STAT2008/STAT6038

Hypothesis tests on partial regression coefficients

Multiple regression – partitioning of variability

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

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

Recall: General Form of ANOVA Table in the Multiple Linear Regression Model

Source	**d.f.**	**Sum of Squares**	**Mean Squares**	**F Statistic**
Regression	K = (p-1)	SSR	MSR=SSR/k =SSR/(p-1)	F= MSR/MSE
Error	n-k-1 = n-p	SSE	MSE= SSE/(n-k-1)= SSE/(n-p)	
Total	**n-1**	**Variation in Y SST**		

Sequential Sums of squares

$SSR = SSR(\beta_1, \beta_2, \ldots, \beta_k|\beta_0) = SSR(\beta_1|\beta_0) + SSR(\beta_2|\beta_1, \beta_0) + SSR(\beta_3|\beta_2, \beta_1, \beta_0)$ $+ \ldots + SSR(\beta_k|\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-2}, \beta_{k-1})$.

$SSR(\beta_2|\beta_1, \beta_0)$ is the amount of the unexplained variability from a simple linear regression on x_1 which is subsequently explained by x_2.

i.e. it is the increase in the sum of squares obtained by adding the predictor x_2 to the model.

Sequential Sums of Squares

For example, if we have a model with 4 predictors, and we wish to investigate whether the last two are adding anything to the explanation of the response, we note:

$$SSR(\beta_1, \beta_2, \beta_3, \beta_4 | \beta_0) = SSR(\beta_1, \beta_2 | \beta_0) + SSR(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2).$$

If the second term on the right-hand side of this equality is only a small proportion of the overall total, then we might wish to conclude that the last two predictors are not adding much!

Tests on subsets of parameters

partition both the parameter vector, β, and the design matrix, X, as:

$$X = [X_{(1)} \ X_{(2)}]; \qquad \beta = \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix},$$

where $X_{(1)}$ is an $n \times p_1$ matrix containing columns for each of the predictors associated with the parameters in the sub-vector $\beta_{(1)}$, and thus $X_{(2)}$ and $\beta_{(2)}$ contain the remaining columns of predictors and parameters, respectively.

$$Y = X\beta + \epsilon = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \epsilon.$$

Hypothesis Tests

$$Y = X\beta + \epsilon = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \epsilon.$$

Let $\beta_{(1)}$ consists of the parameters for the predictors under investigation.

$$H_0 : \beta_{(1)} = 0 \qquad \text{versus} \qquad H_A : \beta_{(1)} \neq 0.$$

i.e. how much of the overall regression sum of squares the predictors associated with $\beta_{(1)}$ are contributing

Significance Testing

Can test two different things

- Significance of the overall regression

- Significance of specific partial regression coefficients.

Hypothesis Tests

If the errors are assumed to be normal, then under H_0, the F-statistic

$$F = \frac{SSR(\beta_{(1)}|\beta_{(2)})/p_1}{s^2}$$

has an F-distribution with p_1 numerator and $n - p$ denominator degrees of freedom.

Significance of the regression

if we want to test the overall significance of the regression; that is, test the null hypothesis that the response variable is not related to ANY of the predictors in the model, then we are in the case where $\beta_{(1)} = (\beta_1, \beta_2, \ldots, \beta_k)$, $\beta_{(2)} = \beta_0$, $p_1 = k = p - 1$, and the test statistic becomes the familiar:

$$F = \frac{SSR(\beta_1, \beta_2, \ldots, \beta_k | \beta_0)/(p-1)}{s^2} = \frac{SSR/(p-1)}{MSE} = \frac{MSR}{MSE},$$

where $MSR = SSR/(p-1)$ is the overall regression sum of squares divided by its appropriate degrees of freedom, also called the mean square for regression.

Significance of the overall regression

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
H_A: not all slopes = 0

Test Statistic: Found in "ANOVA" table

Decision Rule: Compared to an F-distribution with k, (n-k-1 = n-p) degrees of freedom.

If H_0 is rejected, one or more slopes are not zero. Additional tests are needed to determine which slopes are significant.

Significance of specific partial regression coefficients using T

H_0: β_i=0

H_A: $\beta_i \neq 0$

Test Statistic:

$$T = \frac{\hat{\beta} - \beta_i}{SE(\hat{\beta}_i)}$$

Decision Rule: Compared to a t-distribution with (n-k-1) degrees of freedom (i.e. residual d.f. from ANOVA table) [k is the number of predictors being fitted.]

If H_0 is rejected, the slope of the i^{th} variable is significantly different from zero. That is, *once the other variables are considered*, the i^{th} predictor has a significant linear relationship with the response.

Significance of specific partial regression coefficients using F

If we want to test the significance of a single β_j, then we have $\beta_{(1)} = \beta_j$, $p_1 = 1$, and the test statistic becomes

$$F = \frac{SSR(\beta_j | \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{k-1}, \beta_k)}{s_\epsilon^2}.$$

As before we have :

$$F = T^2$$

In general..

$$
\begin{aligned}
& SSR(\beta_{(1)}|\beta_{(2)}) \\
&= SSR(\beta) - SSR(\beta_{(2)}) \\
&= SSR(\text{full model}) - SSR(\text{reduced model}) \\
&= SSE_{reduced} - SSE_{full}.
\end{aligned}
$$

$H_0 : \beta_{(1)} = 0$ is based on the drop in the residual sum of squares from the model without the predictors associated with the parameters being tested to the full model with all the predictors.

If this drop is ”large”, then we will reject the null hypothesis.

Otherwise, we will decide that the predictors associated with these parameters are not significantly adding to the explanatory power