Linear Regression

How does a quantity y, vary as a function of another quantity, or vector of quantities $\boldsymbol{x}$? We are interested in $p(y|\theta, \boldsymbol{x})$ under a model in which n observations (x_i, y_i) are exchangeable.

NOTATION.

- y (continuous) is the *response* or *outcome variable*;
- $\boldsymbol{x} = (x_1, \ldots, x_k)$ (discrete or continuous) are the *explanatory variables*;
- We will denote $\boldsymbol{y} = (y_1, \ldots, y_n)$ the vector of outcomes and $\boldsymbol{X}$ the $n \times k$ matrix of explanatory variables.

Linear Regression

- The *normal linear model* is a model such that the distribution of $\boldsymbol{y}|\boldsymbol{X}$ is a normal whose mean is a linear function of $\boldsymbol{X}$

$$E(y_i|\boldsymbol{\beta}, \boldsymbol{X}) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \quad i = 1:n.$$

Usually $x_{i1} = 1$.

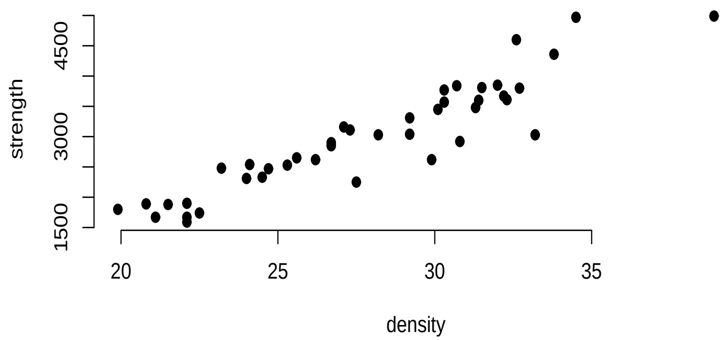
In matrix notation we write

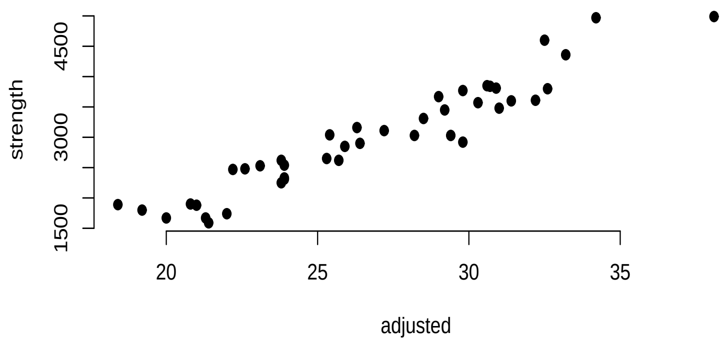
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n\times k}$ and $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{I})$.

Example

42 specimens of radiate pine (Carlin & Chib, 1995 and Williams 1995). For each specimen the maximum compressive strength y_i was measured, with its density x_i and its density adjusted for resin content z_i.





Two models can be considered in this case

$$M_1 := \quad E(y_i|\boldsymbol{\beta}^{(1)}, \boldsymbol{X}) = \beta_1^{(1)} + \beta_2^{(1)} x_i$$

$$M_2 := \quad E(y_i|\boldsymbol{\beta}^{(2)}, \boldsymbol{Z}) = \beta_1^{(2)} + \beta_2^{(2)} z_i$$

For model M_1: $n = 42$, $k = 2$, $x_{1i} = 1$, $x_{2i} = x_i$, $\beta_1 = \beta_1^{(1)}$ and $\beta_2 = \beta_2^{(1)}$.

For model M_2: $n = 42$, $k = 2$, $x_{1i} = 1$, $x_{2i} = z_i$, $\beta_1 = \beta_1^{(2)}$ and $\beta_2 = \beta_2^{(2)}$.

Classical Regression

Consider M_2. If $y_i \sim N(\beta_1^{(2)} + \beta_2^{(2)} z_i, \sigma_2^2)$, the maximum likelihood estimator of $\boldsymbol{\beta}^{(2)}$ is given by the solution of $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{y}$, i.e.

$$\hat{\boldsymbol{\beta}}^{(2)} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}.$$

Furthermore, $\hat{\boldsymbol{\beta}}^{(2)} \sim N(\boldsymbol{\beta}^{(2)}, \sigma_2^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1})$. The MLE of σ_2^2 is given by,

$$\tilde{\sigma}_2^2 = (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}^{(2)})^T(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}^{(2)})/n,$$

however, this estimator is not unbiased, so an unbiased estimator is given by

$$\hat{\sigma}_2^2 = (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}^{(2)})^T(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}^{(2)})/(n-k).$$

Computing the LSE

The goal is to find $\boldsymbol{\beta}$ such that $||\boldsymbol{y} - \boldsymbol{X\beta}||$ is minimized. We obtain the QR decomposition of $\boldsymbol{X}$. So, $\boldsymbol{X} = \boldsymbol{QR}$ where $\boldsymbol{Q}$ is an orthogonal matrix ($\boldsymbol{Q'Q} = \boldsymbol{I}$). and $\boldsymbol{R}$ a rectangular matrix such that only the upper triangle has non 0 entries. Then

$$||\boldsymbol{y} - \boldsymbol{X\beta}|| = ||\boldsymbol{Q'y} - \boldsymbol{Q'QR\beta}|| = ||\boldsymbol{Q'y} - \boldsymbol{R\beta}||$$

Write $\boldsymbol{Q} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2)$, where $\boldsymbol{Q}_1$ corresponds to the first k columns of $\boldsymbol{Q}$. Then

$$||\boldsymbol{Q'y} - \boldsymbol{R\beta}||^2 = ||\boldsymbol{Q}_1'\boldsymbol{y} - \boldsymbol{R\beta}||^2 + ||\boldsymbol{Q}_2'\boldsymbol{y}||^2$$

Thus, the solution to the LSE problem is given by $\boldsymbol{Q}_1'\boldsymbol{y} = \boldsymbol{R}\hat{\boldsymbol{\beta}}$. The residual sum of squares is $||\boldsymbol{Q}_2'\boldsymbol{y}||^2$.

Fitting the linear regression in R

```
>pines.linear<-lm(strength~adjusted)
Call:
lm(formula = strength ~ adjusted)

Residuals:
     Min       1Q   Median       3Q      Max
-623.907 -188.821    4.951  197.334  619.691
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1917.639    252.874  -7.583 2.93e-09 ***
adjusted      183.273      9.304  19.698  < 2e-16 ***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 276.9 on 40 degrees of freedom
```

Distributions

$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. This justifies the following $100(1-\alpha)\%$ C.I. for the regression coefficients $\boldsymbol{\beta}_i$,

$$\hat{\boldsymbol{\beta}}_i \pm t_{\alpha/2,n-k}\hat{\sigma} * \sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$$

A 95% C.I. for $\beta_2^{(2)}$ is given by $(164.5, 202.1)$

We can test the following hypothesis on each β_i

$$H_0: \quad \beta_i = 0 \ \ vs \ \ H_1: \quad \beta_i \neq 0$$

The test statistics is given by

$$t = \frac{\hat{\beta}_i}{\hat{\sigma} * \sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}},$$

F Test

When comparing two nested models we can use the F test.

Let $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ denote the corresponding design matrices and $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_1$ the LSE. If H_0 is correct, then

$$f = \frac{(\hat{\boldsymbol{\beta}}_1^T \boldsymbol{X}_1^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}_0^T \boldsymbol{X}_0^T \boldsymbol{y})/(p-q)}{(\boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}_1^T \boldsymbol{X}_1^T \boldsymbol{y})/(n-p)} \sim F_{p-q,n-p}$$

Therefore, values of f that are large relative to the $F_{p-q,n-p}$ provide evidence against H_0.

Sufficient Statistics

The likelihood for a normal linear model is given by

$$f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{X}) \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}$$

We note that

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + ||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||^2$$

So $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are sufficient statistics for $\boldsymbol{\beta}$ and σ^2. So

$$f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{X}) \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}$$

$$\exp\left\{-\frac{1}{2\sigma^2}(n-k)\hat{\sigma}^2\right\}$$

The Bayesian Approach

We consider the model

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \\ p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}) &\propto \sigma^{-2} \end{aligned}$$

Notice that this model assumes conditionality on $\boldsymbol{X}$. The situation where the regressors are subject to error require a prior distribution for $\boldsymbol{X}$.

The posterior distribution.

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y})$$

Conditional posterior of $\boldsymbol{\beta}$.

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{y} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}\sigma^2)$$

with $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{V}_{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Marginal posterior of σ^2.

$$p(\sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})}{p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})}$$

$$\sigma^2|\boldsymbol{y} \sim IG((n-k)/2, (n-k)\hat{\sigma}^2/2),$$

Marginal posterior of $\boldsymbol{\beta}$.

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \left(1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(n-k)\hat{\sigma}^2}\right)^{-(n-k+k)/2}$$

which corresponds to k-variate student with location $\hat{\beta}$ and scale matrix $\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Checking that the posterior is proper. $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$ is proper if

1. $n > k$
2. the rank of $\boldsymbol{X}$ equals k (i.e. columns of $\boldsymbol{X}$ are l.i.)

Sampling from the Posterior

1. Compute the QR factorization of $\boldsymbol{X}$.
2. Obtain $\hat{\boldsymbol{\beta}}$ as the solution of $\boldsymbol{Q}_1'\boldsymbol{y} = \boldsymbol{R}\hat{\boldsymbol{\beta}}$.
3. Obtain $\hat{\sigma}^2$ as $||\boldsymbol{Q}_2'\boldsymbol{y}||^2/(n-k)$.
4. Sample $\sigma^2 \sim IG((n-k)/2, (n-k)\hat{\sigma}^2/2)$.
5. Note that $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}'\boldsymbol{R}$, so $\boldsymbol{R}$ is a Cholesky factor of $\boldsymbol{X}'\boldsymbol{X}$. So, if $\boldsymbol{z} \sim N_k(0, \boldsymbol{I})$ then $\boldsymbol{R}^{-1}z \sim N_k(0, \boldsymbol{V}_{\boldsymbol{\beta}})$. DON'T compute $\boldsymbol{R}^{-1}$ explicitly! Solve $\boldsymbol{R}\boldsymbol{\beta} = z$, then do $\sigma\boldsymbol{\beta} + \hat{\boldsymbol{\beta}}$.

To make the generation of $\boldsymbol{\beta}$ more efficient you have to avoid computing $\boldsymbol{Q}$ explicitly. Also, when operating with $\boldsymbol{R}$ you have to remember that it is an upper triangular matrices. See `R` routines like `backsolve` and `qr.solve`.