Statistical Theory & Applied Statistics

Part III Courses, Michaelmas 2009

Revision Notes

Daniel Guetta

guetta@cantab.net

© Daniel Guetta, 2010

Based on lectures by Richard Samworth and Susan Pitts, Michaelmas 2009

Preliminary Computing

Directory, Linux & Miscellaneous operations

- Changing password: **`passwd`**
- Creating a new directory: **`mkdir`**
- Changing to directoy **`newDir`**: **`cd newDir`**. The command **`cd ..`** moves up one level in the directoy tree.
- List all files in the current directory: **`ls`**. Can be followed by **`*.tex`**, for example, to only list tex files.
- Delete **`File1`**: **`rm File1`**
- Move **`File1`** to directory **`newDir`**: **`mv File1 newDir`**
- Copy **`File1`** to a file with name **`File2`**: **`cp File1 File2`**
- Examine the content of **`File1`**: **`more File1`**
- Leaving emacs: **`ctrl x, ctrl c`**

Printing

- Looking at the print queue: **`lpq`**
- Looking at my jobs: **`ps x`**

General Splus7 stuff

- Opening Splus for the first time in a given directory: **`Splus7 CHAPTER`**
- Opening Splus7: **`Splus7 -e`** *The –e allows the use of* arrows
- Leaving Splus7: **`q()`**
- Listing all objects created: **`objects()`**
- Removing an object: **`rm(objectName)`**
- Getting help with function `blah`: **`?blah`**
- **`rep(1:5,9)`** repeats the sequence "1 2 3 4 5" 9 times. **`rep(1:5,9)`** produces 9 1s, followed by 9 2s, etc…

Imputing and handling data

- Reading `FileName` into table `dataVar` with column titles `x` and `y`: **`dataVar <- scan("FileName", list(x=0,y=0))`** *The "list" function*

© Daniel Guetta, 2010

simply creates an empty table with two columns; x containing value 0 and y containing value 0. Without this argument, the entire file is read left-to-right into a single list.

- Reading `FileName` into table `dataVar`, when the column titles are already included in the file `FileName`: **`dataVar <- read.table("tumour", header=T)`**
- Inputting a vector directly into `x`: **`x <- scan()`** *End input with a blank line*
- Listing the columns of a table `tableName`: **`names(tableName)`**
- Extracting the `x` from the table `dataVar`: **`dataVector <- dataVar$x`**
- **`x[-11]`** returns vector x without observation 11
- Saving each column in table `dataVar` as a separate vector, with the same name as the column title: **`attach(dataVar)`**
- Joining several columns into a matrix: **`cbind(...)`**
- Many functions (`glm`, `plot`, etc...) can be applied to subsets of the data by adding an argument to the function. For example, the run a function only on data in which `var1` takes the value `A`, add the argument **`subset=(var1=="A")`** to the function.
- **`table(var1,var2)`** will create a contingency table, listing the number of times each possible pair of values occurs.

Simple graphical stuff

- Plotting two vectors against each other: **`plot(xVector, yVector)`**
- Follow the above by: **`identify(xVector, yVector)`** to be able to identify different points on the plot.
- Adding a straight line to the current plot: **`lines(xVector, yVector,type`**$^{\textit{Optional}}$**`)`** *"Type" takes values "p" for points, "l" for a line or "b" for both*
- Fitting two-by-two plots on each page: **`par(mfrow=c(2,2))`**

Outputting stuff to a file

- To output graphics, begin with **`postscript(file='fileName.ps')`** and create all graph using **`plot`** (none of them will appear on screen). When done, use **`dev.off()`**.
- To save output to a file, start with **`sink('fileName')`** – all subsequent output will go to the file. End with **`sink()`**.

© Daniel Guetta, 2010

Simple statistical stuff

- The upper α point of a t_n distribution is given by: **`qt(1-α, n)`**
- The upper α of an $F_{m,n}$ distribution is given by: **`qf(1-α, m, n)`**
- Generating vectors containing a sample of `n` simulated $N(0,1)$ observations: **`rnorm(n)`**

© Daniel Guetta, 2010

Preliminary Mathematics

This section contains preliminary mathematics and proofs that are assumed in these notes.

Projections

- A **idempotent** matrix P on an underlying vector space W satisfies $P^2 = P$ and is a **projection**. The eigenvalues of P can only be 0 or 1, with the following eigenspaces

$$U = \text{Range/Image of } P = \left\{P\boldsymbol{y} : \boldsymbol{y} \in W\right\}$$
$$V = \text{Null space/kernel of } P = \left\{\boldsymbol{x} : P\boldsymbol{x} = 0, \boldsymbol{x} \in W\right\}$$

  We say P projects W onto U.

- $W = U \oplus V$; for any $\boldsymbol{x} \in W$ we can write $\boldsymbol{x} = P\boldsymbol{x} + \left(\boldsymbol{x} - P\boldsymbol{x}\right)$, where the first vector is in U and the second is in W.

  If U and V are orthogonal (see next point), we can show that such a decomposition is unique . Consider

$$\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v} = \boldsymbol{u}' + \boldsymbol{v}'$$

  Then

$$0 = \left\|\boldsymbol{x} - \boldsymbol{x}\right\|^2 = \left\|\left(\boldsymbol{u} - \boldsymbol{u}'\right) + \left(\boldsymbol{v} - \boldsymbol{v}'\right)\right\|^2 = \left\|\left(\boldsymbol{u} - \boldsymbol{u}'\right)\right\|^2 + \left\|\left(\boldsymbol{v} - \boldsymbol{v}'\right)\right\|^2$$

  Since $\left\{\boldsymbol{u} - \boldsymbol{u}' \in U\right\} \perp \left\{\boldsymbol{v} - \boldsymbol{v}' \in V\right\}$. Thus, we must have $\boldsymbol{u} = \boldsymbol{u}', \boldsymbol{v} = \boldsymbol{v}'$.

- A projection is **orthogonal** if $V = U^{\perp} = \left\{\boldsymbol{y} \in W : \boldsymbol{u}^T\boldsymbol{y} = 0, \forall \boldsymbol{u} \in U\right\}$. Consider $\boldsymbol{x}, \boldsymbol{y} \in W$ and $\boldsymbol{u} = P\boldsymbol{x} \in U$, $\boldsymbol{v} = \boldsymbol{y} - P\boldsymbol{y} \in V$. We then have

$$\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{x}^T P^T \left(\boldsymbol{y} - P\boldsymbol{y}\right) = \boldsymbol{x}^T \left(P^T - P\right)\boldsymbol{y}$$

  We therefore have that a projection is orthogonal if and only if $P^T = P$.

- If P and Q are projections and $PQ = 0$, then the range of each of these vectors are orthonormal.

- For the orthogonal projection onto a vector $\boldsymbol{u}$, $P = \boldsymbol{u}\boldsymbol{u}^T$.

Generating Functions

- Let X be a real-valued random variable. The **moment generating function** of X is given by

© Daniel Guetta, 2010

$$M(t) = \mathbb{E}\left(e^{tX}\right) \qquad \forall t \text{ s.t. } \mathbb{E}\left(e^{tX}\right) < \infty$$

- Note that, by an extension of linearity

$$M(t) = \mathbb{E}\left(\sum_{k=0}^{\infty} \frac{t^k X^k}{k!}\right) = \sum_{k=0}^{\infty} \mathbb{E}\left(\frac{t^k X^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}\left(X^k\right)$$

$$\boxed{\left.\frac{\mathrm{d}^n M}{\mathrm{d}t^n}\right|_{t=0} = \mathbb{E}\left(X^n\right)}$$

- If $X \sim N\left(\mu, \sigma^2\right)$, then

$$M(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

- The **cumulant generating function** is given by

$$K(t) = \log M(t)$$

We can write K as a power series

$$K(t) = \log M(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}$$

κ_n is called the $\boldsymbol{n^{\text{th}}}$ **cumulant**. Clearly

$$\kappa_r = \left.\frac{\mathrm{d}^r K_Y(t)}{\mathrm{d}t^r}\right|_{t=0}$$

Note also that we can use the expansion $\log\left(1+z\right) = z - \frac{1}{2}z^2 + \cdots$ on the series expansion $M(t) = 1 + t\mathbb{E}\left(X\right) + \frac{1}{2}t^2\mathbb{E}\left(X^2\right) + \cdots$ to get

$$\begin{aligned} K(t) &= \left\{t\mathbb{E}\left(X\right) + \tfrac{1}{2}t^2\mathbb{E}\left(X^2\right)\right\} - \tfrac{1}{2}\left\{t\mathbb{E}\left(X\right) + \tfrac{1}{2}t^2\mathbb{E}\left(X^2\right)\right\}^2 \\ &= \left\{t\mathbb{E}\left(X\right) + \tfrac{1}{2}t^2\mathbb{E}\left(X^2\right)\right\} - \tfrac{1}{2}\left\{t^2\left[\mathbb{E}\left(X\right)\right]^2 + \cdots\right\} \\ &= \left\{\mathbb{E}\left(X\right)\right\}t + \left\{\mathbb{E}\left(X^2\right) - \left[\mathbb{E}\left(X\right)\right]^2\right\}\frac{t^2}{2} \\ &= \mu t + \sigma^2 \frac{t^2}{2} \end{aligned}$$

This confirms that $\kappa_1 = \mu$ and $\kappa_2 = \sigma^2$.

Distributions

- If $Z_i \sim N\left(0,1\right)$ where $i = 1, \cdots, n$

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

Where χ_n^2 is a χ^2 distribution on n degrees of freedom.

$$\mathbb{E}\left(\chi_n^2\right) = n$$

© Daniel Guetta, 2010

- If $Z \sim N(0,1)$ and $Y \sim \chi^2_n$ and Z and Y are independent, then

$$\frac{Z}{\sqrt{Y/n}} \sim t_n$$

Where t_n is a t distribution on n degrees of freedom, which is symmetric with heaver tails than the normal distribution, and converges in distribution to $N(0,1)$ as $n \to \infty$.

$$\mathbb{E}(t_n) = 0$$

- If $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ and X and Y are independent

$$\frac{X/m}{Y/n} \sim F_{m,n}$$

Where $F_{m,n}$ is an F distribution on m and n degrees of freedom. Note that

$$\mathbb{E}(F_{m,n}) = \frac{n}{n-1}$$

- If $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, I)$, then $U = \boldsymbol{Z}^T\boldsymbol{Z}$ has a non-central χ^2 distribution $\chi^2_n(\delta)$ with n degrees of freedom and non-centrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu}$.

$$\mathbb{E}(\chi^2_n(\delta)) = n + \delta$$

- If $W_1 \sim \chi^2_{n_1}(\delta)$ and $W_2 \sim \chi^2_{n_2}(\delta)$ and W_1 and W_2 are independent, then $\frac{W_1/n_1}{W_2/n_2}$ has a non-central F distribution $F_{n_1,n_2}(\delta)$ with n_1 and n_2 degrees of freedom and non-centrality parameter δ.

$$\mathbb{E}(F_{n_1,n_2}(\delta)) = \frac{n_2(n_1+\delta)}{n_1(n_2-2)}$$

The Multivariate Normal

Y has a univariate normal distribution with mean μ and variance $\sigma^2 \in (0,\infty)$, and we write $Y \sim N(\mu,\sigma^2)$ if its density is

$$f_y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \qquad y \in \mathbb{R}$$

It's also convenient to define a degenerate normal distribution $Y \sim N(\mu,0)$ in which $\mathbb{P}(y=\mu)=1$.

> **Definition**: We say a random vector $\boldsymbol{Y} \in \mathbb{R}^n$ has an ***n*-variate normal distribution** if for every vector $\boldsymbol{t} \in \mathbb{R}^n$, the random variable $\boldsymbol{t}^T\boldsymbol{Y}$ has a (possibly degenerate) univariate distribution. We write $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu},\Sigma)$.

© Daniel Guetta, 2010

We write $\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{Y}\right)$ for the mean vector and $\Sigma = \operatorname{cov}\left(\boldsymbol{Y}\right)$ for the covariance matrix[1] of $\boldsymbol{Y}$, so that

$$\Sigma_{ij} = \mathbb{E}\left\{(Y_i - \mu_i)(Y_j - \mu_j)\right\} = \mathbb{E}\left(Y_i Y_j\right) - \mu_i \mu_j$$

The entire dependence structure of $\boldsymbol{Y}$ is determined by Σ. Components of $\boldsymbol{Y}$ are independent if and only if Σ is diagonal[2].

In general, Σ is **symmetric** and **non-negative definite**, because $\boldsymbol{t}^T \Sigma \boldsymbol{t} = \operatorname{var}\left(\boldsymbol{t}^T \boldsymbol{Y}\right)$ and variances are non-negative.

If, in addition, Σ is **positive definite**, then none of its eigenvalues are 0 and it is **invertible**. We then have

$$f_{\boldsymbol{y}}\left(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma\right) = \frac{1}{\sqrt{\left(2\pi\right)^n \left\|\Sigma\right\|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right)\right\} \qquad \boldsymbol{y} \in \mathbb{R}^n$$

Even if Σ is only non-negative definite, its MGF is

$$M_{\boldsymbol{y}}(\boldsymbol{t}) = \mathbb{E}\left(e^{\boldsymbol{t}^T \boldsymbol{Y}}\right) = e^{\boldsymbol{t}^T \boldsymbol{\mu} + \frac{1}{2} \boldsymbol{t}^T \Sigma \boldsymbol{t}}$$

We finally note that if $\boldsymbol{Y} \sim N_n\left(\boldsymbol{\mu}, \Sigma\right)$ and A is a $p \times n$ matrix, then $A\boldsymbol{Y} \sim N_p\left(A\boldsymbol{\mu}, A\Sigma A^T\right)$[3].

Laws of Large Numbers

Chebyshev's Inequality: Let X be a random variable with $\mathbb{E}(X) = \mu$ and $\mathbb{V}\mathrm{ar}(X) = \sigma^2$. Then

$$\mathbb{P}\left(\left|X - \mu\right| \geq \alpha\right) \leq \frac{\sigma^2}{\alpha^2}$$

---

[1] The covariance matrix is given by $\operatorname{cov}\left(\boldsymbol{X}\right) = \Sigma = \mathbb{E}\left(\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T\right)$, and note that $\operatorname{cov}\left(A\boldsymbol{X}\right) = A \operatorname{cov}\left(\boldsymbol{X}\right) A^T$.

[2] Note that in general, $\operatorname{cov}[u, v] \neq 0$ does **not** imply u and v are independent unless they are jointly normal.

[3] To prove it's p-variate normal, choose $t \in \mathbb{R}^p$, and note that $t^T A\boldsymbol{Y} = \left(A^T t\right)^T \boldsymbol{Y}$ where $\left(A^T t\right)^T \in \mathbb{R}^n$. For the mean, note that $\mathbb{E}\left(A\boldsymbol{Y}\right) = A\mathbb{E}\left(\boldsymbol{Y}\right) = A\boldsymbol{\mu}$ and for the variance

$$\operatorname{cov}\left((A\boldsymbol{Y})_i, (A\boldsymbol{Y})_j\right) = \operatorname{cov}\left(A_{i\alpha} Y_\alpha, A_{j\beta} Y_\beta\right) = A_{i\alpha} A_{j\beta} \operatorname{cov}\left(Y_\alpha, Y_\beta\right) = A_{i\alpha} A_{j\beta} \Sigma_{\alpha\beta} = \left(A\Sigma A^T\right)_{ij}$$

© Daniel Guetta, 2010

Proof:

$$\begin{aligned}
\mathbb{P}\Big(\big|X-\mu\big| \geq \alpha\Big) &= \mathbb{E}\Big(\mathbb{I}\Big\{\big|X-\mu\big| \geq \alpha\Big\}\Big) \\
&= \mathbb{E}\left(\mathbb{I}\left\{\left(\tfrac{X-\mu}{\alpha}\right)^2 \geq 1\right\}\right) \\
&\leq \mathbb{E}\left(\left(\tfrac{X-\mu}{\alpha}\right)^2\right) \\
&= \frac{1}{\alpha^2}\mathbb{E}\Big\{(X-\mu)^2\Big\} \\
&= \frac{\sigma^2}{\alpha^2}
\end{aligned}$$

As required. ■

We define the following modes of convergence

We say that a sequence of random vectors $\left(Y_n\right)$ converges **almost surely** to Y if

$$\mathbb{P}\Big(\lim_{n\to\infty}(Y_n) = Y\Big) = 1$$

or equivalently, for every $\varepsilon > 0$

$$\mathbb{P}\left(\sup_{m\geq n}\big|Y_n - Y\big| > \varepsilon\right) \to 0 \qquad \text{as } n \to \infty$$

And we write $Y_n \xrightarrow{a.s.} Y$.

We say that a sequence of random vectors $\left(Y_n\right)$ converges **in probability** to Y if for every $\varepsilon > 0$

$$\mathbb{P}\Big(\big|Y_n - Y\big| > \varepsilon\Big) \to 0 \qquad \text{as } n \to \infty$$

And we write $Y_n \xrightarrow{p} Y$.

We say that a sequence of random vectors $\left(Y_n\right)$ converges **in distribution** to Y if

$$\mathbb{E}\Big\{f(Y_n)\Big\} \to \mathbb{E}\Big\{f(Y)\Big\}$$

for all bounded, continuous, real-valued functions f. In fact, it is enough that the convergence occurs when f is bounded and Lipschitz (ie: $\exists\ L > 0$ such that $\big|f(x) - f(y)\big| \leq L\big\|x - y\big\|$). Equivalently, if and only if

$$\mathbb{P}\Big(Y_n \leq y\Big) \to \mathbb{P}\Big(Y \leq y\Big)$$

© Daniel Guetta, 2010

at all points where the distribution function of Y is continuous.

We write $Y_n \xrightarrow{d} Y$.

$$Y_n \xrightarrow{a.s.} Y \Rightarrow Y_n \xrightarrow{p} Y \Rightarrow Y_n \xrightarrow{d} Y$$

For any h continuous and real-valued:

$$Y_n \xrightarrow{?} Y \;\Rightarrow f(Y_n) \xrightarrow{?} f(Y)$$

In fact, h may have a set of discontinuities D_h provided that $\mathbb{P}\left(Y \in D_h\right) = 0$.

We are now ready to prove the *weak law of large numbers*

Weak Law of Large Numbers: Let $X_1, X_2, \cdots$ be an infinite sequence of IID random variables with $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mu < \infty$. Define

$$\overline{X}_n = \tfrac{1}{n}\left(X_1 + \cdots + X_n\right)$$

Then assuming $\mathbb{V}\mathrm{ar}(X_1) = \mathbb{V}\mathrm{ar}(X_2) = \cdots = \sigma^2 < \infty$ [4]

$$\overline{X}_n \xrightarrow{\quad p \quad} \mu \qquad \text{as } n \to \infty$$

Proof: By linearity and independence of the X_i, we have that

$$\mathbb{E}(\overline{X}_n) = \tfrac{1}{n}\mathbb{E}(X_1 + \cdots + X_n) = \mu$$
$$\mathbb{V}\mathrm{ar}(\overline{X}_n) = \tfrac{1}{n^2}\mathbb{V}\mathrm{ar}(X_1 + \cdots + X_n) = \sigma^2 / n$$

Then by Chebyshev's Inequality

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \text{as } n \to \infty$$

Thus, by the definition of convergence in probability, we have proved our theorem. ■

Strong Law of Large Numbers: With all the definitions above,

$$\overline{X}_n \xrightarrow{\quad a.s. \quad} \mu \qquad \text{as } n \to \infty$$

[4] The theorem is still true if this condition fails to hold, but the proof is more involved.

© Daniel Guetta, 2010

Uniform Law of Large Numbers: Suppose $f(x,\theta)$ is a function defined for $\theta \in \Theta$. Further suppose that

1. Θ is compact (closed and bounded).
2. $f(x,\theta)$ is continuous at each $\theta \in \Theta$ for almost all x.
3. There exists a dominating function $g(x)$ such that $\mathbb{E}\Big\{g(X)\Big\} < \infty$ and
$$\Big|f(x,\theta)\Big| \leq g(x)\, \forall \theta \in \Theta$$

Then $\mathbb{E}\Big\{f(X,\theta)\Big\}$ is continuous in θ and
$$\sup_{\theta\in\Theta}\left\|\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i,\theta)\right\} - \mathbb{E}\Big\{f(X,\theta)\Big\}\right\| \xrightarrow{\;p\;} 0$$

Slutsky's Theorem

Slutsky's Theorem: If $\boldsymbol{Y}_n$ and $\boldsymbol{Z}_n$ be sequences of random vectors with $\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{Y}$ and $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{c}$ where $\boldsymbol{c}$ is constant, then
$$\Big(\boldsymbol{Y}_n, \boldsymbol{Z}_n\Big) \xrightarrow{\;d\;} \Big(\boldsymbol{Y}, \boldsymbol{c}\Big)$$
and for any continuous real-valued function g
$$g\Big(\boldsymbol{Y}_n, \boldsymbol{Z}_n\Big) \xrightarrow{\;d\;} g\Big(\boldsymbol{Y}, \boldsymbol{c}\Big)$$

© Daniel Guetta, 2010

Proof: Let f be a bounded, Lipschitz function with Lipschitz constant L.[5]

Given $\varepsilon > 0$, let $\delta = \varepsilon / \{3(L+1)\}$.

Since $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{c}$ we can choose $n_0 \in \mathbb{N}$ such that[6]

$$\mathbb{P}\left(\left\|\boldsymbol{Z}_n - \boldsymbol{c}\right\| > \delta\right) \leq \frac{\varepsilon}{6\left(\left\|f\right\|_\infty + 1\right)} \text{ for all } n \geq n_0$$

Since $\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{Y}$ and $f(\cdot, \boldsymbol{c})$ is bounded and continuous, we can also choose $n_1 \in \mathbb{N}$ such that

$$\left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{c}\right\} - \mathbb{E}\left\{f(\boldsymbol{Y}, \boldsymbol{c})\right\}\right| < \frac{\varepsilon}{3} \text{ for all } n \geq n_1$$

We then have that for $n = \max\{n_0, n_1\}$

$$\begin{aligned}
&\left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{Z}_n)\right\} - \mathbb{E}\left\{f(\boldsymbol{Y}, \boldsymbol{c})\right\}\right| \\
&\quad = \left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{Z}_n)\right\} - \mathbb{E}\left\{f(\boldsymbol{Y}, \boldsymbol{c})\right\} - \mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{c})\right\} + \mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{c})\right\}\right| \\
&\quad = \left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{Z}_n) - f(\boldsymbol{Y}_n, \boldsymbol{c})\right\} - \mathbb{E}\left\{f(\boldsymbol{Y}, \boldsymbol{c})\right\} + \mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{c})\right\}\right| \\
&\quad \leq \left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{Z}_n) - f(\boldsymbol{Y}_n, \boldsymbol{c})\right\} \mathbb{I}_{\{\|\boldsymbol{Z}_n - c\| \leq \delta\}}\right| \\
&\qquad\qquad + \left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{Z}_n) - f(\boldsymbol{Y}_n, \boldsymbol{c})\right\} \mathbb{I}_{\{\|\boldsymbol{Z}_n - c\| > \delta\}}\right| \\
&\qquad\qquad + \left|\mathbb{E}\left\{f(\boldsymbol{Y}_n, \boldsymbol{c})\right\} - \mathbb{E}\left\{f(\boldsymbol{Y}, \boldsymbol{c})\right\}\right|
\end{aligned}$$

As required. ■

---

[5] Lipschitz continuity is a smoothness condition for functions which is stronger than regular continuity; it limits how fast a function can change. If a function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L, then

$$\frac{\left|f(x_1) - f(x_2)\right|}{\left|x_1 - x_2\right|} \leq L \qquad \forall x_1 \neq x_2$$

Intuitively, a line joining any two points on the graph of f will never have a slop steeper than L; so there is a double white cone whose vertex can be translated along the graph so that the graph always remains entirely outside the code.

[6] For a bounded function f defined over the set S, $\left\|f\right\|_\infty = \left\|f\right\|_{\infty, S} = \sup\left\{\left|f(x)\right| : x \in S\right\}$.

© Daniel Guetta, 2010

Linear Models

Introduction and MLEs

A general linear model in which we make ***n* observations** and we want to fit ***p* parameters** takes the following form:

$$\underbrace{\boldsymbol{Y}}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1}$$

Where

- X is an $n \times p$ matrix of known **covariates** or **explanatory variables** (with the **first column** consisting of 1s, for the intercept). Also called the **design matrix**.

  It is usual to assume that this matrix has **full rank** so that $X^T X$ is **positive definite**[7].
- $\boldsymbol{Y}$ is an $n \times 1$ vector of **observations** or **response variables**.
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of IID **errors** $\boldsymbol{\varepsilon} \sim N_n\left(0, \sigma^2\right)$.
- $\boldsymbol{\beta}$ is an unknown **vector** of **regression coefficients**.

$\boldsymbol{Y}$ is a linear combination of constants and normal variables, and in fact

$$\boldsymbol{Y} \sim N_n\left(X\boldsymbol{\beta}, \sigma^2 \mathbb{I}\right)$$

In this parametric model, the unknown parameter is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)^T$. The **maximum likelihood estimators** of $\boldsymbol{\beta}$ and σ can be worked out[8]

---

[7] If X had linearly dependent rows, there would be a vector $\boldsymbol{z}$ such that $\boldsymbol{z}X^T X\boldsymbol{z} = ||X\boldsymbol{z}||^2 = 0$, which would imply $X^T X$ were not positive definite.

[8] …by maximising, with respect to both $\boldsymbol{\beta}$ and σ, the loglikelihood

$$\ell\left(\boldsymbol{Y}; \boldsymbol{\beta}, \sigma^2\right) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\|\boldsymbol{Y} - X\boldsymbol{\beta}\right\|^2 + \text{constants}$$

Specifically,

- In the $\boldsymbol{\beta}$ case, this amounts to minimising the following, with respect to $\boldsymbol{\beta}$

  $$\left\|\boldsymbol{Y} - X\boldsymbol{\beta}\right\| = \left(\boldsymbol{Y} - X\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y} - X\boldsymbol{\beta}\right) = \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T\boldsymbol{Y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$

  Differentiate using $\frac{\partial(\boldsymbol{x}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{\partial(\boldsymbol{\beta}\boldsymbol{x})}{\partial\boldsymbol{\beta}} = \boldsymbol{x}$, $\frac{\partial(\boldsymbol{x}\boldsymbol{\beta}^T)}{\partial\boldsymbol{\beta}} = \frac{\partial(\boldsymbol{\beta}^T\boldsymbol{x})}{\partial\boldsymbol{\beta}} = \boldsymbol{x}^T$ and $\frac{\partial(\boldsymbol{\beta}^T A\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \boldsymbol{\beta}^T\left(A^T + A\right)$, then transpose the lot. For distribution, note that $\boldsymbol{\beta}$ is a linear transformation of $\boldsymbol{Y}$.
- In the σ^2 case, it's a straight differentiation with respect to σ^2 . σ is **not** a linear transformation of $\boldsymbol{Y}$, and so the distribution requires Cochran's Theorem – start with

© Daniel Guetta, 2010

$$\boxed{\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \boldsymbol{Y} \sim N_p\left(\boldsymbol{\beta}, \left(X^T X\right)^{-1} \sigma^2\right)} \qquad \boxed{\hat{\sigma}^2 = \frac{1}{n}\left\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\right\|^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}}$$

And these are **independent of each other**[9]. (Note that the $\left\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\right\|^2$ term in the expression for $\hat{\sigma}^2$ is just the residual sum of squares, RSS $= \hat{\varepsilon}^T\hat{\varepsilon}$, and that it can also be written as $\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 = \left\|\left(I - H\right)\boldsymbol{Y}\right\|^2$).

Note that

- $\hat{\boldsymbol{\beta}}$ is an **unbiased estimator**, but its different components may be correlated, since the covariance matrix is not, in general, diagonal.

$$\hat{\sigma}^2 = \frac{1}{n}\left\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\right\|^2 = \frac{1}{n}\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 = \frac{1}{n}\left\|\left(I - H\right)\boldsymbol{Y}\right\|^2 = \frac{1}{n}\boldsymbol{Y}^T\left(I - H\right)^T\left(I - H\right)\boldsymbol{Y}$$
$$= \frac{1}{n}\boldsymbol{Y}^T\left(I - H\right)\boldsymbol{Y} = \frac{1}{n}\left(\boldsymbol{Y} - X\boldsymbol{\beta}\right)^T\left(I - H\right)\left(\boldsymbol{Y} - X\boldsymbol{\beta}\right)$$

(The last step is non-trivial and requires some working – start with the result and work backwards). We also have that $I = H + \left(I - H\right)$ with both H and $I - H$ symmetric. All we now need to show is that $\text{rank}(H) + \text{rank}(I - H) = n$. We do that by noting that $H^2 = H$ which means that all the eigenvalues of H are 0 or 1, and we can find Q such that $Q^T Q = I$ and $Q^T H Q = D$ where D is diagonal and contains only 0s or 1s. Thus

$$\text{rank}(H) = \text{rank}(D) = \text{tr}(D) = \text{tr}(DQ^TQ) = \text{tr}(QDQ^T) = \text{tr}(H) = \text{tr}\overbrace{\left((X^TX)^{-1}X^TX\right)}^{I_p} = p$$

And $I_n - H = I_n - QDQ^T = Q(I_n - D)Q^T$, so

$$\text{rank}(I_n - H) = \text{rank}\left(I_n - D\right) = n - p$$

[9] From the penultimate line in the previous footnote,

$$(I_n - H) = Q\left(I_n - D\right)Q^T = Q\begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix}Q^T = LL^T \qquad\qquad L = Q\begin{pmatrix} I_{n-p} \\ 0_{p,n-p} \end{pmatrix}$$

Now consider the vector $\boldsymbol{T} = \begin{pmatrix} B \\ L^T \end{pmatrix}\boldsymbol{Y}$ with $B = \left(X^TX\right)^{-1}X^T$. Note that $B\boldsymbol{Y} = \hat{\boldsymbol{\beta}}$ and that $\hat{\sigma}^2 = Y^T\left(I - H\right)Y = Y^TLL^TY = \left\|L^TY\right\|^2$. So if we can show the two components of the vector are independent, we have shown what we wanted. Consider the covariance matrix of $\boldsymbol{T}$.

$$\Sigma = \sigma^2\begin{pmatrix} B \\ L^T \end{pmatrix}\begin{pmatrix} B^T & L \end{pmatrix} = \sigma^2\begin{pmatrix} BB^T & BL \\ L^TB^T & L^TL \end{pmatrix}$$

And finally, note that

$$L^TL = \begin{pmatrix} I_{n-p} & 0_{n-p,p} \end{pmatrix}Q^TQ\begin{pmatrix} I_{n-p} \\ 0_{p,n-p} \end{pmatrix} = I_{n-p}$$
$$BL = BLL^TL = B\left(I - H\right)L = X\left(X^TX\right)^{-1}X^T\left(I - X\left(X^TX\right)^{-1}X^T\right)L = 0$$

The covariance matrix is therefore diagonal, and so the components are independent.

© Daniel Guetta, 2010

- $\mathbb{E}\left(\hat{\sigma}^2\right) = \frac{n-p}{n}\sigma^2$, and so $\hat{\sigma}^2$ is a slightly biased estimator. An unbiased estimator for σ^2 is

$$\boxed{s^2 = \frac{1}{n-p}\left|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\right|^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}}$$

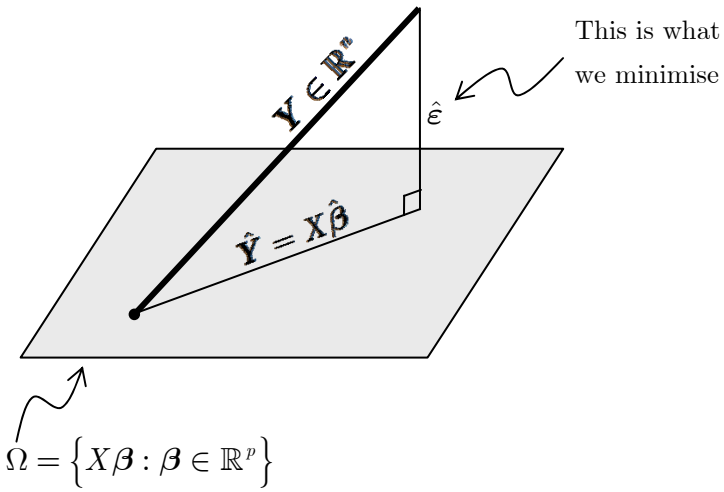
  This is the estimator to use in working out the covariance of $\boldsymbol{\beta}$, which is therefore given by $\operatorname{cov}\left(\hat{\boldsymbol{\beta}}\right) = \left(X^T X\right)^{-1} s^2$.

Geometrical interpretation

The **vector of fitted values** contains our best guess of what $\boldsymbol{Y}$ should be and is given by

$$\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}} = X\left(X^T X\right)^{-1} X^T \boldsymbol{Y} = H\boldsymbol{Y}$$

We note that H is **symmetric** and **idempotent** ($HH = H$, which also implies that $(I - H)(I - H) = I - H$). These are the properties of a **projection matrix**; indeed H **projects** $\boldsymbol{Y}$ onto the space spanned by the columns of X, which contains all $X\boldsymbol{\beta}$. Thus, H effectively finds the closest fit in $\Omega = \left\{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\right\}$ for $\boldsymbol{Y}$. (The residuals are given by $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \left(I - H\right)\boldsymbol{Y} \sim N_n\left(0, (I - H)\sigma^2\right)$):



Hypothesis tests on $\boldsymbol{\beta}$

For each component of $\hat{\boldsymbol{\beta}}$, we have that

- $\operatorname{var}\left(\hat{\beta}_i\right) = \sigma^2 (X^T X)^{-1}_{ii}$
- Our best estimator of this is $\operatorname{se}\left(\hat{\beta}_i\right) = \sqrt{s^2 (X^T X)^{-1}_{ii}}$

© Daniel Guetta, 2010

So:

$$\begin{aligned}\frac{\hat{\beta}_i}{\text{se}\left(\hat{\beta}_i\right)} &= \frac{\hat{\beta}_i}{\sqrt{s^2(X^TX)^{-1}_{ii}}} \\ &= \frac{\hat{\beta}_i \,/\, \sqrt{\sigma^2(X^TX)^{-1}_{ii}}}{\sqrt{\frac{n-p}{n-p}}\sqrt{s^2(X^TX)^{-1}_{ii}} \,/\, \sqrt{\sigma^2(X^TX)^{-1}_{ii}}} \\ &= \frac{\hat{\beta}_i \,/\, \sqrt{\sigma^2(X^TX)^{-1}_{ii}}}{\sqrt{\frac{1}{n-p}\frac{s^2(n-p)}{\sigma^2}}}\end{aligned}$$

Since $\hat{\beta}_i \sim N\left(\beta_i, \sigma^2(X^TX)^{-1}_{ii}\right)$ and $s^2 = \frac{\sigma^2}{n-p}\chi^2_{n-p}$, we can write

$$\sim \frac{N\left(\beta_i, 1\right)}{\sqrt{\frac{1}{n-p}\chi^2_{n-p}}}$$

If $\beta_i = 0$, then this ratio is a t distribution with $n - p$ degrees of freedom:

$$H_0 : \mathbb{P}_{\beta_i=0,\sigma^2}\left(\frac{\hat{\beta}_i}{\text{se}\left(\hat{\beta}_i\right)} \le t_{n-p}(\alpha)\right) = 1-\alpha$$

So a $(1-\alpha)$-level confidence set for β_i **assuming** $\beta_i = 0$ is

$$\boxed{\left\{\beta_i \in \mathbb{R} : \frac{\hat{\beta}_i}{\text{se}\left(\hat{\beta}_i\right)} \le t_{n-p}(\alpha)\right\}}$$

We can also find a confidence set for $\boldsymbol{\beta}$. First, notice that

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim N_p(0, (X^TX)^{-1}\sigma^2)$$

this implies that[10]

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T X^TX\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \sigma^2\chi^2_p$$

we also know that

$$s^2 = \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

This implies that the ratio of these two quantities has an F distribution:

$$\mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\frac{\frac{1}{p}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T X^TX\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{s^2} \le F_{p,n-p}(\alpha)\right) = 1-\alpha$$

So a $(1-\alpha)$-level confidence set for $\boldsymbol{\beta}$ is

[10] If $\boldsymbol{X} \sim N(0,\Sigma)$, then $\Sigma^{-1/2}\boldsymbol{X} \sim N(0,I)$ and so $\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X} \sim \chi^2$. In this case, $\Sigma = \sigma^2(X^TX)^{-1}$.

© Daniel Guetta, 2010

$$\boxed{\left\{\boldsymbol{\beta}\in\mathbb{R}^p : \frac{\frac{1}{p}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T X^T X\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}{s^2} \le F_{p,n-p}(\alpha)\right\}}$$

[Note: imagine $p=1$, so that we only have 1 β to test, and that we're testing whether $\beta=0$, this becomes

$$\left\{\beta\in\mathbb{R}^p : \frac{X^T X\hat{\beta}^2}{s^2} \le F_{1,n-1}(\alpha)\right\} \Rightarrow \left\{\beta\in\mathbb{R}^p : \frac{\hat{\beta}^2}{s^2(X^TX)^{-1}} \le F_{1,n-1}(\alpha)\right\}$$

it so happens that $F_{1,n-1}(\alpha)=t^2_{n-1}(\alpha)$, so this is precisely a t-test].

Diagnostic Tests

A number of diagnostic tests can be carried out to check whether a model we have fitted is appropriate:

- We can check whether a larger model is needed (ie: more columns of X) by adding extra terms (like x^2) and seeing whether they're needed using the t-test above.
- We can use the fact that $\hat{\varepsilon}$ and $\hat{\boldsymbol{Y}}$ are uncorrelated under the MLE fit[11], as are $\hat{\varepsilon}$ and $\hat{\boldsymbol{\beta}}$. This means that a plot of errors against fitted values should show no systematic relationship.

  If it did, then a transformation of the Y_i might make sense. To understand why, imagine a situation in which the variance and mean of Y *are* correlated, and a "well-behaved" transformation $f(\cdot)$

$$Y_i \sim N\left(\mu,\sigma^2(\mu)\right)$$
$$f(Y)=f(\mu)+\left(Y-\mu\right)f'\left(\mu\right) \overset{\text{approx}}{\sim} N\left(f(\mu),\sigma^2(\mu)\left[f'(\mu)\right]^2\right)$$

  We can then choose an f such that $\sigma^2(\mu)\left[f'(\mu)\right]^2$ is approximately constant. We try to choose f to **stabilise the variance**. A standard example is to try $\log Y_i=\alpha+\boldsymbol{\beta}\boldsymbol{x}_i+\varepsilon$.

[11] This follows from the fact that

$$\operatorname{cov}\left(\hat{\varepsilon},\hat{\boldsymbol{Y}}\right)=\operatorname{cov}\left(\left(I-H\right)\boldsymbol{Y},H\boldsymbol{Y}\right)=\left(I-H\right)H\operatorname{cov}\left(\boldsymbol{Y},\boldsymbol{Y}\right)=0$$

and the fact that the variables are normally distributed.

© Daniel Guetta, 2010

Another example is to use a **Box-Cox transformation**, in which each y is transformed to

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

This family of transformations combines power and log transformations, is parameterised by λ and is continuous in λ. We wish to find the "best λ". If we have $\mathbb{E}\left(Y^{(\lambda)}\right) = X\beta$, then the likelihood of λ given the observed Y is[12]

$$\frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}^{(\lambda)} - X\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}^{(\lambda)} - X\boldsymbol{\beta}\right)\right)\prod_{i=1}^{n} Y_i^{\lambda-1}$$

Fitting the linear model gives $\hat{\boldsymbol{\beta}}(\lambda) = \left(X^T X\right)^{-1} X^T \boldsymbol{Y}^{(\lambda)}$ and $\hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{Y}^{(\lambda)}\right)^T\left(I - H\right)\boldsymbol{Y}^{(\lambda)} = RSS(\lambda)/n$, and the log-likelihood becomes

$$L(\lambda) = \text{const} - \tfrac{n}{2}\log\sigma^2 + (\lambda - 1)\sum_{i=1}^{n}\log\left(Y_i\right)$$

$$L(\lambda) = \text{const} - \tfrac{n}{2}\log\left(RSS(\lambda)/n\right) + (\lambda - 1)\sum_{i=1}^{n}\log\left(Y_i\right)$$

We can plot this function and pick an MLE $\hat{\lambda}$ that is closest to its maximum. The corresponding likelihood ratio test rejects the hypothesis $H_0 : \lambda = \lambda_0$ if

$$2\left(L\left(\hat{\lambda}\right) - L\left(\lambda_0\right)\right) > \chi_1^2(\alpha)$$

Where $\chi_1^2(\alpha)$ is the upper $100\alpha\%$ point of a χ_1^2-distribution.

- Our model assumes that $\mathbb{P}\left(\varepsilon_i \leq x\right) = \Phi\left(x/\sigma\right)$. We can approximate this by the **empirical distribution function** of the residuals

$$\hat{F}_n = \frac{\#\left\{i : \hat{\varepsilon}_i \leq x\right\}}{n}$$

If our assumptions are acceptable, then

$$\Phi^{-1}\left(\hat{F}_n(x)\right) \approx \frac{x}{\sigma}$$

A **Q-Q Plot** $\hat{\varepsilon}_i$ against $\Phi^{-1}\left(\hat{F}_n(\varepsilon_i)\right)$. If everything is as it should be, the points should, roughly, lie on a straight line.

[12] Note that if $u = u(x)$, then $P_u\,\mathrm{d}u = P_x\,\mathrm{d}x$, and so $P_u = P_x\,\mathrm{d}x/\mathrm{d}u$, where the last term is called the **Jacobian**. This is where the last term comes from.

© Daniel Guetta, 2010

We note that since $\text{var}\left(\hat{\varepsilon}_i\right) = \left(1 - H_{ii}\right)\sigma^2$, we could also use $\hat{\varepsilon}_i / \sigma\sqrt{1 - H_{ii}}$ or $\hat{\varepsilon}_i / s\sqrt{1 - H_{ii}}$, to ensure all variables has variance approximately equal to 1.

- We can check if any points appear to be **influential points** (pints whose deletion causes a large change in the analysis) by calculating the **Cook Distance**, D_i for each point i. Let $\hat{\boldsymbol{Y}}_{(i)}$ be the fitted values from a model omitting point i. Then

$$D_i = \frac{\left(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)}\right)^T \left(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)}\right)}{s^2 p}$$

Where p is the number of parameters in β. Large values could indicate an influential observation.

Nested Models

Sometimes, we want to test whether *a number of* the β in our model are redundant. In other, we want to test the following hypotheses

$$\Omega : \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad\qquad \omega_1 : \boldsymbol{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \Rightarrow \boldsymbol{\beta}_2 = \boldsymbol{0}$$

Where

$$\underset{n \times \tilde{p}}{X} = \underset{n \times p_1}{X_1} \,\Big|\, \underset{n \times p_2}{X_2} \qquad\qquad \underset{\tilde{p} \times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \hline \boldsymbol{\beta}_2 \end{pmatrix} \begin{matrix} \swarrow\, p_1 \times 1 \\ \\ \nwarrow\, p_2 \times 1 \end{matrix}$$

[Note even though $\boldsymbol{\beta}$ partitions into $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, it is **not** always the case that $\hat{\boldsymbol{\beta}}$ partitions into $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. Similarly, $\hat{\sigma}$ is not necessarily equal to $\hat{\sigma}_1$].

The alternative, model, in fact, corresponds to a projection of $\boldsymbol{Y}$ onto a smaller subspace spanned by the columns of X_1 only, $\omega_1 = \left\{X_1\boldsymbol{\beta}_1 : \boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}\right\}$. $= \left\{H_{\omega_1}\boldsymbol{Y} : \boldsymbol{Y} \in \mathbb{R}^n\right\}$. The projection matrices are[13]

[13] It is useful to note that

$$\mathbb{R}^n = \overbrace{\underbrace{\omega_1 \oplus \left(\Omega \cap \omega_1^{\perp}\right)}_{\text{Partition } \Omega} \oplus \Omega^{\perp}}^{\text{Partition } \mathbb{R}^n}$$

To show this is true, consider the fact that any vector $\boldsymbol{y} \in \mathbb{R}^n$ may be written as

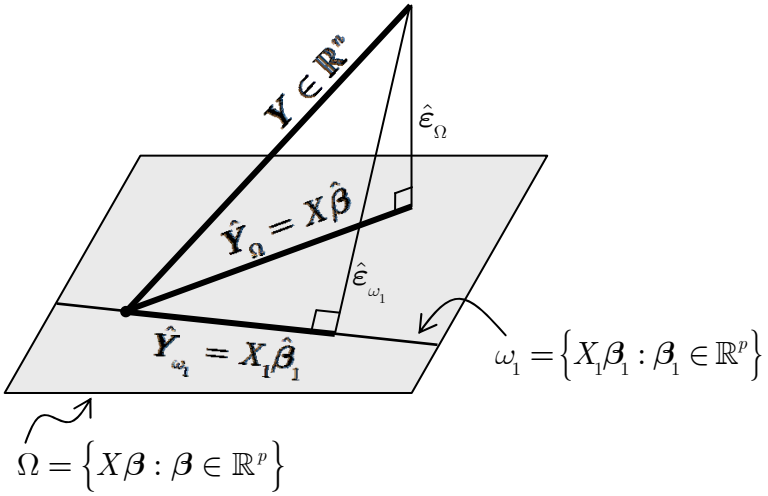
$$\boldsymbol{y} = I\boldsymbol{y} = \left(H_{\omega_1} + (H_{\Omega} - H_{\omega_1}) + (I - H_{\Omega})\right)\boldsymbol{y} = H_{\omega_1}\boldsymbol{y} + \left(H_{\Omega} - H_{\omega_1}\right)\boldsymbol{y} + \left(I - H_{\Omega}\right)\boldsymbol{y}$$

And that

© Daniel Guetta, 2010

$$H_{\omega_1} = X_1\left(X_1^T X_1\right)^{-1} X_1^T \qquad H_\Omega = X\left(X^T X\right)^{-1} X^T$$

$\omega_1 = \left\{X_1\boldsymbol{\beta}_1 : \boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}\right\} \subseteq \left\{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\right\} = \Omega$ because any vector in the LHS set can be obtained by setting some entries of $\boldsymbol{\beta}$ in the RHS set to 0.



Clearly, $\hat{\varepsilon}_{\omega_1} > \hat{\varepsilon}_\Omega$. What will determine whether to accept or reject ω_1 is exactly how much bigger $\hat{\varepsilon}_{\omega_1}$ is. If it's not much bigger, then we might as well drop all the extra superfluous parameters. We can formalise this in terms of a likelihood-ratio test. The log-likelihood is as usual in the linear model

---

- $\boxed{H_{\omega_1}\boldsymbol{y} \in \omega_1}$ (obviously!)
- $\boxed{\left(H_\Omega - H_{\omega_1}\right)\boldsymbol{y} \in \left(\Omega \cap \omega_1^\perp\right)}$. To see why, we must show it is both in Ω and $\omega_1^\perp$
  - Clearly $(H_\Omega - H_{\omega_1})\boldsymbol{y} \in \Omega$ because $H_\Omega\boldsymbol{y} \in \Omega$ and $H_{\omega_1}\boldsymbol{y} \in \omega_1 \subseteq \Omega$.
  - Consider a $\boldsymbol{z} \in \omega_1$. Then $[(H_\Omega - H_{\omega_1})\boldsymbol{y}]^T \cdot \boldsymbol{z} = \boldsymbol{y}^T(H_\Omega - H_{\omega_1})\boldsymbol{z} = \boldsymbol{y}^T(H_\Omega\boldsymbol{z} - H_{\omega_1}\boldsymbol{z}) = 0$, since $\boldsymbol{z} \in \omega_1 \subseteq \Omega \Rightarrow H_\Omega\boldsymbol{z} = H_{\omega_1}\boldsymbol{z} = \boldsymbol{z}$. Thus, $(H_\Omega - H_{\omega_1})\boldsymbol{y} \in \omega_1^\perp$.
- $\boxed{\left(I - H_\Omega\right)\boldsymbol{y} \in \Omega^\perp}$. Consider a $\boldsymbol{z} \in \Omega$. Then $\left[\left(I - H_\Omega\right)\boldsymbol{y}\right]^T \cdot \boldsymbol{z} = \boldsymbol{y}^T\left(z - H_\Omega\boldsymbol{z}\right) = 0$, since $\boldsymbol{z} \in \Omega \Rightarrow H_\Omega\boldsymbol{z} = \boldsymbol{z}$.

Furthermore, we note that

- $\boxed{\omega_1 \perp \left(\Omega \cap \omega_1^\perp\right)}$, because $\left(\Omega \cap \omega_1^\perp\right) \subseteq \omega_1^\perp$
- $\boxed{\left(\Omega \cap \omega_1^\perp\right) \perp \Omega^\perp}$, because $\left(\Omega \cap \omega_1^\perp\right) \subseteq \Omega$
- $\boxed{\omega_1 \perp \Omega^\perp}$, because $\omega_1 \subseteq \Omega$

© Daniel Guetta, 2010

$$\begin{aligned}\ell\left(\hat{\boldsymbol{\beta}},\hat{\sigma}^2\right) &= -\frac{n}{2}\log\hat{\sigma}^2 - \frac{1}{2\sigma^2}\left\|\boldsymbol{Y}-X\hat{\boldsymbol{\beta}}\right\|^2 \\ &= -\frac{n}{2}\log\left(\frac{1}{n}\left\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right\|^2\right) - \frac{1}{2\sigma^2}n\sigma^2 \\ &= -\frac{n}{2}\log\left(\frac{1}{n}\left\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right\|^2\right) - \frac{n}{2}\end{aligned}$$

And the log-likelihood ratio statistic for ω_1 is

$$\begin{aligned}W\left(\omega_1\right) &= 2\left\{\ell\left(\hat{\boldsymbol{\beta}},\hat{\sigma}^2_{\Omega}\right) - \ell\left(\hat{\boldsymbol{\beta}}_1,\hat{\sigma}^2_{\omega_1}\right)\right\} \\ &= 2\left\{-\frac{n}{2}\log\left(\frac{1}{n}\left\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\Omega}\right\|^2\right) - \frac{n}{2} + \frac{n}{2}\log\left(\frac{1}{n}\left\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\omega_1}\right\|^2\right) + \frac{n}{2}\right\} \\ &= n\log\left(\varepsilon^2_{\omega_1}\Big/\varepsilon^2_{\Omega}\right)\end{aligned}$$

By looking at the diagram above and using Pythagoras' Theorem[14] we get

$$\begin{aligned}\varepsilon^2_{\omega_1} &= \varepsilon^2_{\Omega} + \left\|\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega_1}\right\|^2 \\ \Rightarrow \left\|\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega_1}\right\|^2 &= \varepsilon^2_{\omega_1} - \varepsilon^2_{\Omega}\end{aligned}$$

This allows us to write

$$\frac{\varepsilon^2_{\omega_1}}{\varepsilon^2_{\Omega}} = 1 - \frac{\left\|\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega_1}\right\|^2}{\varepsilon^2_{\Omega}}$$

Now, by the extension of Cochran's Theorem[15]:

---

[14] And the fact that ε^T_{Ω} and $\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega}$ are perpendicular, because

$$\begin{aligned}\varepsilon^T_{\Omega}\left(\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega_1}\right) &= \left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\Omega}\right)^T\left(\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega_1}\right) = \boldsymbol{Y}^T\left(I-H_{\Omega}\right)\left(H_{\Omega}-H_{\omega_1}\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}^T\left(H_{\Omega}-H_{\omega_1}-H_{\Omega}+H_{\Omega}H_{\omega_1}\right)\boldsymbol{Y} = \boldsymbol{Y}^T\left(H_{\Omega}H_{\omega_1}\boldsymbol{Y}-H_{\omega_1}\boldsymbol{Y}\right) = \boldsymbol{0}\end{aligned}$$

In the last step, we use the fact that $H_{\omega_1}\boldsymbol{Y}\in\omega_1\subseteq\Omega \Rightarrow H_{\Omega}H_{\omega_1}\boldsymbol{Y} = H_{\omega_1}\boldsymbol{Y}$.

[15] First note that $\boldsymbol{Y}\sim N\left(X\boldsymbol{\beta},\sigma^2 I\right)$. Consider that

$$\begin{aligned}\left\|\boldsymbol{Y}\right\|^2 &= \left\|H_{\omega_1}\boldsymbol{Y}\right\|^2 + \left\|\left(H_{\Omega}-H_{\omega_1}\right)\boldsymbol{Y}\right\|^2 + \left\|\left(I-H_{\Omega}\right)\boldsymbol{Y}\right\|^2 \\ &= \left\|H_{\omega_1}\boldsymbol{Y}\right\|^2 + \left\|\hat{\boldsymbol{Y}}_{\Omega}-\hat{\boldsymbol{Y}}_{\omega_1}\right\|^2 + \varepsilon^2_{\Omega} \\ &= \boldsymbol{Y}^T H_{\omega_1}\boldsymbol{Y} + \boldsymbol{Y}^T\left(H_{\Omega}-H_{\omega_1}\right)\boldsymbol{Y} + \boldsymbol{Y}^T\left(I-H_{\Omega}\right)\boldsymbol{Y}\end{aligned}$$

Note that:

- $I = H_{\omega_1} + \left(H_{\Omega}-H_{\omega_1}\right) + \left(I-H_{\Omega}\right)$
- $\dim\left(\omega_1\right) = p_1$ and $\dim\left(\Omega\right) = \tilde{p}$, so $\dim\left(\Omega^{\perp}\right) = n-\tilde{p}$ and $\dim\left(\Omega\cap\omega_1^{\perp}\right) = \tilde{p}-p_1$.

  Therefore, $\operatorname{rank}(H_{\omega_1}) = p_1$, $\operatorname{rank}(H_{\Omega}-H_{\omega_1}) = \tilde{p}-p_1$ and $\operatorname{rank}(I-H_{\Omega}) = n-\tilde{p}$.

We can therefore apply the extension to Cochran's Theorem to find the second two terms in the expansion above. The relevant non-centrality parameters δ are (note that $X\boldsymbol{\beta}\in\Omega$):

$$\sigma^2\delta_{\varepsilon^2_{\Omega}} = \boldsymbol{\beta}^T X^T\left(I-H_{\Omega}\right)X\boldsymbol{\beta} = \boldsymbol{\beta}^T X^T\left(X\boldsymbol{\beta}-X\boldsymbol{\beta}\right) = 0$$

© Daniel Guetta, 2010

- $\varepsilon_{\Omega}^2 \sim \sigma^2 \chi_{n-\tilde{p}}^2$
- $\left\| \hat{\boldsymbol{Y}}_{\Omega} - \hat{\boldsymbol{Y}}_{\omega_1} \right\|^2 \sim \sigma^2 \chi_{\tilde{p}-p_1}^2 \left(\frac{1}{\sigma^2} \left\| (I - H_{\omega_1}) X\beta \right\|^2 \right)$.

Therefore

$$\boxed{F = \frac{\frac{1}{\tilde{p}-p_1} \left\| \hat{\boldsymbol{Y}}_{\Omega} - \hat{\boldsymbol{Y}}_{\omega_1} \right\|^2}{\frac{1}{n-\tilde{p}} \varepsilon_{\Omega}^2} = \frac{\frac{1}{\tilde{p}-p_1} \left(\varepsilon_{\omega_1}^2 - \varepsilon_{\Omega}^2 \right)}{\frac{1}{n-\tilde{p}} \varepsilon_{\Omega}^2} \sim F_{\tilde{p}-p_1, n-\tilde{p}} \left(\frac{1}{\sigma^2} \left\| (I - H_{\omega_1}) X\beta \right\|^2 \right)}$$

Under ω_1, $X\beta$ lies in ω_1 already and so $(I - H_{\omega_1})X\beta = (X\beta - X\beta) = 0$ and so we expect a standard ***F*-distribution**.

Intuitively, the F statistic is as follows

$$F = \frac{\frac{1}{\tilde{p}-p_1} \left(\varepsilon_{\omega_1}^2 - \varepsilon_{\Omega}^2 \right)}{\frac{1}{n-\tilde{p}} \varepsilon_{\Omega}^2}$$

Error per extra degree of freedom added by going from $\omega_1 \to \Omega$

Error per degree of freedom present in Ω anyway

Under Ω, this will be large, because all the components of $\boldsymbol{\beta}$ are important; indeed, we get a non-central F distribution, which is stochastically increasing in the non-centrality parameter. Thus, we **reject ω_1 for large values of *F***.

The data used to carry out F tests is often summarised in a table. Consider the following nested models:

ω : $\boldsymbol{Y} = \mu\boldsymbol{1}$	[ie: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = 0$]	1
ω_1 : $\boldsymbol{Y} = \mu\boldsymbol{1} + X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$	[ie: $\boldsymbol{\beta}_2 = 0$]	$p_1 + 1$
Ω : $\boldsymbol{Y} = \mu\boldsymbol{1} + X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \varepsilon$		$p_1 + p_2 + 1$

Our table is then

Source	**Degrees of freedom**	**Sum of squares**	**Mean sum of squares**	***F*-value**
Due to β_1 [ie: $\omega \to \omega_1$]	p_1	$SS_{\beta_1 \mid \mu} = \varepsilon_{\omega}^2 - \varepsilon_{\omega_1}^2$	$S_{\beta_1 \mid \mu} \,/\, p_1$	$\frac{SS_{\beta_1 \mid \mu} / p_1}{\varepsilon_{\Omega}^2 / (n - p_1 - p_2 - 1)}$
Due to β_2 [ie: $\omega_1 \to \Omega$]	p_2	$SS_{\beta_2 \mid \mu, \beta_1} = \varepsilon_{\omega_1}^2 - \varepsilon_{\Omega}^2$	$S_{\beta_2 \mid \mu, \beta_1} \,/\, p_2$	$\frac{SS_{\beta_2 \mid \mu, \beta_1} / p_2}{\varepsilon_{\Omega}^2 / (n - p_1 - p_2 - 1)}$
Residual [in Ω anyway]	$n - p_1 - p_2 - 1$	ε_{Ω}^2	$\frac{\varepsilon_{\Omega}^2}{n - p_1 - p_2 - 1}$	
Total	$n - 1$	$\varepsilon_{\omega}^2 = \sum \left(Y_i - \bar{Y} \right)^2$		

$$\sigma^2 \delta_{\left\| \hat{\boldsymbol{Y}}_{\Omega} - \hat{\boldsymbol{Y}}_{\omega_1} \right\|^2} = \boldsymbol{\beta}^T X^T \left(H_{\Omega} - H_{\omega_1} \right) X\boldsymbol{\beta} = \boldsymbol{\beta}^T X^T \left(I - H_{\omega_1} \right) X\boldsymbol{\beta} = \left\| \left(I - H_{\omega_1} \right) X\boldsymbol{\beta} \right\|$$

© Daniel Guetta, 2010

Note that the expected value of each of the F values in the table above is

$$\frac{n-p_1-p_2-1}{\left(n-p_1-p_2-1\right)-2}$$

What we're effectively doing with this table is

- Taking the **total variation in** $\boldsymbol{Y}$, given by $\varepsilon_\omega^2 = \sum\left(Y_i - \bar{Y}\right)^2$
- Seeing how much of it remains unexplained even in our full model ε_Ω^2
- Seeing how much we "lose out" each time we reduce the size of our model.

[Note, of course, that unless only factor is present, each of these F values will be different to the F value testing the drop in variance from $\omega \to \Omega$].

As we mentioned above, even though $\boldsymbol{\beta}$ partitions into $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, it is **not** always the case that $\hat{\boldsymbol{\beta}}$ partitions into $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. This adds extra complications when calculating sum-of-squares, in that $SS_{\boldsymbol{\beta}_2|\boldsymbol{\beta}_1} = SS_{\boldsymbol{\beta}_2}$.

To investigate, consider the following models

$$\begin{aligned}
\omega &: \boldsymbol{Y} = \boldsymbol{\varepsilon} \\
\omega_1 &: \boldsymbol{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \\
\omega_2 &: \boldsymbol{Y} = X_2\boldsymbol{\beta}_2 + \varepsilon \\
\Omega &: \boldsymbol{Y} = X\boldsymbol{\beta} + \varepsilon
\end{aligned} \qquad X = \left(X_1 \,\middle|\, X_2\right) \qquad \boldsymbol{\beta} = \left(\frac{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}\right)$$

We say that the models are **orthogonal** if $X_1^T X_2 = 0$.

Let's first consider the MLEs $\hat{\boldsymbol{\beta}}_i$. If the models are orthogonal, then

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(X^T X\right)^{-1} X^T \boldsymbol{Y} = \left(\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix}\begin{pmatrix} X_1 & X_2 \end{pmatrix}\right)^{-1}\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix}\boldsymbol{Y} = \begin{pmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{pmatrix}^{-1}\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix}\boldsymbol{Y} \\
&= \begin{pmatrix} (X_1^T X_1)^{-1} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{pmatrix}\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix}\boldsymbol{Y} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}
\end{aligned}$$

Thus, the coefficients partition neatly if and only if the models are orthogonal.

Next, consider that

© Daniel Guetta, 2010

$$\begin{aligned} SS_{\beta_2|\beta_1} &= \varepsilon_{\omega_1}^2 - \varepsilon_\Omega^2 \\ &= \boldsymbol{Y}^T\left(I - H_1\right)\boldsymbol{Y} - \boldsymbol{Y}^T\left(I - H\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}^T\left(H - H_1\right)\boldsymbol{Y} \\ SS_{\beta_2} &= \varepsilon_\omega^2 - \varepsilon_{\omega_2}^2 \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\left(I - H_2\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}^T H_2 \boldsymbol{Y} \end{aligned}$$

If the models are orthogonal, then by the calculation above

$$H = X\left(X^TX\right)^{-1}X^T = \begin{pmatrix} X_1 & X_2 \end{pmatrix}\begin{pmatrix} (X_1^TX_1)^{-1} & 0 \\ 0 & (X_2^TX_2)^{-1} \end{pmatrix}\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix} = H_1 + H_2$$

So the above becomes

$$SS_{\beta_2|\beta_1} = \boldsymbol{Y}^T H_2 \boldsymbol{Y} = SS_{\beta_2}$$

Thus, the order in which parameters are added to a model doesn't matter if and only if the models are orthogonal.

Coefficient of Determination

Consider the two models

$$\begin{aligned} \omega &: \boldsymbol{Y} = \mu\boldsymbol{1} \\ \Omega &: \boldsymbol{Y} = \mu\boldsymbol{1} + X\boldsymbol{\beta} + \varepsilon \end{aligned}$$

Clearly,

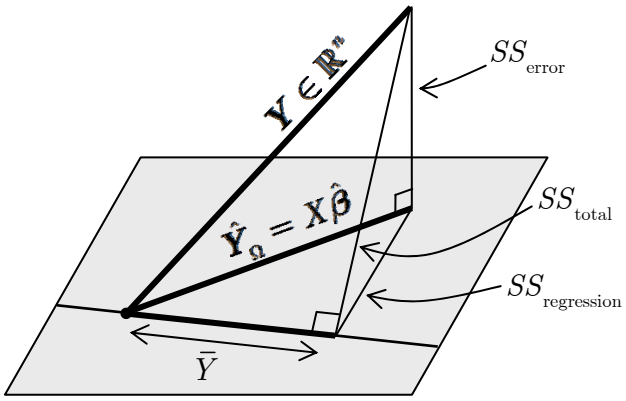
- $\varepsilon_\omega^2 = \left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\omega\right\|^2 = \left\|\boldsymbol{Y} - \bar{Y}\boldsymbol{1}\right\|^2 = SS_{\text{total}}$ is the **total variability in the sample**.
- $\varepsilon_\Omega^2 = \left\|\hat{\boldsymbol{Y}}_\Omega - \boldsymbol{Y}\right\|^2 = SS_{\text{error}}$ is the **variability not explained by** Ω.
- $\left\|\hat{\boldsymbol{Y}}_\Omega - \bar{Y}\boldsymbol{1}\right\|^2 = \left\|\hat{\boldsymbol{Y}}_\Omega - \hat{\boldsymbol{Y}}_\omega\right\|^2 = SS_{\text{regression}}$ is the variability that **is** explained by Ω.

We know, however, from the previous section that $\left\|\hat{\boldsymbol{Y}}_\Omega - \hat{\boldsymbol{Y}}_\omega\right\|^2 = \varepsilon_\omega^2 - \varepsilon_\Omega^2$, and this implies that[16] the total variability can be split as follows:

$$SS_{\text{total}} = SS_{\text{(Explained by) regression}} + SS_{\text{(Explained by) error}}$$

[16] Reminder: because $SS_{\text{total}} = \left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\omega\right\|^2 = \left\|\left(\hat{\boldsymbol{Y}}_\Omega - \hat{\boldsymbol{Y}}_\omega\right) + \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\right)\right\|^2 = \left\|\left(\hat{\boldsymbol{Y}}_\Omega - \hat{\boldsymbol{Y}}_\omega\right) + \varepsilon_\Omega\right\|^2$, the first term squared is $SS_{\text{regression}}$, and see previous footnote for proof $\varepsilon_\Omega^T\left(\hat{\boldsymbol{Y}}_\Omega - \hat{\boldsymbol{Y}}_\omega\right) = 0$.

© Daniel Guetta, 2010



The **coefficient of determination**, $\boldsymbol{R^2}$ is given by the **ratio** between the **error explained by regression** and the **total error**:

$$\boxed{R^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}}}$$

It gives us an idea of how well our model explains the variability in the $\boldsymbol{Y}$.

ANOVA

ANOVA is a model for the analysis of categorical data, in which our observation depends on one or more discrete factors.

We first begin with a simple example in which one factor is involved. If Y_{ij} is the j^{th} observation at level i of the factor, our model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad\qquad \varepsilon_{ij} = \text{NID}\left(0, \sigma^2\right)$$

This model can be expressed in least squares form. Consider the example in which the factor can take 3 values, and each group contains 2 observations:

$$\boldsymbol{\beta} = \begin{pmatrix} \mu & \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix}^T$$

$$X = \begin{pmatrix} 1 & 1 & & \\ 1 & 1 & & \\ 1 & & 1 & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & 1 \end{pmatrix}$$

The easiest way to find the MLE is to use the equation $X^T X\boldsymbol{\beta} = X^T\boldsymbol{Y}$ directly. So in the case above, we get

© Daniel Guetta, 2010

$$\begin{pmatrix} 10 & 2 & 2 & 2 \\ 2 & 2 & & \\ 2 & & 2 & \\ 2 & & & 2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} Y_{++} \\ Y_{1+} \\ Y_{2+} \\ Y_{3+} \end{pmatrix}$$

This model is over-defined, because we can add a fixed amount to μ and remove that amount from every α. We therefore need to impose an additional constraint on the parameterisation. Two common choices are:

- **Sum-to-zero constraint**: $\boxed{\sum \alpha_i = 0}$. The MLE is then
$$\hat{\mu} = \bar{Y}_{\bullet\bullet} \qquad \hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$$
Using these constraints, every score is compared to the *overall* mean.
- **Corner-point constraint**: $\boxed{\alpha_1 = 0}$. The MLE is then
$$\hat{\mu} = \bar{Y}_{1\bullet} \qquad \hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{1\bullet}$$
Using these constraints, every score is compared to the mean of group 1.

(Where the $\bar{Y}$ imply an average of Y over $\bullet$ indices)

In **two-way ANOVA**, two factors are involved. If Y_{ijk} is the k^{th} observation at level i of factor A and level j of factor B, we can fit two models:

- Model with **no interaction between factors**
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad (*)$$
- Model **with interaction between factors**
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad (\#)$$
The corner point constraints for this model are $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$.

In the second model, the effect of changing factor A depends on the precise value of factor B.

This can also be extended to models mixing "factor" and "non-factor" in the same model. An example might be the weight W of a child depending on their age A (non-factor) and sex S (factor). Two possible models can then be fit:

- Model with **no interaction**
$$W_{ij} = \mu + \alpha_i + \beta A_{ij} + \varepsilon_{ij} \qquad \left[i \in \{\text{M},\text{F}\}, \alpha_{\text{M}} = 0\right]$$
- Model **with interaction**
$$W_{ij} = \mu + \alpha_i + \left(\beta + \gamma_i\right) A_{ij} + \varepsilon_{ij} \qquad \left[i \in \{\text{M},\text{F}\}, \alpha_{\text{M}} = \gamma_{\text{M}} = 0\right]$$

© Daniel Guetta, 2010

{**Proof of Cochran's Theorem**}

I have been asked to remove all notes from the Part III Statistical Theory class from this website. Please email me if you have any questions.

Computer Code

Carrying out regression

- Regressing one variable on another: **`linearModelName <- lm(yVariable~xVariable)`**
- To carry out the regression on two covariates, use **`linearModelName <- lm(yVariable~firstCovariate+secondCovariate)`**
- To fit orthogonal polynomials such as $P_0 = a_0$, $P_1 = a_1 + b_1 x$, $P_2 = a_2 + b_2 x + c_2 x^2$, etc…, use
  `linearModelName <- lm(yVariable~Poly(Covariate,2))`

Interpreting a regression

- Selected components of the variable `linearModelName`
  - `coefficients`
  - `residuals`
  - `fitted.values`

  Each of those can be extracted using the **`linearModelName$coefficients`**, for example.
- Seeing a summary of the linear model: **`summary(linearModelName)`**.
  - The "residual standard error" given in the output is s.
  - Remember that only $|t|$ counts; the sign matters not.

  Add **`cor=F`** as an argument to ommitt correlation of coefficients.
- **`deviance(linearModelName)`** gives the residual sum of squares. It can be used to carry out an F test directly.
- Carrying out successive F-tests on nested models in which each term in the regression is added sequentially: **`anova(linearModelName)`**
  - The last F statistic tested in the output tests the null model $\boldsymbol{Y} = \mu \boldsymbol{1} + \varepsilon_i$ against the full model.

Model selection (*see later for theory*)

- *Backwards elimination*
  - Fit successive models, find t values using **`summary`** removing the covariate with lowest $|t|$ each time.
  - **`library(MASS)`** followed by **`dropterm(linearModelName, test="F")`** gives the effect of dropping each variable in turn, including F values. These, however, are just the square of the t values given by **`summary`**.

© Daniel Guetta, 2010

- *Forward selection*
  - **library(MASS)** followed by **addterm(linearModelName, covariate1+covariate2)** gives the effect of adding each variable in turn to `linearModelName`. Choose the one with the largest reduction in RSS each time.
- *Best subset regression*
  - **library(MASS)** followed by **stepAIC(linearModel)** successively removes the covariate that decreases the AIC the most, until there is no way to reduce the AIC further.
  - **library(MASS)** followed by

```
stepAIC(NulllinearModelName,
        list(upper=~1+Covt1+Covt2+Covt3
             lower=~1),
        direction="forward")
```

    starts with an empty model and successively adds the covariate that increases the AIC least.

    Omitting the "direction" causes the program to go **stepwise** – at each step, it checks whether adding back a covariate that was previously removed could decrease the AIC

Diagnostic plots

- **plot(linearModelName, ask=T)** lists and shows all possible plots.
- **res <- linearModelName$residuals** followed by **qqnorm(res)** and **qqline(res)** gives a Q-Q plot.
- **library(MASS)** followed by **boxcox(linearModel)** gives a Box-Cox plot.

Categorical data – single factor

- Best illustrated by example

Control	Treat A	Treat B
1	2	3
4	5	6
7	8	9

Should be input as two vectors

```
Result = 1 2 3 4 5 6 7 8 9
treat  = 1 2 3 1 2 3 1 2 3
```

  Followed by the command **Treat <- factor(treat)**
- A box-and-whisker plot of the data can be obtained using **plot(factorName, resultName)**

© Daniel Guetta, 2010

- To specify the ANOVA constraint, use
  - Corner-point constraints: **`options(contrasts = c("contr.treatment","contr.poly"))`**
  - Sum-to-zero constraints: **`options(contrasts = c("contr.sum","contr.poly"))`**
- The linear model is then fitted using **`linearModelName <- lm(Result~Factor)`**.
- Alternatively, **`linearModelName <- aov(Result~Factor)`** will also fit the model, but then the summary needs to be produced using **`summary.lm(linearModelName)`**.
- Note:
  - **`tapply(Result,Factor,mean)`** will return a table with the mean of each group. Alternatively, if an `aov`-style model was fitted, **`model.tables(LinearModelName,type="means", se=T)`** will automatically calculate means for every factor in the model (and the standard errors that appear in `summary`).
  - **`tapply(Result,Factor,function(x) sqrt(var(x)/ length(x)))`** will return a table with the standard error of each group.

Categorical data – multiple factors

- Consider running

```
var1 <- c(1,2); var2 <- 1:5; var3 <- 1:4
grid <- expand.grid(var1,var2,var3)
```

  this creates a table with every possible permutation of the variables, in that order. We can then apply

```
Var1 <- factor(grid[,1])
Var2 <- factor(grid[,2])
```

- To fit a box-and-whisker diagram involving two factors

```
ourData <- data.frame(Var1, Var2, Observations)
plot.design(ourData)
```

- To create an interaction plot

```
interaction.plot(Var1,Var2,Observations)
```

- When fitting more than one factor
  - No interaction: `linearModelName <- lm(Result~Factor1+Factor2)`
  - Interaction: `linearModelName <- lm(Result~Factor1*Factor2)`

© Daniel Guetta, 2010

{**Likelihood Theory**}

I have been asked to remove all notes from the Part III Statistical Theory class from this website. Please email me if you have any questions.

Condition (7) allows us to write

$$\begin{aligned} 0 &= \bar{U}(\hat{\theta}_n)\mathbb{1}_{A_n} \\ &= \left(\bar{U}(\theta_0) - \bar{j}(\tilde{\theta}_n)^T(\hat{\theta}_n - \theta_0)\right)\mathbb{1}_{A_n} \end{aligned}$$

Where $\tilde{\theta}_n$ lies on the line segment between θ_0 and $\hat{\theta}_n$:

$$\tilde{\theta}_n \in \left\{t\theta_0 + (1-t)\hat{\theta}_n : t \in [0,1]\right\}$$

Now, by the triangle rule:

$$\begin{aligned} \left\|\bar{j}(\tilde{\theta}_n)\mathbb{1}_{A_n} - i^{(1)}(\theta_0)\right\| &\le \left\|\bar{j}(\tilde{\theta}_n)\mathbb{1}_{A_n} - i^{(1)}(\tilde{\theta}_n)\mathbb{1}_{A_n}\right\| \\ &\quad + \left\|i^{(1)}(\tilde{\theta}_n)\mathbb{1}_{A_n} - i^{(1)}(\theta_0)\mathbb{1}_{A_n}\right\| \\ &\quad + \left\|i^{(1)}(\theta_0)\mathbb{1}_{A_n} - i^{(1)}(\theta_0)\right\| \\ &\le \sup_{\theta\in\Theta}\left\|\bar{j}(\tilde{\theta}_n) - i^{(1)}(\tilde{\theta}_n)\right\| \\ &\quad + \left\|i^{(1)}(\tilde{\theta}_n)\mathbb{1}_{A_n} - i^{(1)}(\theta_0)\mathbb{1}_{A_n}\right\| \\ &\quad + \left\|i^{(1)}(\theta_0)\right\|\mathbb{1}_{A_n^c} \end{aligned}$$

Generalised Linear Models

- Consider a family of distributions with respect to a σ-finite measure of the form[22]

$$f\left(y;\mu,\sigma^2\right) = a\left(\sigma^2,y\right)\exp\left[\frac{1}{\sigma^2}\left\{\theta(\mu)y - k\left(\theta(\mu)\right)\right\}\right] \qquad \begin{array}{c} y \in E_1 \subseteq \mathbb{R} \\ \mu \in \mathrm{M} \subseteq \mathbb{R} \\ \sigma^2 \in \Phi \subseteq (0,\infty) \end{array}$$

  Where $a\left(\sigma^2,y\right)$ is a known positive function. Such a family is called an **exponential dispersion family** and σ^2 is called the **dispersion parameter**.

- If y is of exponential dispersion family form, then its **moment generating function** is

$$\begin{aligned} M(t) &= \int_{E_1} a\left(\sigma^2,y\right)\exp\left[t\cdot y + \tfrac{1}{\sigma^2}\left\{\theta(\mu)y - k\left(\theta(\mu)\right)\right\}\right]\mathrm{d}\mu(y) \\ &= \exp\left[\tfrac{1}{\sigma^2}\left\{k\left(\sigma^2 t + \theta(\mu)\right) - k\left(\theta(\mu)\right)\right\}\right] \\ &\qquad \int_{E_1} a\left(\sigma^2,y\right)\exp\left[\tfrac{1}{\sigma^2}\left\{\left(\sigma^2 t + \theta(\mu)\right)\cdot y - k\left(\sigma^2 t + \theta(\mu)\right)\right\}\right]\mathrm{d}\mu(y) \end{aligned}$$

[22] Note: in applied statistics, the distribution is given in the form

$$f\left(y;\mu,\sigma^2\right) = \exp\left[\frac{y\mu - k(\mu)}{\sigma^2} + c\left(y,\sigma^2\right)\right]$$

In other words, it assumes $\theta(\cdot) = \cdot$ and $a\left(y,\sigma^2\right) = \exp\left(c(y,\sigma^2)\right)$.

© Daniel Guetta, 2010

The integral is 1, because it is the sum over all space of another exponential distribution family. Thus

$$M(t) = \exp\left[\frac{1}{\sigma^2}\left\{k\left(\sigma^2 t + \theta(\mu)\right) - k\left(\theta(\mu)\right)\right\}\right]$$

And the **cumulant generating function** (ie: the logarithm of the mgf) is

$$K_Y\left(t;\mu,\sigma^2\right) = \frac{1}{\sigma^2}\left\{k\left(\sigma^2 t + \theta(\mu)\right) - k\left(\theta(\mu)\right)\right\}$$

- The discussion above allows to us to find

$$\mathbb{E}_{\mu,\sigma^2}\left(y\right) = \kappa_1 = k'\left(\theta(\mu)\right)$$
$$\mathrm{var}_{\mu,\sigma^2}\left(y\right) = \kappa_2 = \sigma^2 k''\left(\theta(\mu)\right) = \sigma^2 V\left(\mu\right)$$

The function $V\left(\mu\right)$ is called the **variance function**. (*Note*: to find it, differentiate k with respect to θ - not with respect to μ).

- A different derivation of these results uses the loglikelihood

$$\ell(y;\mu,\sigma^2) = \frac{\theta(\mu)y - k\left(\theta(\mu)\right)}{\sigma^2} + \log\left\{a\left(\sigma^2, y\right)\right\}$$

Note that

$$\frac{\partial \ell}{\partial \theta} = \frac{y - k'\left(\theta(\mu)\right)}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \theta^2} = -\frac{k''\left(\theta(\mu)\right)}{\sigma^2}$$

It now suffices to use two results derived in the previous section:

$$\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \Rightarrow \mathbb{E}\left(\frac{y - k'\left(\theta(\mu)\right)}{\sigma^2}\right) = 0$$
$$\boxed{\mathbb{E}\left(y\right) = k'\left(\theta\left(\mu\right)\right)}$$
$$\mathbb{E}_\theta\left\{\left(\frac{\partial \ell}{\partial \theta}\right)^2\right\} = -\mathbb{E}_\theta\left\{\frac{\partial^2 \ell}{\partial^2 \theta}\right\} \Rightarrow \frac{k''\left(\theta(\mu)\right)}{\sigma^2} = \mathbb{E}\left(\frac{\left(y - \mathbb{E}(y)\right)^2}{\sigma^4}\right) = \frac{\mathrm{var}\left(y\right)}{\sigma^4}$$
$$\boxed{\mathrm{var}\left(y\right) = \sigma^2 k''\left(\theta(\mu)\right)}$$

As above.

- An exponential dispersion family is completely characterised by $\left(V(\mu), \mathrm{M}, \Phi\right)$. We may therefore write $Y \sim \mathrm{ED}\left(\mu, \sigma^2 V(\mu)\right)$ with $\mu \in \mathrm{M}$ and $\sigma^2 \in \Phi$ to mean that Y is of exponential family form with mean μ and variance $\sigma^2 V(\mu)$.
- The linear we have thus far been studying can be specified as follows:
  - **Distribution**: $Y_i \sim N\left(\mu_i, \sigma^2\right)$
  - **Link function**: $\mathbb{E}\left(Y_i\right) = \mu_i = \boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}$

© Daniel Guetta, 2010

this, however, is only a subset of **general linear models** (GLMs). A GLM is a model for **independent** responses $Y_1,\cdots,Y_n$ in which

- **Distribution**: $Y_i \sim \text{ED}\left(\mu_i,\sigma_i^2 V(\mu_i)\right)$, $\mu_i \in \text{M}$, $\sigma_i^2 = \sigma^2 a_i$ where σ^2 is an **unknown dispersion parameter** and $a_1,\cdots,a_n$ are **known constants**.

  For example, if $Y_i \sim \frac{1}{n_i}\text{bin}\left(n_i,\mu_i\right)$, we have $\sigma_i^2 = {}^1\!/_{n_i}$, so we can take $a_i = {}^1\!/_{n_i}$ and $\sigma^2 = 1$.
- **Link function**: The i^{th} component of the **linear predictor** $\eta_i = \boldsymbol{X}_{i\bullet}^T\boldsymbol{\beta}$ are related through $g\left(\mu_i\right) = \eta_i$ where
  - $\boldsymbol{X}_{i\bullet}$ is a vector of known explanatory variables
  - β is an unknown vector of regression coefficients
  - g is a strictly increasing twice differentiable function called the **link function**.

  (Note that for the standard linear model, $g(\mu) = \mu$).

- The choice $g\left(\mu\right) = \theta\left(\mu\right)$ is called the **canonical link function** and it simplifies calculations in certain cases.

  For example, consider the density of $\boldsymbol{y} = \left(y_1,\cdots,y_n\right)^T$

$$f_Y(\boldsymbol{y};\mu,\sigma^2) = \left\{\prod_{i=1}^n a\left(\sigma^2,y_i\right)\right\}\exp\left\{\sum_{i=1}^n\left(\frac{\theta\left(\mu_i\right)y_i}{\sigma_i^2} - \frac{K\left(\theta(\mu_i)\right)}{\sigma_i^2}\right)\right\}$$

  However, if we use the canonical link function, $\mu_i = g^{-1}\left(x_{i\bullet}^T\beta\right) = \theta^{-1}\left(x_{i\bullet}^T\beta\right)$

$$f_Y(\boldsymbol{y};\mu,\sigma^2) = \left\{\prod_{i=1}^n a\left(\sigma^2,y_i\right)\right\}\exp\left\{\beta^T\sum_{i=1}^n\left(\frac{x_{i\bullet}y_i}{\sigma_i^2} - \frac{K\left(x_i^T\beta\right)}{\sigma_i^2}\right)\right\}$$

  In that case, we see that the vector

$$\sum_{i=1}^n\left(\frac{x_{i\bullet}y_i}{\sigma_i^2}\right) = \left(\sum_{i=1}^n\frac{x_{i1}y_i}{a_i},\cdots,\sum_{i=1}^n\frac{x_{ip}y_i}{a_i}\right)^T$$

  is sufficient for β, for each fixed value of σ^2.

- In general, there is no closed-form expression for the MLE $\hat{\beta}$, but we can use a Newton-Raphson type of algorithm (**Fisher Scoring**) to find a sequences converging to $\hat{\beta}$. Moreover, under mild conditions on $\boldsymbol{X}_{i\bullet}$, we can apply the result of section 2.2 to deduce that

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{\quad d\quad} N_p\left(0,i^{(1)}(\beta)^{-1}\right)$$

© Daniel Guetta, 2010

This result can be used to estimate the standard deviation of components of $\hat{\beta}$, or to test hypotheses about β.

- Alternatively, tests can be based on the **deviance**, which is closely related to the likelihood ratio statistic.

  Let ω_s be the **saturated model** with as many parameters as observations. This achieves a perfect fit with $\hat{Y}_i^{(s)} = Y_i$. Let the corresponding maximised loglikelihood be $\ell_{\max}^{(s)}$. Let $\omega_f \subset \omega_s$ be a fitted model, with maximised loglikelihood $\ell_{\max}^{(f)} < \ell_{\max}^{(s)}$.

  The **deviance** of the fitted model is then given by

  $$D\left(\omega_f, \omega_s\right) = 2\phi\left(\ell_{\max}^{(s)} - \ell_{\max}^{(f)}\right)$$

  The **scaled deviance** is given by

  $$S\left(\omega_f, \omega_s\right) = 2\left(\ell_{\max}^{(s)} - \ell_{\max}^{(f)}\right)$$

  Good models will have a **small** scaled deviance.

  To compare two models $\omega_1 \subset \omega_2 \subset \omega_s$, we consider the drop in deviance:

  $$S\left(\omega_1, \omega_s\right) - S\left(\omega_2, \omega_s\right) = \frac{D\left(\omega_1, \omega_s\right) - D\left(\omega_2, \omega_2\right)}{\phi} \approx \chi^2_{p_2 - p_1} \qquad \text{if } \omega_1 \text{ true}$$

  where $p_i = \dim(\omega_i)$.

  This can be used in two contexts

  - Checking whether it is worth adding an extra term in the model, by comparing the drop in deviance to $\chi^2_{p_2 - p_1}$.
  - Testing whether a model is a "good fit" – in other words, whether it approximates the saturated model correctly. If it does, we would expect S for the model to be distributed as χ^2_{n-p}. We are effectively using the test above with $\omega_1 =$ our model and $\omega_2 = \omega_s$, which implies that $S(\omega_2, \omega_s) = 0$.

  When ϕ is known (for example, Binomial and Poisson), this can be used to test whether the drop in deviance is significant. Otherwise, we can use F-tests.

GLMs ~ Normal data

- If $Y \sim N\left(\mu, \sigma^2\right)$. The density is

© Daniel Guetta, 2010

$$\begin{aligned} f\left(y\right) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(y-\mu\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y^2}{2\sigma^2}+\frac{y\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y^2}{2\sigma^2}\right)\exp\left(\frac{1}{\sigma^2}\left(y\mu-\tfrac{1}{2}\mu^2\right)\right) \end{aligned}$$

This is of exponential distribution family form with

- $a\left(\sigma^2,y\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y^2}{2\sigma^2}\right)$
- $\theta(\mu) = \mu$
- $k(\mu) = \frac{1}{2}\mu^2$

We note that $V(\mu) = 1$.

- Now, consider variables $Y_1,\cdots,Y_n \sim N\left(\mu_i,\sigma^2\right)$. The loglikelihood is given by

$$\ell = -\frac{1}{2\sigma^2}\left\|\boldsymbol{Y}-\boldsymbol{\mu}\right\|^2 - \frac{n}{2}\log\left(2\pi\sigma^2\right)$$

We then have

- ω_s has fitted values $\hat{\mu}_i^{(s)} = Y_i$
- ω_f has fitted values $\hat{\mu}_i^{(f)} = \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_{\bullet i}$

The scaled deviance is then

$$\begin{aligned} S\left(\omega_f,\omega_s\right) &= 2\left(\ell_{\max}^{(s)} - \ell_{\max}^{(f)}\right) \\ &= -\frac{1}{\sigma^2}\left(\left\|\boldsymbol{Y}_i-\boldsymbol{Y}_i\right\|^2 - \left\|\boldsymbol{Y}_i - X\boldsymbol{\beta}\right\|\right) \\ &= \frac{\mathrm{RSS}_f}{\sigma^2} = \frac{\varepsilon_f^2}{\sigma^2} \\ &\sim \chi^2_{n-p_f} \qquad \text{if } \omega_f \text{ is true} \end{aligned}$$

GLMs ~ Binomial Data

- If $Z \sim \mathrm{bin}\left(n,p\right)$. Let $Y = Z/n$ be the proportion of successes. We note that $\mathbb{E}(Y) = p$. Now:

$$\begin{aligned} \mathbb{P}\left(Y=y\right) &= \mathbb{P}\left(Z=ny\right) \\ &= {}^nC_{ny}p^{ny}\left(1-p\right)^{n-ny} \\ &= {}^n\,C_{ny}\exp\left\{\frac{y\log\left(\frac{p}{1-p}\right)+\log\left(1-p\right)}{1/n}\right\} \end{aligned}$$

This is, once again, of exponential distribution family form, with

- $\sigma^2 = 1/n$

© Daniel Guetta, 2010

- $\mu = p$ and $\theta(p) = \log\left(\frac{p}{1-p}\right) \Rightarrow p = \frac{e^\theta}{1+e^\theta}$
- $k(\theta(p)) = -\log\left(1-p\right)$
- $V\left(p\right) = p\left(1-p\right)$

- Now consider variables $Y_i \sim \text{bin}\left(n_i, p_i\right)$.
  - The canonical link is given by

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i} \qquad p_i = p_i\left(\boldsymbol{\beta}\right) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}}}{1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}}}$$

    This is known as a **logit link function**.

    Note that it can be written as:

$$\text{Odds} = e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}}$$

    and so the interpretation of the model is that an increase in a given covariate i will result in the odds being increase by a factor of $\exp\left(\beta_i\right)$.
  - $\sigma_i^2 = \frac{1}{n_i}\sigma^2$, with $\sigma^2 = 1$.
  - The likelihood is given by

$$L\left(\boldsymbol{Y}\right) = \prod_{i=1}^{n} {}^{n_i}C_{y_i} p_i^{y_i} \left(1-p_i\right)^{n_i - y_i}$$

  - The loglikelihood is given by

$$\begin{aligned}
\ell(\boldsymbol{Y}) &= \sum_{i=1}^{n} \left\{ Y_i \log p_i + \left(n_i - Y_i\right)\log\left(1-p_i\right) + \log\left({}^{n_i}C_{y_i}\right)\right\} \\
&= \sum_{i=1}^{n} \left\{ Y_i \log \frac{p_i}{1-p_i} + n_i \log\left(1-p_i\right) + \log\left({}^{n_i}C_{y_i}\right)\right\} \\
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ Y_i \boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet} + n_i \log\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}\right) + \log\left({}^{n_i}C_{y_i}\right)\right\} \\
&= \boldsymbol{\beta}^T \left(\sum_{i=1}^{n} Y_i \boldsymbol{X}_{i\bullet}\right) - \sum_{i=1}^{n} n_i \log\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}\right) + \sum_{i=1}^{n} \log\left({}^{n_i}C_{y_i}\right)
\end{aligned}$$

  - As such, the MLE is found using

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}\boldsymbol{\beta}} &= \sum_{i=1}^{n} Y_i \boldsymbol{X}_{i\bullet} - \sum_{i=1}^{n} n_i \boldsymbol{X}_{i\bullet} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}}{1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}} = 0 \\
\sum_{i=1}^{n} Y_i \boldsymbol{X}_{i\bullet} &= \sum_{i=1}^{n} n_i \boldsymbol{X}_{i\bullet} \underbrace{\frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}}{1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}}}_{p_i}
\end{aligned}$$

    (The equation basically says that the observed value should be equal to the expected value, and can be solved using an iterative method).

© Daniel Guetta, 2010

- We also note that

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} n_i \boldsymbol{X}_{i\bullet} \boldsymbol{X}_{i\bullet}^T \left(\frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}}{1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}} - \frac{e^{2\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}}{\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}\right)^2} \right)$$

$$= \sum_{i=1}^{n} n_i \boldsymbol{X}_{i\bullet} \boldsymbol{X}_{i\bullet}^T \frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}}{\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{X}_{i\bullet}}\right)^2}$$

$$= \sum_{i=1}^{n} n_i \boldsymbol{X}_{i\bullet} \boldsymbol{X}_{i\bullet}^T p_i \left(1-p_i\right)$$

So

$$\mathbb{E}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) = V\left(\boldsymbol{\beta}\right)^{-1} = \sum_{i=1}^{n} n_i \boldsymbol{X}_{i\bullet} \boldsymbol{X}_{i\bullet}^T p_i \left(1-p_i\right) = i(\boldsymbol{\beta})$$

And so

$$\hat{\boldsymbol{\beta}} \approx N_p\left(\boldsymbol{\beta}, i(\hat{\boldsymbol{\beta}})^{-1}\right)$$

- We now look at deviance. Note that
  - For the saturated model ω_s, we fit $\hat{p}_i^{(s)} = Y_i / n_i$, so

$$\ell_{\max}^{(s)} = \sum_{i=1}^{n} \left\{ Y_i \log \frac{Y_i}{n_i} + \left(n_i - Y_i\right) \log\left(\frac{n_i - Y_i}{n_i} \right) + \log\left({}^{n_i}C_{y_i} \right) \right\}$$

  - For the fitted model ω_f, $\hat{p}_i^{(f)} = p_i\left(\hat{\boldsymbol{\beta}}\right) = e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{X}_{\bullet i}} \left(1+e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{X}_{\bullet i}}\right)^{-1}$

$$\ell_{\max}^{(f)} = \sum_{i=1}^{n} \left\{ Y_i \log\left[p_i(\hat{\boldsymbol{\beta}})\right] + \left(n_i - Y_i\right) \log\left[1 - p_i(\hat{\boldsymbol{\beta}})\right] + \log\left({}^{n_i}C_{y_i} \right) \right\}$$

  The deviance (=scaled deviance since $\sigma^2 = 1$) is given by

$$D\left(\omega_f, \omega_s\right) = 2\sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{Y_i}{n_i \hat{p}_i} \right) + \left(n_i - Y_i\right) \log\left(\frac{n_i - Y_i}{n_i - n_i \hat{p}_i} \right) \right\}$$

  As usual, we asses the fit by comparing D to χ^2_{n-p} where $p = \dim \omega_f$. If ω_f is a bad fit, D will be **larger** than this.

- Expanding the logs, we find that that

$$D\left(\omega_f, \omega_s\right) = \sum_{i=1}^{n} \left\{ \frac{\left(Y_i - n_i \hat{p}_i\right)^2}{n_i \hat{p}_i} + \frac{\left\{\left(n_i - Y_i\right) - n_i\left(1-\hat{p}_i\right)\right\}^2}{n_i\left(1-\hat{p}_i\right)} \right\}$$

  This is **Pearson's chi-squared statistic**[23]**!**

- Logit is the most commonly used link functions. Other possibilities:

[23] It is of the form $\sum_{\substack{\text{all} \\ \text{variables}}} \sum_{\substack{\text{2 possibilities} \\ \text{for each variable}}} \frac{(O-E)^2}{E}$.

© Daniel Guetta, 2010

- **Probit link**: $g\left(p_i\right) = \Phi^{-1}\left(p_i\right)$ where Φ denotes the $N(0,1)$ CDF.
- **Complimentary log-log**: $g\left(p_i\right) = \log\left(-\log\left(1-p_i\right)\right)$

GLMs ~ Poisson Data

- If $Y \sim \text{poi}\left(\lambda\right)$, then

$$\mathbb{P}\left(Y=y\right) = \exp\left\{y\log\lambda - \lambda\right\}\frac{1}{y!}$$

Once again of exponential distribution family form with

- $\sigma^2 = 1$
- $\mu = \lambda$ and $\theta(\lambda) = \log\lambda$
- $k(\theta(\lambda)) = \lambda = e^{\theta}$
- $V(\lambda) = \lambda$

- Now consider variables $Y_i \sim \text{Po}\left(\lambda_i\right)$.
  - The canonical link is given by

$$\log(\lambda) = \boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i} \qquad\qquad \lambda_i = e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}} = \lambda_i\left(\boldsymbol{\beta}\right)$$

  - The likelihood is given by

$$\begin{aligned}\ell\left(\boldsymbol{Y}\right) &= \sum_{i=1}^{n}\left\{-\lambda_i + Y_i\log\lambda_i - \log\left(Y_i!\right)\right\} \\ \ell\left(\boldsymbol{\beta}\right) &= -\sum_{i=1}^{n} e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}} + \boldsymbol{\beta}^T \sum_{i=1}^{n} Y_i \boldsymbol{X}_{\bullet i} - \sum_{i=1}^{n}\log\left(Y_i!\right)\end{aligned}$$

  - As such, the MLE is found using

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= -\sum_{i=1}^{n} \boldsymbol{X}_{\bullet i} e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}} + \sum_{i=1}^{n} Y_i \boldsymbol{X}_{\bullet i} = 0 \\ \sum_{i=1}^{n} Y_i \boldsymbol{X}_{\bullet i} &= \sum_{i=1}^{n} \boldsymbol{X}_{\bullet i} e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}}\end{aligned}$$

  - Also

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} \boldsymbol{X}_{\bullet i} \boldsymbol{X}_{\bullet i}^T e^{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i}} = i(\boldsymbol{\beta})$$

and so

$$\hat{\boldsymbol{\beta}} \approx N_p\left(\boldsymbol{\beta}, i(\hat{\boldsymbol{\beta}})^{-1}\right)$$

  - We now look at deviance. Note that
    - For the saturated model ω_s, we fit $\hat{\lambda}_i^{(s)} = Y_i$, so

$$\ell_{\max}^{(s)} = \sum_{i=1}^{n}\left\{Y_i\log Y_i - Y_i - \log\left(Y_i!\right)\right\}$$

    - For the fitted model ω_f, $\hat{\lambda}_i^{(f)} = \lambda_i^{(f)}(\hat{\boldsymbol{\beta}}) = e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{X}_{\bullet i}}$

© Daniel Guetta, 2010

$$\ell_{\max}^{(f)} = \sum_{i=1}^{n}\left\{Y_i \log\left[\lambda_i(\hat{\boldsymbol{\beta}})\right] - \lambda_i(\boldsymbol{\beta}) - \log\left(Y_i!\right)\right\}$$

The deviance (=scaled deviance since $\sigma^2 = 1$) is given by

$$D = 2\left\{\ell_{\max}^{(s)} - \ell_{\max}^{(f)}\right\} = 2\left\{\sum_{i=1}^{n} Y_i \log\left(\frac{Y_i}{\lambda_i\left(\hat{\boldsymbol{\beta}}\right)}\right) - \sum_{i=1}^{n}\left(Y_i - \lambda_i\left(\hat{\boldsymbol{\beta}}\right)\right)\right\}$$

As usual, we asses the fit by comparing D to χ^2_{n-p} where $p = \dim \omega_f$. If ω_f is a bad fit, D will be **larger** than this.

- Expanding the logs using $s\log(\frac{s}{t}) \approx (s-t) + \frac{(s-t)^2}{2t}$, we find that that

$$D \approx \sum_{i=1}^{n} \frac{\left(Y_i - \lambda_i(\hat{\boldsymbol{\beta}})\right)^2}{\lambda_i(\hat{\boldsymbol{\beta}})} \left(= \sum \frac{O-E}{E}\right)$$

Once again, this is **Pearson's chi-squared statistic**!

- Suppose $Y_1, \cdots, Y_n$ are counts for different "exposures" $m_1, \cdots, m_n$, and

$$Y_i \sim \mathrm{Po}\left(\lambda_i\right) \qquad \lambda_i = m_i\theta_i$$

where θ_i is a **rate**, and the interest lies in modelling how this rate depends on the covariates. Our model is then

$$\log\left(\lambda_i\right) = \log\left(m_i\theta_i\right) = \log\left(m_i\right) + \underbrace{\log\left(\theta_i\right)}_{\boldsymbol{\beta}^T \boldsymbol{X}_{\bullet i} = \log\theta_i}$$

In this type of situation, $\log\left(m_i\right)$ is called an **offset**, and its coefficient is forced to be 1.

- Note that if we fit a model of the kind $\log\left(\lambda_i\right) = \alpha + \beta x_i$, we are effectively saying that an increase of 1 in x results in an increase of $\exp\beta$ in the mean.

GLMs ~ Contingency Tables

- Consider an $r \times c$ contingency table containing n individuals, where the probability of being in cell (i, j) is p_{ij} and the number of items in cell (i, j) is Y_{ij}. Let $\boldsymbol{Y} = \left(Y_{11}, Y_{12}, \cdots, Y_{rc}\right)^T$. Then

$$\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}\right) = \frac{n!}{\prod_{i,j}\left(y_{ij}!\right)} \prod_{i,j} p_{ij}^{y_{ij}}$$

To see, why, consider that

- $n!$ is the total number of ways of arranging n items ignoring bins.

© Daniel Guetta, 2010

- Each bin can be arranged in $y_{ij}!$ ways, so we have $\prod\left(y_{ij}!\right)$ repeats.
- Thus, $n!\Big/\prod\left(y_{ij}!\right)$ is the total number of ways of getting a given arrangement.
- $\prod p_{ij}^{y_{ij}}$ is the probability of any one such arrangement.

So $\boldsymbol{Y}$ ~ Multinomial$\left(n,\boldsymbol{p}\right)$, where $\boldsymbol{p}=\left(p_{11},p_{12},\cdots,p_{rc}\right)^T$

- Unfortunately, the GLM framework does not admit a multinomial distribution. However, it can be shown that if Y_{ij} ~ Po$\left(\mu_{ij}\right)$

$$\boldsymbol{Y}\,|\left(\sum_{i,j}Y_{ij}=n\right)\sim\text{Multinomial}\left(n,\left(\frac{\mu_{11}}{\sum_{i,j}\mu_{ij}},\cdots,\frac{\mu_{rc}}{\sum_{i,j}\mu_{ij}}\right)^T\right)$$

So we can simply use a Poisson-like GLM, with the sum of all Y constrained to be n.

Computer Code

Generalised Linear Models

- The error estimates given in **`summary(glmName)`** come from the asymptotic normality of MLEs.
- **`anova(glmName,test="Chisq")`** will return an analysis of the deviance for the GLM, with relevant chi-squared values.

Binomial Data

- Best illustrated by example

Covariate	Num Trials	Num "yes"
1	50	10
2	51	11
3	52	10

Should be input as three vectors

```
Covariate  = 1  2  3
numTrials  = 50 51 52
numYes     = 10 11 10
```

The model is then fit by setting **`pYes <- numYes/numTrials`** followed by

```
glmName <- glm(pYes~Covariate,family=binomial,weights=numTrials)
```

- To use a **probit** link function instead, use

```
glmName <- glm(pYes~Covariate,family=binomial(link=probit),weights=numTrials)
```

- It is worth noting that another way to find these models is

```
glmName <- glm(dataTable~Covariate,family=binomial)
```

© Daniel Guetta, 2010

where `dataTable` contains two columns; the first being the "successes" and the second being the "fails". The table can be generated using, for example, `cbind`.

- For categorical data, every point made previously regarding design plots, interaction plots, and fitting applies here as well.
- Note that if our binomial data is binary (ie: only values are 0 and 1), there is no need to specify a weight in the `glm` function.
- In a model with only one observation for each combination of factors, the fit will be perfect. The maximum loglikelihood of the saturated model will therefore be 0.

Poisson Data

- To fit the original model, use **`glmName <- glm(Observations~Covariates,poisson)`**
- If **`Observations`** are counts for different amount of exposures, stored in **`Exposures`**, fit a rate model using **`glmName <- glm(Observations~ offset(log(Exposures)) + Covariates,poisson)`**
- To display a different kind of residual, use **`residuals(glmName, type="pearson")`**

© Daniel Guetta, 2010

{**High Dimensional Problem**}

{**Multiple Testing**}

I have been asked to remove all notes from the Part III Statistical Theory class from this website. Please email me if you have any questions.

Non-Parametric Statistics

Two Independent Samples

- Consider two samples $X_1,\cdots,X_m$ ~ IID F_X and $Y_1,\cdots,Y_n$ ~ IID F_Y
- Assume that $X_i = e_i$ and $Y_j = e_{m+j} + \Delta$, where the e are IID from the **same** continuous distribution.
- We are interested in testing

$$H_0 : \Delta = 0 \qquad H_1 : \Delta > 0$$

- The **Wilcoxon rank sum test** combines both samples, and assigns a rank $S_1,\cdots,S_n$ to each of the $Y_1,\cdots,Y_n$. We then **reject H_0** if

$$T = \sum_{i=1}^{n} S_i > c$$

where c is such that

$$\mathbb{P}\left(T > c \mid H_0\right) = \alpha$$

- We can work c out because under H_0, each of the ${}^{m+n}C_n$ arrangements of in the pooled sample are equally likely and have probability $1 / {}^{m+n}C_n$.
- For large samples, approximations need to be used.

Matched Pairs

- Consider a set of matched pairs $\left(X_1,Y_1\right),\cdots,\left(X_n,Y_n\right)$.
- The differences are given by $Z_i = Y_i - X_i$, and we assume $X_i = \theta + e_i$, where the e are IID from a continuous distribution symmetric about 0.
- We are interested in testing

$$H_0 : \theta = 0 \qquad H_1 : \theta > 0$$

- The **Wilcoxon rank sum test** assigns a rank S_i to each of the $\left|Z_i\right|$ and then adds a sign to that rank; positive if Z_i is positive, negative otherwise. We then **reject H_0** if

$$W = \sum_{\substack{\text{only positive} \\ \text{signed ranks}}} S_i > c$$

where c is such that

$$\mathbb{P}\left(W > c \mid H_0\right) = \alpha$$

- We can work c out because under H_0, each of the 2^n possible assignments of signs to the ranks is equally likely, with probability $1/2^n$.

© Daniel Guetta, 2010

Based on lectures by Richard Samworth and Susan Pitts, Michaelmas 2009

- For large samples, approximations need to be used.

Computer Code

- For an unpaired Wilcoxon test on vectors **`x`** and **`y`** in which we want to test whether **`x < y`**, use **`wilcox.test(x,y,alternative="less")`**
- For a paired Wilcoxon test on vectors **`x`** and **`y`**, use **`wilcox.test(x,y,paired=T)`**. This function will try to run the test directly and will otherwise use an approximation and return a warning. Adding an extra argument **`exact=F`** uses an approximation directly and mutes the warning.

© Daniel Guetta, 2010