Linear models

L1 Basic Ideas

L1.1 Statistical models

One of the tasks of a statistician is the analysis of data. Statistical analysis usually involves one or more of the following.

1. Summarising data
2. Estimation
3. Inference
4. Prediction

In general, we statistically model the relationship between two or more random variables by considering models of the form:

$$Y = f(\boldsymbol{X}, \boldsymbol{\beta}) + \epsilon,$$

where,

Y is the response variable,
f is some mathematical function,
$\boldsymbol{X}$ is some matrix of predictor (input) variables,
$\boldsymbol{\beta}$ are the model parameters,
ϵ is the random error term.

If we assume that $E[\epsilon] = 0$, then $E[Y] = f(\boldsymbol{X}, \boldsymbol{\beta})$ if $\boldsymbol{X}$ is assumed to be non-random. Otherwise $E[Y|X] = f(\boldsymbol{X}, \boldsymbol{\beta})$.

Examples

1. We observe the number of cars passing an intersection over a one minute interval. We want to estimate the average rate at which cars pass this intersection.

If X is the number of cars passing the intersection over a one minute interval, then X is likely to have a Poisson distribution with mean λ. We want to estimate λ.

2. In economics, the production of an industry, Y, is modelled to be a function of the amount of labour available, L, and the capital input, K. In particular, the Cobb-Douglas Production Function is given to be $Y = C_0 L^\alpha K^\beta$.

Furthermore if $\alpha + \beta = 1$, then an industry is said to operate under constant returns to scale, i.e. if capital and labour increase by a factor of t, then production also increases by a factor of t.

As a consultant to an economic researcher, you collect production, labour and capital data for a specific industry and want to estimate the functional relationship and test whether $\alpha + \beta = 1$ in this industry.
Theoretical model: $Y = C_0 L^\alpha K^\beta$.
Stat. model: $\log Y = C^* + \alpha \log L + \beta \log K + \epsilon$.
Estimate: C^*, α and β.
Test: $\alpha + \beta = 1$.

3. Suppose we are interested in studying what factors affect a person's blood pressure. Proposed model:

Blood pressure $= f$(age, weight, gender, activity level, personality type, time of day, genetic predisposition) $+ \epsilon$.

We want to estimate the functional relationship f and potentially test which of the factors has a significant influence on a person's blood pressure.

L1.2 The linear model

Assume

$$Y_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}) + \epsilon_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

for all $i = 1, \ldots, n$. A matrix representation of the model is $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Note that $\boldsymbol{Z}$ is called the *design matrix* and in models with a constant term includes a column of ones as well as the *data matrix* $\boldsymbol{X}$ and possibly functions of $\boldsymbol{X}$. If no constant or functions are included in the model, $\boldsymbol{Z}$ and $\boldsymbol{X}$ are equivalent.

A model is considered to be *linear* if it is linear in its *parameters*. For example,
1. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$ is linear.
2. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^2 + \epsilon_i$ is linear.
3. $Y = C_0 L^\alpha K^\beta + \epsilon$ is linear since we can transform the model into $\log Y = C^* + \alpha \log L + \beta \log K + \epsilon$.
4. $Y = \frac{\beta_1}{\beta_1 - \beta_2}\left[e^{-\beta_2 X} - e^{\beta_1 X}\right] + \epsilon$ is non-linear.

The assumptions of the linear model are:
1. Model form: $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i$ for all i.
2. $E[\epsilon_i] = 0$ for all i.
3. $\text{Var}(\epsilon_i) = \sigma^2$ for all i.
4. $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

L1.3 The Normal (Gaussian) linear model

The Normal linear model assumes:
1. Model form: $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i$ for all i.
2. $\epsilon_i \sim$ i.i.d.$N(0, \sigma^2)$.

There are two implications of these assumptions:

1. $Y_i \sim N(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}, \sigma^2)$ for all $i = 1, \ldots, n$. In equivalent matrix form, $\boldsymbol{Y} \sim N_n\left(\boldsymbol{Z\beta}, \sigma^2 \boldsymbol{I}_n\right)$ where N_n is the n-dimensional multivariate normal distribution. Note that if two random variables X and Y are normally distributed, then Cov$(X, Y) = 0$ if and only if X and Y are independent.

2. Since Cov$(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, then Cov$(Y_i, Y_j) = 0$ for all $i \neq j$ and $Y_1, Y_2, \ldots, Y_n$ are independent.

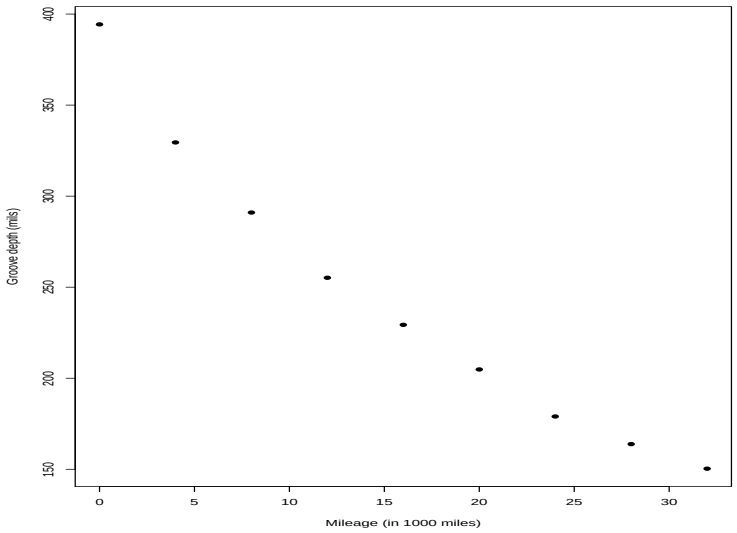
L1.4 Example

A laboratory tests tyres for tread wear by conducting an experiment where tyres from a particular manufacturer are mounted on a car. The tyres are rotated from wheel to wheel every 1000 miles, and the groove depth is measured in mils (0.001 inches) initially and after every 4000 miles giving the following data (Tamhane and Dunlap, 2000):

Mileage (1000 miles)	Groove Depth (mils)
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

Firstly we have to determine which is the response variable and which is the predictor (or controlled) variable. Secondly, we have to hypothesise a functional relationship between the two variables, using either theoretical relationships or exploratory data analysis.

Let the response variable, Y, be groove depth and let the predictor variable, X, be mileage. A plot of mileage vs. depth:



Note that as mileage increases, the groove depth decreases.

Model 1: straight line model

The simple linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

for $i = 1, \ldots, n$ (i.e. $E[Y] = \alpha + \beta X$).

Questions to explore:
1. What are the values of α and β?
2. Is $\beta < 0$?
3. What groove depth do we expect if the tyres have travelled 15000 miles?
4. Could another model be more useful in explaining the relationship between X and Y?

Definition Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimated values of α and β, respectively. We call y_i the ith *observed value* and $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ the ith *fitted value*.

Definition $D = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is called the *model deviance*.

For the estimated line to fit the data well, one wants an estimator of α and β that minimises the model deviance. So, choose $\hat{\alpha}$ and $\hat{\beta}$ to minimise

$$D = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

To minimise D,

$$\frac{\partial D}{\partial \alpha} = -2 \sum_{i=1}^{n} \left(y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right) = 0 \qquad (1)$$

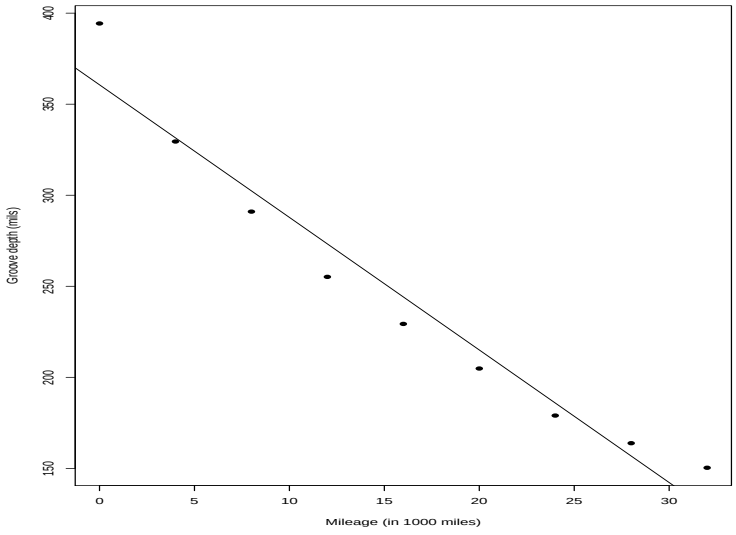
$$\frac{\partial D}{\partial \beta} = -2 \sum_{i=1}^{n} x_i \left(y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right) = 0 \qquad (2)$$

Equations (1) and (2) are called the *normal equations* and solving them we get,

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{(n-1)s_{xy}}{(n-1)s_x^2} = \frac{s_{xy}}{s_x^2},\\ \hat{\alpha} &= \bar{y}-\hat{\beta}\bar{x}.\end{aligned}$$

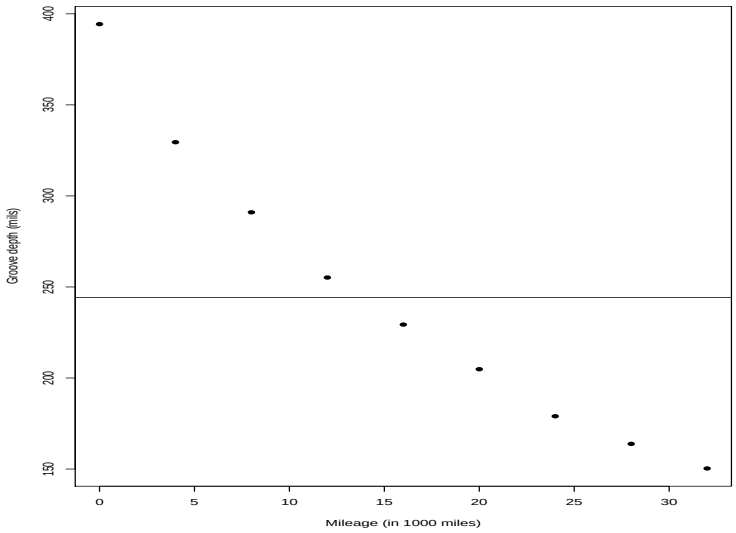
Note that $\hat{\alpha}$ and $\hat{\beta}$ are called *least squares estimators* of α and β. We did not use the normality assumption in our derivation, so the least squares estimators are invariate to the choice of distribution for the error terms. (Although the properties of the estimators may change depending on the underlying distribution.)

If we include the assumption of normality, then it can be shown that $\hat{\beta} = \dfrac{s_{xy}}{s_x^2}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ are also the MLEs of α and β.



Model 2: horizontal line model

$y_i = \mu + \epsilon_i$ for $i = 1, \ldots, n$. (i.e. $E[Y] = \mu$)



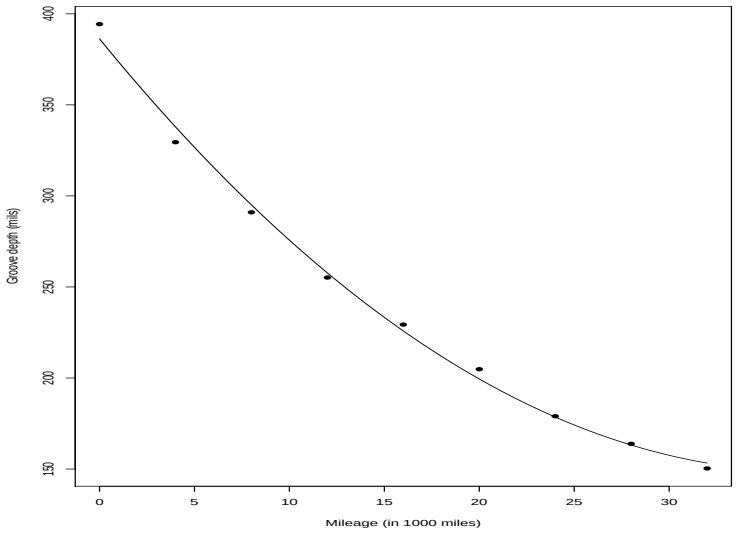
To estimate μ by least squares we minimise $D = \sum_{i=1}^{n}(y_i - \mu)^2$. Setting $\frac{\partial D}{\partial \mu} = 0$ and solving, we get $\hat{\mu} = \bar{y}$ and model deviance, $D = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

In this model we assume the predictor variable has no ability to explain the variance in the response variable.

If D_1 is the deviance of the straight line model and D_2 is the deviance of the horizontal line model, then the linear model does a better job of explaining the variance in Y if $D_1 \leq D_2$. Hence we say that the linear model "fits" the data better.

Model 3: quadratic model

$y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i$



To estimate α, β and γ by least squares we minimise $D = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i + \gamma x_i^2)\right)^2$.

If we let $D_3 = \sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i + \hat{\gamma} x_i^2)\right)^2$, then the quadratic model "fits" the data better than the linear model if $D_3 \leq D_1$.

L1.5 Example

Suppose we are interested in the effect a certain drug has on the weight of an organ. An experiment is designed in which rats are randomly assigned to different treatment groups in which each group receives the drug at one of 7 different levels (e.g. 0 mg, 100 mg, 200 mg, 300 mg, etc.) Upon completion of the treatment, the organs are harvested from the rats and weighed.

Let Y_{ij} be the weight of the organ (response variable) of the jth rat in the ith treatment,

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, 7 \quad j = 1, \ldots, J.$$

We want to test whether $\mu_1 = \mu_2 = \cdots = \mu_7$.

The model is linear in the parameters $\mu_1, \mu_2, \ldots, \mu_7$, so we can estimate using least squares to minimise

$$D = \sum_{i=1}^{7} \sum_{j=1}^{J} (y_{ij} - \mu_i)^2.$$

The least squares estimators are given by:

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{J} \sum_{i=1}^{J} y_{ij}.$$

L2 Least squares estimation

L2.1 Linear algebra review

Definition Let $\boldsymbol{M}$ be any $n \times m$ matrix. Then the *rank* of $\boldsymbol{M}$ is the maximum number of linearly independent column vectors of $\boldsymbol{M}$.

Definition If $\boldsymbol{M} = (m_{ij})$, then $\boldsymbol{M}^T = (m_{ji})$ is said to be the *transpose* of the matrix $\boldsymbol{M}$.

Definition Suppose $\boldsymbol{A}$ is a square $n \times n$ matrix, then
(a) $\boldsymbol{A}$ is *symmetric* if and only if $\boldsymbol{A}^T = \boldsymbol{A}$,
(b) $\boldsymbol{A}^{-1}$ is the *inverse* of $\boldsymbol{A}$, if and only if $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$,
(c) The matrix $\boldsymbol{A}$ is *nonsingular* if and only if rank$(\boldsymbol{A}) = n$,
(d) $\boldsymbol{A}$ is *orthogonal* if and only if $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$,
(e) $\boldsymbol{A}$ is *idempotent* if and only if $\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$,
(f) $\boldsymbol{A}$ is *positive definite* if and only if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all non-zero vectors $\boldsymbol{x}$.

Note that (i) $\boldsymbol{A}$ has an inverse if and only if $\boldsymbol{A}$ is nonsingular, i.e. the rows and columns are linearly independent; (ii) $\boldsymbol{A}^T\boldsymbol{A}$ is positive definite if $\boldsymbol{A}$ has an inverse.

Computational results

1. Let $\boldsymbol{N}$ be an $n \times p$ matrix and $\boldsymbol{P}$ a $p \times n$ matrix, then $(\boldsymbol{N}\boldsymbol{P})^T = \boldsymbol{P}^T\boldsymbol{N}^T$.

2. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are two invertible $n \times n$ matrices, then $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

3. We can write the sum of squares $\sum_{i=1}^n x_i^2 = \boldsymbol{x}^T\boldsymbol{x}$, where $\boldsymbol{x}^T = [x_1, x_2, \ldots, x_n]$ is a $1 \times n$ vector.

Calculus of matrices

Given n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$, we define

$$\frac{d\boldsymbol{y}}{d\boldsymbol{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}.$$

Then,

1. $\frac{d}{d\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^T$, where $\boldsymbol{A}$ is a matrix of constants.

2. $\frac{d}{d\boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}$ whenever $\boldsymbol{A}$ is symmetric.

3. If $f(\boldsymbol{x})$ is a function of several variables the necessary condition to maximise or minimise $f(\boldsymbol{x})$ is

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = 0.$$

If we let $\boldsymbol{H} = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T}$ be the Hessian of f, i.e. the matrix of second derivatives, then a maximum will occur if $\boldsymbol{H}$ is negative definite, a minimum will occur if $\boldsymbol{H}$ is positive definite.

Expectation and variance of matrices Let $\boldsymbol{A}$ be a matrix of constants and $\boldsymbol{Y}$ be a random vector, then

(a) $E[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}E[\boldsymbol{Y}]$,
(b) $\mathrm{Var}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{Y})\boldsymbol{A}^T$.

L2.2 Deriving the least squares estimator

Recall the linear model in matrix form: $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{(p-1)1} \\ 1 & X_{12} & \cdots & X_{(p-1)2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{(p-1)n} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

such that,
$\boldsymbol{Y}$ is an $n \times 1$ column vector of observations of the response variable,
$\boldsymbol{Z}$ is the $n \times p$ design matrix whose first column is a column of 1's, if there is a constant in the model. The other columns are the observations on the explanatory variables $(X_1, X_2, \ldots, X_{p-1})$,
$\boldsymbol{\beta}$ is a $p \times 1$ column vector of the unknown parameters,
$\boldsymbol{\epsilon}$ is an $n \times 1$ column vector of the random error terms.

Assumptions The general linear regression model assumes,

1. $E[\boldsymbol{\epsilon}] = \mathbf{0}$,
2. $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

Aim We want to estimate the unknown vector of parameters, $\boldsymbol{\beta}$, by choosing the value of $\boldsymbol{\beta}$ which minimises the model deviance,

$$\begin{aligned} D &= \sum_{i=1}^{n} (\boldsymbol{y}_i - (\boldsymbol{Z}\boldsymbol{\beta})_i)^2 = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{Z}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta} \end{aligned}$$

Taking the derivative of D with respect to $\boldsymbol{\beta}$ and noticing that $\boldsymbol{Z}^T\boldsymbol{Z}$ is a symmetric matrix, we get,

$$\begin{aligned}\frac{\partial D}{\partial \boldsymbol{\beta}} &= (-2\boldsymbol{y}^T\boldsymbol{Z})^T + 2\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta} \\ &= -2\boldsymbol{Z}^T\boldsymbol{y} + 2\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta}.\end{aligned}$$

Therefore $\hat{\boldsymbol{\beta}}$ will be the least squares estimator of $\boldsymbol{\beta}$ if $-2\boldsymbol{Z}^T\boldsymbol{y} + 2\boldsymbol{Z}^T\boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{0}$. This system of equations are the normal equations for the general linear regression model.

Now solving for $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}2\boldsymbol{Z}^T\boldsymbol{Z}\hat{\boldsymbol{\beta}} &= 2\boldsymbol{Z}^T\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{Z}\hat{\boldsymbol{\beta}} &= \boldsymbol{Z}^T\boldsymbol{y}.\end{aligned}$$

To be able to isolate $\hat{\boldsymbol{\beta}}$ it is necessary for $\boldsymbol{Z}^T\boldsymbol{Z}$ to be invertible; therefore we need $\boldsymbol{Z}$ to be of full rank, i.e. rank$(\boldsymbol{Z}) = p$. If rank$(\boldsymbol{Z}) = p$, then

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

For $\hat{\boldsymbol{\beta}}$ to minimise D, we need to check the Hessian to see that it is positive definite. If $\boldsymbol{Z}$ has full rank, then

$$\boldsymbol{H} = \frac{\partial^2 D}{\partial \boldsymbol{\beta}^2} = (2\boldsymbol{Z}^T\boldsymbol{Z})^T = 2\boldsymbol{Z}^T\boldsymbol{Z},$$

and $\boldsymbol{Z}^T\boldsymbol{Z}$ is positive definite. Hence, $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$.

It can also be shown that

$$s^2 = \frac{1}{n-p}(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})$$

is an unbiased estimator of σ^2.

L2.3 Examples

Example L2.1 Suppose we have two observations such that

$$y_1 = \theta + \epsilon \qquad y_2 = 2\theta + \epsilon.$$

Derive the least squares estimator of θ.

Choose $\boldsymbol{Z} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ then $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} = \frac{1}{5}$ and

$$\hat{\theta} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Y} = \frac{1}{5}(y_1 + 2y_2).$$

Example L2.2 Suppose we have our simple regression model, $Y = \alpha + \beta X + \epsilon$, then in matrix terms $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where,

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Derive the least squares estimator of $\boldsymbol{\beta}$.

Then the least squares estimators of $\boldsymbol{\beta}$ will be given by,

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

where,

$$\begin{aligned}
\boldsymbol{Z}^T\boldsymbol{Z} &= \begin{bmatrix} 1 & 1 & \cdots 1 \\ x_1 & x_2 & \cdots x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\
&= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \\
\boldsymbol{Z}^T\boldsymbol{y} &= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}.
\end{aligned}$$

Therefore,

$$(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

and so

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y} \\
&= \frac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum x_i^2 \sum y_i - \bar{x}\sum x_i y_i \\ -\bar{x}\sum y_i + \sum x_i y_i \end{bmatrix} \\
&= \frac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i \\ \sum y_i (x_i - \bar{x}) \end{bmatrix} \\
&= \frac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum (x_i - \bar{x})^2 - \bar{x}\sum (x_i - \bar{x})(y_i - \bar{y}) \\ \sum (x_i - \bar{x})(y_i - \bar{y}) \end{bmatrix} \\
&= \begin{bmatrix} \bar{y} - \bar{x}\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}.
\end{aligned}$$

So,

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}\bar{x} \\ \frac{s_{xy}}{s_x^2} \end{bmatrix}.$$

L2.4 Properties of $\hat{\boldsymbol{\beta}}$

1. Unbiasedness

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}] = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T E[\boldsymbol{y}] \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T E[\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\beta} + E[\boldsymbol{\epsilon}]) \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\beta} + \mathbf{0}) = \boldsymbol{I}_p\boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

2. Variance

$$\begin{aligned} \mathsf{Var}(\hat{\boldsymbol{\beta}}) &= \mathsf{Var}((\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}) \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathsf{Var}(\boldsymbol{y})((\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T)^T \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathsf{Var}(\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon})\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathsf{Var}(\boldsymbol{\epsilon})\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \\ &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\sigma^2\boldsymbol{I}_n\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \\ &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \\ &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}. \end{aligned}$$

Note that $\mathsf{Var}(\hat{\boldsymbol{\beta}})$ is the $p \times p$ variance-covariance matrix of the vector $\hat{\boldsymbol{\beta}}$ where the ith diagonal entry is $\mathsf{Var}(\hat{\beta}_i)$ and the i, jth entry is $\mathsf{Cov}(\hat{\beta}_i, \hat{\beta}_j)$.

For example, in the simple linear regression case, we had

$$(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \mathsf{Var}\left(\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}\right) &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}, \end{aligned}$$

and so,

$$\begin{aligned}
\text{Var}(\hat{\alpha}) &= \frac{\sigma^2}{n}\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i-\bar{x})^2}, \\
\text{Var}(\hat{\beta}) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}, \\
\text{Cov}(\hat{\alpha},\hat{\beta}) &= \frac{-\sigma^2\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2}.
\end{aligned}$$

3. If we assume in addition that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, then

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}).$$

This is true since,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\
&= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta} + (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}.
\end{aligned}$$

Hence $\hat{\boldsymbol{\beta}}$ is a linear function of a normally distributed random variable. Consequently $\hat{\boldsymbol{\beta}}$ has a normal distribution with mean and variance as shown above.

Note that since $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1})$, then each of the individual parameters

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2((\boldsymbol{Z}^T\boldsymbol{Z})^{-1})_{ii}),$$

but the individual $\hat{\beta}_i$ are not independent.

4. Let $\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}}$ be the $n \times 1$ vector of *fitted values* of $\boldsymbol{y}$.

Note that $\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$. If we let $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$, then we can write

$$\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}.$$

$\boldsymbol{P}$ is therefore often referred to as the *hat matrix* and is symmetric and idempotent because $\boldsymbol{P}^T = \boldsymbol{P}$ and $\boldsymbol{P}^2 = \boldsymbol{P}$.

L2.5 Gauss-Markov Theorem

Theorem If $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$, then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the *unique linear unbiased estimator* of $\boldsymbol{a}^T\boldsymbol{\beta}$ with minimum variance.

Proof

1. Suppose $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$. Hence,

$$\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y} = \boldsymbol{C}\boldsymbol{y},$$

where $\boldsymbol{C} = \boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$ and $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$.

2. $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$ because,

$$\begin{aligned} E[\boldsymbol{a}^T\hat{\boldsymbol{\beta}}] &= E[\boldsymbol{C}\boldsymbol{y}] = \boldsymbol{C}E[\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\ &= \boldsymbol{C}\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{C}E[\boldsymbol{\epsilon}] \\ &= \boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{0} \\ &= \boldsymbol{a}^T\boldsymbol{\beta}. \end{aligned}$$

3. Suppose there exists another linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$, say $\boldsymbol{b}^T\boldsymbol{y}$, then

$$E[\boldsymbol{b}^T\boldsymbol{y}] = \boldsymbol{a}^T\boldsymbol{\beta}$$

and

$$E[\boldsymbol{b}^T\boldsymbol{y}] = \boldsymbol{b}^T E[\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \boldsymbol{b}^T\boldsymbol{Z}\boldsymbol{\beta}.$$

Therefore, $\boldsymbol{b}^T\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, so

$$\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{Z}.$$

4. Consider,

$$\begin{aligned}\text{Var}(\boldsymbol{b}^T\boldsymbol{y}) &= \boldsymbol{b}^T\text{Var}(\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\epsilon})\boldsymbol{b} = \boldsymbol{b}^T\text{Var}(\boldsymbol{\epsilon})\boldsymbol{b} \\ &= \boldsymbol{b}^T\sigma^2\boldsymbol{I}_n\boldsymbol{b} = \sigma^2\boldsymbol{b}^T\boldsymbol{b}\end{aligned}$$

and

$$\begin{aligned}\text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) &= \text{Var}(\boldsymbol{C}\boldsymbol{y}) = \boldsymbol{C}\text{Var}(\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\epsilon})\boldsymbol{C}^T \\ &= \boldsymbol{C}\text{Var}(\boldsymbol{\epsilon})\boldsymbol{C}^T = \boldsymbol{C}\sigma^2\boldsymbol{I}_n\boldsymbol{C}^T = \sigma^2\boldsymbol{C}\boldsymbol{C}^T \\ &= \sigma^2(\boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T)(\boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T)^T \\ &= \sigma^2\boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{a} \\ &= \sigma^2\boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{a}.\end{aligned}$$

But recall from (3) that $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{Z}$, therefore we can rewrite

$$\begin{aligned}\text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{b}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}(\boldsymbol{b}^T\boldsymbol{Z})^T \\ &= \sigma^2\boldsymbol{b}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{b}.\end{aligned}$$

Comparing $\text{Var}(\boldsymbol{b}^T\boldsymbol{y})$ and $\text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}})$, we get

$$\begin{aligned}\text{Var}(\boldsymbol{b}^T\boldsymbol{y}) - \text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{b}^T\boldsymbol{b} - \sigma^2\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T(\boldsymbol{I}_n - \boldsymbol{P})^2\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T(\boldsymbol{I}_n - \boldsymbol{P})^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{b} \\ &= \sigma^2\boldsymbol{D}^T\boldsymbol{D},\end{aligned}$$

where $\boldsymbol{D} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{b}$. Therefore,

$$\text{Var}(\boldsymbol{b}^T\boldsymbol{y}) - \text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{D}^T\boldsymbol{D} \geq 0,$$

so $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ has the smallest variance of any other linear unbiased estimator.

5. Suppose that $\boldsymbol{b}^T\boldsymbol{y}$ is another linear unbiased estimator such that $\text{Var}(\boldsymbol{b}^T\boldsymbol{y}) = \text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}})$, then $\text{Var}(\boldsymbol{b}^T\boldsymbol{y}) - \text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{D}^T\boldsymbol{D} = 0$ implies $\boldsymbol{D} = \boldsymbol{0}$.

If $\boldsymbol{D} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{b} = \boldsymbol{0}$, then

$$\boldsymbol{b} = \boldsymbol{P}\boldsymbol{b} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{b} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{a}$$

and $\boldsymbol{b}^T = \boldsymbol{a}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$, so

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

Therefore $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the unique linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$.

Corollary If $\boldsymbol{a}^T = (0, 0, \ldots, 1, 0, \ldots, 0)$ where the 1 is in the ith position, then $\hat{\beta}_i$ is the *best linear unbiased estimator* (BLUE) of β_i.

L3 Basic hypothesis tests

L3.1 Tests on a single parameter

Suppose we are given the linear model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1} X_{(p-1)i} + \epsilon_i,$$

where $\epsilon_i \sim$ i.i.d. $N(0, \sigma^2)$.

We want to test $H_0 : \beta_j = b$ vs. $H_1 : \beta_j \neq b$ at level α where b is some constant.

The **decision rule** is to reject H_0 if

$$|T| = \left| \frac{\hat{\beta}_j - b}{\mathsf{SE}(\hat{\beta}_j)} \right| > t_{n-p,\alpha/2},$$

where $\mathsf{SE}(\hat{\beta}_j) = \sqrt{\mathsf{Var}(\hat{\beta}_j)}$ is the standard error of the parameter. Recall from Sections L2.2 and L2.4 that $\mathsf{Var}(\hat{\beta}_j) = s^2((\boldsymbol{Z}^T\boldsymbol{Z})^{-1})_{jj}$.

A special case of the above test occurs when we choose $b = 0$. The test $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ at level α has the **decision rule** to reject H_0 if

$$|T| = \left| \frac{\hat{\beta}_j}{\mathsf{SE}(\hat{\beta}_j)} \right| > t_{n-p,\alpha/2}.$$

Note that if we reject $H_0 : \beta_j = 0$ we are claiming that the explanatory variable X_j is useful in predicting the response variable Y when all the other variables are included in the model.

The test statistic $|T| = \left| \frac{\hat{\beta}_j}{\mathsf{SE}(\hat{\beta}_j)} \right|$ is often reported in the output from statistical software such as R.

Example L3.1 A dataset considers fuel consumption for 50 US states plus Washington DC ($n = 51$ observations). The response "fuel" is fuel consumption measured in gallons per person. The predictors considered are "dlic", the percentage of licensed drivers, "tax", motor fuel tax in US cents per gallon, "inc", income per person in $1,000s and "road", the log of the number of miles of federal highway. R fits a linear model of the form,

$$\text{fuel} = \beta_0 + \beta_1 \text{dlic} + \beta_2 \text{tax} + \beta_3 \text{inc} + \beta_4 \text{road}$$

and the output was

	Estimate	Std. Error
Constant	154.193	194.906
dlic	4.719	1.285
tax	-4.228	2.030
inc	-6.135	2.194
road	26.755	9.337

Test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ at $\alpha = 0.05$.

The decision rule is to reject H_0 if

$$|T| = \left| \frac{\hat{\beta}_2}{\text{SE}(\hat{\beta}_2)} \right| = \left| \frac{-4.228}{2.030} \right| = |-2.083| > t_{46,0.025} = 2.013.$$

So we reject H_0 and conclude that the "tax" variable is useful for prediction of "fuel" after having included the other variables.

L3.2 Confidence intervals for parameters

Recall that

$$|T| = \left| \frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \right| \sim t_{n-p}.$$

It follows that a $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\mathsf{SE}(\hat{\beta}_j)$$

where $\mathsf{SE}(\hat{\beta}_j) = s\sqrt{((\boldsymbol{Z}^T\boldsymbol{Z})^{-1})_{jj}}$.

Example L3.2 Consider example L3.1 regarding fuel consumption. Construct a 95% confidence interval for β_2.

A 95% confidence interval for β_2 is

$$\begin{aligned}\hat{\beta}_2 \pm t_{46,0.025}\mathsf{SE}(\hat{\beta}_j) &= -4.228 \pm 2.013 \times 2.030 \\ &= (-8.31, -0.14)\end{aligned}$$

L3.3 Tests for the existence of regression

We want to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ vs. $H_1 : \beta_j \neq 0$ for some j at level α.

Note that if we reject H_0 we are saying that the model $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1}X_{(p-1)i}$ has some ability to explain the variance that we are observing in Y. (i.e. There exists a linear relationship between the explanatory variables and the response variable.)

If D_0 is the model deviance under the null hypothesis and D_1 is the model deviance under the alternative hypothesis, then the **decision rule** is reject H_0 if

$$F = \frac{(D_0 - D_1)/(p-1)}{D_1/(n-p)} > F_{p-1,n-p,\alpha}.$$

Example L3.3 For the data in Example L3.1 models

$$M_1 : \mathsf{fuel} = \beta_0 + \beta_1\mathsf{dlic} + \beta_2\mathsf{tax} + \beta_3\mathsf{inc} + \beta_4\mathsf{road}$$

and

$$M_0 : \text{fuel} = \beta_0$$

were fitted with residual sum of squares $D_1 = 193700$ and $D_0 = 395694.1$, respectively.

Test $H_0 : \beta_1 = \cdots = \beta_4 = 0$ vs. $H_1 : \beta_j \neq 0$ for some $j = 1, \ldots, 4$ at level $\alpha = 0.05$.

The decision rule is to reject H_0 if

$$\begin{aligned} F &= \frac{(395694.1 - 193700)/(5-1)}{193700/(51-5)} = \frac{50498.525}{4210.870} \\ &= 11.99 > F_{4,46,0.05} = 2.574. \end{aligned}$$

Therefore, we reject H_0 and can say that the linear model has some power in explaining the variability in fuel.

L4 ANOVA tables and F tests

L4.1 The residuals

Consider the linear model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Recall,
1. We *assume* $E[\boldsymbol{\epsilon}] = \boldsymbol{0}$; Var$(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.
2. $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$ is the *LSE* of $\boldsymbol{\beta}$.
3. $\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}}$ is the $n \times 1$ vector of *fitted values*.
4. $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$, where $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$.

Let $\boldsymbol{r} = \hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ be the $n \times 1$ vector of *residuals*. Note that

$$\boldsymbol{r} = \hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y},$$

where $\boldsymbol{I}_n - \boldsymbol{P}$ is also symmetric idempotent and trace $(\boldsymbol{I}_n - \boldsymbol{P}) = \text{rank}(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} = n - p$.

Theorem The vector of fitted values is orthogonal to the vector of residuals. i.e.

$$\hat{\boldsymbol{y}}^T\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{y}} = 0.$$

Proof

$$\begin{aligned}
\hat{\boldsymbol{y}}^T\hat{\boldsymbol{\epsilon}} &= (\boldsymbol{P}\boldsymbol{y})^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{P}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} \\
&= \boldsymbol{y}^T\boldsymbol{P}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} = 0,
\end{aligned}$$

since $\boldsymbol{P}$ is orthogonal (i.e. $\boldsymbol{P}^T = \boldsymbol{P}$) and idempotent (i.e. $\boldsymbol{P}^2 = \boldsymbol{P}$).

Therefore, $\boldsymbol{y}$ can be written as a linear combination of orthogonal vectors. i.e.

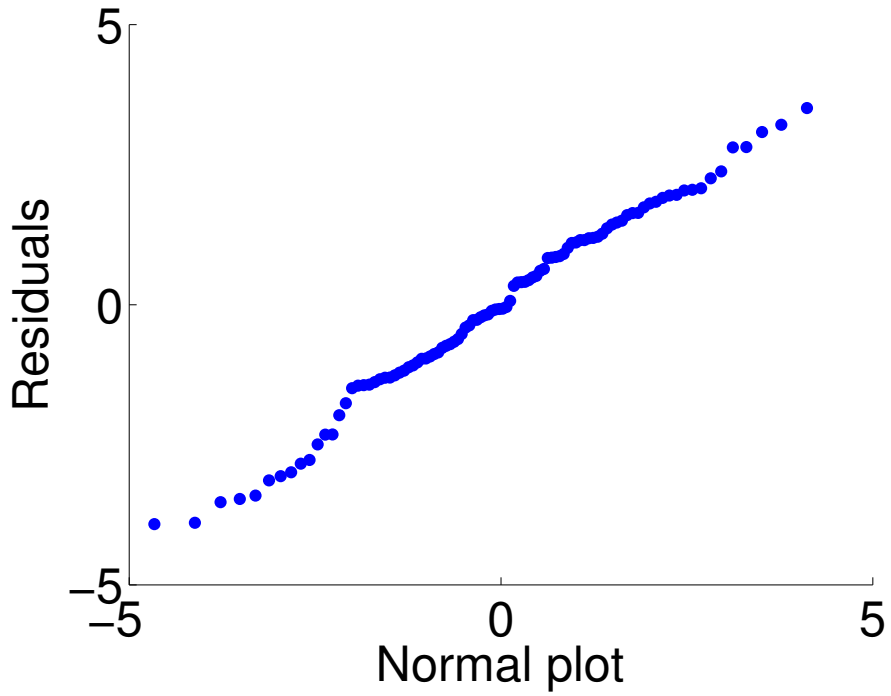
$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\epsilon}}.$$

The normal linear model assumes $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. We would expect the sample residuals, $\hat{\boldsymbol{\epsilon}}$ to exhibit many of the properties of the error terms.

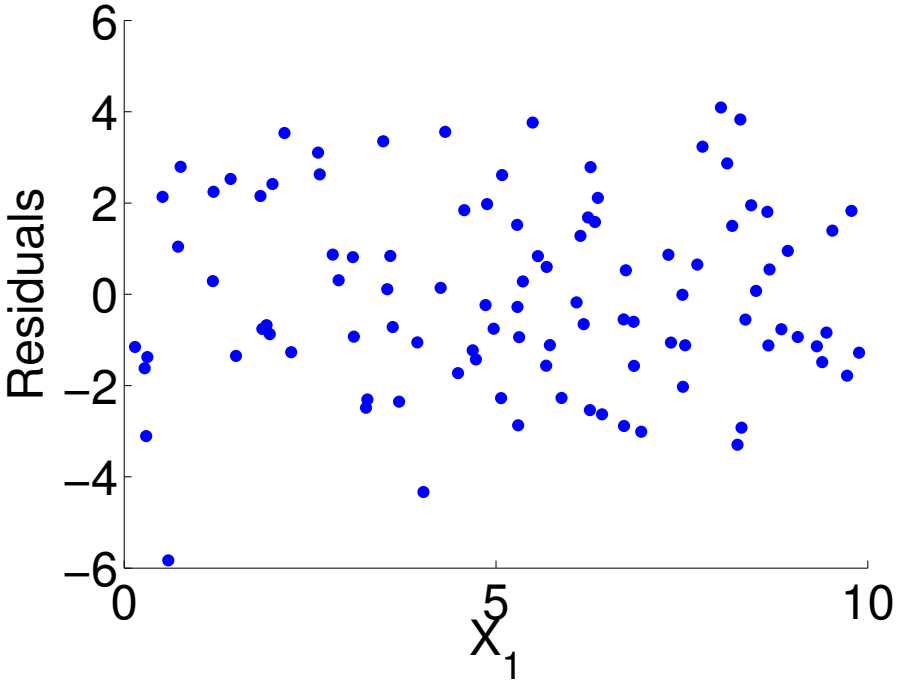
We can use the following graphical methods of detect model violations:

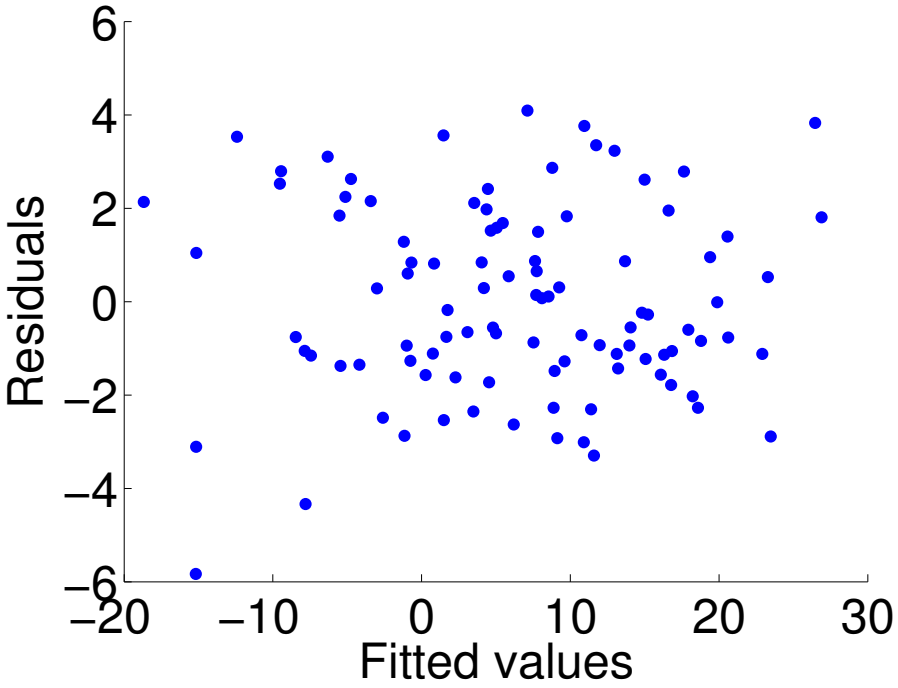
1. Scatter plot of residuals vs. each predictor variable.
2. Scatter plot of residuals vs. the fitted values.
3. Normal probability or quantile plot of residuals.

This is an example of a QQ plot for a correctly fitted model:

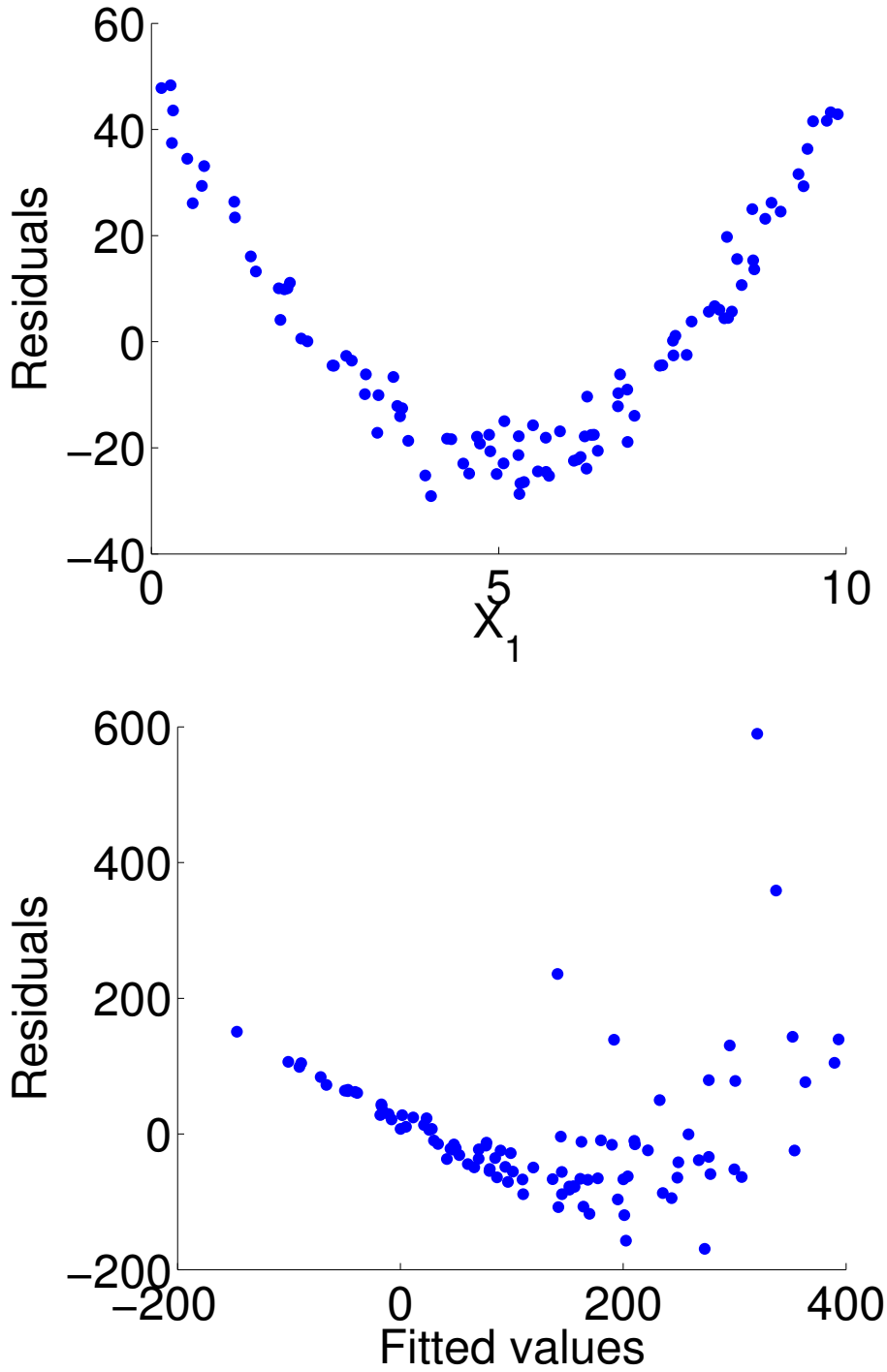


These are examples of scatter plots for a correctly fitted model:

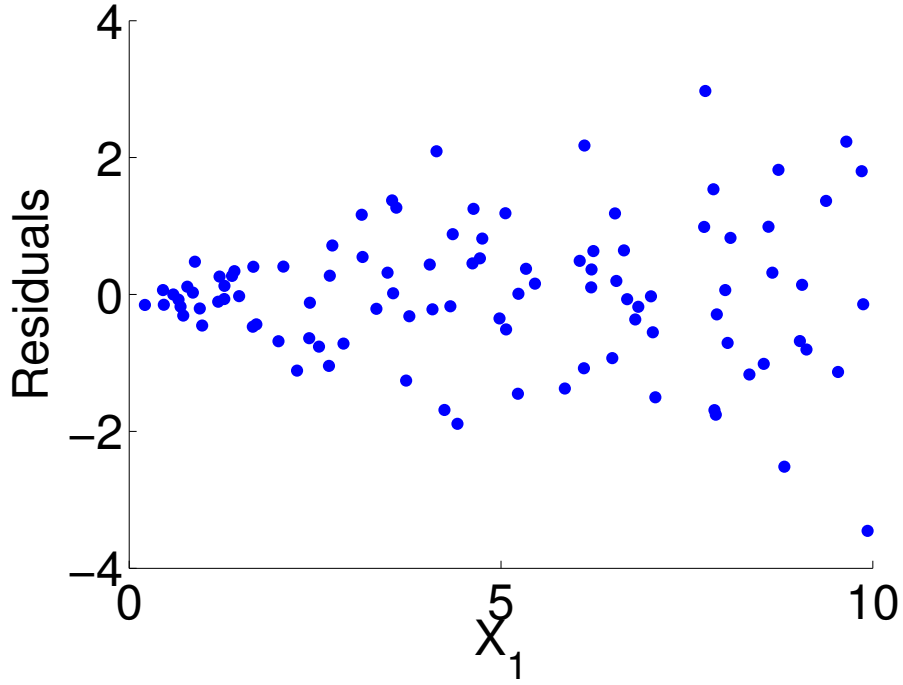


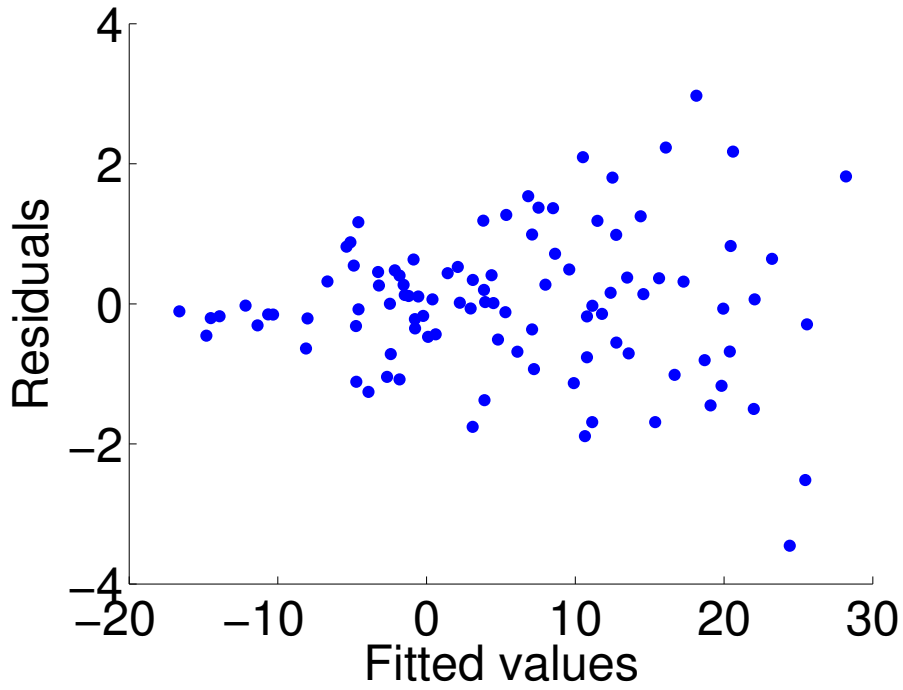


These are examples for a model where the X or Y variable needs transforming:



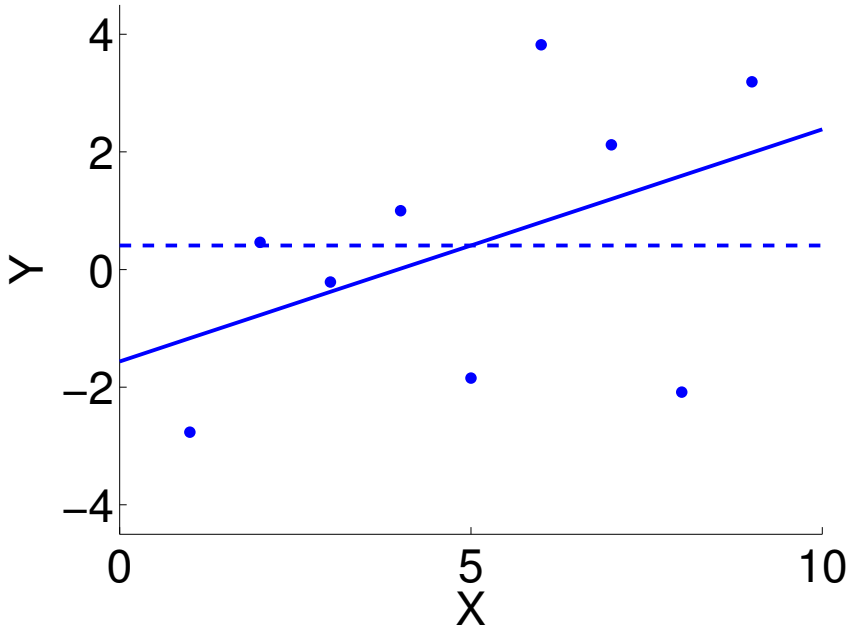
These are examples for a model where the variance is not constant:





L4.2 Sums of squares

Let y_i be the ith observation, $\hat{y}_i$ be the ith fitted value and $\bar{y}$ be the mean of the observed values.



Then,

$$
\begin{aligned}
(y_i - \bar{y}) &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}), \\
(y_i - \bar{y})^2 &= [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}), \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&\quad +2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).
\end{aligned}
$$

Now,

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i-\sum_{i=1}^{n}(y_i-\hat{y}_i)\bar{y}\\ &= \sum_{i=1}^{n}\hat{\epsilon}_i\hat{y}_i-\bar{y}\sum_{i=1}^{n}\hat{\epsilon}_i\\ &= \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{y}}-\bar{y}\sum_{i=1}^{n}\hat{\epsilon}_i=0-0=0.\end{aligned}$$

since $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{y}}$ are orthogonal and $\sum_{i=1}^{n}\hat{\epsilon}_i=0$ is one of the normal equations. Therefore,

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

1. We call SStot $=\sum_{i=1}^{n}(y_i-\bar{y})^2$ the *total sum of squares*. This is proportional to the total variability in y since SStot $=(n-1)\text{Var}(y)$. It does not depend on the choice of predictor variables in $\boldsymbol{Z}$.

2. We call SSres $=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ the *residual sum of squares*. This is a measure of the amount of variability in y the model was unable to explain. This is equivalent to the deviance of the model (SSres $=D$).

3. We call SSreg $=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ the *regression sum of squares*. This is the difference between SStot and SSres and is a measure of the amount of variability in y the model was able to explain.

From our above derivations, we get

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\\ \text{SStot} &= \text{SSres}+\text{SSreg}\end{aligned}$$

Definition The *coefficient of determination* is

$$R^2 = \frac{\text{SSreg}}{\text{SStot}}.$$

It measures the proportion of variability explained by the regression.

Notes

1. $0 \leq R^2 \leq 1$.
2. $R^2 = 1 - \dfrac{\text{SSres}}{\text{SStot}}$.
3. R^2 is often used as a measure of how well the regression model fits the data: the larger the R^2, the better the fit. One needs to be careful in interpreting how large is large.

Definition The *adjusted* R^2 is

$$R^2_{\text{adj}} = 1 - \frac{\text{SSres}/(n-p)}{\text{SStot}/(n-1)}.$$

It is often used to compare the fit of models with different numbers of parameters.

Under the null model, $Y_i = \beta_0 + \epsilon_i$, $\bar{y} = \hat{y}_i$, so in this special case, SStot = SSres = D, SSreg = 0 and $R^2 = R^2_{\text{adj}} = 0$.

L4.3 Analysis of Variance (ANOVA)

Recall that the F statistic used in the test for the existence of regression is

$$F = \frac{(D_0 - D_1)/(p-1)}{D_1/(n-p)}$$

where D_1 and D_0 are the model deviance or SSres under the alternative and null hypotheses respectively. We noted above that D_0, the deviance under the null hypothesis, is equivalent to SStot (under any model).

1. We call the numerator in the F statistic,

$$\text{MSreg} = \frac{D_0 - D_1}{p-1} = \frac{\text{SStot} - \text{SSres}}{p-1} = \frac{\text{SSreg}}{p-1}$$

the *mean square regression*.

2. We call the denominator in the F statistic,

$$\text{MSres} = \frac{D_1}{n-p} = \frac{\text{SSres}}{n-p}$$

the *mean square residual* and it is an unbiased estimator of σ^2. Similarly the *residual standard error*, $\text{RSE} = \sqrt{\text{MSres}}$ is an unbiased estimate of σ.

The quantities involved in the calculation of the F statistic are usually displayed in an ANOVA table:

Source	Df	Sum Sq	Mean Sq	F
Regression	$p-1$	SSreg	$\text{MSreg} = \frac{\text{SSreg}}{p-1}$	$F = \frac{\text{MSreg}}{\text{MSres}}$
Residual	$n-p$	SSres	$\text{MSres} = \frac{\text{SSres}}{n-p}$	
Total	$n-1$	SStot		

Example L4.1 For the data in Example L3.1 the model

$$\text{fuel} = \beta_0 + \beta_1 \text{dlic} + \beta_2 \text{tax} + \beta_3 \text{inc} + \beta_4 \text{road}$$

was fitted to the $n = 51$ observations with residual standard error, $\text{RSE} = 64.8912$. Summary statistics show $\text{Var}(\text{fuel}) = 7913.88$. Complete an ANOVA table and compute R^2 for the fitted model.

1. Note $p - 1 = 4$, $n - p = 46$ and $n - 1 = 50$.
2. $\text{SStot} = (n-1)\text{Var}(\text{fuel}) = 50 \times 7913.88 = 395694$.
3. $\text{MSres} = \text{RSE}^2 = 64.8912^2 = 4210.87$.
4. $\text{SSres} = (n-p)\text{MSres} = 46 \times 4210.87 = 193700$.

5. SSreg $=$ SStot $-$ SSres $= 395694 - 193700 = 201994$.
6. MSreg $=$ SSreg$/(p-1) = 201994/4 = 50498.50$.
7. $F =$ MSreg/MSres $= 50498.5/4210.87 = 11.99$.

Hence the completed ANOVA table is

Source	Df	Sum Sq	Mean Sq	F
Regression	4	201994	50498.50	11.99
Residual	46	193700	4210.87	
Total	50	395694		

Finally, $R^2 = \dfrac{\text{SSreg}}{\text{SStot}} = \dfrac{201994}{395694} = 0.5105.$

L4.4 Comparing models

Consider two models, M_1 and M_2, where M_2 is a simplification of M_1. For example,

$$\begin{aligned} M_1: &\quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \\ M_2: &\quad Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \epsilon. \end{aligned}$$

The residual sum of squares from model M_1 will always be less then M_2, but we can test

$$H_0 : \beta_1 = \beta_3 = 0 \text{ vs. } H_1 : \beta_1 \neq 0, \beta_3 \neq 0$$

at level α to test if removing these terms significantly increases the residual sum of squares.

Let $D_1 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ be the model deviance (or SSres) for model M_1.

Let $D_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ be the model deviance (or SSres) for model M_2.

The **decision rule** is to reject H_0 if

$$F = \frac{(D_2 - D_1)/q}{D_1/(n-p)} > F_{q,n-p,\alpha},$$

where n is the number of observations,
p is the number of parameters in M_1 and
q is the number of parameters fixed to reduce M_1 to M_2.

For the example above, $p = 5$ and $q = 2$.

Example L4.2 Let model 1 be

$$\text{fuel} = \beta_0 + \beta_1 \text{dlic} + \beta_2 \text{tax} + \beta_3 \text{inc} + \beta_4 \text{road}$$

and let model 2 be

$$\text{fuel} = \beta_0 + \beta_1 \text{dlic} + \beta_3 \text{inc}$$

For the fuel data in Example L3.1, the residual sum of squares was 193700 for model 1 and 249264 for model 2. Test which model fits the data better.

Equivalently, we test:

$$H_0 : \beta_2 = \beta_4 = 0 \text{ vs. } H_1 : \beta_2 \neq 0, \beta_4 \neq 0$$

at $\alpha = 0.05$. The decision rule is to reject H_0 if

$$F = \frac{(D_2 - D_1)/q}{D_1/(n-p)} > F_{q,n-p,\alpha} = F_{2,46,0.05} = 3.20.$$

Substituting in the data gives,

$$F = \frac{(249264 - 193700)/2}{193700/(51-5)} = 18.47.$$

Consequently, we will reject H_0. Model 1 fits the data better at $\alpha = 0.05$.

Let's consider the more general case where the basic model M_1 is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon.$$

We denote

$$\text{SSreg}(M_1) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = R(\beta_1, \beta_2, \ldots, \beta_{p-1} | \beta_0),$$

assuming there is a constant in the model.

Goal We want to build a regression model which "best" describes the response variable. Hence we would like to explain as much of the variance in Y as possible, yet keep the model as simple as possible (Principle of Parsimony). Consequently we want to determine which explanatory variables are worthwhile to include in the final model.

Idea Explanatory variables should be included in the model if the extra portion of the regression sum of squares (called the extra sum of squares) which arises from their inclusion in the model is relatively large compared to the unexplained variance in the model (residual sum of squares).

Consider a second model M_2 which is a simplification of M_1, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \epsilon,$$

where $k < p$. Then

$$\text{SSreg}(M_2) = R(\beta_1, \beta_2, \ldots, \beta_{k-1} | \beta_0).$$

Definition The *extra sum of squares* due to the inclusion of the terms $\beta_k X_k + \cdots + \beta_{p-1} X_{p-1}$ in the model is

$$\text{SSreg}(M_1) - \text{SSreg}(M_2).$$

It is denoted

$$R(\beta_k, \ldots, \beta_{p-1}|\beta_0, \beta_1, \ldots, \beta_{k-1}) = \\ R(\beta_1, \beta_2, \ldots, \beta_{p-1}|\beta_0) - R(\beta_1, \beta_2, \ldots, \beta_{k-1}|\beta_0).$$

The extra sum of squares has $q = p - k$ degrees of freedom where q is the number of parameters on the left of the bar, i.e. number of explanatory variables added to the reduced model to make the full model.

We can test
H_0 : Reduced model, M_2, best describes the data
H_1 : Full model, M_1, best describes the data

The **decision rule** is to reject H_0 if

$$F = \frac{R(\beta_k, \ldots, \beta_{p-1}|\beta_0, \ldots, \beta_{k-1})/q}{\text{SSres}(M_1)/(n-p)} > F_{q,n-p,\alpha}.$$

Rejecting H_0 implies the full model describes the data better, so we should include the variables $X_k, \ldots, X_{p-1}$ (jointly) in our model.

The test for the existence of regression is a special case of this type of test, where $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (i.e. the reduced model is $Y = \beta_0 + \epsilon$). Note that $\text{SSreg}(M_1) = R(\beta_1, \beta_2, \ldots, \beta_{p-1}|\beta_0)$ is the extra sum of squares in this case.

L4.5 Sequential sum of squares

Definition The *sequential sum of squares* for each j is

$$R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}) = R(\beta_1, \beta_2, \ldots, \beta_j|\beta_0) - R(\beta_1, \beta_2, \ldots, \beta_{j-1}|\beta_0)$$

and is the extra sum of squares that one incurs by adding the explanatory variable X_j to the model given that $X_1, \ldots, X_{j-1}$ are already present.

The sequential sum of squares is often given in addition to the basic ANOVA table.

Source	Df	Sum Sq	Mean Sq	F
X_1	df_1	SSseq_1	$\mathsf{MSseq}_1 = \frac{\mathsf{SSseq}_1}{\mathsf{df}_1}$	$F = \frac{\mathsf{MSseq}_1}{\mathsf{MSres}}$
X_2	df_2	SSseq_2	$\mathsf{MSseq}_2 = \frac{\mathsf{SSseq}_2}{\mathsf{df}_2}$	$F = \frac{\mathsf{MSseq}_2}{\mathsf{MSres}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{p-1}	df_{p-1}	SSseq_{p-1}	$\mathsf{MSseq}_{p-1} = \frac{\mathsf{SSseq}_{p-1}}{\mathsf{df}_{p-1}}$	$F = \frac{\mathsf{MSseq}_{p-1}}{\mathsf{MSres}}$
Residuals	$n-p$	SSres	$\mathsf{MSres} = \frac{\mathsf{SSres}}{n-p}$	

Note that given the sequential sum of squares, one can calculate

$$R(\beta_j, \beta_{j+1}, \ldots, \beta_k | \beta_0, \beta_1, \ldots, \beta_{j-1}) = \sum_{i=j}^{k} \mathsf{SSseq}_i.$$

However, one cannot calculate the nonsequential sums of squares in this manner, for example, $R(\beta_1, \beta_3, \beta_5 | \beta_0, \beta_2, \beta_4)$.

Example L4.3 The output from R for the fuel data in Example L3.1 is:

Source	df	Sum Sq	Mean Sq
dlic	1	86854	86854
tax	1	19159	19159
inc	1	61408	61408
road	1	34573	34573
Residuals	46	193700	4211

Test:

$$H_0: \quad Y = \beta_0 + \beta_1 \mathsf{dlic} + \beta_2 \mathsf{tax} + \epsilon$$
$$H_1: \quad Y = \beta_0 + \beta_1 \mathsf{dlic} + \beta_2 \mathsf{tax} + \beta_3 \mathsf{inc} + \beta_4 \mathsf{road} + \epsilon$$

The decision rule is to reject H_0 if

$$F = \frac{R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2)/2}{\mathsf{SSres}/(n-p)} > F_{2,n-p,0.05}$$

where

$$\begin{aligned} R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2) &= R(\beta_3 | \beta_0, \beta_1, \beta_2) + R(\beta_4 | \beta_0, \beta_1, \beta_2, \beta_3) \\ &= 61408 + 34573 = 95981. \end{aligned}$$

Hence,

$$F = \frac{95981/2}{4211} = 11.40 > F_{2,46,0.05} = 3.20.$$

Therefore we will reject H_0 at $\alpha = 0.05$. Including the variables "inc" and "road" significantly improves the model. An approximate p value is $p = P(F_{2,46} > 11.40) < 0.001$.

Note that

SSreg $= R(\beta_1|\beta_0)+R(\beta_2|\beta_0, \beta_1)+R(\beta_3|\beta_0, \beta_1, \beta_2)+R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3)$.

Definition The *partial sum of squares* for each j is

$$\begin{aligned} &R(\beta_j | \beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{p-1}) \\ &\quad = R(\beta_1, \beta_2, \ldots, \beta_{p-1} | \beta_0) - R(\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{p-1} | \beta_0) \end{aligned}$$

and is the extra sum of squares that one incurs by adding the explanatory variable X_j to the model given that $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{p-1}$ are already present.

Note that the F test for testing

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

at level α, is equivalent to the t test for the individual parameter since $t^2_{n-p} = F_{1,n-p}$