilde{\boldsymbol{X}}_3, ..., \tilde{\boldsymbol{X}}_M) \\ &SS(\tilde{\boldsymbol{X}}_3|\boldsymbol{X}_2, \tilde{\boldsymbol{X}}_4, ..., \tilde{\boldsymbol{X}}_M) \\ &\vdots \\ &SS(\tilde{\boldsymbol{X}}_M|\boldsymbol{X}_{M-1}). \end{aligned} \tag{9-181}$$

Hence we see that the two partitioning will agree for the last effect, but may differ for all other effects. Even though there are situations where the Type I partitioning is relevant, we recommend the Type III partitioning, in some of the situation covered here the two partitioning actually agree for all levels.

9.10 Simple and multiple linear regression as a LM

The simple linear regression problem can be formulated in vector-matrix notation as or

$$\begin{aligned} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon_i \sim N(0, \sigma^2) \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \end{aligned} \tag{9-182}$$

hence directly in the notation of the general linear model, and all the results we have seen so far apply here. Further it is straight forward to generalize the result to multiple linear regression

$$\begin{aligned} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon_i \sim N(0, \sigma^2) \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \end{aligned} \tag{9-183}$$

hence again in the notation of the general linear model and the results related to test for model reduction (Type I or III) also apply here and also the central estimator for the variance apply.

The standard error of the parameter estimates are constructed from the variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}}_\beta = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{9-184}$$

In summaries (from statistical software) results from a multiple linear regression model usually present the partial t-test ($H_0 : \beta_i = \beta_{i,0}$), the general constructed is

$$\frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{(\hat{\boldsymbol{\Sigma}}_\beta)_{ii}}} \sim t(n-2), \tag{9-185}$$

and p-values are usually reported for $\beta_{i,0} = 0$, the partial t-test correspond to a specific Type III partitioning.

Theorem 9.35 Partial t-test and Type III partitioning of variation

The partial t-test for the hypothesis $\beta_{i,0} = 0$ and the Type III ANOVA test are equivalent in the sense that is if $\tilde{\boldsymbol{X}}_i$ is a vector then

$$t^2_{obs,i} = F_i \tag{9-186}$$

where F_i is the F-test statistics using the Type III partitioning, and both test statistics should be compared to an F-distribution with 1 and $n-1$ degrees of freedom.

Proof

Without loss of generality we can set $\tilde{\boldsymbol{X}} = [\boldsymbol{X} \quad \boldsymbol{v}] \in \mathbb{R}^{n\times p}$ and use Lemma 9.6 to write t_{obs} as

$$t_{obs} = \frac{\hat{\beta}_p}{\hat{\sigma}/\sqrt{k}}, \tag{9-187}$$

with k as in Corollary 9.7 and hence we have

$$t^2_{obs} = \frac{\hat{\beta}^2_p}{\hat{\sigma}^2/k}. \tag{9-188}$$

Type III F-test can be written as

$$F_p = \frac{\boldsymbol{Y}(\tilde{\boldsymbol{H}} - \boldsymbol{H})\boldsymbol{Y}}{\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/(n-p)} = \frac{\boldsymbol{Y}(\tilde{\boldsymbol{H}} - \boldsymbol{H})\boldsymbol{Y}}{\hat{\sigma}^2} \tag{9-189}$$

also, using Corollary 9.7, we have

$$\tilde{\boldsymbol{H}} - \boldsymbol{H} = \frac{1}{k}(\boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H} - \boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H} - \boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H} + \boldsymbol{v}\boldsymbol{v}^T). \tag{9-190}$$

Now we rewrite $\hat{\beta}_p$ in terms of $\boldsymbol{H}$ and $\boldsymbol{v}$, with $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} =& (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y} \\ =& \begin{bmatrix} \boldsymbol{A} + \frac{\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{k} & \frac{-\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}}{k} \\ \frac{-\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{A}}{k} & \frac{1}{k} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{v}^T \end{bmatrix} \boldsymbol{Y} \\ =& \begin{bmatrix} \boldsymbol{A}\boldsymbol{X}^T + \frac{\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H}}{k} - \frac{\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}\boldsymbol{v}^T}{k} \\ -\frac{\boldsymbol{v}^T\boldsymbol{H}}{k} + \frac{\boldsymbol{v}^T}{k} \end{bmatrix} \boldsymbol{Y},\end{aligned} \tag{9-191}$$

and therefore

$$\hat{\beta}_p = \frac{\boldsymbol{v}^T}{k}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{9-192}$$

and since $\hat{\beta}_p$ is a scalar ($\hat{\beta}_p^2 = \hat{\beta}_p^T\hat{\beta}_p$) we can write

$$\hat{\beta}_p^2 = \frac{1}{k^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v}\boldsymbol{v}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \tag{9-193}$$

and hence

$$t_{obs}^2 = \frac{\frac{1}{k}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v}\boldsymbol{v}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{\hat{\sigma}^2} \tag{9-194}$$

and using (9-190) we get

$$\frac{1}{k}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v}\boldsymbol{v}^T(\boldsymbol{I} - \boldsymbol{H}) = \frac{1}{k}\left(\boldsymbol{v}\boldsymbol{v}^T - \boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H} + \boldsymbol{H}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H}\right) = \tilde{\boldsymbol{H}} - \boldsymbol{H} \tag{9-195}$$

we have shown that $F_p = t_{obs}^2$ and the proof is completed.

■

Test for total homogeneity

Often a test for total homogeneity will be reported along with the partial t-test a discussed above, referring to (9-183) this correspond to the test

$$\beta_1 = \cdots = \beta_p = 0 \tag{9-196}$$

against the alternative that at least one variable have a significant effect (i.e. reduce the sum of squares) in the output.

Example 9.36 Temperature anomali

As an example we look at the so-called global temperature anomali, which is defined as the global average temperature of a year minus the average global temperature over the period 1900-2000. In the data the period covered is 1850-2023. The result of a simple linear regression model is given below.

```
fitTemp = smf.ols('Anomaly ~ Year',data = GlobalTemp).fit()
fitTemp.summary(slim=True)


<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                Anomaly   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.603
No. Observations:                 174   F-statistic:                     263.2
Covariance Type:            nonrobust   Prob (F-statistic):           1.65e-36
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -12.0355      0.745    -16.146      0.000     -13.507     -10.564
Year           0.0062      0.000     16.224      0.000       0.005       0.007
==============================================================================

"""
```

From the summary it is clear that there is a significant increase of temperature, according to the model the increase is around 0.0062 degrees per year. We will get back to the validity of the model in the following sections. The p-values for intercept and slope are both reported as 0 (of course it just mean that the are very small). The test statistics for total homogeneity is 263.2, since this is a simple linear regression model is equals the squared rest statistics for the slope ($16.224^2 = 263.2$), and in this case the numerical value of the p-value is actually given ($1.65 \cdot 10^{-36}$).

9.10.1 Linear transformation of regressors (input)

The LM is invariant to linear transformation of the design matrix, suppose for example that we have collected some output under different temperature con-

ditions, and hence hat the design matrix

$$\boldsymbol{X}_C = [\mathbf{1} \quad \boldsymbol{t}], \tag{9-197}$$

where $\boldsymbol{t}$ is a collection of temperature measurements (measured in degrees Celsius) and associated with some outcome to be modeled, someone now ask for the same model but with temperature given in degrees Fahrenheit, i.e. the design matrix

$$\boldsymbol{X}_F = [\mathbf{1} \quad \boldsymbol{f}], \tag{9-198}$$

where $\boldsymbol{f}$ is the temperatures measured in degrees Fahrenheit, the conversion is

$$\boldsymbol{f}_i = 32 + 1.8\boldsymbol{t}_i, \tag{9-199}$$

and hence we can write

$$\boldsymbol{X}_F = \boldsymbol{X}_C \begin{bmatrix} 1 & 32 \\ 0 & 1.8 \end{bmatrix}. \tag{9-200}$$

Hence the models are equivalent as long as the intercept is included. This property (that model are invariant to linear transformations) is also the reason that it is usually not recommended to remove the intercept in model selection steps, and in the above example the models would not be equivalent if the intercept would have been removed as part of a model selection procedure.

|||| Example 9.37 Temperature anomali

In the temperature example above it seems reasonable to use either the mid-point of the years ($(1850 + 2023)/2 = 1936$), or the the midpoint of the reference period (1950) as reference. If we denote that point (i.e. either 1936 or 1950) as x_{ref}, then the transformation matrix would be

$$\boldsymbol{X}_{ref} = \boldsymbol{X} \begin{bmatrix} 1 & -x_{ref} \\ 0 & 1 \end{bmatrix}, \tag{9-201}$$

here $\boldsymbol{X}$ is a matrix with the first column a vector of ones and the second column a vector with the years. If $x_{ref} = 1936$ then the parametrization is orthogonal and otherwise it is not.

9.10.2 Residual analysis

Even though the raw residuals

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} \tag{9-202}$$

are often used for residual analysis, it is more common to use some standardized version. First we note that even though the residual errors (ϵ_i) are iid. with constant variance, then the observed residuals are not. The distribution of the observed residuals is

$$\boldsymbol{r} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})) \tag{9-203}$$

and hence $V[r_i] = \sigma^2(1 - h_{ii})$ where h_{ii} is the i'th diagonal element of $\boldsymbol{H}$. In that light it is natural to define standardized residuals

|||| Definition 9.38 Standardized residuals[a]

Standardized residuals are defined as

$$r_i^{rs} = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}. \tag{9-204}$$

[a] Standardized residuals are sometimes (e.g. in some Python packages) referred to as internally Studentized residuals.

The standardized residuals are widely used and have the advantage that the variance is constant ($V[r_i^{rs}] = V[r_j^{rs}]$), for all (i, j) if the model assumption is correct. Hence the standardized residuals are well suited for assessing the assumption of variance homogeneity, however the enumerator and denominator are not independent, this imply that the standardized residuals have a very complicated distribution, and hence for more precise assessment the Studentized residuals are often used

|||| Definition 9.39 Studentized residuals

Studentized residuals are defined as

$$r_i^{rt} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}, \tag{9-205}$$

where $\hat{\sigma}^2_{(i)}$ is the estimate of the variance, excluding the i'th observation.

One advantage of the Studentized residuals is that normalization factor is not inflated by large values of r_i, which may be a problem in the standardized version. Further the distribution of the Studentized residuals is simpler.

Theorem 9.40 Distribution of studentized residuals

r_i and $\hat{\sigma}^2_{(i)}$ are independent and

$$r_i^{rt} \sim t(n-p-1). \tag{9-206}$$

Proof

We have already established that $\frac{n-p-1}{\sigma^2}\hat{\sigma}^2_{(i)} \sim \chi^2(n-p-1)$ and further from the discussion in this section we also have that $\frac{r_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$, and also

$$\frac{\frac{r_i}{\sigma\sqrt{1-h_{ii}}}}{\sqrt{\frac{n-p-1}{\sigma^2}\hat{\sigma}^2_{(i)}\frac{1}{n-p-1}}} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i^{rt}, \tag{9-207}$$

hence if r_i and $\hat{\sigma}_{(i)}$ are independent then the proof is done. To that end it is enough to show that r_i and $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}_{-i}$ are independent.

For the independence we denote a specific row (i) of a matrix by $\boldsymbol{A}_{i,\cdot}$, and also all rows except row i by $\boldsymbol{A}_{-i,\cdot}$. With this we can write

$$\begin{aligned} r_i &= (\boldsymbol{I}-\boldsymbol{H})_{i,\cdot}\boldsymbol{Y} = Y_i - \boldsymbol{H}_{i,\cdot}\boldsymbol{Y} \\ \hat{\sigma}^2_{(i)} &= \boldsymbol{Y}^T_{-i}(\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{Y}_{-i}, \end{aligned} \tag{9-208}$$

where $\tilde{\boldsymbol{H}}$ is the projection matrix for the model excluding the i'th observation. Hence it suffice to show that the covariance between r_i and $(\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{Y}_{-1}$ is zero

$$\begin{aligned} Cov[r_i, (\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{Y}_{-i}] &= Cov[Y_i, (\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{Y}_{-i}] - Cov[\boldsymbol{H}_{i,\cdot}\boldsymbol{Y}, (\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{Y}_{-i}] \\ &= \boldsymbol{0} - \boldsymbol{H}_{i,\cdot}Cov[\boldsymbol{Y}, \boldsymbol{Y}_{-i}](\boldsymbol{I}-\tilde{\boldsymbol{H}}), \end{aligned} \tag{9-209}$$

now note that we can write $\boldsymbol{Y}_{-i}$ as $\boldsymbol{I}_{-i,\cdot}\boldsymbol{Y}$, and since $\boldsymbol{I}^T_{-i,\cdot} = \boldsymbol{I}_{\cdot,-i}$, we have

$$\begin{aligned} Cov[r_i, (\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{Y}_{-i}] &= -\boldsymbol{H}_{i,\cdot}\boldsymbol{I}_{\cdot,-i}(\boldsymbol{I}-\tilde{\boldsymbol{H}}) \\ &= -\boldsymbol{H}_{i,-i}(\boldsymbol{I}-\tilde{\boldsymbol{H}}). \end{aligned} \tag{9-210}$$

Hence we need to show that $\boldsymbol{H}_{i,-i} = \boldsymbol{H}_{i,-i}\tilde{\boldsymbol{H}}$, for that purpose write the two matrices

$$\begin{aligned} \boldsymbol{H}_{i,-i} &= \boldsymbol{X}_{i,\cdot}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T_{-i,\cdot} \\ \tilde{\boldsymbol{H}} &= \boldsymbol{X}_{-i,\cdot}(\boldsymbol{X}^T_{-i,\cdot}\boldsymbol{X}_{-i,\cdot})^{-1}\boldsymbol{X}^T_{-i,\cdot}, \end{aligned} \tag{9-211}$$

and form the product

$$\begin{aligned} \boldsymbol{H}_{i,-i}\tilde{\boldsymbol{H}} &= \boldsymbol{X}_{i,\cdot}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T_{-i,\cdot}\boldsymbol{X}_{-i,\cdot}(\boldsymbol{X}^T_{-i,\cdot}\boldsymbol{X}_{-i,\cdot})^{-1}\boldsymbol{X}^T_{-i,\cdot} \\ &= \boldsymbol{X}_{i,\cdot}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T_{-i,\cdot} = \boldsymbol{H}_{i,-i}, \end{aligned} \tag{9-212}$$

which complete the proof.

■

From the definition it seems that one would have to re-estimate the model n times in order to find the Studentized residuals, there does however exist solutions for calculating the Studentized residual directly from the standardized (or raw) residuals allowing fast computation.

|||| Example 9.41 Temperature anomali

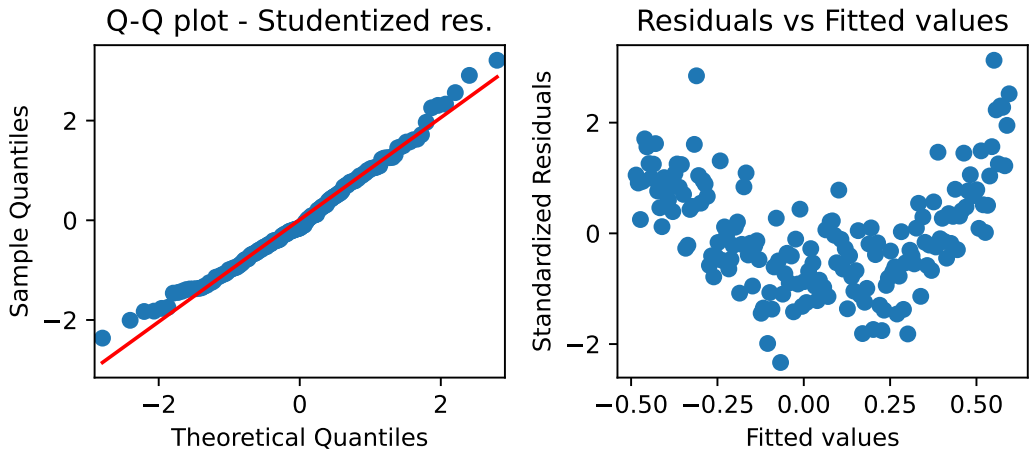
The standardized and Studentized residuals can be calculated in Python by

```
n = len(GlobalTemp["Year"])
X = np.array([np.repeat(1,n), GlobalTemp["Year"]]).T
H = X @ np.linalg.inv(X.T @ X) @ X.T
h = H.diagonal(0)
r = fitTemp.resid
sigma = np.sqrt(fitTemp.scale)
rstandard = r / (sigma * (np.sqrt(1 - h)))
rstudent = fitTemp.outlier_test()
rstudent
```

```
     student_resid   unadj_p  bonf(p)
0         1.053760  0.293480      1.0
1         0.908978  0.364641      1.0
2         0.249237  0.803477      1.0
3         0.938548  0.349287      1.0
4         1.716483  0.087884      1.0
..             ...       ...      ...
169       2.330454  0.020950      1.0
170       2.305061  0.022365      1.0
171       1.222946  0.223033      1.0
172       1.966563  0.050852      1.0
173       2.561582  0.011283      1.0

[174 rows x 3 columns]
```

The `outlier_test` method include p-values for each of the residuals, either based directly on the t-distribution of a Bonferroni adjusted version (in this case we do 174 tests). Here we focus on the visual inspection, implying residual vs. fitted and qq-plot of the residuals

```
ypred = fitTemp.predict(GlobalTemp)
fig, ax = plt.subplots(1,2)
fig = sm.qqplot(rstudent["student_resid"], stats.t,
                distargs=(174-2-1,),line="q",a=1/2,ax=ax[0])
ax[0].set_title("Q-Q plot - Studentized res.")
ax[1].scatter(ypred, rstandard)
ax[1].set_xlabel("Fitted values")
ax[1].set_ylabel("Standardized Residuals")
ax[1].set_title("Residuals vs Fitted values")
plt.tight_layout()
plt.show()
```



It is clear that the fit is not satisfactory, the Studentized residual does not follow a t-distribution with 171 degrees of freedom and is seems that at least a quadratic term is needed.

Influential observation

The residuals analysis is used for verifying the model assumptions, this imply checking the distribution and variance homogeneity assumptions. As we have discussed above the raw residuals does not have variance homogeneity even when the iid. assumption is true. Therefore it is better to use standardized or Studentized residuals for residual analysis. Further for verifying the distributional assumption the Studentized residual have an advantage. We will however also note that in most situations the adjustment made by $\sqrt{1-h_{ii}}$ is small and conclusions in well designed problems will not be greatly affected by which type of residuals we use.

A more important part of the residual analysis is to identify influential observa-

tions. Observation with large residuals have a high impact on the loss function (RSS), and as such these may have a large impact on the parameter estimates. Using the Studentized also allow us to determine what a large residual is in absolute terms (i.e. compare with a specific distribution function). Even though large residuals are in violation of the distribution assumption of the model, it may not have a very large impact on the mean value parameters.

Besides being far away from the model, prediction an observation can also be unusual in the sense that the experimental condition are far away from other experiential conditions. This is measured by leverage, which is defined as the diagonal elements of $\boldsymbol{H}$, in order to understand this consider the derivative of the fitted vales wrt. the observations

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}} = \boldsymbol{H}, \tag{9-213}$$

hence it is a measure of the change in the fitted vales for a unit chance in the observation. This imply that an observation with a high leverage has the potential of being very influential, and we should keep an extra eye on high leverage points. This does not imply that we should avoid such points as they are helpful in spanning the space of possible outcomes, however as they have the potential of greatly impacting the parameter values we should pay attention to those points.

Hence when assessing the model assumptions we should

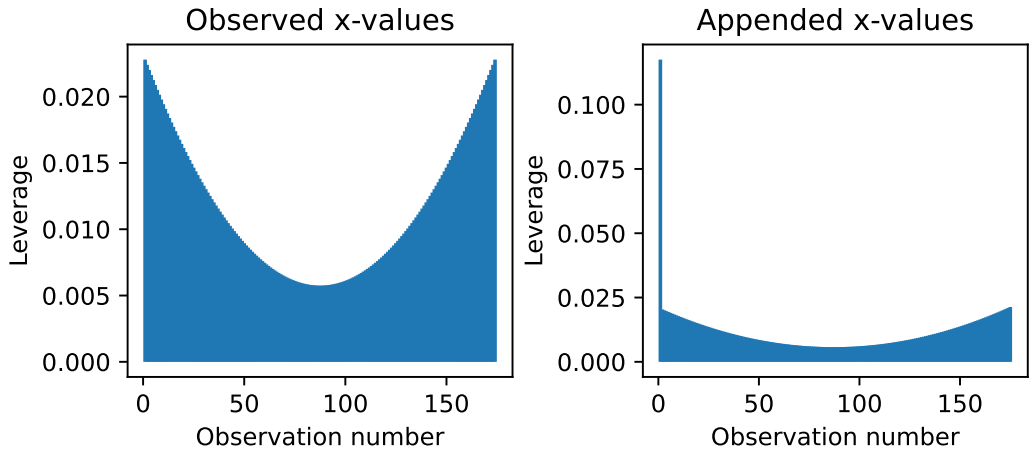
- check normality using standardized or Studentized residuals (qq-plot)
- check variance homogeneity using standardized or Studentized residuals (residuals vs. fitted)
- keep an eye on leverage (e.g. plotting the leverage as a function of obs number)
- check independence (when relevant), e.g. autocorrelation using using standardized or Studentized residuals

Observation that have the largest influence on the model are those with high leverage and a large absolute value of the residual, and the two are sometimes combined in Cook's distance (which wwe will not discuss here).

|||| Example 9.42 Temperature anomali

The leverage corresponding to the explanatory variable (year) in the temperature data is plotted below (left). We see that the leverage is smaller for observation close

to the center of the observed explanatory variables and somewhat higher at the endpoints in the interval. The right plot is constructed imagining we have an observation of the temperature anomaly in year 1700, this is a quite extreme value compared to the other observed years, and resulting in a very high leverage, and hence an observation there would have the potential to greatly influence the model.



9.10.3 Multicollinarity

Multicollinarity is linked to high empirical correlation between columns in the design matrix, and can often be identified through visual analysis of pairwise plots of the regressors. Another way is to consider the correlation between parameters, or even more generally consider properties of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ in particular the condition number can indicate if the matrix is close to singular.

|||| Example 9.43 An ill-conditioned problem

To illustrate the multicollinarity problem consider the data

```
x1 = np.array([5.5, 6.5, -2.5, -6.5, -0.5, 1.5, -3.5, -1.5, -5.5,  4.5,
               2.5, -4.5, 3.5, 0.5])
x2 = np.array([1.5, -3.5, 3.5, -1.5, 4.5, -5.5, -0.5, 6.5, -2.5, -4.5,
               5.5, 2.5, 0.5, -6.5])
x3 = np.array([7.0, 3.0, 1.0, -8.0, 4.01, -4.01, -4.0, 5.0, -8.0, 0.0,
               8.0, -2.0, 4.0, -6.0])
y = np.array([32.43, 15.54, 2.50, -15.62, 5.45, -8.61, -2.50, 33.33,
              -39.56, -27.26,  14.44,  -0.50, 22.66, -18.53])
datIll = pd.DataFrame({'x1': x1, 'x2': x2, 'x3': x3, 'y': y})
```

`x1`, `x2`, and `x3` are constructed such that the average of each of them is zero, and hence the correlation between (not to be confused with the correlation between the parameters) them can be calculate by

```
X = pd.DataFrame({'x1': x1, 'x2': x2, 'x3': x3})
C = X.T @ X
Cd = np.diag(np.sqrt(np.diag(C)))
np.linalg.inv(Cd) @ C @ np.linalg.inv(Cd)
```

```
          0         1         2
0  1.000000 -0.164835  0.646002
1 -0.164835  1.000000  0.646410
2  0.646002  0.646410  1.000000
```

here there are no very strong correlation, however the condition number is

```
np.linalg.cond(C)
```

```
np.float64(10968536.72257104)
```

which is extremely large. In this case the result of Type I and Type III partitioning of variation will also be very different.

```
fit = smf.ols('y ~ x1 + x2 + x3',data = datIll).fit()
## Type I
sm.stats.anova_lm(fit, typ = 1)
```

```
            df       sum_sq      mean_sq          F    PR(>F)
x1         1.0  1193.490794  1193.490794   7.513585  0.020794
x2         1.0  3139.890575  3139.890575  19.767085  0.001243
x3         1.0   202.736604   202.736604   1.276322  0.284954
Residual  10.0  1588.443948   158.844395        NaN       NaN
```

```
## Type III
sm.stats.anova_lm(fit, typ = 3)
```

```
                sum_sq    df          F    PR(>F)
Intercept    13.543779   1.0  0.085264  0.776254
x1          203.773250   1.0  1.282848  0.283797
x2          204.077732   1.0  1.284765  0.283458
x3          202.736604   1.0  1.276322  0.284954
Residual   1588.443948  10.0       NaN       NaN
```

hence we see that from the Type I analysis we should remove x3 (because it was entered last), while the Type III analysis show that we can actually remove any of the 3 regressors.

The example illustrate that there might be big differences in conclusion depending on the chosen partitioning and a natural question is if there are situations where conclusions is aligned, the answer is given in the next theorem.

Theorem 9.44 Orthogonal parameters and and partioning

With an orthogonal parametrization (see Definition 9.25) then Type I and Type III partitioning is equivalent.

Proof

An orthogonal parametrization imply that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}_{ii} = \lambda_i$ and $\boldsymbol{\Lambda}_{ij} = 0$ if $i \neq j$. Hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{\Lambda}^{-1}$ with $(\boldsymbol{\Lambda}^{-1})_{ii} = 1/\lambda_i$ and zero otherwise. Now let the columns of $\boldsymbol{X}$ be denoted by $\boldsymbol{x}_i$, the orthogonality imply that

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i}\boldsymbol{x}_i\boldsymbol{x}_i^T, \tag{9-214}$$

testing using Type I partitioning we would have

$$\boldsymbol{H}_i - \boldsymbol{H}_{i-1} = \sum_{j=1}^{i} \frac{1}{\lambda_j}\boldsymbol{x}_j\boldsymbol{x}_j^T - \sum_{j=1}^{i-1} \frac{1}{\lambda_j}\boldsymbol{x}_j\boldsymbol{x}_j^T = \frac{1}{\lambda_i}\boldsymbol{x}_i\boldsymbol{x}_i^T, \tag{9-215}$$

and in the Type III set up we would have

$$\boldsymbol{H}_p - \boldsymbol{H}_{-i} = \sum_{j=1}^{p} \frac{1}{\lambda_j}\boldsymbol{x}_j\boldsymbol{x}_j^T - \sum_{j\neq i} \frac{1}{\lambda_j}\boldsymbol{x}_j\boldsymbol{x}_j^T = \frac{1}{\lambda_i}\boldsymbol{x}_i\boldsymbol{x}_i^T, \tag{9-216}$$

hence exactly the same projection matrix and therefore also the same test-statistics, p-values and so on.

■

The example above highlight some multicollinarity problems in addition to the Type I and III partitioning not agreeing large changes in the parameter values will also be present when reducing the model (see Exercise 16). Further Theorem 9.44 state that we do not have to worry abut such problem if we already

have an orthogonal parametrization. For the problems we consider here ($\boldsymbol{X}^T\boldsymbol{X}$) is invertible, and it is always possible to transform a multicollinarity problem to an orthogonal parametrization (see Section 9.13.2). The price to is pay is interpretability of the parameters.

9.10.4 Polynomial and basis function regression

Polynomial regression is often used as a way of modeling otherwise non-linear relationships, it is well known that any continuous function can be approximated by its Taylor expansion, hence if we assume that

$$Y_i = f(x_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2), \tag{9-217}$$

then Y_i can be approximated by

$$Y_i \approx \sum_{j=0}^{p} \frac{1}{j!} f^{(j)}(x_0)(x_j - x_0)^j + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2), \tag{9-218}$$

and as $p \to \infty$ the approximation becomes better. When used in statistical modeling, we do not known the coefficients $\left(\frac{1}{i!} f^{(i)}(x_0)\right)$, and hence the statistical model would be

$$Y_i = \sum_{j=0}^{p} \beta_j \cdot (x_i - x0)^j + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2), \tag{9-219}$$

here we can choose the expansion point (x_0) as we please. The choice of x_0 will however affect the parameter correlation and thereby the multicollinarity of the problem, often the problem is actually casted as

$$Y_i = \sum_{j=0}^{p} \beta_j x_i^j + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2) \tag{9-220}$$

such an parametrization may lead to strong multicollinarity and often the model is formulated as

$$Y_i = \sum_{j=0}^{p} \beta_j \cdot p_j(x_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2) \tag{9-221}$$

where $p_j(x_i)$ is a j'th order polynomial, chosen such that

$$\sum_{i=1}^{n} p_k(x_i) p_l(x_i) = 0; \quad \text{for } k \neq l. \tag{9-222}$$

resulting in an orthogonal parametrization. Often the extra constraint $\sum_i p_k(x_i)^2 = 1$ for $k > 0$ is added.

|||| Example 9.45 Temperature anomali

The analysis in Example 9.41 suggest that at least a quadratic term should be included. As a starting point we might included a forth order polynomial, in the summary below `pj_raw` is short for $(Year_i/max(Year))^j$, from the partial t-test it seems that none of coefficient are significant. However it is also clear from the test of total homogeneity that at least one of the terms are significant. Further it is noted in the summary that the smallest eigenvalue is $2 \cdot 10^{-13}$ indicating very strong multicollinarity.

```
fitTemp4 = smf.ols('Anomaly ~ p1_raw + p2_raw+ p3_raw+p4_raw',
                   data = GlobalTemp).fit()
fitTemp4.summary(slim=True)


<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                Anomaly   R-squared:                       0.839
Model:                            OLS   Adj. R-squared:                  0.835
No. Observations:                 174   F-statistic:                     220.0
Covariance Type:            nonrobust   Prob (F-statistic):           7.31e-66
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5915.1051   4.01e+04      0.148      0.883   -7.32e+04     8.5e+04
p1_raw     -2.707e+04   1.68e+05     -0.162      0.872   -3.58e+05    3.04e+05
p2_raw       4.64e+04   2.63e+05      0.177      0.860   -4.72e+05    5.65e+05
p3_raw     -3.531e+04   1.83e+05     -0.193      0.847   -3.97e+05    3.26e+05
p4_raw      1.007e+04   4.78e+04      0.211      0.833   -8.43e+04    1.04e+05
==============================================================================

[2] The smallest eigenvalue is 2.01e-13. This might indicate that there are
"""
```

In addition to the notes made above we also see very large coefficient (the output is plus minus a few degrees and the coefficient are above 10^4). Of course we can in this case just check third and second degree order polynomials

```
sm.stats.anova_lm(fitTemp2,fitTemp3,fitTemp4)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	171.0	4.821795	0.0	NaN	NaN	NaN
1	170.0	4.559946	1.0	0.261849	9.707168	0.002155
2	169.0	4.558750	1.0	0.001196	0.044341	0.833474

It is clear that the model can be reduced to a third degree polynomial, but should probably not be reduced further.

As illustrated in the above example care should be taken when constructing polynomial regression models. Strong multicollinarity might be introduced if polynomials are naively formulated, below we will discuss how orthogonal polynomials can be formulated.

Construction of orthogonal polynomials

The definitions discussed above might seems a bit abstract and difficult to handle in practice, it is however quite simple to set up recursive algorithms for the construction. Start by setting $p_0(x_i) = 1$, and define

$$p_1(x_i) = a_{10} + x_i \tag{9-223}$$

the orthogonality constraint imply

$$\sum_i p_0(x_i)p_1(x_i) = \sum_i a_{10} + x_i = na_{10} + n\bar{x} = 0 \tag{9-224}$$

or $a_{10} = -\bar{x}$. For the normalization set

$$\tilde{p}_1(x_i) = a_{11}(a_{10} + x_i) \tag{9-225}$$

and hence the normalization imply

$$\sum \tilde{p}_1(x_i)^2 = a_{11}^2(a_{10} + x_i)^2 = 1 \tag{9-226}$$

or $a_{11} = 1/\sqrt{\sum(a_{10} + x_i)^2} = 1/\sqrt{\sum(x_i - \bar{x})^2}$ and hence

$$\tilde{p}_1(x_i) = -\frac{\bar{x}}{\sqrt{\sum(x_i - \bar{x})^2}} + \frac{x_i}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{9-227}$$

In order to simplify notation we will set $p_{ki} = p_k(x_i)$ (i.e. the k'th order polynomial applied to x_i), for the second order polynomial ($p_{2i} = a_{20} + a_{21}x_i + x_i^2$) we

have

$$\begin{aligned}\sum \tilde{p}_{0i}p_{2i} &= \sum_i a_{20}\tilde{p}_{0i} + a_{21}\tilde{p}_{0i}x_i + x_i^2\tilde{p}_{0i} = a_{20}n\overline{\tilde{p}_0} + a_{21}n\overline{\tilde{p}_0 x} + n\overline{x^2\tilde{p}_0} = 0\\ \sum \tilde{p}_{1i}p_{2i} &= \sum_i a_{20}\tilde{p}_{1i} + a_{21}\tilde{p}_{1i}x_i + x_i^2\tilde{p}_{1i} = a_{20}n\overline{\tilde{p}_1} + a_{21}n\overline{\tilde{p}_1 x} + n\overline{x^2\tilde{p}_{1i}} = 0,\end{aligned} \tag{9-228}$$

where the "bar" notation simply means the average of what is under the bar (e.g. $\overline{p_1 x} = \frac{1}{n}\sum_i p_{1i}x_i$). This define a set of linear equations

$$\begin{bmatrix}\overline{\tilde{p}_0} & \overline{\tilde{p}_0 x}\\ \overline{\tilde{p}_1} & \overline{\tilde{p}_1 x}\end{bmatrix}\begin{bmatrix}a_{20}\\ a_{21}\end{bmatrix} = \begin{bmatrix}-\overline{x^2\tilde{p}_0}\\ -\overline{x^2\tilde{p}_1}\end{bmatrix} \tag{9-229}$$

which is easily solved numerically, finally the polynomial can be normalized by

$$a_{22} = \frac{1}{\sqrt{\sum_i p_{2i}^2}} \tag{9-230}$$

and setting $\tilde{a}_{20} = a_{22}a_{20}$ and $\tilde{a}_{21} = a_{22}a_{21}$ by the same factor to get the polynomial

$$\tilde{p}_{2i} = \tilde{a}_{20} + \tilde{a}_{21}x_i + a_{22}x_i. \tag{9-231}$$

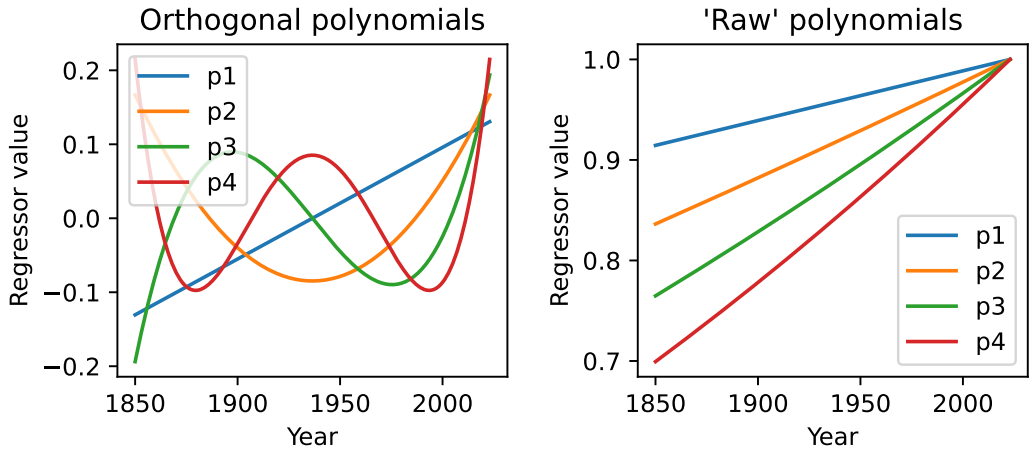
In general we can calculate the the first k coefficient of the k'th order orthogonal, based in the previous polynomials as the solution to

$$\begin{bmatrix}\overline{\tilde{p}_0} & \overline{\tilde{p}_0 x} & \cdots & \overline{\tilde{p}_0 x^{k-1}}\\ \vdots & \vdots & & \vdots\\ \overline{\tilde{p}_{k-1}} & \overline{\tilde{p}_{k-1}x} & \cdots & \overline{\tilde{p}_{k-1}x^{k-1}}\end{bmatrix}\begin{bmatrix}a_{k0}\\ \vdots\\ a_{k,k-1}\end{bmatrix} = \begin{bmatrix}-\overline{x^k\tilde{p}_0}\\ \vdots\\ -\overline{x^k\tilde{p}_{k-1}}\end{bmatrix} \tag{9-232}$$

which can again be normalized as in the case of the second degree polynomial.

|||| Example 9.46 Temperature anomali

The figure below show the orthogonal and the "raw" polynomials (Example 9.45), the "raw" polynomials all seems linear on this scale. This apparent linearity leads to the large multicollinarity problems that was evident in Example 9.45. On the other hand it is clear orthogonal polynomials are well separated and able to take care of different shapes in the resulting models.



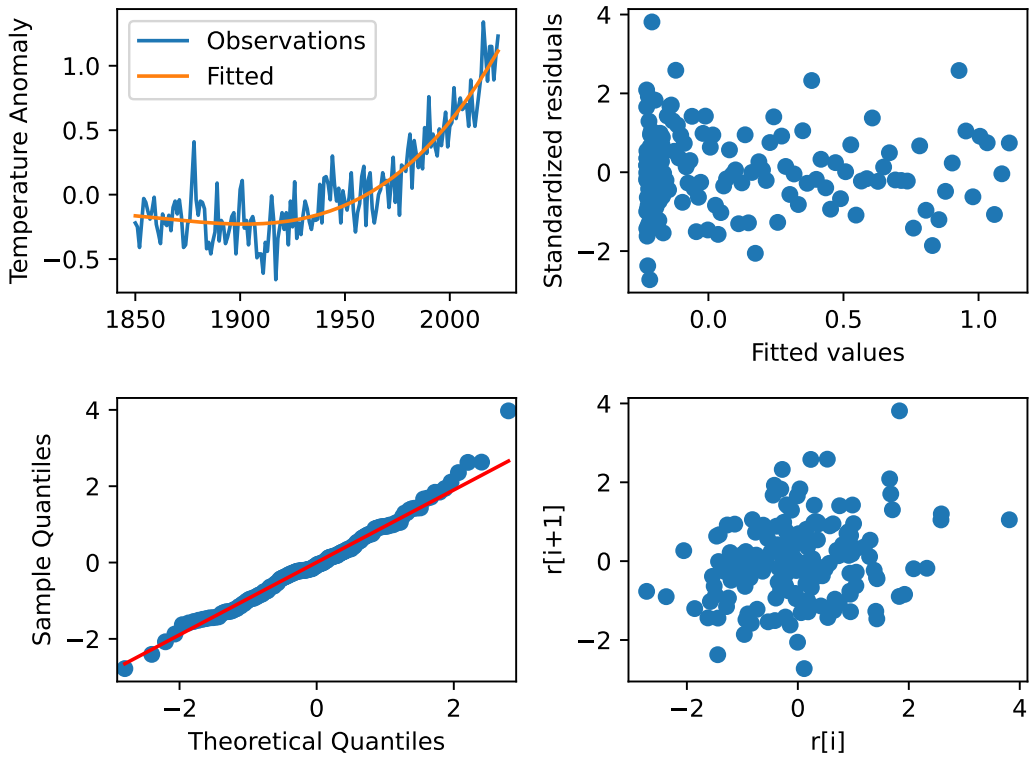
The result of fitting the 4'th order orthogonal polynomials to data is given in the summary table below, the overall statistics (test for total homogeneity, and R^2) are the same, but we can now directly from the output see that the 3'rd order polynomial should be included (using the usual 5%) level, but that the 4'th order should not. Also the extreme values of the parameters are no longer present.

```
fitTemp4ort = smf.ols('Anomaly ~ p1 + p2 + p3 + p4',data=GlobalTemp).fit()
fitTemp4ort.summary(slim=True)


<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                Anomaly   R-squared:                       0.839
Model:                            OLS   Adj. R-squared:                  0.835
No. Observations:                 174   F-statistic:                     220.0
Covariance Type:            nonrobust   Prob (F-statistic):           7.31e-66
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0545      0.012      4.380      0.000       0.030       0.079
p1             4.1365      0.164     25.186      0.000       3.812       4.461
p2             2.5217      0.164     15.354      0.000       2.197       2.846
p3             0.5117      0.164      3.116      0.002       0.187       0.836
p4             0.0346      0.164      0.211      0.833      -0.290       0.359
==============================================================================

"""
```

For completeness we include a more complete residual and model analysis of the

final 3'rd order polynomial regression model. The figure below show that the model follow the data quite well, and there are no systematic behavior in the standardized residuals vs. fitted values (of course there are many observations of small fitted values, but that is the nature of data). Also the qq-plot of Studentized residuals does not raise any concerns, there is one quite large Studentized residual of about 4, which is caused by the unusually high temperature around the year 1880.

The last plot is used for assessing the independence assumption and is based on the standardized residuals. The data is given as a time-series and therefore it is reasonable to check the correlation between observations at time t and time $t+1$, even though weak there seem to be some positive temporal correlation in the residuals.



For a more precise statement on the correlation between r_i^{rs} and r_{i+1}^{rs} we can calculate it in Python by

```
n = len(rstandard)
r1 = rstandard[0:(n-1)]
r2 = np.roll(rstandard,-1)[0:(n-1)]
res = pd.DataFrame({"r1" : r1, "r2" : r2})
res.corr()
```

```
          r1        r2
r1  1.000000  0.234337
r2  0.234337  1.000000
```

hence an estimated correlation of about 0.234, which by (9-130) (on page 31), should be compared with a $N(0, 1/(n-1))$ distribution, the resulting test statistics is $z = 0.234/\sqrt{1/173} = 3.08$, and hence there is a significant autocorrelation in the residuals. Even though there is a significant autocorrelation it is small in this case and not expected to affect the estimation results greatly in this case.

In the example above we saw that including orthogonal polynomial gave more reasonable results and in that light it is important. However simpler methods will often be enough, e.g. subtraction the average of the regressors usually make polynomial regression much more robust (even though not completely orthogonal). In addition variants of polynomial basis functions, like Legendre polynomials, will often also do very good (when implemented appropriate ways). Hence simpler measures can be taken that greatly improve the condition number without making everything completely orthogonal.

Other basis functions

Before using polynomial regression one should carefully consider if it is the right choice, for example if there is a natural periodicity (e.g. hour of day) it is better to use Fourier series expansion, i.e. replace $\beta_j \cdot p_j(x_i)$ by $\beta_{1j} sin(j2\pi x_i/P) + \beta_{2j} cos(j2\pi x_i/P)$ where P is the period (e.g. 24 hours). Finally more local basis functions (e.g. spline basis functions) are often used.

Predictions using basis function

Extrapolation the results of linear regression models should always be done with care, this is especially true if polynomial type basis functiuons are used. The behaviour of the resulting functions may be quite extreme in areas where there are no data.

9.11 One-way ANOVA as a LM

The one-way ANOVA model can be written as

$$Y_{ij} = \beta_i + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{9-233}$$

in the following we will assume that the vector of observations is organized as $\boldsymbol{y} = [y_{11}, y_{12}, ..., y_{1n_1}, y_{21}, ..., y_{2n_2}, ..., y_{Kn_K}]$, with that convention the design matrix for the one-way ANOVA model can be written as

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & & \ddots & \vdots \\ \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \dots & \mathbf{1}_{n_K} \end{bmatrix}, \tag{9-234}$$

in this case the parameters are the group means. The standard encoding, in e.g Python, is

$$\boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & & \ddots & \vdots \\ \mathbf{1}_{n_K} & \mathbf{0}_{n_K} & \dots & \mathbf{1}_{n_K} \end{bmatrix}, \tag{9-235}$$

in which case the first parameters is the mean of group 1 and the remaining parameters is the difference between mean in group 1 and and group i, $\boldsymbol{\beta} = [\mu_1, \mu_2 - \mu_1, \dots, \mu_K - \mu_1]^T$. Again we can write $\boldsymbol{X}_2$ as

$$\boldsymbol{X}_2 = \boldsymbol{X}\boldsymbol{T}, \tag{9-236}$$

with

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \dots & 1 \end{bmatrix}, \tag{9-237}$$

and hence the two models are equivalent.

In Chapter 8 we considered the model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{9-238}$$

such a model is over-parameterized and in Chapter 8 this over-parametrization was dealt with (even though not explicitly mentioned) by the linear constraints

$$\sum_{i=1}^{K} n_i \alpha_i = 0. \tag{9-239}$$

We can now choose an arbitrary reference level, e.g. group K, and write

$$\alpha_K = -\sum_{i=1}^{K-1} \frac{n_i}{n_K}\alpha_i \tag{9-240}$$

and with $\mu_i = \mu + \alpha_i$ we can write

$$\boldsymbol{X}_3 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \dots & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & \vdots & & \ddots & \vdots \\ \mathbf{1}_{n_{K-1}} & \mathbf{0}_{n_{K-1}} & \dots & & \mathbf{1}_{n_{K-1}} \\ \mathbf{1}_{n_K} & -\frac{n_1}{n_K}\mathbf{1}_{n_K} & \dots & \dots & -\frac{n_{K-1}}{n_K}\mathbf{1}_{n_K} \end{bmatrix}, \tag{9-241}$$

which again can be written as

$$\boldsymbol{X}_3 = \boldsymbol{X}\boldsymbol{T}, \tag{9-242}$$

for appropriate choice of $\boldsymbol{T}$.

9.11.1 Orthogonal design: Helmert-transform

The formulation (9-234) is an orthogonal parametrization, however there is not one parameter for the over all mean value, but rather one parameter for the mean in each group. In the balanced case another orthogonal transformation is the Helmert transformation, defined by

$$\boldsymbol{T}_H = \begin{bmatrix} 1 & -1 & -1 & -1 & \dots & -1 \\ 1 & 1 & -1 & -1 & \dots & -1 \\ 1 & 0 & 2 & -1 & \dots & -1 \\ 1 & 0 & 0 & 3 & \dots & -1 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 1 & 0 & \dots & & 0 & k-1 \end{bmatrix}, \tag{9-243}$$

if $\boldsymbol{T}_H$ is "normalized" by a diagonal matrix $\boldsymbol{D}$ with $\boldsymbol{D}_{ii} = 1/i$ (i.e. $\boldsymbol{T}_{HN} = \boldsymbol{T}\boldsymbol{D}$), the interpretation of the parameters is (Exercise 17)

$$\begin{aligned} \hat{\beta}_1 &= \bar{y} \\ \hat{\beta}_i &= \bar{y}_{i+1} - \frac{1}{i}\sum_{j=1}^{i} \bar{y}_j, \quad \text{for } i = 1, 2, ..., k-1. \end{aligned} \tag{9-244}$$

Hence the difference between group i and the average of the preceding groups. Orthogonality imply that variances of linear combinations of parameters are easily calculated, and also as discussed above that Type I and Type III are equivalent.

9.11.2 Statistical tests

Statistical test are preformed exactly as we have seen in the previous sections, compared to the linear regression the only difference is that usually the performed test is the test for total homogeneity (all mean values are equal), and hence no discussion about the order. The post hoc analysis (i.e. when the null hypothesis is rejected), does include a decision on the partitioning.

9.11.3 Contrasts

The matrix $\boldsymbol{T}$ define so-called contrasts, we will no go further into that subject here, just mentioned that the transformation defined by (9-237) is often called treatment-coding, while the formulation (9-241) is (at least in the balanced case $(n_i = n_j)$) called sum-coding.

9.11.4 Partial tests and post hoc analysis

If we are interested in a particular quantity (e.g. $\mu_i - \mu_j$ for fixed (i, j)), then we can simply formulate the model such that the difference is a parameter and use the usual partial t-test. In more generality, if we are interested in all pairwise comparisons (as in Method 8.9), it correspond to a Type III partitioning of variation.

|||| Theorem 9.47 Post hoc comparison and Type III

The post hoc comparison in Methods 8.9 and 8.10, is equivalent to comparing the model

$$Y_{ij} = \beta_i + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{9-245}$$

to a model where $\beta_l = \beta_h$ using a Type III partitioning of variation.

Proof

Method 8.10 state that under the hypothesis that $\mu_l = \mu_h$ then

$$t_{obs} = \frac{\bar{Y}_l - \bar{Y}_h}{\sqrt{MSE\left(\frac{1}{n_l} + \frac{1}{n_h}\right)}} \sim t(n-k) \tag{9-246}$$

implying that $t_{obs}^2 \sim F(1, n-k)$. Hence we need to show that

$$\frac{\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/df_{SSE}} = t_{obs}^2, \tag{9-247}$$

where $\boldsymbol{H}$ is the projection matrix corresponding to the full model and $\boldsymbol{H}_0$ is the projection matrix corresponding to the null hypothesis. First note that $MSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/df_{SSE}$, and hence we need to show that

$$\frac{(\bar{Y}_l - \bar{Y}_h)^2}{\frac{1}{n_l} + \frac{1}{n_h}} = \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y}. \tag{9-248}$$

The projection matrix for the model is

$$\boldsymbol{H} = \begin{bmatrix} \frac{1}{n_1}\boldsymbol{E}_{n_1 n_1} & \boldsymbol{0}_{n_1 n_2} & \dots & \boldsymbol{0}_{n_1 n_k} \\ \boldsymbol{0}_{n_2 n_1} & \frac{1}{n_2}\boldsymbol{E}_{n_2 n_2} & \dots & \boldsymbol{0}_{n_2 n_k} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0}_{n_K n_1} & \boldsymbol{0}_{n_K n_2} & \dots & \frac{1}{n_k}\boldsymbol{E}_{n_K n_k} \end{bmatrix} \tag{9-249}$$

where $\boldsymbol{E}_{n_i n_j}$ is an n_i by n_j matrix of ones. Without loss of generality we can consider $l = 1$ and $h = 2$, in that case the null hypothesis correspond to the design matrix

$$\boldsymbol{X}_0 = \begin{bmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{0}_{n_1} & \dots & \boldsymbol{0}_{n_1} \\ \boldsymbol{1}_{n_2} & \boldsymbol{0}_{n_2} & \dots & \boldsymbol{0}_{n_2} \\ \boldsymbol{0}_{n_3} & \boldsymbol{1}_{n_3} & \dots & \boldsymbol{0}_{n_3} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0}_{n_K} & \boldsymbol{0}_{n_K} & \dots & \boldsymbol{1}_{n_K} \end{bmatrix} \tag{9-250}$$

resulting in

$$\boldsymbol{H}_0 = \begin{bmatrix} \frac{1}{n_1+n_2}\boldsymbol{E}_{n_1+n_2, n_1+n_2} & \boldsymbol{0}_{n_1+n_2, n_3} & \dots & \boldsymbol{0}_{n_1+n_2, n_k} \\ \boldsymbol{0}_{n_3, n_2+n_1} & \frac{1}{n_3}\boldsymbol{E}_{n_3 n_3} & \dots & \boldsymbol{0}_{n_3 n_k} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0}_{n_k, n_1+n_2} & \boldsymbol{0}_{n_k n_2} & \dots & \frac{1}{n_k}\boldsymbol{E}_{n_k n_k} \end{bmatrix} \tag{9-251}$$

and hence

$$\boldsymbol{H} - \boldsymbol{H}_0 = \begin{bmatrix} \left(\frac{1}{n_1} - \frac{1}{n_1+n_2}\right)\boldsymbol{E}_{n_1,n_1} & -\frac{1}{n_1+n_2}\boldsymbol{E}_{n_1,n_2} & \boldsymbol{0} \\ -\frac{1}{n_1+n_2}\boldsymbol{E}_{n_2,n_1} & \left(\frac{1}{n_2} - \frac{1}{n_1+n_2}\right)\boldsymbol{E}_{n_2 n_2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \tag{9-252}$$

now since $\boldsymbol{Y}_i^T \boldsymbol{E}_{n_i,n_i} \boldsymbol{Y}_i = n_i^2 \bar{Y}_i^2$, and $\boldsymbol{Y}_1^T \boldsymbol{E}_{n_1,n_2} \boldsymbol{Y}_2 = n_1 n_2 \bar{Y}_1 \bar{Y}_2$ we get

$$\begin{aligned}
\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y} = &\left(\frac{1}{n_1} - \frac{1}{n_1+n_2}\right) n_1^2 \bar{Y}_1^2 + \left(\frac{1}{n_2} - \frac{1}{n_1+n_2}\right) n_2^2 \bar{Y}_2^2 - \\
&\frac{2}{n_1+n_2} n_1 n_2 \bar{Y}_1 \bar{Y}_2 \\
=&n_1 \bar{Y}_1^2 + n_2 \bar{Y}_2^2 - \frac{1}{n_1+n_2}(n_1 \bar{Y}_1 + n_2 \bar{Y}_2)^2 \\
=&\frac{1}{n_1+n_2}(n_1(n_1+n_2)\bar{Y}_1^2 + n_2(n_1+n_2)\bar{Y}_2^2 - (n_1 \bar{Y}_1 + n_2 \bar{Y}_2)^2) \quad (9\text{-}253) \\
=&\frac{1}{n_1+n_2}(n_1 n_2 \bar{Y}_1^2 + n_2 n_1 \bar{Y}_2^2 - 2 n_1 n_2 \bar{Y}_1 \bar{Y}_2) \\
=&\frac{n_1 n_2}{n_1+n_2}(\bar{Y}_1 - \bar{Y}_2)^2 \\
=&\frac{(\bar{Y}_1 - \bar{Y}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}
\end{aligned}$$

which is (9-248).

■

Of course the comments on multiple testing still apply and the significance level might be adjusted accordingly. As a more general remark it also imply that when using Type III partitioning the risk of over parametrization should always be taken into account, in particular if a high number of hypothesis are tested during model development.

9.12 Two-way ANOVA as a LM

The two-way anova model can be written as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{9-254}$$

as we will see below the model is easily written as an LM, we start by showing the equivalence between a specific two-way anova and the paired t-test.

9.12.1 Paired t-test as an LM

The paired t-test can be written as a two-way anova model as

$$Y_{1j} = \mu_1 + \beta_j + \epsilon_{1j}; \quad \epsilon_{1j} \sim N(0, \sigma^2) \tag{9-255}$$

if the observation is from group 1 and

$$Y_{2j} = \mu_2 + \beta_j + \epsilon_{2j}; \quad \epsilon_{2j} \sim N(0, \sigma^2) \tag{9-256}$$

if the observation is from group 2. In the paired t-test set up we consider

$$\begin{aligned} D_j = Y_{1j} - Y_{2j} =& \mu_1 - \mu_2 + \epsilon_{1j} - \epsilon_{2j} \\ =& \mu_D + \tilde{\epsilon}_j; \quad \tilde{\epsilon}_j \sim N(0, \tilde{\sigma}^2), \end{aligned} \tag{9-257}$$

note that the assumption of equal variance is not formally a part of the paired t-test as the method only "see" the difference ($\tilde{\epsilon}_j$), actually $\epsilon_{1,j}$ and $\epsilon_{2,j}$ does not even have to be independent or normally distributed.

The system described in (9-255)-(9-256) is over parameterized (we cannot identify μ_1, μ_2 and $\beta_1, ..., \beta_n$), as discussed in the previous section there are a number of ways to solve this, one is to parameterized by

$$\begin{aligned} Y_{1,j} =& \frac{1}{2}\mu_D + \beta_j + \epsilon_{1,j} \\ Y_{2,j} =& -\frac{1}{2}\mu_D + \beta_j + \epsilon_{2,j}, \end{aligned} \tag{9-258}$$

with the design matrix

$$\boldsymbol{X} = \begin{bmatrix} \frac{1}{2}\mathbf{1} & \boldsymbol{I} \\ -\frac{1}{2}\mathbf{1} & \boldsymbol{I} \end{bmatrix}, \tag{9-259}$$

which is an orthogonal parametrization (see Exercise 18). The parameters are $\boldsymbol{\beta} = [\mu_D, \beta_1, ..., \beta_n]^T$. The estimator for $\hat{\mu}_D$ is the average difference is (see Exercise 18)

$$\hat{\mu}_D = \bar{Y}_1 - \bar{Y}_2 = \bar{D}, \tag{9-260}$$

and we known from Chapter 2 that the usual paired t-test is

$$t_{obs} = \frac{\bar{D}}{s_D/\sqrt{n}} \sim t(n-1), \tag{9-261}$$

hence equivalence between the two-way anova setup and the paired t-test correspond to $s_D^2/n = SSE/df_{SSE}(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11}$, it can be shown that (Exercise 18)

$$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \frac{1}{2}\sum_{i=1}^{n}(D_i - \bar{D})^2 \tag{9-262}$$

and that $(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii} = \frac{2}{n}$, and further $df_{SSE} = n - 1$. Hence

$$\frac{SSE}{df_{SSE}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{s_D^2}{n}, \tag{9-263}$$

showing the equivalence. An added benefit of the anova approach is that the effect of "subjects" ($\hat{\beta}_j$) is estimated as part of the procedure.

9.12.2 Two-way anova as an LM

In the general case of two way anova a direct approach for the design matrix could be

$$\boldsymbol{X}_0 = \begin{bmatrix} \mathbf{1}_b & \mathbf{0}_b & \dots & \mathbf{0}_b & \boldsymbol{I} \\ \mathbf{0}_b & \mathbf{1}_b & \dots & \mathbf{0}_b & \boldsymbol{I} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0}_b & \mathbf{0}_b & \dots & \mathbf{1}_b & \boldsymbol{I} \end{bmatrix}, \tag{9-264}$$

where b is the number of "blocks" and the number of treatments is k. The model is over-parameterized ($rank(\boldsymbol{X})$ is $b+k-1$ not $b+k$), as e.g. the first column can be written as the sum of the last b columns minus column 2 through k. Hence one column should be removed, e.g. by replacing $\boldsymbol{I}$ with

$$\tilde{\boldsymbol{I}} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{I}_{b-1} \end{bmatrix}. \tag{9-265}$$

In this case the first k parameters are the mean value for the k treatments when observing from block 1, and the remaining $b-1$ parameters describe the derivation from those values due to different block effects.

Hence one encoding of the two-way anova is

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_b & \mathbf{0}_b & \dots & \mathbf{0}_b & \tilde{\boldsymbol{I}} \\ \mathbf{0}_b & \mathbf{1}_b & \dots & \mathbf{0}_b & \tilde{\boldsymbol{I}} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0}_b & \mathbf{0}_b & \dots & \mathbf{1}_b & \tilde{\boldsymbol{I}} \end{bmatrix}. \tag{9-266}$$

A more common encoding is

$$\boldsymbol{X}_1 = \begin{bmatrix} \mathbf{1}_b & \mathbf{0}_b & \dots & \mathbf{0}_b & \tilde{\boldsymbol{I}} \\ \mathbf{1}_b & \mathbf{1}_b & \dots & \mathbf{0}_b & \tilde{\boldsymbol{I}} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{1}_b & \mathbf{0}_b & \dots & \mathbf{1}_b & \tilde{\boldsymbol{I}} \end{bmatrix}, \tag{9-267}$$

in this case the first parameter is the expected value for an observation in treatment 1 and block 1. And the transformation between the two formulation can be done by the matrix

$$\boldsymbol{T} = \begin{bmatrix} 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{b-1} & \boldsymbol{I}_{b-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{b-1} \end{bmatrix}. \tag{9-268}$$

Finally we considered the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{9-269}$$

in Chapter 8, and the implicit constraints are

$$\sum_{i=1}^{k} \alpha_i = 0; \quad \sum_{j=1}^{l} \beta_j = 0, \tag{9-270}$$

and with the same arguments as in the one-way ANOVA model the design matrix can be written as (see Exercise 19)

$$\boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1}_l & \mathbf{1}_l & \dots & \mathbf{0}_l & \boldsymbol{B} \\ \mathbf{1}_l & \mathbf{0}_l & \dots & \mathbf{1}_l & \boldsymbol{B} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{1}_l & -\mathbf{1}_l & \dots & -\mathbf{1}_l & \boldsymbol{B} \end{bmatrix}, \tag{9-271}$$

with

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{bmatrix}. \tag{9-272}$$

The transformation between the encoding is a bit more complicated in the case. Regardless of the particular parametrization, then testing in the two-way anova model in situations as describe above is the same regardless of the used separation of variation (type I or II). In order to be able to make the precise statement we need the concept of balanced design.

|||| Definition 9.48 Balanced design

A design matrix is said to be balanced if the number of observations for any given combination of factors is the same fixed number.

In a two-way ANOVA there are 2 factors each on a number of levels, further in the development we have presented here it is assumed that the number of observations is exactly one for each combination. Of course the definition hint to the fact that we could have more than one, but then the design matrix is only balanced if there are exactly the same number of observations for each combination. We can now make the precise statement about equivalence of the tests.

Theorem 9.49 Equivalence between Type I and Type III

For two-way ANOVA with balanced design, the Type I and Type III partitioning of variation is equivalent.

In the proof below some steps are skipped, as these are much better done using Kronecker products, and the point is mostly which matrices that should be compared.

Proof

We consider the following design matrices

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{0}_b & \cdots & \boldsymbol{0}_b & \boldsymbol{I} \\ \boldsymbol{1}_b & & \boldsymbol{0}_b & \boldsymbol{I} \\ & \ddots & \vdots & \vdots \\ \boldsymbol{0}_b & \cdots & \boldsymbol{1}_b & \boldsymbol{I} \end{bmatrix}; \quad \boldsymbol{X}_{Tr} = \begin{bmatrix} \boldsymbol{1}_b & \cdots & \boldsymbol{0}_b \\ \vdots & \ddots & \vdots \\ \boldsymbol{0}_b & \cdots & \boldsymbol{1}_b \end{bmatrix}; \quad \boldsymbol{X}_{Bl} = \begin{bmatrix} \boldsymbol{I} \\ \vdots \\ \boldsymbol{I} \end{bmatrix}; \quad \boldsymbol{X}_0 = \boldsymbol{1}, \tag{9-273}$$

and projection matrices based on each of these design matrices. The Type I partitioning would be

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}_{Tr} - \boldsymbol{H}_0)\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_{Tr})\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{9-274}$$

or

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}_{Bl} - \boldsymbol{H}_0)\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_{Bl})\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{9-275}$$

depending on which effect (treatment or "block") that entered the model last. Hence we are done if we can show that $\boldsymbol{H} - \boldsymbol{H}_{Bl} = \boldsymbol{H}_{Tr} - \boldsymbol{H}_0$ and $\boldsymbol{H} - \boldsymbol{H}_{Tr} = \boldsymbol{H}_{Bl} - \boldsymbol{H}_0$. By direct matrix multiplications it can be shown that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} b\boldsymbol{I} & \boldsymbol{E}_{k-1,b} \\ \boldsymbol{E}_{b,k-1} & k\boldsymbol{I} \end{bmatrix} \tag{9-276}$$

and it is easy to check that (using that $\boldsymbol{E}_{k-1,b}\boldsymbol{E}_{b,k-1} = b\boldsymbol{E}_{k-1,k-1}$)

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \begin{bmatrix} \frac{1}{b}(\boldsymbol{I} + \boldsymbol{E}_{k-1,k-1}) & -\frac{1}{b}\boldsymbol{E}_{k-1,b} \\ -\frac{1}{b}\boldsymbol{E}_{b,k-1} & \frac{1}{k}\left(\boldsymbol{I} + \frac{k-1}{b}\boldsymbol{E}_{bb}\right) \end{bmatrix} \tag{9-277}$$

which imply that (and here we leave out some of the details, but see Exercise 20) the projection matrix can be written as

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \cdots & \boldsymbol{H}_{1k} \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}_{k1} & \cdots & \boldsymbol{H}_{kk} \end{bmatrix}, \tag{9-278}$$

with

$$\boldsymbol{H}_{ii} = \frac{1}{k}\boldsymbol{I} + \frac{k-1}{kb}\boldsymbol{E}_{bb}; \quad \text{and} \quad \boldsymbol{H}_{ij} = \frac{1}{k}\boldsymbol{I} - \frac{1}{kb}\boldsymbol{E}_{bb}, \quad \text{for } i \neq j. \tag{9-279}$$

Now since $\boldsymbol{X}_{Tr}^T\boldsymbol{X}_{Tr} = b\boldsymbol{I}$ and $\boldsymbol{X}_{Bl}^T\boldsymbol{X}_{Bl} = k\boldsymbol{I}$, we can write the corresponding elements of the other projection matrices as

$$\begin{aligned} \boldsymbol{H}_{Tr,ii} &= \frac{1}{b}\boldsymbol{E}_{bb}; \quad \boldsymbol{H}_{Tr,ij} = \boldsymbol{0}, \quad \text{for } i \neq j. \\ \boldsymbol{H}_{Bl,ij} &= \frac{1}{k}\boldsymbol{I}; \quad \text{for all } (i,j) \\ \boldsymbol{H}_{0,ij} &= \frac{1}{bk}\boldsymbol{E}_{bb}; \quad \text{for all } (i,j) \end{aligned} \tag{9-280}$$

and hence

$$\begin{aligned} \boldsymbol{H}_{ii} - \boldsymbol{H}_{Tr,ii} &= \frac{1}{k}\boldsymbol{I} + \frac{k-1}{kb}\boldsymbol{E}_{bb} - \frac{1}{b}\boldsymbol{E}_{bb} = \frac{1}{k}\boldsymbol{I} - \frac{1}{kb}\boldsymbol{E}_{bb} = \boldsymbol{H}_{Bl,ii} - \boldsymbol{H}_{0,ii} \\ \boldsymbol{H}_{ij} - \boldsymbol{H}_{Tr,ij} &= \frac{1}{k}\boldsymbol{I} - \frac{1}{kb}\boldsymbol{E}_{bb} - \boldsymbol{0} = \frac{1}{k}\boldsymbol{I} - \frac{1}{kb}\boldsymbol{E}_{bb} = \boldsymbol{H}_{Bl,ij} - \boldsymbol{H}_{0,ij} \end{aligned} \tag{9-281}$$

showing that $\boldsymbol{H} - \boldsymbol{H}_{Tr} = \boldsymbol{H}_{Bl} - \boldsymbol{H}_0$, and further

$$\begin{aligned} \boldsymbol{H}_{ii} - \boldsymbol{H}_{Bl,ii} &= \frac{1}{k}\boldsymbol{I} + \frac{k-1}{kb}\boldsymbol{E}_{bb} - \frac{1}{k}\boldsymbol{I} = \frac{k-1}{kb}\boldsymbol{E}_{bb} = \boldsymbol{H}_{Tr,ii} - \boldsymbol{H}_{0,ii} \\ \boldsymbol{H}_{ij} - \boldsymbol{H}_{Bl,ij} &= \frac{1}{k}\boldsymbol{I} - \frac{1}{kb}\boldsymbol{E}_{bb} - \frac{1}{k}\boldsymbol{I} = -\frac{1}{kb}\boldsymbol{E}_{bb} = \boldsymbol{H}_{Tr,ij} - \boldsymbol{H}_{0,ij} \end{aligned} \tag{9-282}$$

showing that $\boldsymbol{H} - \boldsymbol{H}_{Bl} = \boldsymbol{H}_{Tr} - \boldsymbol{H}_0$ and completing the proof.

■

Theorem 9.49 show that in the case of two-way ANOVA with a balanced design, we do not have to worry about differences in how we test. This is a unique property of balanced design and it is usually not present in regression type models. Further it is not unusual that there are missing data in a factorial experiment, and then the two test strategies will differ. In general the Type III partitioning of variation is simpler to understand, but of course observing mass significance (and adjust significance levels), if many tests are conducted.

9.13 Further generalizations

Clearly one can imagine endless generalizations of the general linear model, here we have selected a few that we will briefly cover without going into many details of the modeling aspects. Instead focusing on the general model set up in each of the cases.

9.13.1 Multiple factors, interactions and regression

The one- and two way anova models that we have covered so far can be generalized to more than two factors in a fairly obvious way, so that we have measurements a associated treatments on a number of different levels, e.g. the yield from of some crop depending on the field (`field`), fertilized (`fer`), and pesticides (`pes`), a simple model would be

$$Y_i = \beta_0 + \beta_1(\texttt{field}_i) + \beta_2(\texttt{fer}_i) + \beta_2(\texttt{pes}_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2),$$

where each of the parameters (e.g. β_1) are actually vectors (e.g. with four fields then $\beta_1 \in \mathbb{R}^3$). In such a setup we can have more than one observation for each combination of field, pesticide, and fertilizer. Clearly we can have an arbitrary number of factors

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_p(\texttt{fac}_{ji}) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

In these cases the design matrix may be parameterized by zeros and ones. All comments on the design matrix that we have covered in the previous also hold in this situation. Further interaction effects are often considered, corresponding to the model

$$\begin{aligned} Y_i =& \beta_0 + \beta_1(\texttt{field}_i) + \beta_2(\texttt{fer}_i) + \beta_2(\texttt{pes}_i) + \beta_4(\texttt{field}_i, \texttt{fer}_i) + \\ & \beta_5(\texttt{field}_i, \texttt{pes}_i) + \beta_5(\texttt{fer}_i, \texttt{pes}_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2), \end{aligned}$$

this is referred to as a two-way interaction model, and of course we could imaging three or four way interaction models. The number of parameters grow quite fast and considerations on that should be taken. Again the comment on test still apply, though higher order interactions are usually tested before main effects (and lower order interactions), this is in essence what is referred to as Type II partitioning of variation.

Regression analysis and factor analysis can also easily be implemented as an LM, with one factor (on p levels) the model would be

$$Y_i = \beta_0(\texttt{fac}_i) + \beta_1(\texttt{fac}_i)x_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2) \tag{9-283}$$

essentially implying that the slope is different in different groups, and it can of course be combined with multiple factor and multiple regressors. With increasing complexity of the models the choices of model reduction strategies also become more important and some thoughts have to be out into that.

9.13.2 Orthogonal parametrization: PCR

We have previously seen that multicollinarity should be dealt with if it occur. In Chapter 6 we discussed very simple way to deal with it, in this section we will briefly explain one way of removing multicollinarity all together, the price to pay is that the interpretation of the parameters become much more difficult. First note that the parameters are orthogonal (independent) if

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{\Lambda}, \tag{9-284}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix.

Assume that we have a design matrix

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \dots & \boldsymbol{x}_p \end{bmatrix}, \tag{9-285}$$

the first column is independent from the remaining columns if $\bar{x}_i = 0$ for all i, to see this consider

$$(\boldsymbol{X}^T\boldsymbol{X})_{1,i} = \mathbf{1}^T\boldsymbol{x}_i = \sum_j x_{ij} = n\bar{x}_i. \tag{9-286}$$

Hence defining the transformation matrix

$$\boldsymbol{T} = \begin{bmatrix} 1 & -\bar{x}_1 & -\bar{x}_2 & \dots & -\bar{x}_p \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 1 \end{bmatrix}, \tag{9-287}$$

we have

$$\boldsymbol{X}\boldsymbol{T} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X}_c \end{bmatrix}, \tag{9-288}$$

where

$$(\boldsymbol{X}\boldsymbol{T})^T\boldsymbol{X}\boldsymbol{T} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_c^T\boldsymbol{X}_c \end{bmatrix}, \tag{9-289}$$

if we denote the collection of eigen-vectors of $\boldsymbol{X}_c^T\boldsymbol{X}_c$ by $\boldsymbol{W}$, then by definition

$$\boldsymbol{W}^{-1}\boldsymbol{X}_c^T\boldsymbol{X}_c\boldsymbol{W} = \boldsymbol{\Lambda}, \tag{9-290}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with diagonal elements equal the eigenvalues of $\boldsymbol{X}_c^T\boldsymbol{X}_c$, further as $\boldsymbol{X}_c^T\boldsymbol{X}_c$ is symmetric, we also have

$$\boldsymbol{W}^{-1} = \boldsymbol{W}^T \tag{9-291}$$

by Lemma 9.3. And hence

$$(\boldsymbol{X}_c\boldsymbol{W})^T\boldsymbol{X}_c\boldsymbol{W} = \boldsymbol{\Lambda}, \tag{9-292}$$

and hence setting

$$\boldsymbol{T}_w = \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W} \end{bmatrix} \tag{9-293}$$

then with

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{T}\boldsymbol{T}_w \tag{9-294}$$

the parameters are orthogonal, i.e. $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ is a diagonal matrix. The price to pay is that each parameter refer to linear combinations of regressors, and hence difficult to interpret.

9.13.3 Estimation correlation structures

The general linear model can be written as

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{9-295}$$

so far we have considered cases where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, but is is natural to ask what happens if $\boldsymbol{\Sigma} \neq \sigma^2\boldsymbol{I}$, or rather what happens if the observations are not independent?

Actually the first question might be why the observations would not be independent. Here the answer would be in data collection procedure, if data is collected as a time series it is natural to assume serial dependence. This would lead to time series models, and we will not go into any details here but just mention the simplest model

$$\epsilon_i = \phi\epsilon_{i-1} + u_i; \quad u_i \sim N(0, \sigma^2), \tag{9-296}$$

with $|\phi| < 1$, such a model is called an autoregresive model of order 1 (AR(1)), and the resulting structure of the covariance matrix is

$$\boldsymbol{\Sigma}_{ij} = \frac{\sigma^2\phi^{|i-j|}}{1-\phi^2}, \tag{9-297}$$

hence an exponential decay of the covariance as a function of distance in time ($|i-j|$). Here we have one extra parameter (ϕ) that needs to be estimated.

Another group of models that lead to non-diagonal covariance structures is the linear mixed effect model, where we have multiple observations from each subject, and subjects are treated as random variables, in its simplest form the model is

$$Y_{ij} = \beta_0 + u_i + \epsilon_{ij}; \quad u_i \sim N(0, \sigma_u^2), \quad \epsilon_{ij} \sim N(0, \sigma^2). \tag{9-298}$$

With both u_i and ϵ_{ij} iid and independent of each other. This structure lead to a block diagonal structure where observations from different subjects have covariance zero, while different observations from the same subject have covariance σ_u^2 and the variance of the observations is $\sigma^2 + \sigma_u^2$. Again we get an extra parameter (σ_u^2) to describe the covariance structure.

For estimating parameters in general covariance structures we will need more general objective functions than the RSS, namely yhe so-called likelihood function. The models considered in this section can be written as

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma}(\boldsymbol{\psi})) \tag{9-299}$$

where $\boldsymbol{\psi}$ is the parameters of the covariance function (in our examples $\boldsymbol{\psi} = [\sigma^2, \phi]$ or $\boldsymbol{\psi} = [\sigma^2, \sigma_u^2]$).

The idea of likelihood estimation is to maximize the probability density function wrt. the parameters, $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{\psi}]$, formally with $L(\boldsymbol{\theta}) = f(\boldsymbol{y}; \boldsymbol{\theta})$, the likelihood estimate is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} L(\boldsymbol{\theta}), \tag{9-300}$$

usually the log-likelihood function $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ have better numerical properties, and therefore the optimization problem is usually formulated as

$$\hat{\boldsymbol{\theta}} = \texttt{arg}\max_{\theta} l(\boldsymbol{\theta}). \tag{9-301}$$

In the cases we have considered here the probability density function can be written as

$$f(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})} \tag{9-302}$$

which result in the log-likelihood

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{9-303}$$

where additive constants (related to 2π) have been omitted. Notice that in the case where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$ then

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS(\boldsymbol{\beta}) \end{aligned} \tag{9-304}$$

and hence in that case the estimation of $\boldsymbol{\beta}$ is not affected by $\boldsymbol{\psi}$, and maximizing $l(\sigma^2, \boldsymbol{\beta})$, wrt. $\boldsymbol{\beta}$ is the same a minimizing RSS. In the general case the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ however have to the done jointly, and in some cases specific method are available for specific models (e.g. conditional 1-step ahead distributions for time series) while in other cases one simply have to optimize the log-likelihood directly.

9.14 Exercises

Exercise 9.1 Proof of Theorem 9.14

a) Prove Theorem 9.14, using the definition in equation (9-38).

Exercise 9.2 Independence and correlation

a) Simulate Y_1, X and Y_2 using the setting in Example 9.15.

b) Check that both Y_1 and Y_2 are normal, and plot Y_2 as a function of Y_1.

c) Calculate the correlation between Y_1 and Y_2 and plot Y_2 as a function of Y_1 and comment on the results

Exercise 9.3 Proff of Eq. (9-50)

a) Prove that rowsums of $\boldsymbol{A}$ in (9-49) is equal zero, i.e. that $\boldsymbol{A1} = \mathbf{0}$

b) Prove Eq. (9-48)

c) Prove Eq. (9-50).

Exercise 9.4 Proff of Corollary 9.18

a) Show that when $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then, $\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{V}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, with $\boldsymbol{V}$, and $\boldsymbol{\Lambda}$ as in Lemma 9.3.

b) Prove Corollary 9.18.

Exercise 9.5 Projection matrix

a) Use exercise 3 to show that $\boldsymbol{A}$ in (9-49) is an orthogonal projection matrix.

Exercise 9.6 Proof of Lemma 9.22

a) Use Lemma 9.3, property 1 of Lemma 9.22 and Theorem 9.5 to prove property 2 of Lemma 9.22.

Exercise 9.7 Correlation

a) With $\boldsymbol{r}$ as in (9-51) what is the correlation between r_i and r_j?

Exercise 9.8 Lag-1 autocorrelation

Consider the random variables $\epsilon_i \sim N(0, \sigma^2)$, iid. and $t = \{1, ..., n\}$. Now consider the correlation estimate,

$$\hat{\rho}_\epsilon(1) = \frac{\sum_{t=1}^{n-1} \epsilon_t \epsilon_{t+1}}{\sum_{t=1}^{n} \epsilon_t^2} = \frac{C}{Q}, \tag{9-305}$$

the idea of the questions below is that show that $\hat{\rho}_\epsilon(1) \approx N(0, 1/n)$ by showing that $V[\hat{\rho}_\epsilon(1)] \approx 1/n$. $\hat{\rho}_\epsilon(1)$ is simpler than $\hat{\rho}(1)$ in (9-129), but for n large the behavior is similar.

a) Show that $E[C] = 0$, $E[Q] = n\sigma^2$, $V[C] = (n-1)\sigma^4$, $V[Q] = 2n\sigma^4$, and $Cov[C, Q] = 0$.

b) Use the result from question a) and non-linear error propagation to show that $V[\hat{\rho}_\epsilon(1)] \approx 1/n$, for n large.

|||| Exercise 9.9 Orthogonal projections

a) With $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ as in (9-70), show that $Cov[\boldsymbol{H}_1\boldsymbol{Z}, \boldsymbol{H}_2\boldsymbol{Z}] = \boldsymbol{0}$. Hint: Use Theorem 9.10 and Exercise 5.

|||| Exercise 9.10 Proof of Corollary 9.29

In this exercise we will prove Corollary 9.29 by a series of sub questions.

a) Show that if $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ then

$$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y} \sim \chi^2(n-2). \tag{9-306}$$

Independently of the value of $\boldsymbol{\beta}$

b) Show that if $\boldsymbol{Y} \sim N(\boldsymbol{1}\mu, \sigma^2\boldsymbol{I})$ then

$$\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y} \sim \chi^2(1). \tag{9-307}$$

independently of the value of μ, you may use the the formulation in (9-137) to calculate $\boldsymbol{H}_1$, or simply use the fact that $\boldsymbol{1}\boldsymbol{H}_1 = \boldsymbol{1}^T$ (see Exercise 11).

c) Show that if $\boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ then

$$\boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} \sim \chi^2(1). \tag{9-308}$$

Exercise 9.11 t-test Orthogonal projections

a) Show that the projection matrices in rhs of (9-141) are orthogonal i.e. $\boldsymbol{H}_0(\boldsymbol{H}_1 - \boldsymbol{H}_0) = \boldsymbol{0}$, $\boldsymbol{H}_0(\boldsymbol{I} - \boldsymbol{H}_1) = \boldsymbol{0}$ and $(\boldsymbol{H}_1 - \boldsymbol{H}_0)(\boldsymbol{I} - \boldsymbol{H}_1) = \boldsymbol{0}$. Hint: you may start by showing that $\boldsymbol{X}_0^T\boldsymbol{H}_1 = \boldsymbol{X}_0^T$. You may use the parametrization (9-137).

b) Use the result to show that

$$\begin{aligned} Cov[\boldsymbol{H}_0\boldsymbol{Y}, (\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}] &= \boldsymbol{0} \\ Cov[\boldsymbol{H}_0\boldsymbol{Y}, (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}] &= \boldsymbol{0} \\ Cov[(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}, (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}] &= \boldsymbol{0} \end{aligned} \tag{9-309}$$

and hence that the projected vectors are independent. Also what is the interpretation in trems of fitted values?

Exercise 9.12 t-test $\hat{\sigma}^2$ central

a) Show that $\hat{\sigma}^2$ (in Equation (9-147)) is a central estimator for the variance in the LM, and find $V[\hat{\sigma}^2]$.

Exercise 9.13 t-test Central estmators under Null-hypothesis

Consider the projection matrices for the two sample t-test (equation (9-141)), consider two groups $Y_{1,i} \sim N(\mu_1, \sigma^2)$ and iid., $i = \{1, 2, ..., n_1\}$ and $Y_{2,j} \sim N(\mu_2, \sigma^2)$ and iid., $j = \{1, 2, ..., n_2\}$. Define $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T]^T = [Y_{1,1}, ..., Y_{1,n_1}, Y_{2,1}, ..., Y_{1,n_2}]^T$ and

a) Show that

$$\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y} = \frac{n_1 n_2}{n_1 + n_2}(\bar{Y}_1 - \bar{Y}_2)^2 \tag{9-310}$$

b) Show that $E[\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}] = \frac{n_1 n_2}{n_1 + n_2}(\mu_1 - \mu_2)^2 + \sigma^2$

- Under the assumption $\mu_1 = \mu_2 = \mu$ conclude that $\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}$ is a central estimator for σ^2, find the variance of this estimator, and compare with the estimator (9-147).

|||| Exercise 9.14 Nested projections

Let $\boldsymbol{X}_i$ be as in (9-150), i.e.

$$\boldsymbol{X}_i = [\boldsymbol{X}_{i-1} \quad \tilde{\boldsymbol{X}}_i] \tag{9-311}$$

and condider the projection matrices based on $\boldsymbol{X}_{i-1} \in \mathbb{R}^{n\times p_{i-1}}$, and $\boldsymbol{X}_i \in \mathbb{R}^{n\times(p_i+q_i)}$ $(q_i > 0)$

$$\begin{aligned} \boldsymbol{H}_{i-1} &= \boldsymbol{X}_{i-1}(\boldsymbol{X}_{i-1}^T\boldsymbol{X}_{i-1})^{-1}\boldsymbol{X}_{i-1}^T \\ \boldsymbol{H}_i &= \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T \end{aligned} \tag{9-312}$$

a) Show that $\boldsymbol{X}_i^T\boldsymbol{H}_i = \boldsymbol{X}_i^T$.

b) Set $\boldsymbol{A} = (\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}$, with

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \tag{9-313}$$

with $\boldsymbol{A}_{11} \in \mathbb{R}^{p_i\times p_i}$, $\boldsymbol{A}_{12} = \boldsymbol{A}_{21}^T \in \mathbb{R}^{p_i\times q_i}$, and $\boldsymbol{A}_{22} \in \mathbb{R}^{q_i\times q_i}$, show that $\boldsymbol{A}_{kl}$ solve the equations

$$\begin{aligned} \boldsymbol{X}_{i-1}^T\boldsymbol{X}_{i-1}\boldsymbol{A}_{11} + \boldsymbol{X}_{i-1}^T\tilde{\boldsymbol{X}}_i\boldsymbol{A}_{21} &= \boldsymbol{I} \\ \boldsymbol{X}_{i-1}^T\boldsymbol{X}_{i-1}\boldsymbol{A}_{12} + \boldsymbol{X}_{i-1}^T\tilde{\boldsymbol{X}}_i\boldsymbol{A}_{22} &= \boldsymbol{0} \\ \tilde{\boldsymbol{X}}_i^T\boldsymbol{X}_{i-1}\boldsymbol{A}_{11} + \tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i\boldsymbol{A}_{21} &= \boldsymbol{0} \\ \tilde{\boldsymbol{X}}_i^T\boldsymbol{X}_i\boldsymbol{A}_{12} + \tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i\boldsymbol{A}_{22} &= \boldsymbol{I} \end{aligned} \tag{9-314}$$

c) Use the result above to show that $\boldsymbol{X}_{i-1}^T\boldsymbol{H}_i = \boldsymbol{X}_{i-1}^T$.

Exercise 9.15 t-test parametrization

a) Assuming that $Y_{1,i} \sim N(\mu_1, \sigma^2)$ and $Y_{2,j} \sim N(\mu_2, \sigma^2)$ are iid and $i \in \{1, ..., n_1\}$ and $j \in \{1, ..., n_2\}$ formulate an LM (i.e. parametrize $\boldsymbol{X}$)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}, \tag{9-315}$$

with

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}_{n_1} & a\boldsymbol{1}_{n_1} \\ \boldsymbol{1}_{n_2} & b\boldsymbol{1}_{n_2} \end{bmatrix} \tag{9-316}$$

such that the parametrization is orthogonal and $\hat{\beta}_1 = \frac{1}{n_1+n_2}(n_1\bar{Y}_1 + n_2\bar{Y}_2)$, i.e. the average of all observation, and $\hat{\beta}_2 = \bar{Y}_1 - \bar{Y}_2$.

Exercise 9.16 An ill conditioned problem

a) Using the data from Example 9.43 fit parameters for the full model and parameter for a reduced model and compare the parameters values.

Exercise 9.17 Helmert transformation

a) With reference to (9-243) show that

$$\boldsymbol{T}_{HN} = \begin{bmatrix} 1 & -1/2 & -1/3 & -1/4 & \dots & -1/k \\ 1 & 1/2 & -1/3 & -1/4 & \dots & -1/k \\ 1 & 0 & 2/3 & -1/4 & \dots & -1/k \\ 1 & 0 & 0 & 3/4 & \dots & -1/ \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 1 & 0 & \dots & & 0 & (k-1)/k \end{bmatrix}, \tag{9-317}$$

b) Using $\boldsymbol{X}$ as in (9-234) show that

$$\boldsymbol{X}\boldsymbol{T}_{HN} = \begin{bmatrix} \boldsymbol{1} & -\frac{1}{2}\boldsymbol{1} & -\frac{1}{3}\boldsymbol{1} & \dots & -\frac{1}{k}\boldsymbol{1} \\ \boldsymbol{1} & \frac{1}{2}\boldsymbol{1} & -\frac{1}{3}\boldsymbol{1} & \dots & -\frac{1}{k}\boldsymbol{1} \\ \boldsymbol{1} & \boldsymbol{0} & \frac{2}{3}\boldsymbol{1} & \dots & -\frac{1}{k}\boldsymbol{1} \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{1} & \boldsymbol{0} & \boldsymbol{0} & \dots & \frac{k-1}{k}\boldsymbol{1} \end{bmatrix} \tag{9-318}$$

c) Show that

$$(\boldsymbol{X}_{HN}^T\boldsymbol{X}_{HN})^{-1} = \frac{1}{n}\begin{bmatrix} \frac{1}{k} & 0 & 0 & \dots & 0 \\ 0 & 2 & 0 & \dots & 0 \\ 0 & 0 & \frac{3}{2} & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{k}{k-1} \end{bmatrix} \tag{9-319}$$

d) Use the above to prove (9-244).

|||| Exercise 9.18 Paired t-test

a) Show that the parametrization in (9-259) is an orthogonal parametrization.

b) Find the parameter estimates based on the desing matrix (9-259).

c) Find the projecion matrix corresponding to the desing matrix (9-259).

d) Prove (9-262) (Hint: you may use that $\boldsymbol{Y}_i^T\boldsymbol{E} = n\bar{Y}_i\boldsymbol{1}^T$)

Exercise 9.19 2-way Anova sum-constraint

a) Find a matrix $\boldsymbol{T}$ such that

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{T}\boldsymbol{\beta} \tag{9-320}$$

with $\tilde{\boldsymbol{\beta}} = [\mu, \alpha_1, ...\alpha_k, \beta_1, ..., \beta_l]^T$ and $\boldsymbol{\beta} = [\mu, \alpha_1, ...\alpha_{k-1}, \beta_1, ..., \beta_{l-1}]^T$, such that the constraints (9-270) are fulfilled.

b) Show that the constraints (9-270) can be realized by the desing matrix in (9-271) (hint use the transfromation matrix $\boldsymbol{T}$ and the appropriate (non identifiable) desing matrix corresponding to $\tilde{\boldsymbol{\beta}}$).

Exercise 9.20 Two-way ANOVA*

This porpuse of this exercise is to show equation (9-278), this will rely on Kronecker products, and hence solving the exercise require basic understanding of those.

First note that the (non-unique) design matrices can be written in terms of Kronecker products as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{0}_{b,k-1} & \boldsymbol{I}_b \\ \boldsymbol{I}_{k-1} \otimes \boldsymbol{1}_b & \boldsymbol{1}_{k-1} \otimes \boldsymbol{I}_b \end{bmatrix}; \quad \boldsymbol{X}_{Tr} = \boldsymbol{I}_k \otimes \boldsymbol{1}_b; \quad \boldsymbol{X}_{Bl} = \boldsymbol{1}_k \otimes \boldsymbol{I}_b; \quad \boldsymbol{X}_0 = \boldsymbol{1}_k \otimes \boldsymbol{1}_b$$

and

a) Use the above to write the projection matrices $\boldsymbol{H}_0$, $\boldsymbol{H}_{Tr}$ and $\boldsymbol{H}_{Bl}$ in terms of Kronecker products.

b) Using (9-277) it is staight forward to show that

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{C}_1 + \boldsymbol{C}_2 - \boldsymbol{C}_3 \tag{9-321}$$

with

$$\begin{aligned} \boldsymbol{C}_1 &= \frac{1}{b}\begin{bmatrix} \boldsymbol{I} + \boldsymbol{E}_{k-1,k-1} & -\boldsymbol{E}_{k-1,b} \\ -\boldsymbol{E}_{b,k-1} & \boldsymbol{E}_{bb} \end{bmatrix}; \quad \boldsymbol{C}_2 = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0}_{k-1,b} \\ \boldsymbol{0}_{b,k-1} & \frac{1}{k}\boldsymbol{I} \end{bmatrix}; \\ \boldsymbol{C}_3 &= \begin{bmatrix} \boldsymbol{0}_{k-1,k-1} & \boldsymbol{0}_{k-1,b} \\ \boldsymbol{0}_{b,k-1} & \frac{1}{bk}\boldsymbol{E}_{bb} \end{bmatrix}, \end{aligned} \tag{9-322}$$

show that $\boldsymbol{X}\boldsymbol{C}_1\boldsymbol{X}^T = \boldsymbol{H}_{Tr}$, $\boldsymbol{X}\boldsymbol{C}_2\boldsymbol{X}^T = \boldsymbol{H}_{Bl}$, and $\boldsymbol{X}\boldsymbol{C}_3\boldsymbol{X}^T = \boldsymbol{H}_0$, and hence that $\boldsymbol{H} = \boldsymbol{H}_{Tr} + \boldsymbol{H}_{Bl} - \boldsymbol{H}_0$.

c) Use the above to conclude that $\boldsymbol{H} - \boldsymbol{H}_{Tr} = \boldsymbol{H}_{Bl} - \boldsymbol{H}_0$ and $\boldsymbol{H} - \boldsymbol{H}_{Bl} = \boldsymbol{H}_{Tr} - \boldsymbol{H}_0$.