STAC67H: Regression Analysis

Fall, 2014

Instructor: Dr. Jabed Tomal

Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, ON
Canada

September 22, 2014

1. We have a total of n observations $Y_1, Y_2, \cdots, Y_n$ in a sample. The deviation of the ith observation from its mean is

$$Y_i - \bar{Y}.$$

2. The measure of total variation, denoted by SST, is the sum of the squared deviations:

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2.$$

1. When we utilize the predictor variable X, the deviation of Y_i from its predicted value is defined as

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

2. The sum of the squared deviations is called *error sum of squares* or *SSE*:

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

1. We know that
$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i.$$
2. The sum of the squares of the fitted values $\hat{Y}_i$ from its mean is defined as:
$$\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2.$$
and is called *regression sum of squares*, or *SSR* in short.

1. The total sum of squares in Y's can be partitioned as

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

2. In short, we write

$$SST = SSR + SSE.$$

1. In short, the partition of sum of squres

$$SST = SSR + SSE.$$

2. The corresponding partition of degrees of freedom is following

$$(n-1) = 1 + (n-2).$$

1. The *regression mean square*, MSR in short

$$MSR = \frac{SSR}{1} = SSR.$$

2. The *error mean square*, MSE in short

$$MSE = \frac{SSE}{n-2}.$$

1. In Assignment # 1, you will prove that

$$E(MSE) = \sigma^2.$$

2. Furthermore, it is easy to show that

$$SSR = b_1^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

3. We can prove that

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Simple Linear Regression Model:
Analysis of Variance (ANOVA) Table

Source of Variation	***SS***	***df***	***MS***	$\boldsymbol{E\{MS\}}$
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum \left(X_i - \bar{X}\right)^2$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	SST	$n-1$		

Definition - F-distribution: Let U_1 and U_2 are two independent chi-squared random variables with $n_1 \geq 1$ and $n_2 \geq 1$ degrees of freedoms, respectively. Then the ratio

$$F = \frac{U_1/n_1}{U_2/n_2}$$

follows F-distribution with numerator degrees of freedom n_1 and denominator degrees of freedom n_2.

F-distribution

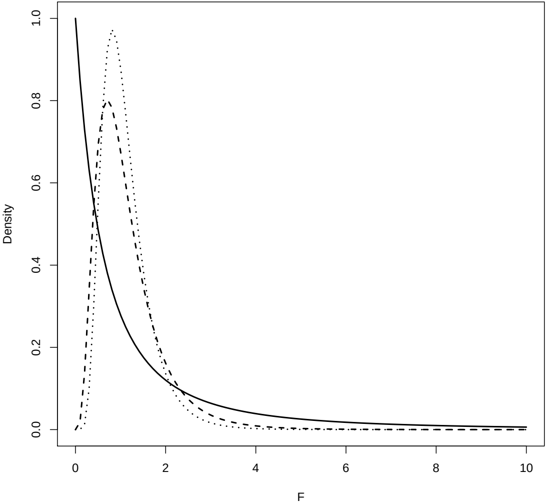

Figure: Density functions for $F(2,3)$ (solid line), $F(10,15)$ (dashed line), and $F(20,20)$ (dotted line) distributions.

Statement: If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and SST is decomposed into k sum of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if:

$$\sum_{r=1}^{k} df_r = n - 1.$$

Simple Linear Regression Model:
F-Test

- We want to test the null hypothesis

$$H_0 : \beta_1 = 0,$$

- against the alternative

$$H_A : \beta_1 \neq 0.$$

Simple Linear Regression Model:
F-Test

- Under the null hypothesis $\beta_1 = 0$ all the observations Y_i have the same mean
$$\mu = \beta_0,$$
- and variance
$$\sigma^2.$$
- Using *Cochran's Theorem*, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables with $n - 2$ and 1 degrees of freedoms, respectively.

- Hence, under the null hypothesis $\beta_1 = 0$

  $$F^* = \frac{MSR}{MSE}$$

  follows F distribution with numerator degrees of freedom 1 and denominator degrees of freedom $n - 2$.

- Decision: Reject the null hypothesis at α level of significance (probability of type I error) if

  $$F^* > F(1 - \alpha; 1, n - 2),$$

  where $F(1 - \alpha; 1, n - 2)$ is the $(1 - \alpha)$th percentile of the F distribution with 1 and $n - 2$ numerator and denominator degrees of freedoms, respectively.

- In a simple linear regression model, the F and t tests to test the hypotheses $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$ are equivalent.
- **Exercise:** Link the *Cochran's Theorem* to derive the t test for the hypotheses $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$.

- The *coefficient of determination* is computed as following

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

- R^2 ranges from

$$0 \leq R^2 \leq 1,$$

while the larger values are preferred.

- R^2 shows the amount of variability in the *response variable* that is explained by the fitted linear regression model.

- A *correlation coefficient* shows the strength of linear association between two random variables X and Y

$$-1 \leq r \leq 1.$$

- In a simple linear regression model the coefficient of determination R^2 can be obtained from the coefficient of correlation r

$$R^2 = \{r\}^2 .$$

- and *vice versa*

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

where the sign of r depends on the sign of b_1.

Simple Linear Regression Model:

Prediction Interval for $Y_{h(new)}$

- Consider you are given a new X, denoted by X_h, using which you want to estimate Y_h.
- The estimate is
$$\hat{Y}_h = b_0 + b_1 X_h.$$
- The prediction error is
$$Y_h - \hat{Y}_h,$$
here Y_h and $\hat{Y}_h$ are independent as the latter used only $(Y_1, X_1), (Y_2, X_2), \cdots, (Y_n, X_n)$ – NOT Y_h.
- The variance of the prediction error is
$$Var(Y_h - \hat{Y}_h) = Var(Y_h) + Var(\hat{Y}_h).$$

- **Exercise:** show that

$$Var(\hat{Y}_h) = \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right] \sigma^2.$$

- Hence, the variance of the prediction error is

$$Var(Y_h - \hat{Y}_h) = \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right] \sigma^2.$$

- The estimated variance of the prediction error is

$$s^2(Y_h - \hat{Y}_h) = \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right] MSE.$$

- The following statistic
$$\frac{Y_h - \hat{Y}_h}{s(Y_h - \hat{Y}_h)}$$
follows a t distribution with $n - 2$ degrees of freedom.

- Hence, the $100(1 - \alpha)\%$ confidence interval of Y_h is
$$\hat{Y}_h \pm t(1 - \alpha/2; n - 2) \times s(Y_h - \hat{Y}_h).$$

Airfreight breakage problem:

Source of Variation	***SS***	***df***	***MS***	***F***
Regression	160	1	160	72.72
Error	17.6	8	2.2	
Total	177.6	9		

- Consider you want to test the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0.$$

- The calculated F statistic is

$$F = 72.73$$

which follows F distribution with 1 and 8 degrees of freedoms.

- At 5% level of significance, the tabulated value of F is

$$F(0.95, 1, 8) = 5.32,$$

  which is smaller than 72.73.
- We reject the null hypothesis $H_0 : \beta_1 = 0$ at 0.05 level of significance.

Simple Linear Regression Model:
Coefficient of Determination R^2

- The coefficient of determination is

$$R^2 = \frac{SSR}{SST} = 0.9009009.$$

- Almost 90.1% variability in the response variable is explained by this fitted linear regression model.

Simple Linear Regression Model:

Coefficient of Correlation r

- The coefficient of correlation is

$$r = \sqrt{0.9009009} = +0.949158.$$

- If we assume that the two variables, the number of transfers from one aircraft to another and the number of ampules broken, are both random variables, then the strength of linear association between the two variables is 0.95.