Recall:

In the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, 2, \ldots, n$$
$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \quad and \quad Cov(\varepsilon_i, \varepsilon_j) = 0 \ for \ all \ i \neq j.$$

Then

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad and \qquad Var(Y_i) = \sigma^2.$$

The point estimates of β_0, β_1 are

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = b_0 = \bar{Y} - b_1\bar{X}$$

$$\hat{\beta}_1 = b_1 = \sum_{i=1}^{n} K_i Y_i, \quad K_i = \frac{\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}, \ \hat{\beta}_0 = \sum_{i=1}^{n} L_i Y_i, \quad L_i = \frac{1}{n} - \bar{X} K_i$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}, \ Var(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right],$$

The unbiased estimate of σ^2 is

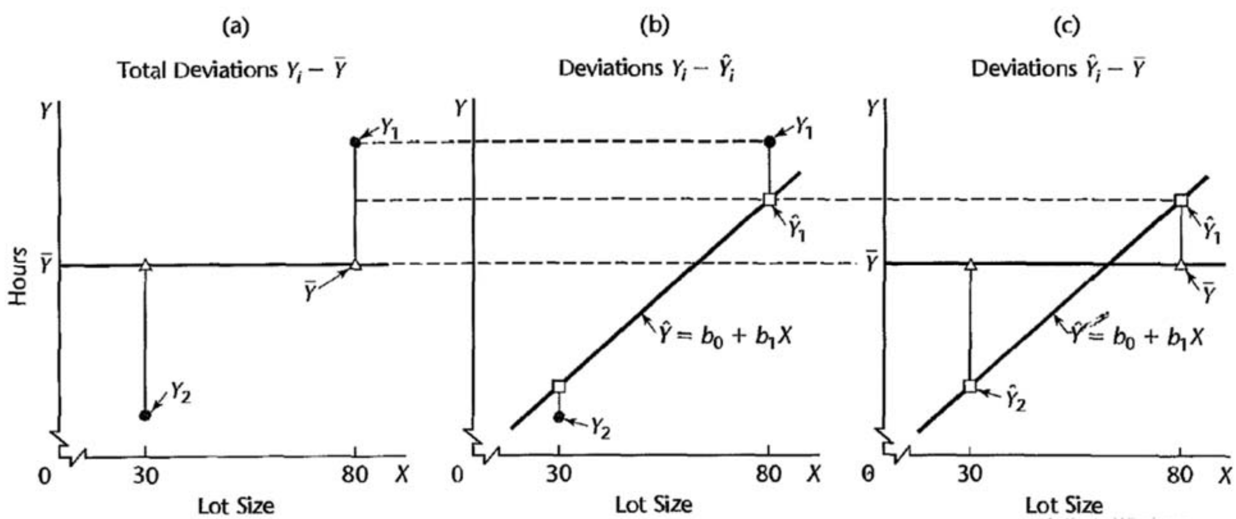
$$s^2 = MSE = \widehat{\sigma^2} = \frac{SSE}{n-2}, \quad SSE = \sum_{i=1}^{n} e_i^2, \quad e_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots$$

Analysis of Variance Approach to Regression Analysis

We now have developed the basic regression model and demonstrated its major uses. At this point, we consider the regression analysis from the perspective of analysis of variance. This new perspective will not enable us to do anything new, but the analysis of variance approach will come into its own when we take up multiple regression models and other types of linear statistical models.

Types of sum of squared Errors

Use the data in Toluca Company example, we show three types of sum of the squared errors as:



1- *SSTO* stands for *total sum of squares*

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

2- SSE* stands for *error sum of squares

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

3- SSR stands for regression sum of squares

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

<u>Lemma:</u>

SSTO=SSR+SSE

Proof.

$$\underbrace{Y_i - \bar{Y}}_{\substack{\text{Total}\\ \text{deviation}}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation}\\ \text{of fitted}\\ \text{regression}\\ \text{value}\\ \text{around mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation}\\ \text{around}\\ \text{fitted}\\ \text{regression}\\ \text{line}}}$$

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= \sum [(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 \\ &= \sum [(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)] \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\end{aligned}$$

But

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) e_i = \sum_{i=1}^{n} \hat{Y}_i e_i - \bar{Y} \sum e_i = 0 - 0 = 0$$

Then

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

$$= SSR + SSE$$

Types of sum of squared Errors

4- *SSTO* stands for *total sum of squares*

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

5- SSE* stands for *error sum of squares

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

6- SSR stands for regression sum of squares

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

Lemma:

$$SSTO=SSR+SSE$$

$$SSR = b_1^2\sum(X_i-\bar{X})^2$$

ANOVA TABLE

Source of Variation	SS	df	MS	
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F_0 = \frac{MSR}{MSE}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

ANOVA tables are widely used; we shall usually utilize the basic type of the tables.

F-test

This test is basically depending on the ANOVA table and it works like t-test (in the simple linear regression model). We summarize the test as given below:

i) Setup the hypotheses

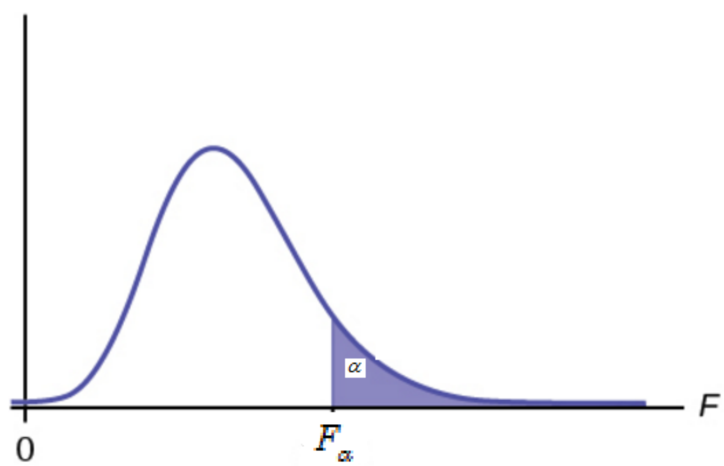
$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

ii) Test statistic under H0

$$F_0 = \frac{MSR}{MSE} \quad \text{(From ANOVA table)}$$

this statistic has F_α distribution with (1,n-2) d.f

iii) Critical region



The shaded area is the rejection Region and the unshaded area is acceptance Region

iv) Decision

When the calculated F_0 belongs to the shaded area, we reject the null hypothesis H0, otherwise Accept H0.

P-value approach

i) Setup the hypotheses

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

ii) Calculate p-value

$$P - value = P(F \geq F_0)$$

Reject H0, otherwise, Accept H0

Remarks:

Since Both of F-test and t-test do the same job for testing

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

So, there is a relation between F_0 *and* T_0 as $(F_0 = T_0^2)$

Example

Consider the Toluca Company example, Use F test (ANOVA) for testing the significance of the linear term in the simple linear regression model.

From the data, we have

$$SSTO=\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = 307203$$

$$SSE=\sum_{i=1}^{n}(\hat{Y}_i-Y_i)^2 =54825$$

$$SSR=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = b_1^2 S_{xx} = SSTO - SSE$$

$$= 307203 - 54825 = 252378$$

Then The ANOVA Table is

Source of Variation	SS	df	MS	F0
Regression	SSR=252378	1	MSR=252378	F_0=105.9
Error	SSR=54825	23	MSE=2384	
SSTO	SSTo=307203	24		

i) Setup the hypotheses

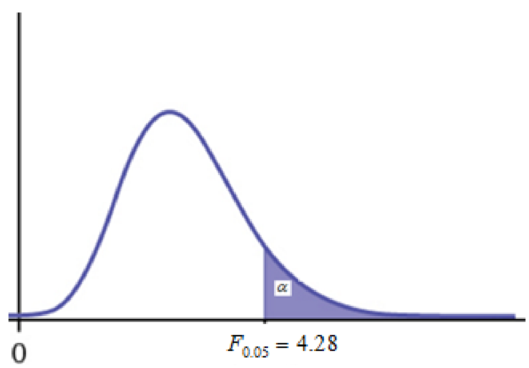
$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

ii) Test statistic under H0

$F_0 = 105.9$ (From ANOVA table)

this statistic has F_α distribution with (1,n-2) d.f

iii) Critical regions



The shaded area is the rejection Region and the unshaded shaded area is acceptance Region

iv) Decision

since the calculated $F_0 = 105.9 > 4.28$ belongs to the shaded area, we reject the null

P-value approach

i) Setup the hypotheses

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

ii) Calculate p-value

$$P - value = P(F \geq F_0) \approx 0.0000 = .0000 < 0.05$$

Then Reject H0

Remarks:

Since Both of F-test and t-test do the same job for testing

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

So, there is a relation between F_0 *and* T_0 as $(105.9 = (10.29)^2)$

```
> anova(model)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 252378  252378  105.88 4.449e-10 ***
Residuals 23  54825    2384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficient of Determination

We saw earlier that *SSTO* measures the variation in the observations Y_i, or the uncertainty in predicting *Y*, when no account of the predictor variable· X is taken. Thus, *SSTO* is a measure of the uncertainty in predicting *Y* when X is not considered. Similarly, *SSE* measures the variation in the Y_i when a regression model utilizing the predictor variable *X* is employed. A natural measure of the effect of X in reducing the variation in *Y*, i.e., in reducing the uncertainty in predicting *Y*, is to express the reduction in variation *(SSTO - SSE = SSR)* as a proportion of the total variation

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

The measure R^2 is called the *coefficient of determination.*

For the Toluca Company example, we obtained $SSTO = 307{,}203$ and $SSR = 252{,}378$. Hence:

$$R^2 = \frac{252{,}378}{307{,}203} = .822$$

Thus, the variation in work hours is reduced by 82.2 percent when lot size is considered.

Coefficient of Correlation

A measure of linear association between Y and X when both Y and X are random is the *coefficient of correlation*. This measure is the signed square root of R^2:

$$r = \pm\sqrt{R^2}$$

A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative. Thus, the range of r is: $-1 \le r \le 1$.

Example

For the Toluca Company example, we obtained $R^2 = .822$. Treating X as a random variable, the correlation coefficient here is:

$$r = +\sqrt{.822} = .907$$

The plus sign is affixed since b_1 is positive.

Remark

The Correlation Coefficient between two variables Y_1 *and* Y_2 is given by

$$r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\left[\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2\right]^{1/2}}$$

```
> summary(model)$r.squared
[1] 0.8215335
```

cor(x,y)

[1] 0.9063848

tt=summary(model)$r.squared

> sqrt(tt)

[1] 0.9063848