Model Fitting: Inference

Given several predictors and a response, we need to figure out whether all are needed.

Consider a large model, Ω, and a smaller model, □□, which consist of a subset of predictors in Ω.

- If there is not much difference in fit, we perfer the smaller model.
- If the larger model has a much better fit, we perfer the larger model

Suppose we have a response Y and a vector p^{l} regressors $X^{l} = (X_1, X_2)$ that we partition into two parts so that:

- X_2 has q regressors
- X_1 has the remaining p^{l} - q

The general hypothesis test we consider is:

$$H_0: E(Y|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = 0) = \boldsymbol{x}_1'\boldsymbol{\beta}_1$$
$$H_1: E(Y|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2$$

The null hypothesis is obtained by setting beta_2 = 0

The reasoning is that if $RSS_{□□}$ - RSS_{Ω} is small, the fit of the smaller model is almost as good as the larger model. On the other hand, if the difference is large the superior fit of the larger model would be perferred.

This suggests $(RSS_{□□} - RSS_{\Omega})/RSS_{\Omega}$ would be a potentially good test statistic where the denominator is used for scaling purpose

F Tests

Suppose the dimensions of Ω is p and that of □□ is q. The general formula for the test is:

$$F = \frac{(RSS_{\omega} - RSS_{\Omega})/(df_{\omega} - df_{\Omega})}{RSS_{\Omega}/df_{\Omega}} = \frac{SSReg/df_{Reg}}{\widehat{\sigma^2}}$$

where df_{Ω} = n - p, and $df_{□□}$ = n - q

Thus, we would reject the null hypothesis if $F > F^{\alpha}_{p - q, n - p}$

Simple Regression

$$H_0: E(Y|X = x) = \beta_0 \text{ v.s. } H_1: E(Y|X = x) = \beta_0 + \beta_1 x$$

Recall the ANOVA table for a simple regression:

Source	df	SS	MS	F	p-value
Regression	1	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n-2$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

Under the null hypothesis:

$$RSS_{NH} = \sum(y_i - \widehat{\beta_0})^2 = \sum(y_i - \bar{y}\ )^2 = SYY$$

The df is the number of n observations minus the number of estimated parameters: (n-1)

Under the alternative hypothesis:

$$F = \frac{(SYY - RSS)/((n-1)-(n-2))}{RSS/(n-2)}$$

Test of All the Predictors

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$RSS_{AH} = \sum(y - X\hat{\beta})'(y - X\hat{\beta}) = e'e$$

$$RSS_{NH} = \sum(y - \bar{y})'(y - \bar{y}) = SYY$$

$$F = \frac{(SYY-RSS)/((n-1)-(n-p-1))}{RSS/(n-p-1)} = \frac{(SYY-RSS)/p}{RSS/(n-p-1)}$$

$$F \sim F_{p,n-p-1}$$

Where p is the number of regressors and n is the sample size.

Source of Variation	Sum of Square	DF	Mean of Square
Regression (model)	SSreg	p	MSReg= SSreg/p
Error	RSS	n - (p+1)	MSE= RSS/(n-p-1)
Total	SYY	n-1	

One Predictor

Can a particular predictor be dropped from the model?

$$H_0: \beta_i = 0$$

A t test can be used with (n - p - 1) degrees of freedom

$$t_i = \hat{\beta}_i / \text{se}(\hat{\beta}_i)$$

The F test may be used as introduced earlier with a df of 1, n-p-1. t_i^2 here is exactly the F-statistic.

Revision #4
Created 22 September 2022 21:58:46 by Elkip
Updated 29 September 2022 14:28:58 by Elkip