

# Chapter 2

## Inferences in Regression Analysis

We will focus on the normal error model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where

1.  $\beta_0$  and  $\beta_1$  are parameters to be estimated,
2.  $X_i$ 's are known constants,
3.  $\epsilon_i$ 's are *iid* (independent and identically distributed)  $N(0, \sigma^2)$ .

For convenience, we often use  $b_0$  and  $b_1$  as the estimators of  $\beta_0$  and  $\beta_1$  respectively, *i.e.*,  $b_0 = \hat{\beta}_0$  and  $b_1 = \hat{\beta}_1$ .

## 2.1 Inferences concerning $\hat{\beta}_1$

### 2.1.1 Distribution of $\hat{\beta}_1$

We want to test

$$H_0 : \beta_1 = 0 \quad \text{and} \quad H_1 : \beta_1 \neq 0.$$

For the hypothesis test above, we have to know the distribution of  $\hat{\beta}_1$  under  $H_0$ .

**Fact.** *If the independent random variable  $Y_i$  has a normal distribution with  $N(\mu_i, \sigma_i^2)$ , then the linear combination  $\sum_{i=1}^n (c_i Y_i + d_i)$  also has a normal with  $N(\sum_{i=1}^n (c_i \mu_i + d_i), \sum_{i=1}^n c_i^2 \sigma_i^2)$ .*

We derived the regression estimator

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n k_i (Y_i - \bar{Y}) = \sum_{i=1}^n k_i Y_i, \quad (2.1)$$

where  $k_i = (X_i - \bar{X})/S_{xx}$  and  $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$ . Then we have

$$\hat{\beta}_1 \sim N\left(\sum_{i=1}^n k_i (\beta_0 + \beta_1 X_i), \sum_{i=1}^n k_i^2 \sigma^2\right).$$

Using  $\sum_{i=1}^n k_i = 0$ ,  $\sum_{i=1}^n k_i X_i = 1$ ,  $\sum_{i=1}^n k_i^2 = 1/S_{xx}$ , we have

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

**Theorem 2.1.** *The regression estimator  $\hat{\beta}_1$  is a BLUE (best linear unbiased estimator) of  $\beta_1$ . The term “best” here is used in the sense of minimum variance.*

*Proof.* Let  $\hat{\beta}_1^*$  be another unbiased linear estimator of the form:

$$\hat{\beta}_1^* = \sum_{i=1}^n c_i Y_i,$$

where the  $c_i$ 's are constants. Since  $\hat{\beta}_1^*$  is unbiased, it should satisfy

$$E(\hat{\beta}_1^*) = \sum_{i=1}^n c_i E(Y_i) = \sum_{i=1}^n c_i (\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i X_i = \beta_1.$$

Hence, we have the following two conditions for the BLUE of  $\beta_1$ :

$$\boxed{\sum_{i=1}^n c_i = 0 \quad \text{and} \quad \sum_{i=1}^n c_i X_i = 1.} \quad (2.2)$$

The variance of  $\hat{\beta}_1^*$  is

$$\text{Var}(\hat{\beta}_1^*) = \sum_{i=1}^n c_i^2 \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^n c_i^2. \quad (2.3)$$

We know  $\text{Var}(\hat{\beta}_1^* - \hat{\beta}_1) \geq 0$ . Let's try to find  $\text{Var}(\hat{\beta}_1^* - \hat{\beta}_1)$ . Then we have

$$\text{Var}(\hat{\beta}_1^* - \hat{\beta}_1) = \text{Var}\left(\sum_{i=1}^n (c_i - k_i) Y_i\right) = \sigma^2 \sum_{i=1}^n (c_i - k_i)^2 = \sigma^2 \sum_{i=1}^n (c_i^2 + k_i^2 - 2c_i k_i).$$

Note that  $\sum_{i=1}^n c_i k_i = \sum_{i=1}^n c_i (X_i - \bar{X})/S_{xx} = (\sum c_i X_i - \sum c_i \bar{X})/S_{xx} = 1/S_{xx}$  and  $\sum_{i=1}^n k_i^2 = \sum_{i=1}^n [(X_i - \bar{X})/S_{xx}]^2 = 1/S_{xx}$ . Hence, we have  $\sum_{i=1}^n c_i k_i = \sum_{i=1}^n k_i^2$ . It is immediate from  $\text{Var}(\hat{\beta}_1^* - \hat{\beta}_1) \geq 0$  that

$$\text{Var}(\hat{\beta}_1^* - \hat{\beta}_1) = \sigma^2 \sum_{i=1}^n (c_i^2 - k_i^2) = \text{Var}(\hat{\beta}_1^*) - \text{Var}(\hat{\beta}_1) \geq 0.$$

Hence,  $\text{Var}(\hat{\beta}_1^*) \geq \text{Var}(\hat{\beta}_1)$ . So  $\hat{\beta}_1$  has minimum variance among all unbiased linear estimators.  $\square$

### 2.1.2 Sampling distribution of $(\hat{\beta}_1 - \beta_1)/\text{SE}(\hat{\beta}_1)$

Since  $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$ , we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1).$$

We need to estimate  $\sigma^2$ , we will use MSE as an estimator of  $\sigma^2$ :

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

Note that

$$\frac{(n-2)\text{MSE}}{\sigma^2} \sim \chi_{n-2}^2 \quad \text{and} \quad \frac{N(0,1)}{\sqrt{\chi_d^2/d}} \sim t(d).$$

Hence, we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \bigg/ \sqrt{\frac{(n-2)\text{MSE}}{\sigma^2}} \bigg/ (n-2) \sim t(n-2),$$

and it follows that

$$\boxed{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{MSE}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \sim t(n-2)}$$

where  $\text{SE}(\hat{\beta}_1) = \sqrt{\text{MSE}/S_{xx}}$ .

### 2.1.3 Confidence interval for $\beta_1$

Since  $(\hat{\beta}_1 - \beta_1)/\text{SE}(\hat{\beta}_1) \sim t(n-2)$ , we have

$$P\left[t\left(\frac{\alpha}{2}; n-2\right) \leq \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \leq t\left(1 - \frac{\alpha}{2}; n-2\right)\right] = 1 - \alpha.$$

Using the symmetry of  $t$  distribution about 0, we have  $t(\frac{\alpha}{2}; n-2) = -t(1 - \frac{\alpha}{2}; n-2)$ .

Hence, the  $1 - \alpha$  confidence limits for  $\beta_1$  are

$$\boxed{\hat{\beta}_1 \pm t\left(1 - \frac{\alpha}{2}; n-2\right) \cdot \text{SE}(\hat{\beta}_1)}.$$

### 2.1.4 Tests for $\beta_1$

- **Two-sided test:**  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$ .

Test statistic:

$$T = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$$

Decision rule (at the significance level  $\alpha$ ):

If  $|T| \leq t(1 - \frac{\alpha}{2}; n - 2)$ , accept  $H_0$ .

If  $|T| > t(1 - \frac{\alpha}{2}; n - 2)$ , reject  $H_0$ .

- **One-sided test:**  $H_0 : \beta_1 \leq 0$  vs.  $H_1 : \beta_1 > 0$ .

Test statistic:

$$T = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$$

Decision rule (at the significance level  $\alpha$ ):

If  $T \leq t(1 - \alpha; n - 2)$ , accept  $H_0$ .

If  $T > t(1 - \alpha; n - 2)$ , reject  $H_0$ .

- **Test when  $H_0 : \beta_1 = \beta_{10}$  or  $H_0 : \beta_1 < \beta_{10}$**

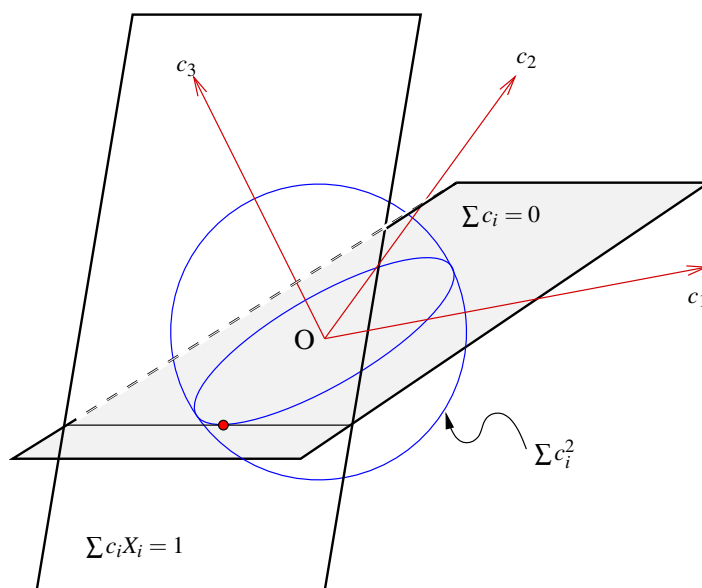
Test statistic:

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{\text{SE}(\hat{\beta}_1)}$$

Decision rule (at the significance level  $\alpha$ ): the same as the above.

### 2.1.5 A geometry of the BLUE of $\beta_1$

The construction of BLUE of  $\beta_1$  can be explained by using the geometric description as in the figure. The goal is to minimize the variance,  $\sigma^2 \sum_{i=1}^n c_i^2$ , in (2.3) together with the conditions,  $\sum_{i=1}^n c_i = 0$  and  $\sum_{i=1}^n c_i X_i = 1$ , in (2.2). This is equivalent to finding the minimization of  $\sum_{i=1}^n c_i^2$  (hyper-sphere in  $\mathbb{R}^n$ ) with the two constraints (two hyper-planes in  $\mathbb{R}^n$ ).

Figure 2.1: Geometry of the BLUE of  $\beta_1$  with  $n = 3$ .

### 2.1.6 BLUE: Optimization with the constraints

We are to minimize  $\sum_{i=1}^n c_i^2$  subject to  $\sum_{i=1}^n c_i = 0$  and  $\sum_{i=1}^n c_i X_i = 1$ . Thus, this minimization can be obtained by using Lagrange multipliers. The auxiliary function with Lagrange multipliers ( $\lambda_1$  and  $\lambda_2$ ) is

$$\Psi = \sum_{i=1}^n c_i^2 + \lambda_1 \sum_{i=1}^n c_i + \lambda_2 \left( \sum_{i=1}^n c_i X_i - 1 \right).$$

It follows from  $\partial\Psi/\partial c_i = 0$  that

$$2c_i + \lambda_1 X_i + \lambda_2 = 0. \quad (2.4)$$

Taking the sum of the above, we have

$$2 \sum_{i=1}^n c_i + \lambda_1 n \bar{X} + n \lambda_2 = 0.$$

It is immediate from the condition  $\sum_{i=1}^n c_i = 0$  that

$$\lambda_1 n \bar{X} + n \lambda_2 = 0. \quad (2.5)$$

Next, multiplying  $c_i$  to (2.4) gives

$$2c_i^2 + \lambda_1 c_i X_i + \lambda_2 c_i = 0.$$

Taking the sum of the above, we have

$$2 \sum_{i=1}^n c_i^2 + \lambda_1 \sum_{i=1}^n c_i X_i + \lambda_2 \sum_{i=1}^n c_i = 0.$$

It is immediate from the two conditions in (2.2) that

$$2 \sum_{i=1}^n c_i^2 + \lambda_1 = 0. \quad (2.6)$$

Solving (2.5) and (2.6) for  $\lambda_1$  and  $\lambda_2$ , we have

$$\lambda_1 = -2 \sum_{i=1}^n c_i^2 \quad \text{and} \quad \lambda_2 = 2 \sum_{i=1}^n c_i^2 \bar{X}.$$

Substituting the above results into (2.4) gives

$$c_i = (X_i - \bar{X}) \sum_{i=1}^n c_i^2. \quad (2.7)$$

Taking the square of (2.7) gives

$$c_i^2 = (X_i - \bar{X})^2 \left( \sum_{i=1}^n c_i^2 \right)^2.$$

After taking the sum of the above, we can solve for  $\sum_{i=1}^n c_i^2$  which results in

$$\sum_{i=1}^n c_i^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{1}{S_{xx}}. \quad (2.8)$$

It is immediate upon substituting (2.8) into (2.7) that

$$c_i = \frac{X_i - \bar{X}}{S_{xx}}.$$

## 2.2 Inferences concerning $\beta_0$

The estimator for  $\beta_0$  is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

The expectation and variance of  $\hat{\beta}_0$  can be shown as follows:

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0 \\ \text{Var}(\hat{\beta}_0) &= \sigma^2 \left[ \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right]. \end{aligned}$$

Let  $k_i = (X_i - \bar{X})/S_{xx}$ . We have studied  $\hat{\beta}_1 = \sum k_i Y_i$  from (2.1). Thus, we have

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \sum_{i=1}^n \frac{1}{n} Y_i - \sum_{i=1}^n k_i Y_i \bar{X} = \sum_{i=1}^n \left( \frac{1}{n} - k_i \bar{X} \right) Y_i. \quad (2.9)$$

Note that  $\epsilon_i$  are independent, so is  $Y_i$ . Hence, we have

$$\begin{aligned} \text{Var}(\hat{\beta}_0) &= \sum_{i=1}^n \left( \frac{1}{n} - k_i \bar{X} \right)^2 \text{Var}(Y_i) \\ &= \sum_{i=1}^n \left( \frac{1}{n^2} - \frac{2}{n} k_i \bar{X} + k_i^2 \bar{X}^2 \right) \sigma^2 \\ &= \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right) \sigma^2. \end{aligned}$$

Hence, an estimator of  $\text{Var}(\hat{\beta}_0)$ , also denoted by  $[\text{SE}(\hat{\beta}_0)]^2$ , is obtained by replacing  $\sigma^2$  by MSE:

$$\widehat{\text{Var}}(\hat{\beta}_0) = [\text{SE}(\hat{\beta}_0)]^2 = \text{MSE} \left[ \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right].$$

Similar to the case of  $\hat{\beta}_1$ , we have the following test statistic:

$$\frac{\hat{\beta}_0 - \beta_0}{\text{SE}(\hat{\beta}_0)} \sim t(n-2).$$

The confidence limits:  $\hat{\beta}_0 \pm t(1 - \frac{\alpha}{2}; n-2) \cdot \text{SE}(\hat{\beta}_0)$ .



## 2.3 Interval estimation of $\mu_{Y_h} = E(Y_h) = \beta_0 + \beta_1 X_h$

Let  $X_h$  denote the level of  $X$  for which we wish to estimate the mean response,  $E(Y_h) = \beta_0 + \beta_1 X_h$ . We want to draw inference about  $E(Y_h)$

Given the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  of the regression parameters in the regression function  $\mu_Y = E(Y) = \beta_0 + \beta_1 X$ , we estimate the regression function as

$$\hat{\mu}_Y = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Thus, when  $X = X_h$ , the estimator of  $\mu_{Y_h} = E(Y_h)$  denoted by  $\hat{Y}_h$  is given by

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h.$$

The estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  all consist of linear combination of  $Y_i$ 's. Hence,  $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$  is also a linear combination of  $Y_i$  and so  $\hat{\beta}_0 + \hat{\beta}_1 X_h$  has a normal distribution. To find the parameters for this normal distribution, we need to find  $E(\hat{Y}_h)$  and  $\text{Var}(\hat{Y}_h)$ . First,  $E(\hat{Y}_h)$  is obtained by

$$E(\hat{Y}_h) = E(\hat{\beta}_0 + \hat{\beta}_1 X_h) = \beta_0 + \beta_1 X_h = \mu_{Y_h}.$$

Next, using  $\hat{\beta}_1 = \sum_{i=1}^n k_i Y_i$  from (2.1) and  $\hat{\beta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - k_i \bar{X}\right) Y_i$  from (2.9), we have

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \sum_{i=1}^n \left\{ \frac{1}{n} + (X_h - \bar{X}) k_i \right\} Y_i.$$

It is immediate that

$$\begin{aligned}
 \text{Var}(\hat{Y}_h) &= \text{Var} \left[ \sum_{i=1}^n \left\{ \frac{1}{n} + (X_h - \bar{X})k_i \right\} Y_i \right] \\
 &= \sum_{i=1}^n \left\{ \frac{1}{n} + (X_h - \bar{X})k_i \right\}^2 \text{Var}(Y_i) \\
 &= \sum_{i=1}^n \left\{ \frac{1}{n^2} + \frac{2}{n}(X_h - \bar{X})k_i + (X_h - \bar{X})^2 k_i^2 \right\} \sigma^2 \\
 &= \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\} \sigma^2,
 \end{aligned}$$

and

$$\widehat{\text{Var}}(\hat{Y}_h) = [\text{SE}(\hat{Y}_h)]^2 = \text{MSE} \left[ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right].$$

Notice that  $\text{SE}(\hat{Y}_h)$  is minimized when  $X_h = \bar{X}$ .

Thus, we have the distribution

$$Z = \frac{\hat{Y}_h - E(\hat{Y}_h)}{\sqrt{\text{Var}(\hat{Y}_h)}} = \frac{\hat{Y}_h - \mu_{Y_h}}{\sqrt{\text{Var}(\hat{Y}_h)}} \sim N(0, 1).$$

Replacing  $\sqrt{\text{Var}(\hat{Y}_h)}$  with  $\sqrt{\widehat{\text{Var}}(\hat{Y}_h)} = \text{SE}(\hat{Y}_h)$ , we have the following test statistic

$$T = \frac{\hat{Y}_h - E(\hat{Y}_h)}{\sqrt{\widehat{\text{Var}}(\hat{Y}_h)}} = \frac{\hat{Y}_h - \mu_{Y_h}}{\text{SE}(\hat{Y}_h)} \sim t(n-2).$$

Converting the above for  $\mu_{Y_h}$ , we have the confidence limits for  $\mu_{Y_h}$ :

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n-2\right) \cdot \text{SE}(\hat{Y}_h).$$

That is,

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n-2\right) \cdot \sqrt{\text{MSE} \cdot \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}}.$$

## 2.4 Prediction Interval of *New* observation, $Y_{h(\text{new})}$

There is often confusion regarding the implication of the word *prediction*. Obviously the statistic  $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$ , the point on the regression line at  $X = X_h$ , serves the dual purpose as the estimate of mean response ( $E(Y_h)$  at  $X = X_h$ ) and the predicted value. However, it is not appropriate for establishing any form of inference on a *future* single observation. Suppose the *mean response* at a fixed  $X = X_h$  is not of interest. Rather, one is interested in some type of bound on a *future* single response observation at  $X = X_h$ .

Consider a *future* single observation at  $X = X_h$ , denoted symbolically by  $Y_{h(\text{new})}$ , independent of  $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$ . We can standardize by considering

$$\text{Var}(Y_{h(\text{new})} - \hat{Y}_h) = \text{Var}(Y_{h(\text{new})}) + \text{Var}(\hat{Y}_h) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}.$$

It immediate from  $E(Y_{h(\text{new})} - \hat{Y}_h) = 0$  that

$$Z = \frac{Y_{h(\text{new})} - \hat{Y}_h - 0}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}}} \sim N(0, 1).$$

Replacing  $\sigma^2$  with MSE, we have

$$T = \frac{Y_{h(\text{new})} - \hat{Y}_h}{\sqrt{\text{MSE}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}}} \sim t(n - 2).$$

Converting the above for  $Y_{h(\text{new})}$ , we have the prediction limits for  $\mu_{Y_h}$ :

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) \cdot \sqrt{\text{MSE} \cdot \left\{ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}}.$$

These limits are the same as

$$\begin{aligned} & \hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) \cdot \sqrt{\text{MSE} + \text{MSE} \cdot \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}} \\ & \hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) \cdot \sqrt{\text{MSE} + \widehat{\text{Var}}(\hat{Y}_h)}. \end{aligned}$$

## 2.5 ANOVA approach to regression analysis

### 2.5.1 Partition of total sum of squares

In any regression, the analyst will observe variation in the response. We want to look at the variation of  $Y_i$  and  $\hat{Y}_i$ . Note that  $\sum Y_i = \sum \hat{Y}_i$ , *i.e.*,  $\bar{Y} = \bar{\hat{Y}}$ .

$$\underbrace{Y_i - \bar{Y}}_{\text{deviation of } Y_i} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{deviation of } \hat{Y}_i} + \underbrace{Y_i - \hat{Y}_i}_{\hat{\epsilon}_i}$$

Then the sum of squares become

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i)^2 \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2 \sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \end{aligned}$$

Considering the last term,

$$\begin{aligned} 2 \sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) &= 2 \sum \hat{Y}_i(Y_i - \hat{Y}_i) - 2 \sum \bar{Y}(Y_i - \hat{Y}_i) \\ &= 2 \sum \hat{Y}_i \hat{\epsilon}_i - 2 \sum \bar{Y} \hat{\epsilon}_i = 0, \end{aligned}$$

We have

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2.$$

We denote this partition as follows:

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

The coefficient of determination ( $R^2$ ) is defined as

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

Table 2.1: ANOVA Table:

Source	df	SS	MS	$F$
Regression	1	SSR	MSR=SSR	$F = \frac{MSR}{MSE}$
Error	$n - 2$	SSE	$MSE = SSE/(n - 2)$	
Total	$n - 1$	SSTO		

This  $R^2$  tells us what proportion of your variation in  $Y$  is explained by the regression on  $X$ . Clearly  $0 \leq R^2 \leq 1$ .

Notice that the df decomposition:  $n - 1 = 1 + (n - 2)$ .

Why SSR has only df=1? Plugging  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$  and  $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$  into  $SSR = \sum(\hat{Y}_i - \bar{Y})^2$ , we have

$$SSR = MSR = \hat{\beta}_1^2 \cdot \sum (X_i - \bar{X})^2 = \hat{\beta}_1^2 \cdot S_{xx}.$$

Here SSR has only one squared quantity ( $\hat{\beta}_1^2$ ).

### 2.5.2 $F$ test

**Fact.** If the errors  $\epsilon_i$  have a iid  $N(0, \sigma^2)$ , then SSE and SSR are independent and have scaled  $\chi^2$  distribution.

Thus, we have

$$1. SSE/\sigma^2 \sim \chi_{n-2}^2.$$

$$2. SSR/\sigma^2 \sim \chi_1^2 \text{ under } H_0 : \beta_1 = 0.$$

If  $\beta_1 \neq 0$ , then  $SSR/\sigma^2$  follows a non-central  $\chi_1^2$  and is usually larger than a  $\chi_1^2$ .

**Fact.** If  $U \sim \chi_a^2$  and  $V \sim \chi_b^2$ , and  $U, V$  are independent. then

$$\frac{U/a}{V/b} \sim F(a, b).$$

Hence, if the errors are  $N(0, \sigma^2)$ , then under  $H_0 : \beta_1 = 0$

$$F^* = \frac{\text{SSR}/\sigma^2}{\text{SSE}/[(n-2)\sigma^2]} = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2)$$

Decision rule:

If  $F^* \leq F(1 - \alpha; 1, n - 2)$ , accept  $H_0 : \beta_1 = 0$ .

If  $F^* > F(1 - \alpha; 1, n - 2)$ , reject  $H_0 : \beta_1 = 0$ .

## 2.6 Case when $X$ is random

In all of the preceding development, we assume throughout that  $X_i$  is a known constant. All the theories developed are based on this assumption. But two other situations occur frequently in practice, however:

1. The variables  $X$  and  $Y$  are both random and are observations from a joint density function.  
 $\Rightarrow$  In this case, generally less interest is associated with prediction than in the case when  $X$  is non-random. Correlation analysis is more suitable
2. The variable  $X$  is a measure with non-ignorable error  
 $\Rightarrow$  Use EIV (Errors-In-Variables) model.