1 *F* Test for Lack of Fit

1.1 Full and Reduced Models

- The ***lack of fit test*** will be used to decide whether or not a **linear model** is appropriate.

  The test requires that for at least one value of X we have multiple Y observations.

- **Notation**:

  - ▷ Let $\boldsymbol{X_1, X_2, \ldots, X_c}$ denote the *distinct*, *unique* observed X levels.
  - ▷ For the jth observed level X_j of X, let $\boldsymbol{n_j}$ denote the number of Y observations at that X_j.
  - ▷ Let $\boldsymbol{Y_{ij}}$, for $i = 1, 2, \ldots, n_j$, denote the ith observed Y value at level X_j of X.
  - ▷ Denote the total sample size by $\boldsymbol{n}$, i.e. $n = \sum_{j=1}^{c} n_j$.

- Define the ***full model*** as

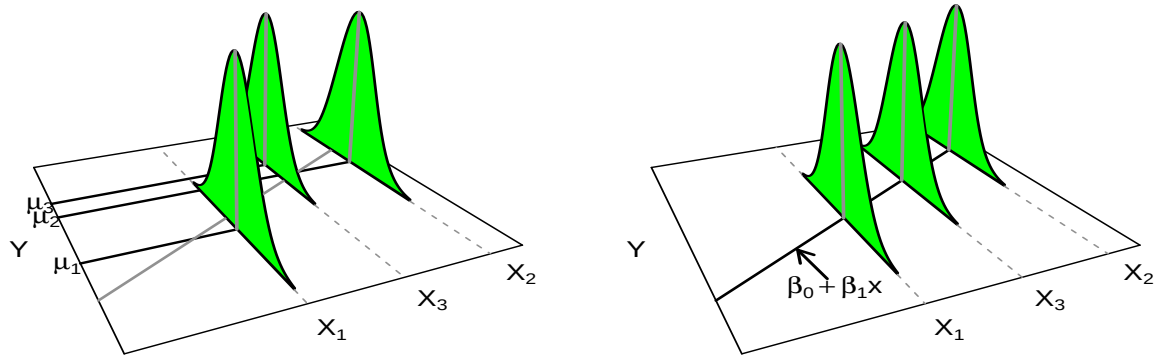
$$Y_{ij} \;=\; \mu_j + \epsilon_{ij} \tag{1}$$

  and the ***reduced model*** to be the usual simple linear regression model

$$Y_{ij} \;=\; \beta_0 + \beta_1 X_j + \epsilon_{ij} \tag{2}$$

  where in both models the ϵ_{ij}'s are iid $N(0, \sigma^2)$.

Full Model **Reduced Model**



- Note that the reduced model is a **special case** of the full model (for which the μ_j's fall on a straight line). The *lack of fit test* can be viewed as a **decision** between these **two models**. It uses the **general linear *F* test approach**.

- We'll test

$$\begin{aligned} H_0: \ \mathrm{E}(Y) &= \beta_0 + \beta_1 X \\ H_a: \ \mathrm{E}(Y) &\neq \beta_0 + \beta_1 X \end{aligned}$$

When H_0 is true, the μ_j's in the full model satisfy

$$\mu_j = \beta_0 + \beta_1 X_j \, .$$

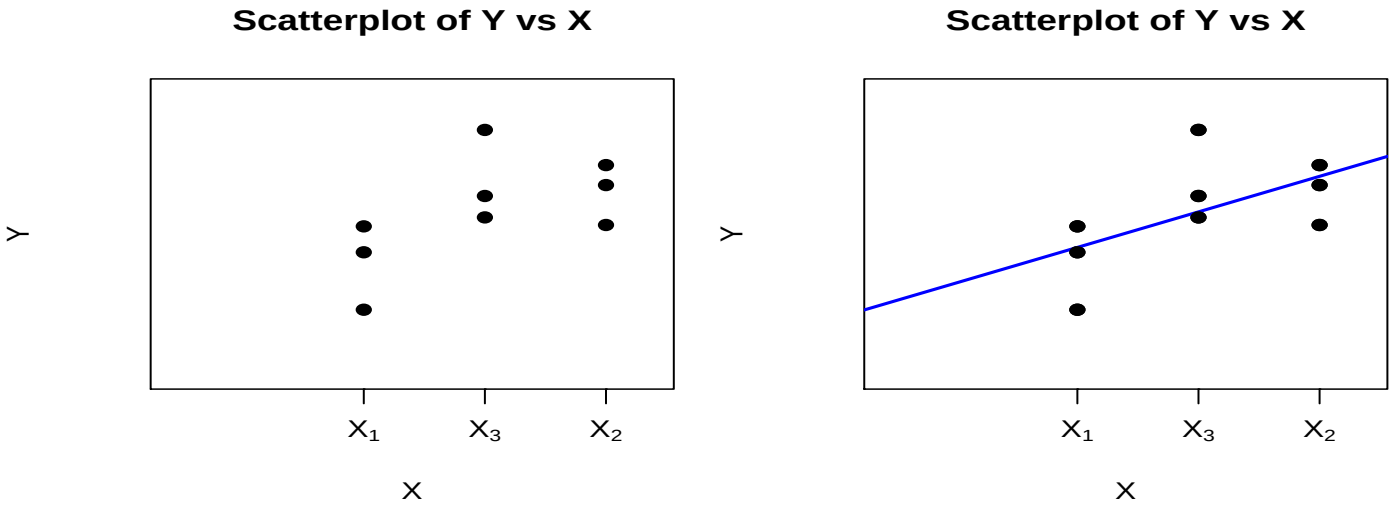


Figure 1: Data for illustrating the lack of fit test.

- Note that the reduced model is a special case of the full model.

1.2 Sums of Squares

- For the **full model** (1), the least squares estimate of μ_j, for $j = 1, 2, \ldots, c$, is the sample mean of the responses associated with X_j,

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} \, .$$

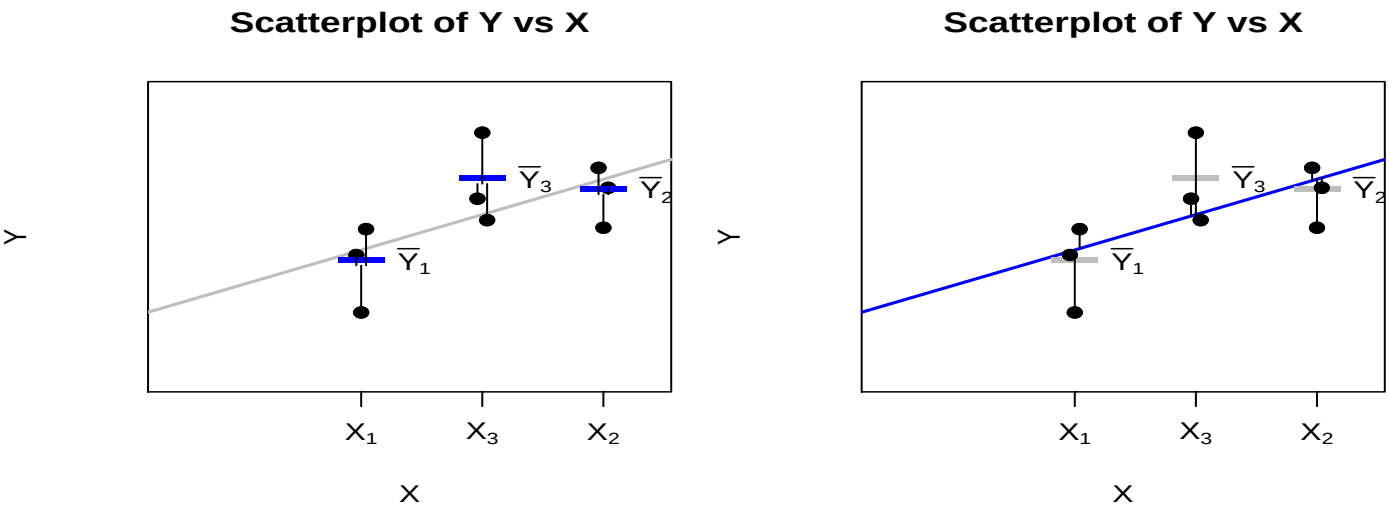


Figure 2: SSE(F) measures the residual variation in the left plot and SSE(R) that in the right plot.

Thus for a given value of X_j, the fitted values $\hat{Y}_{ij}$ are equal to the sample mean of the responses associated with that X_j, and the residuals are deviations away from a sample mean:

$$e_{ij} \;=\; Y_{ij} - \hat{Y}_{ij} \;=\; Y_{ij} - \bar{Y}_j$$

It follows that for the **full model**, the **error sum of squares** is the sum of squared deviations away from the sample means:

$$\text{SSE(F)} \;=\; \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2\,. \tag{3}$$

The **degrees of freedom** for SSE(F) is

$$df_F \;=\; \sum_{j=1}^{c}(n_j - 1) \;=\; \sum_{j=1}^{c} n_j - c \;=\; n - c$$

because the c individual sums of squares $\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$ in (3) each have $n_j - 1$ degrees of freedom.

- For the **reduced model** (2), the fitted values $\hat{Y}_{ij}$ and residuals e_{ij} are just the usual fitted values and residuals after fitting the simple linear regression model, and so the error sum of squares is just the usual SSE:

$$\text{SSE(R)} \;=\; \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - (b_0 + b_1 X_j))^2 \;=\; \text{SSE}$$

where SSE is the usual error sum of squares from simple linear regression.

The **degrees of freedom** for SSE(R) is thus just the usual degrees of freedom for SSE:

$$df_R \;=\; n - 2\,.$$

1.3 *F* Test for Lack of Fit

- The ***F* test statistic for lack of fit** can be viewed in terms of the *general linear test approach*. The test statistic is

Test Statistic: The test statistic for the *lack of fit F test*, in terms of the general linear test approach, is

$$F \;=\; \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_R - df_F)}{\text{SSE(F)}/df_F} \;=\; \frac{(\text{SSE(R)} - \text{SSE(F)})/(c-2)}{\text{SSE(F)}/(n-c)}\,. \quad (4)$$

- In the context of the *lack of fit test*, SSE(F) is sometimes called the ***sum of squares for pure error***, denoted **SSPE**.

Sum of Squares for Pure Error:

$$\text{SSPE} \;=\; \text{SSE(F)} \quad (5)$$

- We define the ***sum of squares for lack of fit***, denoted **SSLF**, to be the difference between SSE(R) and SSE(F):

Sum of Squares for Lack of Fit: We define the sum of squares for lack of fit to be:

$$\text{SSLF} \;=\; \text{SSE(R)} - \text{SSE(F)} \;=\; \text{SSE} - \text{SSPE}\,. \quad (6)$$

Fact 1.1 It can be shown that

$$\text{SSLF} \;=\; \sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(\bar{Y}_j - \hat{Y}_{ij}\right)^2 .$$

- Using these definitions for SSPE and SSLF, the test statistic (4) can be written as

Test Statistic: The test statistic for the *lack of fit F test*, in terms of the sums of squares for lack of fit, is

$$F \;=\; \frac{\text{SSLF}/(c-2)}{\text{SSPE}/(n-c)} \;=\; \frac{\text{MSLF}}{\text{MSPE}}, \tag{7}$$

where

$$\text{MSLF} \;=\; \frac{\text{SSLF}}{c-2} \qquad \text{and} \qquad \text{MSPE} \;=\; \frac{\text{SSPE}}{n-c}$$

are the ***mean square for lack of fit*** and ***mean square for pure error***, respectively.

Fact 1.2 It can be shown that when H_0 is true,

$$F \;\sim\; F(c-2, n-c),$$

i.e. the test statistic (7) follows an F distribution with numerator and denominator degrees of freedom $c-2$ and $n-c$, respectively.

- **Large** values of F (i.e. values substantially greater than one) provide evidence against H_0, i.e. evidence that a straight line doesn't fit the data. The **p-value** is the tail area to the **right** of the observed F value under the $F(c-2, n-2)$ distribution.

1.4 ANOVA Table for the *F* Test for Lack of Fit

- We can divide SSE into parts corresponding to pure error and lack of fit.

Fact 1.3 It can be shown that:

$$\text{SSE} \;=\; \text{SSPE} + \text{SSLF} \tag{8}$$

where SSE is the usual error sum of squares from simple linear regression and

SSPE and SSLF are the sums of squares for pure error and lack of fit given by (5) and (6).

To see why, note that we can decompose a deviation $Y_{ij} - \hat{Y}_j$ away from the fitted regression line as:

$$Y_{ij} - \hat{Y}_j \;=\; Y_{ij} - \bar{Y}_j \;+\; \bar{Y}_j - \hat{Y}_j \tag{9}$$

where

$$\hat{Y}_j \;=\; b_0 + b_1 X_j$$

Squaring both sides, summing over i and j, and using the fact that $\sum_i \sum_j (Y_{ij} - \bar{Y}_j)(\bar{Y}_j - \hat{Y}_j) = 0$, we get (8).

- The sums of squares, degrees of freedom, mean squares, F statistic, and p-value associated with the lack of fit test are usually summarized in a ***lack of fit ANOVA table***. See the book.