Part II — Statistical Modelling

Based on lectures by A. J. Coca

Notes taken by Joseph Tedds using Dexter Chua's header and Gilles Castel's snippets.

Michaelmas 2019

Introduction to the statistical programming language R
Graphical summaries of data, e.g. histograms. Matrix computations. Writing simple functions. Simulation. [2]

Linear Models
Review of least squares and linear models. Characterisation of estimated coefficients, hypothesis tests and confidence regions. Prediction intervals. Model selection. BoxCox transformation. Leverages, residuals, qq-plots, multiple $\mathbb{R}^2$ and Cooks distances. [5]

Overview of basic inferential techniques
Asymptotic distribution of the maximum likelihood estimator. Approximate confidence regions. Wilks theorem. The delta method. Posterior distributions and credible intervals. [3]

Exponential dispersion families and generalised linear models (glm)
Exponential families and meanvariance relationship. Dispersion parameter and generalised linear models. Canonical link function. Iterative solution of likelihood equations. Regression for binomial data; use of logit and other link functions. Poisson regression models, and their surrogate use for multinomial data. Application to 2- and 3-way contingency tables. Hypothesis tests and model selection, including deviance and Akaikes Information Criterion. Residuals and model checking. [8]

Examples in R
Linear and generalised linear models. Interpretation of models, inference and model selection. [6]

Contents

0 Introduction

This course is unusual in that 8 of the lectures are taken as practicals, with the following guidance.

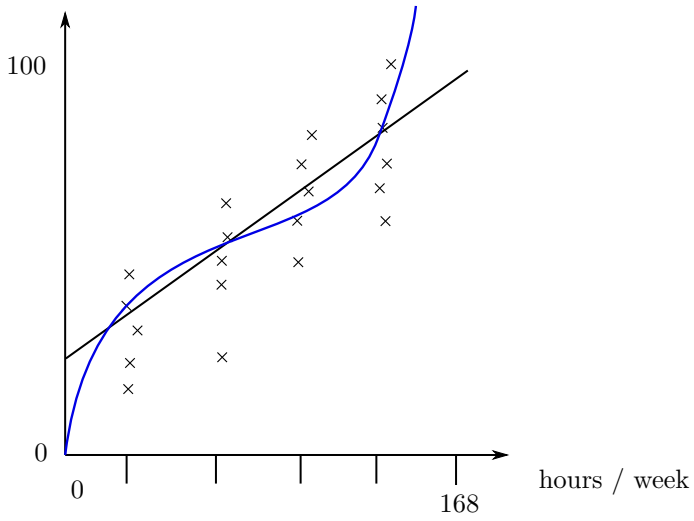
- Ideally use Linux, some things may not work on other operating systems
- Use R and R Studio

We study Data:

- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), i = 1, \ldots, n$, $n =$ sample size.
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ - predictors, covariates , independent or explanatory variables.
- y_i - targets, responses, dependent variables.

Objective: understand the functional relationship relating the y_i's to the $\mathbf{x}_i$'s to develop a regression function.

Example. $x_i =$ number of hours / week student i invests in on statistical modelling, $y_i =$ final grade of student i.



In the next section we model the Y's (note they are now upper-case) as random variables, as $Y_i = f(x_i, \theta) + \varepsilon_i$ independent.

- f is linear in θ
- $\varepsilon_i \approx$ errors / noise with potential causes as measurement errors or our limited understanding of the world.
- $\mathbb{E}[Y_i \mid X_i] = f(x_i, \theta) + \mathbb{E}[\varepsilon_i \mid x_i]$

In the sections thereafter, $\mathbb{E}[Y_i \mid x_i] = f_i(x_i, \theta)$, f_i is not necessarily linear in θ.

Warning. A word of caution, statistical models are not a perfect representation of the world, but they are useful approximations to make decisions.

1 Linear Models

1.1 Ordinary least squares (OLS)

Consider the linear regression model $Y = X\beta + \epsilon$,

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \in \mathbb{R}^p, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

where

(i) $\mathbb{E}[\epsilon_i] = 0$ - does not mean unbiased, but centred

(ii) $\operatorname{var} \epsilon_i = 0$ - homoskedastic

(iii) $\operatorname{cov}(\epsilon_i, \epsilon_j) = 0$ - uncorrelated = linear independence $\neq$ independence

Definition (Design matrix)**.** The design matrix X, unless otherwise stated : $p \leq n$, and $\mathbf{X}$ is full rank i.e. rank $X = p$.

Note, $\theta = \beta$ in the introduction. If we want intercept,

$$X = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}, \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^{p+1}$$

.
If we want higher order terms e.g. quadratic

$$X = \begin{pmatrix} 1 & x_1^T & x_{11}^2 & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^T & x_{n1}^2 & \cdots & x_{np}^2 \end{pmatrix}, \theta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \in \mathbb{R}^{2p+1}.$$

Remember, linear means linear in θ

Definition (Least squares)**.** The *least squares estimator*, $\hat{\beta}$ is defined as

$$\hat{\beta} = \underset{\mathbf{b} \in \mathbb{R}^n}{\operatorname{argmin}} \|Y - X\mathbf{b}\|^2.$$

On the example sheet, we will show that $\hat{\beta} = (X^T X^{-1}) X^T Y$

The fitted values are given by

$$\hat{Y} = X\hat{\beta} = \overbrace{X(X^T X^{-1}) X^T}^{P} Y = PY.$$

We call P the 'hat' matrix and it is an orthogonal projection onto the column space of X.

1.2 Orthogonal projection

Let $V \subseteq \mathbb{R}^n$ be linear. Its orthogonal complement is

$$V^\perp = \{\omega \in \mathbb{R}^n : \omega^T \cdot \mathbf{v} = 0 \ \forall \ \mathbf{v} \in V\}.$$

Fact. (i) $\mathbb{R} \cong V \oplus V^\perp$, so $\forall\ \mathbf{u} \in \mathbb{R}^n\ \exists\ \mathbf{v} \in V, \omega \in V^\perp$ such that $\mathbf{u} = \mathbf{v} + \mathbf{w}$

(ii) $(V^\perp)^\perp = V$

Definition (Orthogonal projection)**.** $\pi \in \mathbb{R}^{n\times n}$ is an *orthogonal projection* onto V if $\pi\mathbf{u} = \mathbf{v}$ whenever $\mathbf{u} = \mathbf{v} + \mathbf{w}, \mathbf{v} \in V, \mathbf{w} \in V^\perp$. π is an orthogonal projection if it is an orthogonal projection onto its column space.

Let π be an orthogonal projection onto V, properties

(i) The column space /range / image/ span of π is V (immediate from the fact above and the definition) so rank $\pi = \dim V$.

(ii) $I - \pi$ is an orthogonal projection onto $V^\perp$. Let $\mathbf{u} = \mathbf{v} + \mathbf{w}, \mathbf{v} \in V, \mathbf{w} \in v^\perp$,

$$(I - \pi)\mathbf{u} = \mathbf{0} + \mathbf{w}.$$

(iii) π is idempotent ($\pi^2 = \pi$) and π is symmetric ($\pi^T = \pi$). The former is by definition, the latter

$$\forall\ \mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n, (\pi\mathbf{u}_1)^T(I - \pi)\mathbf{u}_2 = \begin{cases} 0 \\ \mathbf{u}_1^T((\pi^T - \pi^T\pi)\mathbf{u}_2 \end{cases}.$$

$$\pi^T = \pi^T\pi \iff \pi i = \pi^T\pi = \pi^T.$$

In fact $\pi^2 = \pi = \pi^T$ is an alternative definition for an orthogonal projection. Let $\mathbf{v} \in$ span π.

$$\exists\ \mathbf{u} \in \mathbb{R}^n \pi\mathbf{u} = \mathbf{v} \implies \pi\mathbf{v} = \pi^2\mathbf{u} = \pi\mathbf{u} = \mathbf{v}.$$

Let $\mathbf{v} \in (\text{span } \pi)^\perp$,

$$\pi\mathbf{v} = \pi^T\mathbf{v} = 0.$$

(iv) Orthogonal bases of V and $V^\perp$ are eigenvectors π with eigenvalues 1 or 0. Thus, $\pi = \mathbf{u}D\mathbf{u}^T, \mathbf{u}$ orthonormal ($\mathbf{u}^T\mathbf{u} = \mathbf{u}\mathbf{u}^T = I$), D is diagonal matrix with 1's and 0's

1.3 Analysis of OLS

Recall

$$\hat{\beta} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \|Y - X\mathbf{b}\|^2 = \left(X^TX\right)^{-1} X^TY.$$

and $\hat{Y} = X\hat{\beta} = PY$

$$PX\mathbf{b} = X(X^TX)^{-1}X^TX\mathbf{b} = X\mathbf{b}.$$

If $w \in (\text{span } \pi)^\perp$

$$Pw = X\left(X^TX\right)^{-1}\underbrace{X^T\mathbf{w}}_{\mathbf{0}} = \mathbf{0}.$$

P is an orthogonal projection onto span X, $\hat{Y}$ is a projection of Y onto span X.

The reverse is true: if π is an orthogonal projection onto V, then

$$\pi\mathbf{u} = \arg\min_{\mathbf{v}\in V}\|\mathbf{u}-\mathbf{v}\|^2 \ \forall\ \mathbf{u}\in\mathbb{R}^n.$$

$\hat{\beta}$ is the vector of coefficients of the closest vector in span X to Y as a linear combination of the columns of X Alternative representation of OLS:
Let $X_j = (X_{\cdot j}), X_{-j}$ is X without X_j, $P_{-j} = X_{-j}\left(X_{-j}^T X_{-j}\right)^{-1} X_{-j}^T$

Proposition. Let $X_j^\perp = (I - P_{-j})X_j$. Then

$$\hat{\beta}_j = \frac{\left(X_j^\perp\right)^T Y}{\|X_j^\perp\|^2}.$$

Proof.

$$\begin{aligned}
\left(X_j^\perp\right)^T Y &= X_j^T(I-P_{-j})Y \\
&= X_j^T(I-P_{-j})(PY+(I-P)Y) \\
&= X_j^T(I-P_{-j})PY + X_j^T \underbrace{X_j^T(I-P_{-j})(I-P)Y}_{X_j^T(I-P)Y=0} \\
&= X_j^T(I-P_{-j})PY
\end{aligned}$$

With the reduction of the second term coming from $V_{-k}^\perp \subseteqq V^\perp$

$$\begin{aligned}
\left(X_j^\perp\right)^T &= X_j^T(I-P_{-j})X \\
&= X_j^T \begin{pmatrix} 0 & \cdots & 0 & \underset{j^{\text{th}}\text{ element}}{(I-P_{-j})} & 0\cdots & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & \cdots & 0 & X_j^T(I-P_{-j})^2X_j & 0 & \cdots & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & \cdots & 0 & \|X_j^\perp\|^2 & 0 & \cdots & 0 \end{pmatrix}
\end{aligned}$$

Hence

$$\left(X_j^\perp\right)^T Y = \left(X_j^\perp\right)^T X\hat{\beta} = \begin{pmatrix} 0 & \cdots & 0 & \|X_j^\perp\|^2\beta_j & 0 & \cdots & 0 \end{pmatrix}.$$

□

Recall if $\mathbf{z}_i \in \mathbb{R}^{n_i}, i = 1, 2$

$$\operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2) = \mathbb{E}\left[(\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1])(\mathbf{z}_2 - \mathbb{E}[\mathbf{z}_2]\right].$$

$$(\operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2))_{ij} = \frac{\operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2)_{ij}}{\sqrt{(\operatorname{var}\mathbf{z}_1)_{ii}(\operatorname{var}\mathbf{z}_2)_{jj}}}.$$

$$\forall\ \mathbf{a}_i\in\mathbb{R}^{n_i},\quad \operatorname{cov}(\mathbf{z}_1+\mathbf{a}_1,\mathbf{z}_2+\mathbf{a}_2) = \operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2).$$

and if $A\in\mathbb{R}^{d\times n}, b\in\mathbb{R}^d$

$$\mathbb{E}[\mathbf{b}+A\mathbf{z}_1] = \mathbf{b} + A\mathbb{E}[\mathbf{z}_1].$$

Then,

$$\begin{aligned}\operatorname{var}\hat{\beta}_j &= \frac{1}{\|X_j^{\perp}\|^4}\operatorname{var}\left(X_j^{\perp}\right)^T Y \\ &= \frac{1}{\|X_j^{\perp}\|^4}\operatorname{var}(X_j^{\perp})^T\varepsilon \\ &= \frac{1}{\|X_J^{\perp}\|^4}\left(X_j^{\perp}\right)^T \operatorname{var}\varepsilon\ X_j^{\perp} \\ &= \frac{\sigma^2}{\|X_j^{\perp}\|^2}\end{aligned}$$

.

Now, $\hat{\beta} \in \mathbb{R}^p$, $\hat{\beta}$ is unbiased. Indeed

$$\mathbb{E}_\beta\left[\hat{\beta}\right] = \mathbb{E}\left[(X^TX)^{-1}X^T(X\beta+\epsilon)\right] = \beta.$$

.

$$\begin{aligned}\operatorname{var}(\hat{\beta}) &= \operatorname{var}((X^TX)^{-1}XY) \\ &= \operatorname{var}((X^TX)^{-1}X^T\epsilon) \\ &= (X^TX)^{-1}X^T\underbrace{\operatorname{var}\epsilon}_{\sigma^2 I}X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

.

This is optimal in the following sense.

Theorem (Gauss-Markov)**.** $\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator) i.e. $\forall\ \tilde{\beta}$ linear (in Y) and unbiased estimator

$$\operatorname{var}\tilde{\beta} - \operatorname{var}\hat{\beta} \text{ is positive semidefinite.}$$

Proof. Let $\tilde{\beta} = CY = \hat{\beta} + \underbrace{(C-(X^TX)^{-1}X^T)}_{D}Y$. Note $\mathbb{E}_\beta\left[\tilde{\beta}\right] = \beta\ \forall\ \beta \in \mathbb{R}^p$, so

$$0 = DX\beta \implies DX = 0$$

as this is true $\forall\ \beta$. Then,

$$\operatorname{var}\tilde{\beta} = \operatorname{var}\hat{\beta} + \operatorname{var}DY + 2\operatorname{cov}(\hat{\beta}, DY).$$

$$\operatorname{var}DY = \operatorname{var}D\epsilon = \sigma^2 DD^T,$$

which is positive semi-definite by definition.

.

$$\begin{aligned}\operatorname{cov}(\hat{\beta}, DY) &= \operatorname{cov}((X^TX)^{-1}X^T\epsilon, D\epsilon) \\ &= (X^TX)^{-1}\underbrace{X^TD^t}_{=0}\sigma^2\end{aligned}$$

.

□

Consequently, if x^* is a new observation

Exercise.

$$\mathbb{E}\left[((x^*)^T\hat{\beta} - (x^*)^T\beta)^2\right] \leq \mathbb{E}\left[\left((x^*)^T\tilde{\beta} - (x^*)^T\beta\right)^2\right] \ \forall\ \tilde{\beta} \text{ LUE}.$$

We can also measure the quality of a regression procedure $\tilde{\beta}$ by its mean-squared prediction error:

$$\text{MSPE}(\tilde{\beta}) = \frac{1}{n}\mathbb{E}\left[\|X\tilde{\beta} - X\beta\|^2\right].$$

Proposition.

$$\text{MSPE}(\hat{\beta}) = \sigma^2\frac{p}{n}.$$

Proof. First note that

$$X\hat{\beta} = PY = X\beta + P\epsilon.$$

$$\|X\hat{\beta} - X\beta\|^2 = \|P\epsilon\|^2 = \epsilon^T P\epsilon^T = \text{Tr}(\epsilon^T P\epsilon) = \text{Tr}(P\epsilon^T\epsilon).$$

$$\begin{aligned}\mathbb{E}\left[\text{LHS}\right] &= \text{Tr}(P\mathbb{E}\left[\epsilon^T\epsilon\right] \\ &= \sigma^2\text{Tr}P \\ &= \sigma^2 p\end{aligned}$$

.

□

Lastly, $\hat{\epsilon} = Y - \hat{Y} = (I-P)Y$ is the vector of residuals. This satisfies

$$\begin{aligned}\text{cov}(\hat{\epsilon}, \hat{Y}) &= \text{cov}((I-P)\epsilon, P\epsilon) \\ &= \sigma^2 X(X^TX)^{-1}\underbrace{X^T(I-P)^T}_{=0} = 0\end{aligned}$$

.

So $\hat{\epsilon}$ and $\hat{Y}$ are uncorrelated.

1.4 Normal Errors

1.4.1 Multivariate normal and related distributions

Definition (Multivariate normal)**.** $Z \in \mathbb{R}^d$ is *multivariate normal* if $\forall\ t \in \mathbb{R}^d, t^TZ$ is univariate normal. Thus $\forall\ m \in \mathbb{R}^{k, A\in\mathbb{R}^{k\times d}}, m + AZ$ is (multivariate) normal.

Fact. Normal distributions are uniquely characterised by their mean and variance. So write

$$Z \sim N_d(\mu, \Sigma).$$

if $\mathbb{E}\left[Z\right] = \mu, \text{var}\, Z = \Sigma$

$$\implies m + Az \sim N_k(m + A\mu, A\epsilon A^T).$$

If Σ is invertible , Z has density

$$f(z;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\det\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)\Sigma^{-1}(z-\mu)\right) \ \forall\ z \in \mathbb{R}^d.$$

Proposition. Let Z_1, Z_2 be jointly normal (i.e. (Z_1, Z_2) is multivariate normal).

$$\operatorname{cov}(Z_1, Z_2) = 0 \iff Z_1, Z_2 \text{ are independent } (Z_1 \perp Z_2).$$

Proof. The backward direction is immediate.
In the other direction, let $Z_1' \overset{d}{Z_1}, Z_2' \overset{d}{=} Z_2$. Note

$$\mathbb{E}\left[(Z_1', Z_2')\right] = \left(\mathbb{E}\left[Z_1'\right], \mathbb{E}\left[Z_2'\right]\right) = \mathbb{E}\left[(Z_1, Z_2\right].$$

$$\operatorname{var}(Z_1', Z_2') = \begin{pmatrix} \operatorname{var} Z_1' & \operatorname{cov}(Z_1', Z_2') \\ \operatorname{cov}(Z_1', Z_2') & \operatorname{var} Z_2' \end{pmatrix} = \begin{pmatrix} \operatorname{var} Z_1 & 0 \\ 0 & \operatorname{var} Z_2 \end{pmatrix} = \operatorname{var}(Z_1, Z_2).$$

Also, (Z_1', Z_2') is normal because sums of independent normals is normal. Thus the conclusion follows by the fact above. □

Definition (χ^2 distribution)**.** $X \sim \chi^2_k$ (on k degrees of freedom), if

$$X \overset{d}{=} \sum_{j=1}^{k} Z_j^2, \quad Z_j \overset{\text{iid}}{\sim} N(0,1).$$

Proposition. Let $\pi \in \mathbb{R}^{n\times n}$ be an orthogonal projection with rank k and $\epsilon \sim N_n(0, \sigma^2 I)$. Then

$$\|\pi\epsilon\|^2 \sim \sigma^2 \chi^2_k.$$

Proof. Recall that $\pi = UDU^T$ and noting that $U^T\epsilon \sim N_n(0, \sigma^2 I)$. Then ,

$$\begin{aligned} \|\pi\epsilon\|^2 &= \epsilon^T UDU^T UDU^T \epsilon \\ &= \|Du^T\epsilon\|^2 \\ &\overset{d}{=} \|D\epsilon\|^2 \\ &= \sum_{j:D_{jj}=1} \epsilon_j^2 \\ &\overset{d}{=} \sigma^2 \sum_{j:D_{jj}=1} Z_j^2 \end{aligned}$$

.

□

Definition (t-student distribution)**.** $T \sim t_k$ (on k degrees of freedom) if

$$T \overset{d}{=} \frac{Z}{\sqrt{X/k}}, Z \sim N(0,1), X \sim X_k^2 \text{ independent.}$$

Definition (F distribution)**.** $F \sim F_{k,l}$ (on k, l degrees of freedom) if

$$F \overset{d}{=} \frac{X_1/k}{X_2/l}, X_1 \sim \chi^2_k, X_2 \sim \chi^2_l \text{ independent.}$$

1.4.2 Maximum likelihood estimation

Let $Y \in \mathbb{R}^n$ has density $f(\cdot, \theta), \theta \in \Theta \subseteq \mathbb{R}^p$ unknown (Θ paraneter space unknown). If data y is a realisation of Y, the likelihood function is

$$L(\theta) = f(y : \theta), \theta \in \Theta.$$

Then

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta \in \Theta} L(\theta).$$

It is usually easier to work with the log-likelihood $\ell(\theta) = \log L(\theta)$. Many times we define them up to constants

Example. The t-statistic is given by

$$\frac{\hat{\beta} - \beta}{\sqrt{\text{var}(\hat{\beta})}}.$$

In the second practical we look at

$$\frac{\hat{\beta}_j}{\sqrt{\tilde{\sigma}^2 (X^T X)^{-1}_{jj}}} \sim t_{n-p}.$$

Where

$$\tilde{\sigma^2} = \frac{n}{n-p}\hat{\sigma}^2 = \frac{n}{n-p}\frac{1}{n}\|Y - X\hat{\beta}\|^2 = \frac{1}{n-p}\|(I-P)Y\|^2$$

and $\hat{\sigma}^2$ is the MLE for σ^2.

Assume that $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ or $\varepsilon \sim N_n(0, \sigma^2)$. Then

$$Y = X\beta + \varepsilon \sim N_n(X\beta, \sigma^2 I)$$

and the likelihood is

$$L(\beta, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \exp\left(\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)\right), \beta \in \mathbb{R}^p, \sigma^2 > 0.$$

Thus the log-likelihood (up to constants) is

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left(y_i - Xi^T\beta\right)^2.$$

It follows that

$$\hat{\beta}_{\text{MLE}} = \hat{\beta} = (X^T X)^{-1} X^T Y$$

and

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\|Y - X\hat{\beta}\|^2 = \frac{1}{n}\|(I-P)Y\|^2.$$

Both estimators are intuitive under our model assumptions without $\varepsilon \sim N_n(0, \sigma^2 I)$ necessarily. However, the distributional assumptions on ε induce distributions on the estimators, which will allow us to perform inference.

1.4.3 Inference for the normal linear model

Distributions of $\hat{\beta}_{\text{MLE}}$ and $\hat{\sigma^2}_{\text{MLE}}$:
Note

$$\hat{\beta}_{\text{MLE}} = \beta + (X^T X)^{-1} X^T \varepsilon \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$$

also

$$\hat{\sigma^2}_{MLE} = \frac{1}{n} \|(I - P)\varepsilon\|^2 \sim \frac{\sigma^2}{n} \chi^2_{n-p}.$$

In particular, $\mathbb{E}\left[\hat{\sigma^2}_{\text{MLE}}\right] = \frac{\sigma^2}{n}(n-p) \implies \hat{\sigma^2}_{\text{MLE}}$ is biased (but asymptotically unbiased). So let

$$\tilde{\sigma^2} = \frac{n}{n-p} \hat{\sigma^2}_{\text{MLE}} = \frac{1}{n-p} \|(I - P)Y\|^2 \sim \frac{\sigma^2}{n-p} \chi^2_{n-p}.$$

What can we say about their joint distribution? Recall that PY and $(I-P)Y$ are uncorrelated;

$$\begin{pmatrix} PY \\ (I-P)Y \end{pmatrix} = \begin{pmatrix} P \\ I - P \end{pmatrix} Y$$

is a linear transformation of Y (normal). Then we know by our earlier work that $PY \perp (I-P)Y$. Note $\hat{\beta} = (X^T X)^{-1} X^T PY \implies \hat{\beta} \perp \tilde{\sigma^2}$.
Inference for β :
Note that

$$\frac{\hat{\beta} - \beta}{\sqrt{\tilde{\sigma^2}}} = \frac{N_p(0, (X^T X)^{-1})}{\sqrt{\chi^2_{n-p}/(n-p)}},$$

with the numerator and denominator independent. This is a *pivot* i.e. id does not depend on the unknown parameters (β, σ^2) and hence can be used for inference.

Example.

$$C_j(\alpha) := \{b \in \mathbb{R} : \left| \frac{\hat{\beta}_j - b}{\sqrt{\tilde{\sigma^2}(X^T X)^{-1}_{jj}}} \right| \leq t_{n-p}(\frac{\alpha}{2})\} \quad j = 1. \ldots, p,$$

where if $T \sim t_{n-p}$, then

$$\mathbb{P}\left(-t_{n-p}(\frac{\alpha}{2}) \leq T \leq T_{n-p}(\frac{\alpha}{2})\right) = 1 - \alpha.$$

Since

$$\frac{\hat{\beta}_j - \beta}{\sqrt{\tilde{\sigma^2}(X^T X)^{-1}_{jj}}} \sim t_{n-p},$$

then $\mathbb{P}_{\beta,\sigma^2}(\beta_j \in C_j(\alpha)) = 1 - \alpha$. So $C_j(\alpha)$ is a $(1-\alpha)$-confidence interval for β_j. Note,

$$\prod_{j=1}^{p} C_j(\alpha)$$

is not a $(1-\alpha)$-confidence interval cuboid for β (too small). Though

$$\prod_{j=1}^{p} C_j(\frac{\alpha}{p})$$

has a confidence / coverage of $\geq 1-\alpha$. However, the latter is too large / conservative generally.

A smaller (and exact) alternative. Consider

$$\|X(\hat{\beta}-\beta\|^2 = \|P\varepsilon\|^2 \sim \sigma^2\chi^2_p$$

independent of $\tilde{\sigma^2}$. Let

$$C(\alpha) = \{\mathbf{b} \in R^p : \frac{\|X(\hat{\beta}-\beta\|^2}{\tilde{\sigma^2}} \leq F_{p,n-p}(\alpha)\}.$$

If $F \sim F_{p,n-p}$, then $\mathbb{P}(F \leq F_{p,n-p}(\alpha)) = 1-\alpha$. Then $\mathbb{P}_{\beta,\sigma^2}(\beta \in C(\alpha)) = 1-\alpha$. Note, the same arguments allows us to construct hypotheses tests.

Example. We can test $H_0 : \beta_j = \beta_{j,0}$ v.s. $H_1 : \beta_j \neq \beta_{j,0}$ with

$$\phi_j = 1\{\beta_{j,0} \notin C_j(\alpha)\}.$$

We can also test $H_0 : \beta = \beta_0$ v.s. $\beta \neq \beta_0$ with

$$\phi = 1\{\beta_0 \notin C(\alpha)\}.$$

Prediction Intervals:
Let $\mathbf{x}^*$ be a new observation. Note

$$\mathbf{x}^{*T}(\hat{\beta}-\beta) = \mathbf{x}^{*T}(X^TX)^{-1}X^T \sim N\left(0, \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*\right),$$

and

$$\frac{\mathbf{x}^{*T}(\hat{\beta}-\beta)}{\sqrt{\tilde{\sigma^2}\mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*}} \sim t_{n-p}$$

can be used to perform inference for the regression function at $\mathbf{x}^*$ i.e. $x^{*T}\beta$. Let $Y^* = x^*\beta + \varepsilon^*, \varepsilon \sim N(0,\sigma^2)$ is independent of ε. We can also construct a $(1-\alpha)$-prediction interval for Y^* i.e. a random interval I depending only on Y such that $\mathbb{P}(Y^* \in I) = 1-\alpha$. Indeed,

$$Y^* - X^{*T}\hat{\beta} = \varepsilon^* + x^{*T}(\beta-\hat{\beta}) \sim N(0, \sigma^2(1+x^{*T}(X^TX)^{-1}X^*)).$$

So we can use the pivot

$$\frac{Y^* - X^{*T}\beta}{\sqrt{\tilde{\sigma}^2(1+x^{*T}(X^TX)^{-1}x^*)}} \sim t_{n-p}.$$

Note that the confidence intervals for Y^* will be larger than those for $x^{*T}\beta$ because of the additional variability / uncertainty coming from ϵ^*.

1.4.4 Testing significance of groups of variables

Let

$$X = (X_0.X_1), \beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, X_0 \in \mathbb{R}^{n\times p_0}, X_1 \in \mathbb{R}^{n\times(p-p_0)}, \beta_0 \in \mathbb{R}^{p_0}, \beta_1 \in \mathbb{R}^{p-p_0}.$$

Without loss of generality, we wish to test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$. We can construct a generalised likelihood ratio test : recall if Y has density $f(y, \theta), \theta \in \Theta$ unknown, the likelihood ratio test for $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$, where $\Theta_0 \subseteq \Theta$ reject H_0 for large values of

$$\omega_{LR} := 2 \log \frac{\sup_{\theta' \in \Theta} L(\theta')}{\sup_{\theta' \in \Theta_0} L(\theta')} = 2(\sup_{\theta' \in \Theta} \ell(\theta') - \sup_{\theta' \in \Theta_0} \ell(\theta').$$

Notation. Write $\check{\beta}_0$ and $\check{\sigma}^2$ for the MLEs under the null i.e. $Y = X_0\beta_0 + \varepsilon$

$$\varepsilon \sim N_n(0, \sigma^2 I), \text{ so } \check{\beta}_0 = (X_0^T X_0)^{-1} X_0^T Y, \check{\sigma}^2 = \frac{1}{n}\|Y - X_0\check{\beta}_0\|^2.$$

Then

$$\omega_{LR} = 2(-\frac{n}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\|Y - X\hat{\beta}^2\| + \frac{n}{2}\log\check{\sigma}^2 + \frac{1}{2\check{\sigma}^2}\|Y - X_0\check{\beta}_0\|^2 = n\log(\|(I-P_0)Y\|^2/\|(I-P)Y\|^2).$$

Note $I - P_0 = I - P + P - P_0$ and $PP_0 = P_0$

$$(I - P)(P - P_0) = 0 \implies \|(I - P_0)Y\|^2 = \|(I - P)Y\|^2 + \|(P - P_0)Y\|^2$$

and

$$\frac{\|(I - P_0)Y\|^2}{\|(I - P)Y\|^2} = 1 + \frac{\|(P - P_0)\varepsilon\|^2}{\|(I - P)\varepsilon\|^2}.$$

We claim that $P - P_0$ is an orthogonal projection. Indeed, $P - P_0$ is symmetric and by $P_0P = P_0^T P = (PP_0)^T = P_0^T = P_0$, so

$$(P - P_0)^2 = P - PP_0 - P_0P + P_0 = P - P_0.$$

Also note that if π is an orthogonal projection, then $\text{rank}\pi = \text{tr}\pi$ so,

$$\text{rank}(P - P_0) = \text{tr}(P - P_0) = \text{tr}P - \text{tr}P_0 = \text{rank}P - \text{rank}P_0 = p - p_0$$

and we conclude that $\|(P - P_0)\varepsilon\|^2 \sim \sigma^2\chi^2_{p-p_0}$. Note

$$\text{cov}((I - P)\varepsilon, (P - P_0)\varepsilon) = \sigma^2(I - P)(P - P_0) = 0$$

and

$$\begin{pmatrix} (I - P)\varepsilon \\ (P - P_0)\varepsilon \end{pmatrix}$$

is a linear transformation of ε (normal) so $(I - P)\varepsilon \perp (P - P_0)\varepsilon$. So we take the test

$$\phi = 1\{\frac{\|(P - P_0)Y\|^2/(p - p_0)}{\|(I - P)Y\|^2}/(n - p) \geq F_{p-p_0,n-p}(\alpha)\}.$$

This test has a significance level of α.

1.4.5 Model checking

The validity of our inferential conclusions depends on our assumptions about the distribution of ε being correct. If any of them fail we have the following remarks:

Remark. (i) $\mathbb{E}[\varepsilon_i] = 0$. If not, the coefficients in the linear model should be interpreted with care ($Y = X\beta + \mu + (\varepsilon - \mu)$), $\tilde{\sigma}^2$ is inflated, F-tests will have the correct size (example sheet) but they may lose power.

(ii) $\operatorname{var}\varepsilon_i = \sigma^2$. If not, $\hat{\beta}$ is not as efficient as it could be, and confidence sets and hypothesis test levels may not be correct. If the variance are known up to a multiplicative constant, then do weighted least squares (ex sheet).

(iii) $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i, j$. This usually occurs with temporal / spatial data, confidence sets and levels may be wrong.

(iv) ε is normal. If the first assumptions hold, then the inferential procedures hold asymptotically thanks to the central limit theorem.

Any violation of the above may be checked by looking at

$$\hat{\epsilon} = Y - X\hat{\beta} = (I - P)Y.$$

To check $\mathbb{E}[\varepsilon_i] = 0$ plot $\hat{\varepsilon}$ against $\hat{Y}$ (and sometimes v.s. the covariance), if the assumption holds we should expect to see no trends.
To check $\operatorname{var}\varepsilon_i$ (and $\operatorname{cov}(\varepsilon_i, \varepsilon_j)$) note that $\operatorname{var}\hat{\varepsilon} = \sigma^2(I - P)$ so define the studentised residuals

$$\hat{\eta}_i = \frac{\hat{\varepsilon}_i}{\tilde{\sigma}\sqrt{1 - P_i}},$$

where $P_i = P_{ii}$ is the leverage of the ith observation. Note if $\tilde{\sigma} \leftrightarrow \tilde{\sigma}_{-i}$ (thet corrected MLE when excluding (x_i, Y_i). Then $\hat{\eta}_i \sim t_{n-1-p}$. Note: if $\hat{\varepsilon}_i \neq 0$ a.s., $p_i \neq 1$. Also, assume $n \gg p \implies \tilde{\sigma} \approx \sigma$ with low deviation and $\operatorname{var}(\hat{\eta}_i) \approx 1$. Thus, plot $\sqrt{(\hat{\eta}_i)}$ vs $\hat{Y}$ ($|\cdot|$ avoids the 2-sidedness ; $\sqrt{\cdot}$ brings the studentised residuals closer to 1 if $\operatorname{var}(\hat{\eta}_i) \approx 1$). We expect a flat cloud of points around 1. Furthermore, under our model assumptions, $\hat{\eta}_i$ are approximately $\overset{\text{iid}}{\sim} N(0, 1)$. We check this with a Q-Q (Quantile-Quantile) plot.
Coefficients of determination Popular measure of goodness of fit of a model. It compares the residual sum of squares (RSS) of the model vs that of an intercept only model

$$R^2 = \frac{\|Y - \hat{Y}1_n\|^2 - \|(I - P)Y\|^2}{\|Y - \hat{Y}1_n\|^2}, 1_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n.$$

Note

$$R^2 = \frac{\operatorname{var}_n Y - \hat{\sigma}^2}{\operatorname{var}_n Y} \ \in [0, 1].$$

So R^2 is the proportion of the total variation of the data explained by the model. If we have a larger R^2 value, then we have a better fit, but it always grows whenever we add a new variable. The adjusted

$$\tilde{R}^2 = 1 - \frac{n - 1}{n - p}(1 - R^2).$$

Observation A few observations may not agree with our model assumptions i.e. outliers. So, we go back to the data, they could be very informative or we may have to exclude them.
Leverage $\hat{Y}_i = (PY)_i = \sum_{j=1}^n P_{ij}Y_j$ and $\text{var}(\hat{\varepsilon}_i) = \sigma^2(1-p_i), p_i = P_{ii}$ - the leverage of the ith observable . Thus p_i measures the contribution of Y_i to $\hat{Y}_i$. If $\pi \approx 1$, the model is forced to go through Y_i. It is possible that excluding such an observation does not change $\hat{\beta}$ much, but R^2 and F-tests with H_0 : intercept-only model may be highly affected therefore p_i is a measure of influence. Note $\sum_{i=1}^n p_i = \text{tr}(P) = p$ so our rule of thumb is that if $p_i > 3\frac{p}{n}$ then there is a concern that our ith observation is too influential.
Cook's Distance Defined as

$$D_i = \frac{\|X(\hat{\beta} - \hat{\beta}_{-i}\|^2/p}{\tilde{\sigma}^2},$$

where $\hat{\beta}_{-i}$ is the MLE when excluding (X_i, Y_i). Note

$$D_i = \frac{1}{p}\frac{P_i}{1-p_i}\hat{\eta}_i^2.$$

Recall that

$$\mathbb{P}\left(\frac{\|X(\hat{\beta}-\beta)\|^2/p}{\tilde{\sigma}^2} \leq F_{p,n-p}(\alpha)\right) = 1-\alpha.$$

So we get another rule of thumb, the ith observation's influence may be worrying if $D_i > F_{p,n-p}(0.5)$ i.e. removing (X_i, Y_i) pushes the MLE beyond a 50% confidence interval around $\hat{\beta}$.

1.5 ANOVA & ANCOVA

So far, our predictor is $\in \mathbb{R}$ (continuous variables) can deal with categories (or factors) similarly: e.g. let the responses be the weight loss (WL) under J exercise regimes, with the first being no exercise (control group). Our model is Y_{jk} is the weight loss of the kth participant in regime j (n_j of them)

$$Y_{jk} \overset{\text{iid}}{\sim} N(\mu_j, \sigma^2), j = 1, \ldots, J, k = 1, \ldots, n_j.$$

Equivalently, we can say $Y = X\beta + \varepsilon$ with

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{J1} \\ \vdots Y_{Jn_J} \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}.$$

and we have that $\varepsilon \sim N_{\sum_{j=1}^{J} n_j}(0\sigma^2 I)$. This type of model is called one-way analysis of variace (ANOVA). If $n_i = n_j \ \forall \ i \neq j$ then it is called a balanced one-way ANOVA.

Question. Why do we use ANOVA? Statistics describe the variability in a data set. Fisher proposed the model above to reflect that different groups have different variability, and developed associated F-tests for the means. These are in terms of the variances within and between groups, so the model and / or the tests are called ANOVA.

An alternative parametrisation $Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}$ where the $\varepsilon_{jk} \overset{\text{iid}}{\sim} N(0\sigma^2)$, μ is called the baseline effect, α_j is the jth regime's effect in relation to μ. Note: this model is not identifiable, indeed $\mu + c, \alpha - c$ gives the same model $\forall\ c \in \mathbb{R}$. So we need a constraint, generally we take the corner point constraint (default in R) e.g. $\alpha_1 = 0$ so it is simpler to test with respect to baseline effects / control group. We can further subdivide the dataset e.g. I different food diets so Y_{ijk} is WL of kth participant in diet I and regime j (n_{ij} of them). Our model is now

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2),\ i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij}$$

and the γ_{ij} are the interaction effects. This is called a two-way ANOVA and if we set all γ_{ij} to 0, additive two-way ANOVA. Typically, our (corner-point) constraint is $\alpha_1 = \beta_1 = \gamma_{11} = 0$. we can also include continuous variables e.g. blood pressure, and the resulting normal linear model is called analysis of covariance (ANCOVA).
A word on causal inference and randomised experiments. (non-examinable)
Causal conclusions depend on experiment design. Suppose that given our ANOVA analysis we find that one of our regimes has a positive effect on weight loss. If we allowed the participants to choose their regime it is likely that fitter participants will pick harder routines.

1.6 Model selection

Intuitively, we wish to select "the right model" to focus on the variables of interest. Mathematically, if only β_0 is non-zero in $\beta = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$, then

$$\text{var}(\check{\beta}_{0,j}) \leq \text{var}(\hat{\beta}_j) \text{ (example sheet) and MSPE}(\check{\beta}_0) = \sigma^2 \frac{P_0}{n} \leq \sigma^2 \frac{P}{n}.$$

We have already met two popular model selection techniques F-tests from section 1.2.3 and $\tilde{R}^2$ (or R^2).
Akaike's information criterion (AIC)
Let $\mathcal{M} = \{f_k(\cdot; \theta_k), \theta_k \in \Theta_k\} k = 1, \dots, K$ be a collection of models. Assume $Y_i \overset{\text{iid}}{\sim} g_i, i = 1, \dots, n$ (g may not be in the collection). For simplicity let $g_i i = g \ \forall\ g$ let $\hat{\theta}_k$ be the MLE for $\mathcal{M}_k$ and θ_k and $\hat{f}_k(\cdot) = f_k(\cdot; \hat{\theta}_k)$ AIC minimises $K(g, \hat{f}_k)$ over k where

$$K(g, f) = \int \log \frac{g}{f} g$$

or equivalently minimise $\mathbb{E}_g \left[\log \hat{f}_k\right]$.

Fact.

$$\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_k(Y_i)$$

is an estimator of $\mathbb{E}_g\left[\log \hat{f}_k\right]$ with bias $\approx \dim\frac{\theta_k}{n}$. So AIC minimises

$$\mathrm{AIC}(\mathcal{M}_k) = 2n(\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_k(Y_i) = \frac{\dim\theta_k}{n} = -2(\ell_k\hat{\theta}_k - \dim\theta_k = -2\,(\text{ maximised log-likelihood - no of parame}$$

Note: for the normal linear model $\mathcal{M}_k$

$$\mathrm{AIC}(\mathcal{M}_k) = -2\left(-\log((2\pi\hat{\sigma}_k^2)^{\frac{n}{2}}) - \frac{n}{2} - 2(p_k+1)\right).$$

Note: the models above are general so AIC doesn't require nestedness.

Orthogonality Given p covariates (including intercept), we can form 2^{p-1} models with intercept, and can use e.g. AIC or $\tilde{R}^2$ to select one. This is unfeasible if p is large, unless X has "sufficient orthogonality". Let $X = (X_0, X_1)$ and $\beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$. Then β_0, β_1 are orthogonal sets of parameters if $X_0^T X_1 = 0$. If so, $\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$. Indeed,

$$\begin{aligned}
\hat{\beta} &= (X^TX)^{-1}X^TY\\
&= \left(\begin{pmatrix}X_0^T\\ X_1^T\end{pmatrix}(X_0X_1)\right)^{-1}\begin{pmatrix}X_0^T\\ X_1^T\end{pmatrix}Y\\
&= \begin{pmatrix}X_0^T & 0\\ 0 & X_1^TX_1\end{pmatrix}^{-1}\begin{pmatrix}X_0^TY\\ X_1^TY\end{pmatrix}\\
&= \begin{pmatrix}(X_0^TX_0)^{-1}X_0^TY\\ (X_1^TX_1)^{-1}X_1^TY\end{pmatrix} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}.
\end{aligned}$$

2 or more sets of parameters are mutually orthogonal (MO) if the column spaces of corresponding blocks of X are orthogonal. If so, $\hat{\beta}$ decomposes into the estimators of the blocks, and if there are sufficiently many of these, the strategy above may indeed be feasible. In particular, if all columns of X are orthogonal, and we wish to maximise $\tilde{R}^2$ or R^2 (or minimise the residual sum of squares - RSS) for a fixed p_0; then order $(\hat{\beta}_j\|X_j\|)^2$ (excluding the intercept) increasingly and pick the highest $p_0 - $'1 terms. Indeed, if $S \subseteq \{1, ..., p\}$ and P_s be the orthogonal projection onto the corresponding columns of X,

$$\begin{aligned}
\|(I-P_s)Y\|^2 &= \|Y - \sum_{j\in S}\hat{\beta}_jXj\|^2\\
&= Y^TY - 2\sum\hat{\beta}_jX_j^TY + \sum_{i,j}\hat{\beta}_i\hat{\beta}_jX_i^TX_j\\
&= \|Y\|^2 - \sum\hat{\beta}_j\|X_j\|^2
\end{aligned}$$

.

Exact orthogonality is uncommon, unless we designed X or we transformed it (at the risk of losing interpretability). Orthogonality between intercept and the rest is common, by the common transformation of mean-centring the columns of X.
Forward (fwd) and backward (bwd) selection
If no or little orthogonality, popular strategies are as follows
Forward selection

(i) Fit the intercept only model S_0

(ii) Compute all models with one more parameter and keep the model with the lowest RSS

(iii) Repeat 2 until all (or sufficiently many) covariates are included

(iv) Choose one model from the resulting sequence $S_0 \subset S_1 \subset \cdots$ using e.g. AIC or $\tilde{R}^2$.

Backward selection
The same as above, bur start with all covariates until reaching the intercept-only model (removing one covariate at a time)

1.7 Inference after model selection

Inference from 1.2.3 assumes that the model was fixed prior to data-collection. Therefore, we cannot use the same data for selection and inference. The easy option is to split the data and use each bit for each purpose

2 Exponential families and generalised linear models

2.1 Non-normal responses

Responses not always naturally live in $\mathbb{R}$: e.g. prices (> 0), counting data ($\mathbb{N}$), binary options (yes & no). Thus, the normal linear model may not be appropriate.

Variable transformations

A first approach is to transform the data so that our linear model assumptions are met approximately. If responses > 0 (including $\mathbb{N}$), the Box-Cox transformation is classical:

$$y \mapsto y^{\lambda} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}.$$

One finds λ by MLE regarding $(\lambda, \beta, \sigma^2)$ as the parameter. Drawback, we should at once, achieve approximate normality, variance stabilisation and linearity in β. A more recent and superior strategy is to model these separately by e.g. GLM.

3 Specific regression problems