To Documents

The F-test for Linear Regression

Definitions for Regression with Intercept

- n is the number of observations, p is the number of regression parameters.
- **Corrected Sum of Squares for Model:** SSM = $\sum_{i=1}^{n} (y_i\hat{} - \bar{y})^2$,
  also called sum of squares for regression.
- **Sum of Squares for Error:** SSE = $\sum_{i=1}^{n} (y_i - y_i\hat{})^2$,
  also called sum of squares for residuals.
- **Corrected Sum of Squares Total:** SST = $\sum_{i=1}^{n} (y_i - \bar{y})^2$
  This is the sample variance of the y-variable multiplied by n - 1.
- For multiple regression models, SSM + SSE = SST.
- **Corrected Degrees of Freedom for Model:** DFM = p - 1
- **Degrees of Freedom for Error:** DFE = n - p
- **Corrected Degrees of Freedom Total:** DFT = n - 1
  Subtract 1 from n for the corrected degrees of freedom.
  Horizontal line regression is the null hypothesis model.
- For multiple regression models with intercept, DFM + DFE = DFT.
- **Mean of Squares for Model:** MSM = SSM / DFM
- **Mean of Squares for Error:** MSE = SSE / DFE
  The sample variance of the residuals.
- In a manner analogous to Property 10 of Properties of Random Variables, which states that s^2 is unbiased for σ^2, it can be shown that MSE is unbiased for σ^2 for multiple regression models.
- **Mean of Squares Total:** MST = SST / DFT
  The sample variance of the y-variable.
- In general, a researcher wants the variation due to the model (MSM) to be large with respect to the variation due to the residuals (MSE).
- **Note:** the definitions in this section are not valid for regression through the origin models. They require the use of uncorrected sums of squares.

The F-test

- For a multiple regression model with intercept, we want to test the following null hypothesis and alternative hypothesis:

  H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

  H_1: $\beta_j \neq 0$, for at least one value of j

  This test is known as the overall **F-test for regression**.

- Here are the five steps of the **overall F-test for regression**

  1. State the null and alternative hypotheses:

     H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

     H_1: $\beta_j \neq 0$, for at least one value of j

  2. Compute the test statistic assuming that the null hypothesis is true:

     F = MSM / MSE = (explained variance) / (unexplained variance)

  3. Find a $(1 - \alpha)100\%$ confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.

  4. Accept the null hypothesis if $F \in I$; reject it if $F \notin I$.

  5. Use statistical software to determine the p-value.

- **Practice Problem:** For a multiple regression model with 35 observations and 9 independent variables (10 parameters), SSE = 134 and SSM = 289, test the null hypothesis that all of the regression parameters are zero at the 0.05 level.

  Solution: DFE = n - p = 35 - 10 = 25 and DFM = p - 1 = 10 - 1 = 9. Here are the five steps of the test of hypothesis:

  1. State the null and alternative hypothesis:

     H_0: $\beta_1 = \beta_2 = , \ldots , = \beta_{p-1} = 0$

     H_1: $\beta_j \neq 0$ for some j

  2. Compute the test statistic:

     F = MSM/MSE = (SSM/DFM) / (SSE/DFE) = (289/9) / (134/25) = 32.111 / 5.360 = 5.991

3. Find a $(1 - 0.05)\times 100\%$ confidence interval for the test statistic. Look in the F-table at the 0.05 entry for 9 df in the numerator and 25 df in the denominator. This entry is 2.28, so the 95% confidence interval is [0, 2.34]. This confidence interval can also be found using the R function call qf(0.95, 9, 25).

4. Decide whether to accept or reject the null hypothesis: $5.991 \notin [0, 2.28]$, so reject H_0.

5. Determine the p-value. To obtain the exact p-value, use statistical software. However, we can find a rough approximation to the p-value by examining the other entries in the F-table for (9, 25) degrees of freedom:

Level	**Confidence Interval**	**F-value**
0.100	[0, 0.900]	1.89
0.050	[0, 0.950]	2.28
0.025	[0, 0.975]	2.68
0.010	[0, 0.990]	2.22
0.001	[0, 0.999]	4.71

The F-value is 5.991, so the p-value must be less than 0.005.

- Verify the value of the F-statistic for the Hamster Example.

Technical Details for the Overall F-Test

- If $t_1, t_2, \ldots, t_m$, are independent, $N(0, \sigma^2)$ random variables, then $\boldsymbol{\Sigma}_{i=1}^{m} t_i^2$ is a χ^2 (chi-squared) random variable with m degrees of freedom.

- It can be shown that if H_0 is true and the residuals are unbiased, homoscedastic, independent, and normal:

  1. SSE / σ^2 has a χ^2 distribution with DFE degrees of freedom.

  2. SSM / σ^2 has a χ^2 distribution with DFM degrees of freedom.

  3. SSE and SSM are independent random variables.

- If u is a χ^2 random variable with n degrees of freedom, v is a χ^2 random variable with m degrees of freedom, and u and v are independent, then if $F = (u/n)/(v/m)$ has an **F distribution with (n,m) degrees of freedom**. See the F-tables in the Statistical Tables.

- By the previous information, if H_0 is true, F = [(SSM/σ)/DFM]/[(SSE/σ)/DFE] has an F distribution with (DFM, DFE) degrees of freedom.
- But F = [(SSM/σ)/DFM]/[(SSE/σ)/DFE] = (SSM/DFM)/(SSE/DFE) = MSM/MSE, so F is independent of σ.

The R^2 and Adjusted R^2 Values

- For simple linear regression, R^2 is the square of the sample correlation r_{xy}.
- For multiple linear regression with intercept (which includes simple linear regression), it is defined as r^2 = SSM / SST.
- In either case, R^2 indicates the proportion of variation in the y-variable that is due to variation in the x-variables.
- Many researchers prefer the **adjusted R^2 value** = $\overline{\mathbf{R}}^2$ instead, which is penalized for having a large number of parameters in the model:

  $$\overline{R}^2 = 1 - (1 - R^2)(n - 1) / (n - p)$$

- Here derivation of $\overline{R}^2$: R^2 is defined as 1 - SSE/SST or 1 - R^2 = SSE/SST. To take into account the number of regression parameters p, define the adjusted R-squared value as

  $$1 - \overline{R}^2 = MSE/MST,$$

  where MSE = SSE/DFE = SSE/(n-p) and MST = SST/DFT = SST/(n-1). Thus,

  $$\begin{aligned} 1 - \overline{R}^2 &= [SSE/(n - p)] / [SST/(n - 1)] \\ &= (SSE/SST)(n - 1) / (n - p) \end{aligned}$$

  so

  $$\begin{aligned} \overline{R}^2 &= 1 - (SSE/SST)(n - 1) / (n - p) \\ &= 1 - (1 - R^2)(n - 1) / (n - p) \end{aligned}$$