1 Least Squares Estimation of $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$

- The **least squares estimators** $\boldsymbol{b_0, b_1, \ldots, b_{p-1}}$ of the parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are the values that minimize

$$Q(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{ip-1}X_{ip-1})^2$$

  as a function of $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

- The estimates $b_0, b_1, \ldots, b_{p-1}$ are obtained by setting the partial derivatives of Q (with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$) equal to zero and solving the resulting system of equations (called the ***normal equations***):

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= 0 \\ \frac{\partial Q}{\partial \beta_1} &= 0 \\ &\vdots \\ \frac{\partial Q}{\partial \beta_{p-1}} &= 0 \end{aligned}$$

  It can be shown that the solutions $b_0, b_1, \ldots, b_{p-1}$ this system of equations are the solutions to the system of equations:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y} \tag{1}$$

  where

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

  and $\mathbf{X}$ is the $n \times p$ **design matrix**.

- When the $n \times n$ matrix $\mathbf{X}^T\mathbf{X}$ is **nonsingular** (i.e. when it's columns are linearly independent), which is usually the case, it's **invertible**. Multiplying both sides of (1) by the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ of $\mathbf{X}^T\mathbf{X}$ gives the vector of coefficient estimates $\mathbf{b}$:

Least Squares Estimates of $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$ (Matrix Approach): The

vector of estimated coefficients $\mathbf{b}$ is obtained by:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

- **Comment**: $\mathbf{X}^T\mathbf{X}$ will be **singular**, and therefore **non-invertible**, under either of these conditions:
  - ▷ One of the predictor variables is **constant** (i.e. it's column in $\boldsymbol{X}$ is all one value).
  - ▷ One of the predictor variables is a *multiple* of one of the others (i.e. it's column in $\boldsymbol{X}$ is a multiple of another column in $\boldsymbol{X}$), or more generally, is a *linear combination* of two or more others.

  In this case, there will be **infinitely many solutions** to the normal equations, so $\boldsymbol{b_0, b_1, \ldots, b_{p-1}}$ **won't be uniquely determined**.

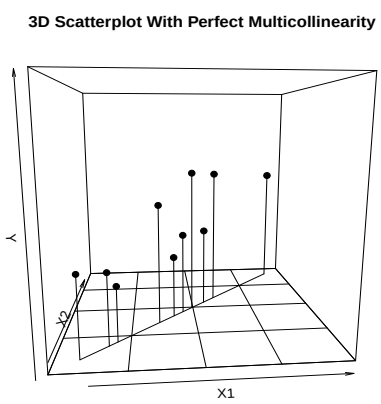
  For example, coefficient estimates for multiple regression *couldn't* be obtained using either of these data sets:

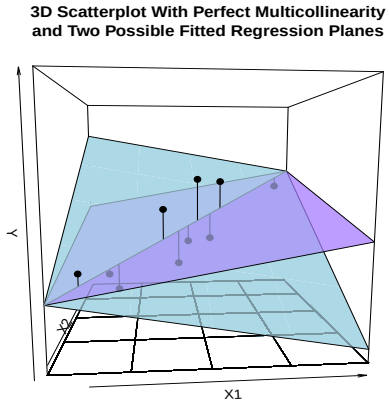
Data Set 1

X_1	X_2	Y
4	7	3
8	7	5
8	7	4
13	7	7
10	7	9
14	7	11

Data Set 2

X_1	X_2	Y
4	8	3
8	16	5
8	16	4
13	26	7
10	20	9
14	28	11





- Once the coefficients have been estimated, the ***fitted multiple regression model*** is

> **Fitted Regression Model**:
>
> $$\hat{Y} \;=\; b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_{p-1} X_{p-1}\,.$$

When there are just two predictors X_1 and X_2, (i.e. $p = 3$), the fitted model defines a *plane*, and when there are more than two ($p > 3$) it defines a *hyperplane*.

2 Fitted Values and Residuals

- The ***fitted values*** (or ***predicted values***), denoted $\hat{\boldsymbol{Y}}_{\boldsymbol{1}}, \hat{\boldsymbol{Y}}_{\boldsymbol{2}}, \ldots, \hat{\boldsymbol{Y}}_{\boldsymbol{n}}$, are the estimates of $\mathrm{E}(Y_i)$ for the *observed* values of the predictors:

> **Fitted (Predicted) Values**:
>
> $$\hat{Y}_i \;=\; b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{ip-1}\,.$$

- In matrix notation, the n fitted values can be written as

$$\hat{\mathbf{Y}} \;=\; \mathbf{X}\mathbf{b}\,, \tag{3}$$

where $\hat{\mathbf{Y}}$ is the $n \times 1$ **vector of fitted values**:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

- Note from (2) and (3) that if we define the ***hat matrix*** $\mathbf{H}$ to be

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{4}$$

then the **fitted values** can be written as

$$\hat{\mathbf{Y}} = \mathbf{HY}.$$

- The ***residuals*** $e_1, e_2, \ldots, e_n$ are defined by

Residuals:

$$e_i = Y_i - \hat{Y}_i$$

- In matrix notation, the n **residuals** can be written as

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}},$$

with $\boldsymbol{e}$ denoting the $n \times 1$ **residual vector**

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

3 Partitioning the Total Variation into Sums of Squares

- As for simple linear regression, we define the ***total sum of squares*** **SSTO** to be

Total Sum of Squares:

$$\text{SSTO} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

SSTO measures the **total variation** in $Y_1, Y_2, \ldots, Y_n$.

- The ***error sum of squares*** **SSE** is defined to be

Error Sum of Squares:

$$\text{SSE} \;=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

SSE measures variation in $Y_1, Y_2, \ldots, Y_n$ due to **random error**.

- The ***regression sum of squares*** **SSR** is again defined as

Regression Sum of Squares:

$$\text{SSR} \;=\; \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

SSR measures variation in $Y_1, Y_2, \ldots, Y_n$ that's **due to the predictors $X_1, X_2, \ldots, X_{p-1}$, i.e. due to the linear relationship between Y and those predictors**.

Partition of the Total Variation in the Responses: It can be shown that

$$\text{SSTO} \;=\; \text{SSR} + \text{SSE}\,.$$

- The corresponding degrees of freedom are

Degrees of Freedom:

$$\begin{aligned} \text{Total df (for SSTO)} &= n-1 \\ \text{df for Regression (SSR)} &= p-1 \\ \text{df for Error (SSE)} &= n-p \end{aligned}$$

Notice that the **degrees of freedom for SSE** is the sample size n **minus** the **number of parameters p** in the model, and that

$$\textbf{Total df} \;=\; \textbf{df for Regression} + \textbf{df for Error}$$

- The ***mean squares*** are again the sums of squares divided by their degrees of freedom.

Mean Squares: The ***mean square for regression*** and ***mean squared error***, denoted **MSR** and **MSE**, respectively, are

$$\text{MSE} = \frac{\text{SSE}}{n-p}$$

$$\text{MSR} = \frac{\text{SSR}}{p-1}$$

4 Regression Model *F* Test

- The ***regression model F test*** *simultaneously* tests whether $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are **all** equal to zero, i.e. tests

$$\begin{aligned} H_0 &: \ \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \\ H_a &: \ \text{Not all } \beta_1, \beta_2, \ldots, \beta_{p-1} \text{ equal } 0 \end{aligned}$$

- It can be shown that

$$\text{E(MSE)} = \sigma^2$$

and

$$\text{E(MSR)} = \sigma^2 + \begin{pmatrix} \text{Non-negative term} \\ \text{involving } \beta_k\text{'s that's} \\ \text{zero when } H_0 \text{ is true} \end{pmatrix}$$

Thus if H_0 is true,

$$\text{E(MSR)} = \text{E(MSE)} = \sigma^2 .$$

On the other hand H_a is true,

$$\text{E(MSR)} > \text{E(MSE)} .$$

- The ***F* test statistic** is

Test Statistic: The test statistic for the *regression model F test* is

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{5}$$

Fact 4.1 Under the multiple linear regression model, with the ϵ_i's independent $N(0, \sigma)$, if $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ is true,

$$F \sim F(p-1, n-p),$$

i.e. the test statistic (5) follows an F distribution with numerator and denominator degrees of freedom $p-1$ and $n-p$, respectively.

- **Large** values of F (i.e. values substantially greater than one) provide evidence against H_0. The **p-value** is the tail area to the **right** of the observed F value under the $F(p-1, n-p)$ distribution.

5 The Regression ANOVA Table

- The sums of squares, degrees of freedom, mean squares, F test statistic, and p-value are usually organized in a ***regression ANOVA table***:

Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**F**	**P-value**
Regression Model	$p-1$	SSR	MSR = SSR/(p-1)	MSR/MSE	p
Error	$n-p$	SSE	MSE = SSE/$(n-p)$		
Total	$n-1$	SSTO			

6 The R^2 and the Adjusted R^2

- We define the ***coefficient of multiple determination***, denoted $\boldsymbol{R^2}$, a measure of the strength of the association between Y and the predictors $X_1, X_2, \ldots, X_{p-1}$, by

Coefficient of Multiple Determination:

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}} \tag{6}$$

R^2 is interpreted as **the proportion of variation in $\boldsymbol{Y_1, Y_2, \ldots, Y_n}$** that can be **explained** by the **model** containing the $\boldsymbol{p}$ predictors $\boldsymbol{X_1, X_2, \ldots, X_{p-1}}$.

- Note that

$$\mathbf{0 \leq R^2 \leq 1}$$

Values of R^2 **close to one** indicate:

- ▷ A large proportion of the variation in Y is explained by $X_1, X_2, \ldots, X_{p-1}$.
- ▷ There's a strong (linear) association between Y and these predictor variables.
- ▷ The regression model fits the data well.

Values of R^2 **close to zero** indicate

- ▷ Only a small proportion of the variation in Y, if any, is explained by $X_1, X_2, \ldots, X_{p-1}$.
- ▷ There's only a weak (linear) association, or none at all.
- ▷ The regression model doesn't fit the data very well.

- Be aware that **adding more predictor variables** to a model ***always*** results in a **larger R^2** since more of the variation in the Y will be explained by the model (and the SSE will be smaller).

For this reason, it's sometimes preferable to use the ***adjusted R^2***, denoted R_a^2 and defined as

Adjusted R^2:

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)}$$

R_a^2 is "adjusted" for the number of predictor variables in the model. Its value can actually *decrease* when another predictor is added to the model if that predictors doesn't explain much of the remaining Y variation not already explained by the other predictors in the model.