15.1. The Overall F-Test

Start with the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,$$

with full rank design matrix (rank($\mathbf{X}$) $= p$). Note that we are assuming the model contains an intercept. Suppose we want to test whether the overall model is significant, i.e.,

$$H : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0.$$

This can be written as

$$H : \mathbf{A}\boldsymbol{\beta} = (\mathbf{0}, \mathbf{I}_{(p-1)\times(p-1)})\boldsymbol{\beta} = \mathbf{0},$$

i.e., all X variables in the model except the intercept can be deleted. The F test for H is

$$\begin{aligned} F &= \frac{(RSS_H - RSS)/(p-1)}{RSS/(n-p)} \\ &\sim F_{p-1,n-p}, \quad \text{if } H \text{ is true} \end{aligned}$$

This is called the *overall F-test statistic* for the linear model. It is sometimes used as a preliminary test of the significance of the model prior to performing model selection to determine *which* variables in the model are important.

15.2. Sample Multiple Correlation Coefficient

The *sample multiple correlation coefficient* is defined as the correlation between the observations Y_i and the fitted values $\hat{Y}_i$ from the regression model:

$$R \equiv \operatorname{corr}(Y_i, \hat{Y}_i) = \frac{\sum_i (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\left[\sum_i (Y_i - \bar{Y})^2 \sum_i (\hat{Y}_i - \bar{\hat{Y}})^2\right]^{1/2}}.$$

For a MVN vector $(X_1, \ldots, X_{p-1}, Y)$, we define

$$\rho_{Y:X_1,\ldots,X_{p-1}} = \operatorname{corr}(Y, \hat{Y})$$

where $\hat{Y}$ in this context means the conditional expectation of Y given $X_1, \ldots, X_{p-1}$. R is a sample estimate of $\rho_{Y:X_1,\ldots,X_{p-1}}$.

15.3. The ANOVA Decomposition for a Linear Model

Theorem 15.3.1:

(i) ANOVA decomposition

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2$$

i.e., Total-SS = RSS + REGRESSION-SS

Proof:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

Show the cross-product term is 0:

$$\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_i (Y_i - \hat{Y}_i)\hat{Y}_i - \bar{Y} \sum_i (Y_i - \hat{Y}_i)$$

The first term is $\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{Y}} = 0$ because the vectors are orthogonal and the second term is $\sum_i \hat{\epsilon}_i = 0$ (midterm).

(ii) R^2

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{\text{REG–SS}}{\text{TOTAL–SS}}.$$

or, equivalently (using (i)),

$$1 - R^2 = \frac{\sum_i (Y_i - \hat{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{RSS}{\text{TOTAL–SS}},$$

Interpretation: R^2 is the proportion of variance in the Y_i explained by the regression model.

15.4. Uses of R^2

Pearson correlation r measures how well two-dimensional data are described by a line with non-zero slope. R^2 is a generalization of r^2 for higher-dimensional data. It indicates how closely the linear model fits the data. If $R^2 = 1$ (the maximum value) then $Y_i = \hat{Y}_i$ and the model is a perfect fit.

The F-test of a hypothesis of the form $H : (\mathbf{0}, \mathbf{A}_1)\boldsymbol{\beta} = \mathbf{0}$ (does not involve the intercept β_0) can also be formulated as a test for a significant reduction in R^2:

$$F = \frac{(R^2 - R_H^2)}{(1 - R^2)} \frac{(n-p)}{q}$$

where R^2 and R_H^2 are the sample multiple correlation coefficients for the full model and the reduced model, respectively.

Note: This shows that R^2 cannot increase when deleting a variable in the model (other than the intercept).

Note: Just as judging the "largeness" of correlation is problematic, so is judging the "largeness" of R^2.

15.5. Goodness of Fit

How can we assess if a linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is appropriate? Do the predictors and the linear model adequately describe the mean of $\mathbf{Y}$? We want something stronger than the overall F test, which tests if the predictors are related to the response.

We can test model adequacy if there are *replicates*, i.e., independent observations with the same values of the predictors (and so the same mean).

Suppose, for $i = 1, \ldots, n$, we have replicates $Y_{i1}, \ldots, Y_{iR_i}$ corresponding to the values $x_{i1}, \ldots, x_{i,p-1}$ of the predictors. The full model is

$$Y_{ir} = \mu_i + \varepsilon_{ir}$$

where the μ_i are any constants. We wish to test whether they have the form

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1},$$

Write $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. We want to test the hypothesis

$$H : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

15.6. The F Test for Goodness of Fit

We now apply the general F test to H. The RSS under the full model is

$$RSS = \sum_{i=1}^{n} \sum_{r=1}^{R_i} (Y_{ir} - \bar{Y}_i)^2$$

and for the reduced model

$$RSS_H = \sum_{i=1}^{n} \sum_{r=1}^{R_i} (Y_{ir} - \hat{\beta}_{0H} - \hat{\beta}_{1H} x_{i1} - \ldots - \hat{\beta}_{p-1,H} x_{i,p-1})^2.$$

It can be shown that in the case of equal replications $(R_i = R)$ the estimates under the reduced model are

$$\hat{\boldsymbol{\beta}}_H = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z},$$

where $Z_i = \bar{Y}_i = \sum_{r=1}^{R} Y_{ir}/R$ (Seber & Lee, p. 116).

The F statistic is

$$F = \frac{(RSS_H - RSS)/(n-p)}{RSS/(N-n)} \sim F_{n-p,N-n},$$

where $N = \sum_{i=1}^{n} R_i$.

This test is sometimes called the *goodness-of-fit test* and sometimes called the *lack-of-fit test.*