Lecture 12. The linear model

Introduction to linear models

- Linear models can be used to explain or model the relationship between a *response*, or *dependent*, variable and one or more *explanatory variables*, or *covariates* or *predictors*.
- For example, how do motor insurance claims depend on the age and sex of the driver, and where they live?

  Here the claim rate is the response, and age, sex and region are explanatory variables, assumed known.

- In the *linear model*, we assume our n observations (responses) are $Y_1, .., Y_n$ are modelled as

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

  where
- $\beta_1, .., \beta_p$ are unknown parameters, $n > p$
- $x_{i1}, .., x_{ip}$ are the values of p covariates for the ith response (assumed known)
- $\varepsilon_1, .., \varepsilon_n$ are independent (or possible just uncorrelated) random variables with mean 0 and variance σ^2.

From (1),

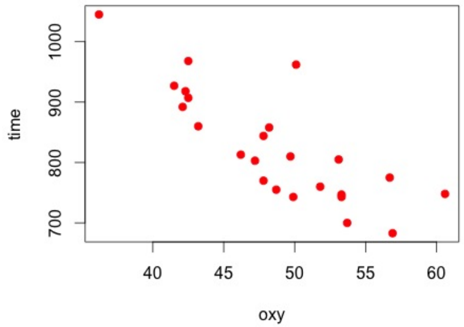
- $\mathbb{E}(Y_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$
- $\text{var}(Y_i) = \text{var}(\varepsilon_i) = \sigma^2$
- $Y_1, .., Y_n$ are independent (or uncorrelated).

Note that (1) is linear in the parameters $\beta_1, .., \beta_p$ (there are a wide range of more complex models which are non-linear in the parameters).

Example 12.1

For each of 24 males, the maximum volume of oxygen uptake in the blood and the time taken to run 2 miles (in minutes) were measured. Interest lies on how the time to run 2 miles depends on the oxygen uptake.

```
oxy=c(42.3,53.1,42.1,50.1,42.5,42.5,47.8,49.9,
      36.2,49.7,41.5,46.2,48.2,43.2,51.8,53.3,
      53.3,47.2,56.9,47.8,48.7,53.7,60.6,56.7)
time=c(918, 805, 892, 962, 968, 907, 770, 743,
      1045, 810, 927, 813, 858, 860, 760, 747,
       743, 803, 683, 844, 755, 700, 748, 775)
plot(oxy, time)
```



- For individual i, let Y_i be the time to run 2 miles, and x_i be the maximum volume of oxygen uptake, $i = 1, ..., 24$.
- A possible model is

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \dots, 24,$$

where ε_i are independent random variables with variance σ^2, and a and b are constants.

Matrix formulation

The linear model may be written in matrix form. Let

$$\underset{n\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ \cdot \\ \cdot \\ Y_n \end{pmatrix}, \quad \underset{n\times p}{X} = \begin{pmatrix} x_{11} & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & x_{np} \end{pmatrix}, \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix}, \quad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix},$$

Then from (1),

$$\begin{aligned} \mathbf{Y} &= X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2) \\ \mathbb{E}(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ \mathrm{cov}(\mathbf{Y}) &= \sigma^2 I \end{aligned}$$

We assume throughout that X has full rank p.

We also assume the error variance is the same for each observation: this is the *homoscedastic* case (as opposed to *heteroscedastic*).

Example 12.1 continued

- Recall $Y_i = a + bx_i + \varepsilon_i, \ \ i = 1, .., 24$.
- In matrix form:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \cdot \\ \cdot \\ Y_{24} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_{24} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \varepsilon_{24} \end{pmatrix},$$

- Then

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Least squares estimation

- In a linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the *least squares estimator* $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimises

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) \\ &= \sum_{i=1}^n (Y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 \end{aligned}$$

- So

$$\left.\frac{\partial S}{\partial \beta_k}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0, \ \ k = 1, .., p.$$

- So $\quad -2\sum_{i=1}^n x_{ik}(Y_i - \sum_{j=1}^p x_{ij}\hat{\beta}_j) = 0, \ \ k = 1, .., p.$
- i.e. $\quad \sum_{i=1}^n x_{ik}\sum_{j=1}^p x_{ij}\hat{\beta}_j = \sum_{i=1}^n x_{ik}Y_i, \ \ k = 1, .., p.$
- In matrix form,

$$X^T X\hat{\boldsymbol{\beta}} = X^T\mathbf{Y} \qquad (3)$$

the *least squares equation*.

- Recall we assume X is of full rank p.
- This implies
$$\mathbf{t}^T X^T X \mathbf{t} = (X\mathbf{t})^T (X\mathbf{t}) = \|X\mathbf{t}\|^2 > 0$$
for $\mathbf{t} \neq \mathbf{0}$ in $\mathbb{R}^p$.
- i.e. $X^T X$ is *positive definite*, and hence has an inverse.
- Hence
$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} \qquad (4)$$
which is linear in the Y_i's.
- We also have that
$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1} X^T \mathbb{E}(\mathbf{Y}) = (X^T X)^{-1} X^T X \boldsymbol{\beta} = \boldsymbol{\beta}$$
so $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.
- And
$$\text{cov}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1} X^T \text{cov}(\mathbf{Y}) X (X^T X)^{-1} = (X^T X)^{-1} \sigma^2 \qquad (5)$$
since $\text{cov}(\mathbf{Y}) = \sigma^2 I$.

Simple linear regression using standardised x's

- The model
$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, n,$$
can be reparametrised to
$$Y_i = a' + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n, \qquad (6)$$
where $\bar{x} = \sum x_i / n$ and $a' = a + b\bar{x}$.
- Since $\sum_i (x_i - \bar{x}) = 0$, this leads to simplified calculations.

- In matrix form, $X = \begin{pmatrix} 1 & (x_1 - \bar{x}) \\ . & . \\ 1 & (x_{24} - \bar{x}) \end{pmatrix}$, so that $X^T X = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$,
where $S_{xx} = \sum_i (x_i - \bar{x})^2$.
- Hence
$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix},$$
so that
$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} = \begin{pmatrix} \bar{Y} \\ \frac{S_{xY}}{S_{xx}} \end{pmatrix},$$
where $S_{xY} = \sum_i Y_i (x_i - \bar{x})$.

- We note that the estimated intercept is $\hat{a}' = \bar{y}$, and the estimated gradient $\hat{b}$ is
$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2)}} \times \sqrt{\frac{S_{yy}}{S_{xx}}}$$
$$= r \times \sqrt{\frac{S_{yy}}{S_{xx}}}$$
- Thus the estimated gradient is the *Pearson product-moment correlation coefficient* r, times the ratio of the empirical standard deviations of the y's and x's.

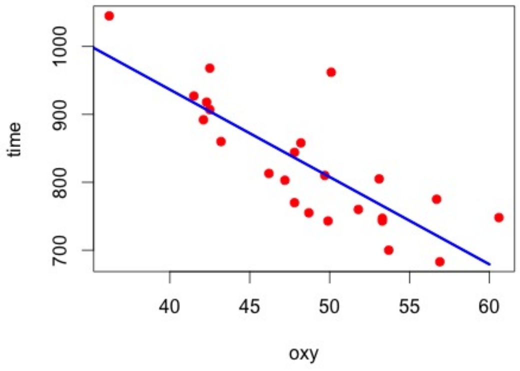
  (Note this estimated gradient is the same whether the x's are standardised to have mean 0 or not.)
- From (5), $\text{cov}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1} \sigma^2$, and so
$$\text{var}(\hat{a}') = \text{var}(\bar{Y}) = \frac{\sigma^2}{n}; \qquad \text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}};$$
- These estimators are uncorrelated.

All these results are obtained without any explicit distributional assumptions.

Example 12.1 continued

$n = 24$, $\hat{a}' = \bar{y} = 826.5$.

$S_{xx} = 783.5 = 28.0^2$, $S_{xy} = -10077$, $S_{yy} = 444^2$, $r = -0.81$, $\hat{b} = -12.9$.



Line goes through $(\bar{x}, \bar{y})$.

'Gauss Markov' theorem

Theorem 12.2

In the full rank linear model, let $\hat{\beta}$ be the least squares estimator of β and let β^* be any other unbiased estimator for β which is linear in the Y_i's.
Then $var(\mathbf{t}^T\hat{\beta}) \le var(\mathbf{t}^T\beta^*)$ for all $\mathbf{t} \in \mathbb{R}^p$.
We say that $\hat{\beta}$ is the Best Linear Unbiased Estimator of β (BLUE).

Proof:

- Since β^* is linear in the Y_i's, $\beta^* = A\mathbf{Y}$ for some $\underset{p\times n}{A}$.
- Since β^* is unbiased, we have that $\beta = \mathbb{E}(\beta^*) = AX\beta$ for all $\beta \in \mathbb{R}^p$, and so $AX = I_p$.
- Now

$$\begin{aligned}
\text{cov}(\beta^*) &= \mathbb{E}\left((\beta^* - \beta)(\beta^* - \beta)^T\right) \\
&= \mathbb{E}\left((AX\beta + A\varepsilon - \beta)(AX\beta + A\varepsilon - \beta)^T\right) \\
&= \mathbb{E}\left(A\varepsilon\varepsilon^T A^T\right) \quad \text{since } AX\beta = \beta \\
&= A(\sigma^2 I)A^T = \sigma^2 AA^T
\end{aligned}$$

- Now $\beta^* - \hat{\beta} = (A - (X^TX)^{-1}X^T)\mathbf{Y} = \underset{p\times n}{B}\mathbf{Y}$, say.
- And $BX = AX - (X^TX)^{-1}X^TX = I_p - I_p = 0$.
- So

$$\begin{aligned}
\text{cov}(\beta^*) &= \sigma^2(B + (X^TX)^{-1}X^T)(B + (X^TX)^{-1}X^T)^T \\
&= \sigma^2(BB^T + (X^TX)^{-1}) \\
&= \sigma^2 BB^T + \text{cov}(\hat{\beta})
\end{aligned}$$

- So for $\mathbf{t} \in \mathbb{R}^p$,

$$\begin{aligned}
\text{var}(\mathbf{t}^T\beta^*) &= \mathbf{t}^T\text{cov}(\beta^*)\mathbf{t} = \mathbf{t}^T\text{cov}(\hat{\beta})\mathbf{t} + \mathbf{t}^T BB^T\mathbf{t}\,\sigma^2 \\
&= \text{var}(\mathbf{t}^T\hat{\beta}) + \sigma^2\|B^T\mathbf{t}\|^2 \\
&\ge \text{var}(\mathbf{t}^T\hat{\beta}).
\end{aligned}$$

- Taking $\mathbf{t} = (0, .., 1, 0, .., 0)^T$ with a 1 in the ith position, gives

$$\text{var}(\hat{\beta}_i) \le \text{var}(\beta_i^*).$$

□

Fitted values and residuals

Definition 12.3

- $\hat{\mathbf{Y}} = X\hat{\beta}$ is the vector of *fitted values*.
- $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the vector of *residuals*.
- The *residual sum of squares* is $\text{RSS} = \|\mathbf{R}\|^2 = \mathbf{R}^T\mathbf{R} = (\mathbf{Y} - X\hat{\beta})^T(\mathbf{Y} - X\hat{\beta})$

- Note $X^T\mathbf{R} = X^T(\mathbf{Y} - \hat{\mathbf{Y}}) = X^T\mathbf{Y} - X^TX\hat{\beta} = 0$ by (3).
- So $\mathbf{R}$ is orthogonal to the column space of X.
- Write $\hat{\mathbf{Y}} = X\hat{\beta} = X(X^TX)^{-1}X^T\mathbf{Y} = P\mathbf{Y}$, where $P = X(X^TX)^{-1}X^T$.
- P represents an orthogonal projection of $\mathbb{R}^n$ onto the space spanned by columns of X. We have $P^2 = P$ (P is idempotent) and $P^T = P$ (symmetric).