Lecture 1: Overview

- General linear model introduction[1];
- matrix-vector notation with examples;
- theoretical results needed for data analysis;
- Interactions and identifiability.

[1] The General linear model is a generalisation of the the simple linear model to multiple predictors as covered in 'Statistical Methodology'.

The general linear model

We can generalise the simple linear model by allowing the response variable to depend on multiple predictor variables (plus an additive constant). These extra predictor variables can themselves be transformations of the original predictors.

Examples

1. $\mu_i = \beta_0 + x_i\beta_1, \quad Y_i = \mu_i + \epsilon_i,$
   is a straight line relationship between *y* and predictor variable, *x*.
2. $\mu_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3, \quad Y_i = \mu_i + \epsilon_i$
   is a cubic model of the relationship between *y* and *x*.
3. $\mu_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + \log(x_i z_i)\beta_3, \quad Y_i = \mu_i + \epsilon_i$
   is a model in which *y* depends on predictor variables *x* and *z* and on the log of their product.

The y_i, is treated as an observation on a random variable, Y_i, where $\mathbb{E}(Y_i) \equiv \mu_i$, $\epsilon_i \sim N(0, \sigma^2)$, β_j are model parameters, the values of which are unknown and will need to be estimated using data.

Re-writing in matrix-vector form

Writing out the μ_i equations for all n pairs, (x_i, y_i), results in a large system of linear equations. For example model 1:

$$\begin{array}{rcl}
\mu_1 & = & \beta_0 + x_1\beta_1 \\
\mu_2 & = & \beta_0 + x_2\beta_1 \\
\mu_3 & = & \beta_0 + x_3\beta_1 \\
. & & . \\
. & & . \\
\mu_n & = & \beta_0 + x_n\beta_1
\end{array}$$

which can be re-written in matrix-vector form as

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

i.e. the expected value vector $\boldsymbol{\mu}$ is given by a **model matrix** (also known as a design matrix), **X**, multiplied by a **parameter vector**, $\boldsymbol{\beta}$. Of course since $y_i = \mu_i + \epsilon_i$, we can also write the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Model 2: $\mu_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3, \quad Y_i = \mu_i + \epsilon_i$

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ . \\ . \\ \mu_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ . & . & . & . \\ . & . & . & . \\ 1 & x_n & x_n^2 & x_n^3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Model 3: $\mu_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + \log(x_i z_i)\beta_3, \quad Y_i = \mu_i + \epsilon_i$

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ . \\ . \\ \mu_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & z_1 & \log(x_1 z_1) \\ 1 & x_2 & z_2 & \log(x_2 z_2) \\ 1 & x_3 & z_3 & \log(x_3 z_3) \\ . & . & . & . \\ . & . & . & . \\ 1 & x_n & z_n & \log(x_n z_n) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Models with factor variables

Example: 9 laboratory rats are fed too much, so that they divide into 3 groups of 3: ‘fat’, ‘very fat’ and ‘enormous’. Their blood insulin levels are then measured 10 minutes after being fed a standard amount of sugar.
Question: Relationship between insulin levels and the factor ‘rat size’.

Hence a model could be set up in which the predictor variable is the factor ‘rat size’, with the three levels ‘fat’, ‘very fat’ and ‘enormous’. Writing y_i for the i^{th} insulin level measurement, a suitable model might be:

$$\mathbb{E}(Y_i) \equiv \mu_i = \begin{cases} \beta_0 & \text{if rat is fat} \\ \beta_1 & \text{if rat is very fat} \\ \beta_2 & \text{if rat is enormous} \end{cases}$$

Rewrite in matrix-vector notation using a dummy predictor variable for each level of the factor:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Theory

Here a summary of useful results is given. For the derivation of the results please see Section 7.1 in Wood, SN. 2015. Core Statistics.

You can also read the relevant chapters of the other books on the reading list.

Theory: Summary

The parameters, $\boldsymbol{\beta}$, of the linear model

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{I}_n\sigma^2)$$

can be estimated by least squares, i.e. by minimising the residual sums of squares

$$\mathcal{S} = \|\mathbf{y} - \boldsymbol{\mu}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

w.r.t. $\boldsymbol{\beta}$ and this results in the unbiased estimator

$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ with distribution $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\sigma^2)$.

Further the result $(\hat{\beta}_i - \beta_i)/\hat{\sigma}_{\hat{\beta}_i} \sim t_{n-p}$ can be derived.
Why is this useful?
It enables confidence intervals for β_i to be found, and is the basis for hypothesis tests about individual β_i's (for example $\mathrm{H}_0 : \beta_i = 0$). Note this relies on independence of **y.**

Theory: Summary - F-ratio results

Needed for testing simultaneous equality to zero of several model parameters,
e.g. for making inferences about factor variables and their interactions, since each factor (or interaction) is typically represented by several elements of $\boldsymbol{\beta}$.

Suppose that we want to test

$$\mathrm{H}_0 : \boldsymbol{\mu} = \mathbf{X}_0\boldsymbol{\beta}_0 \text{ against } \mathrm{H}_1 : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}_0$ is 'nested' within $\mathbf{X}$ (meaning that $\mathbf{X}\boldsymbol{\beta}$ can exactly match any $\mathbf{X}_0\boldsymbol{\beta}_0$, but the reverse is not true).

Theory: Summary - F-ratio results

Assume without loss of generality that $\mathbf{X} = [\mathbf{X}_0 : \mathbf{X}_1]$ (always possible to re-parameterize so that this is the case).

Suppose that $\mathbf{X}_0$ and $\mathbf{X}_1$ have $p - q$ and q columns with $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, the corresponding sub-vectors of $\boldsymbol{\beta}$.

Can re-write null hypothesis: $\mathrm{H}_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.

Assuming H_0 is we can form the F-ratio statistic

$$
\begin{aligned}
F &= \frac{(\|\mathbf{y}-\mathbf{X}_0\hat{\boldsymbol{\beta}}_0\|^2 - \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2)/\{\dim(\boldsymbol{\beta})-\dim(\boldsymbol{\beta}_0)\}}{\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2/\{n-\dim(\boldsymbol{\beta})\}} \\
&\sim \frac{\chi^2_q/q}{\chi^2_{n-p}/(n-p)} \sim F_{q,n-p}
\end{aligned}
$$

When do we use this result?
For comparing models using hypothesis testing, where one model is a smaller version of a bigger model.

Theory: Summary - influence matrix (or hat matrix)

The *influence matrix* (or *hat matrix*) of a linear model is the matrix which yields the fitted value vector, $\hat{\boldsymbol{\mu}}$, when post-multiplied by the data vector, $\mathbf{y}$.

$$\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{A}\mathbf{y}$$

i.e. the matrix $\mathbf{A}$ is the influence (hat) matrix such that $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$.

The trace of the influence matrix is the number of (identifiable) parameters in the model. Also, $\mathbf{A}\mathbf{A} = \mathbf{A}$, a property known as *idempotency*.

Why is it useful?

For model checking and for deriving properties of the fitted values, $\hat{\boldsymbol{\mu}}$, and residuals, $\hat{\epsilon}$.

Theory: Summary - fitted values $\hat{\mu}$, and $\hat{\epsilon}$ residuals

The influence matrix is helpful in deriving properties of the fitted values, $\hat{\boldsymbol{\mu}}$, and residuals, $\hat{\boldsymbol{\epsilon}}$.

$\hat{\boldsymbol{\mu}}$ is unbiased, since $\mathbb{E}(\hat{\boldsymbol{\mu}}) = \mathbb{E}(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$.
The covariance matrix of the fitted values is obtained from the fact that $\hat{\boldsymbol{\mu}}$ is a linear transformation of the random vector $\mathbf{y}$, which has covariance matrix $\mathbf{I}_n\sigma^2$, so that

$$\mathbf{V}_{\hat{\boldsymbol{\mu}}} = \mathbf{A}\mathbf{I}_n\mathbf{A}^\mathsf{T}\sigma^2 = \mathbf{A}\sigma^2,$$

by the idempotence (and symmetry) of $\mathbf{A}$. The distribution of $\hat{\boldsymbol{\mu}}$ is degenerate multivariate normal.

Summary - residuals, $\hat{\epsilon}$, and fitted values, $\hat{\mu}$

Similar arguments apply to the residuals.

$$\hat{\epsilon} = (\mathbf{I} - \mathbf{A})\mathbf{y},$$

so

$$\mathbb{E}(\hat{\boldsymbol{\epsilon}}) = \mathbb{E}(\mathbf{y}) - \mathbb{E}(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0}.$$

As in the fitted value case, we have

$$\mathbf{V}_{\hat{\epsilon}} = (\mathbf{I}_n - \mathbf{A})\mathbf{I}_n(\mathbf{I}_n - \mathbf{A})^{\mathsf{T}}\sigma^2 = (\mathbf{I}_n - 2\mathbf{A} + \mathbf{A}\mathbf{A})\,\sigma^2 = (\mathbf{I}_n - \mathbf{A})\,\sigma^2.$$

Again, the distribution of the residuals will be degenerate normal.

Why is are these results useful?

The results for the residuals are useful for model checking, since they allow the residuals to be standardized, so that they should have constant variance, if the model is correct.

Interactions and identifiability

Fat rat example: Want to formulate the model in terms of an overall mean insulin level, α, and deviations from that level, β_j, associated with each level of the factor.

$$\mu_i = \alpha + \beta_j \text{ if rat } i \text{ is rat size level } j$$

where *j* is 0, 1 or 2, corresponding to ‘fat’, ‘very fat’ or ‘enormous’.

Problem?

This model is not ‘identifiable’ because any constant *c* could be added to α and simultaneously subtracted from each element of $\boldsymbol{\beta}$ without changing the value of $\boldsymbol{\mu}$.

Interactions and identifiability

Diagnose identifiability directly from model matrix:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

The columns of the model matrix are not independent - matrix does not have full column rank. Means formulae for finding the least squares parameter estimates breaks down.

Identifiable models have model matrices of full column rank; unidentifiable ones are column rank deficient.

Solution - impose linear constraints on the model parameters, so that the model becomes identifiable.

Typically the constraint is to set one of the unidentifiable parameters to zero. e.g. set α to zero, and recover the original identifiable model.

Multiple factors

Fat rat example:

$$\mu_i = \alpha + \beta_j + \gamma_k \text{ if rat } i \text{ is rat size level } j \text{ and sex } k$$

where k is 0 or 1 for male or female. Written out in full (assuming the rats are MMFFFMFMM):

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \gamma_0 \\ \gamma_1 \end{pmatrix}.$$

The model matrix is of column rank 4, i.e. there are 4 linearly independent columns (col 5 = col 1 - col 6 and col 2 = col 1 - (col 3 + col 4), implying that two constraints are required to make the model identifiable.

Multiple factors

An obvious pair of constraints would be to set $\beta_0 = \gamma_0 = 0$, so that the full model is

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \gamma_1 \end{pmatrix}.$$

In R, identifiability constraints are automatically imposed, and by default these constraints will be that the parameter for the 'first' level of each factor is zero.

Note that 'first' is essentially arbitrary here — the order of levels of a factor is not important. However, if you need to change which level is 'first', in order to make parameters more interpretable, see the `relevel` function.

Interactions

Fat rat example: suppose that how insulin level depends on size varies with sex.

$$\mu_i = \alpha + \beta_j + \gamma_k + \delta_{jk} \text{ if rat } i \text{ is rat size level } j \text{ and sex } k,$$

where the δ_{jk} terms are the parameters for the interaction of rat size and sex. This model is unidentifiable! Also we have more parameters than data.

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \gamma_0 \\ \gamma_1 \\ \delta_{00} \\ \delta_{01} \\ \delta_{10} \\ \delta_{11} \\ \delta_{20} \\ \delta_{21} \end{pmatrix}.$$

Interactions

The default constraint in R is:
$\beta_0 = \gamma_0 = \delta_{00} = \delta_{01} = \delta_{10} = \delta_{20} = 0$.
The resulting model can still produce any fitted value vector that the full model can produce, but all the columns of its model matrix are independent, so that the model is identifiable:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \delta_{11} \\ \delta_{21} \end{pmatrix}.$$

Factor continuous interactions

Fat rat example: Suppose we had simply measured the weight, w_i, of rats, rather than classified them into three groups:

$$\mu_i = \alpha + \gamma_j + \beta w_i + \delta_j w_i \quad \text{if rat } i \text{ is sex } j$$

Insulin levels vary linearly with weight, but in a different way for male and female rats. Here γ_j is the 'main effect' of sex, βw_i the 'main effect' of weight and $\delta_j w_i$ is the weight-sex 'interaction'.

An identifiable version of the model is:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & w_1 & 0 \\ 1 & 0 & w_2 & 0 \\ 1 & 1 & w_3 & w_3 \\ 1 & 1 & w_4 & w_4 \\ 1 & 1 & w_5 & w_5 \\ 1 & 0 & w_6 & 0 \\ 1 & 1 & w_7 & w_7 \\ 1 & 0 & w_8 & 0 \\ 1 & 0 & w_9 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma_1 \\ \beta \\ \delta_1 \end{pmatrix}.$$

Interactions: Summary

- An interaction is generated in a model when the parameter for one predictor variable depends on another predictor variable.
- For example the slope of a regression on weight itself depends on the factor variable sex.
- The model matrix columns associated with an interaction are given by all possible pairwise products of the model matrix columns for the effects that make up the interaction.
- If those effects are identifiable then the interactions are also identifiable.
- This is assuming that the data are sufficent to estimate the effects.

Notation summary

- The *response variable* vector $\mathbf{y}$, has *expected value* vector $\boldsymbol{\mu} = \mathbb{E}(\mathbf{y})$.
- According to the linear model $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is the *model matrix* and $\boldsymbol{\beta}$ the *parameter vector* (or *coefficient vector*).
- Hence the linear model can also be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a *random error* vector.
- Once the model is estimated, the *parameter estimates* are denoted $\hat{\boldsymbol{\beta}}$. This notation will also be used for the parameter *estimator*, the context indicating which is meant.
- The *fitted value* vector is $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.
- The *residual* vector is $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$.