Statistical inference for multiple regression

The model for multiple regression

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

makes the following assumptions:

i) Linearity: $E(\varepsilon_i) = 0,$ ii) Constant variance: $V(\varepsilon_i) = \sigma^2_\varepsilon,$ iii) Normality: $\varepsilon_i \tilde{} N(0, \sigma^2_\varepsilon),$ iv) Independence: $\varepsilon_i, \varepsilon_j$ are independent for $i \neq j,$ v) Fixed X's or X's measured without error and independent of ε.

Under these assumptions the least-squares estimators of $\alpha, \beta_1, \beta_2, ..., \beta_k$ are:

i) linear functions of the data, ii) unbiased, iii) maximally efficient among unbiased estimators, iv) maximum-likelihood estimators, and v) normally distributed.

Coefficient standard errors and confidence Intervals

The standard error of a slope estimate B_j is:

$$SE(B_j) = \frac{1}{\sqrt{1 - R^2_j}} \frac{S_E}{\sqrt{\sum (x_{ij} - \overline{x}_j)^2}},$$

where R^2_j is from the regression to predict X_j from all of the other X_i's. These standard errors can be used to define confidence intervals for the parameters β_j as:

$$B_j \pm t_{\alpha/2, n-k-1} SE(B_j)$$

Hypothesis Tests

General tests of interest for the multiple regression model can be formulated as setting some subset of the β_j coefficients to zero:

$$H_0 : \beta_1 = \beta_2 = \ ... = \beta_q = 0 \text{ versus } H_a : \text{not all } \beta_1, \beta_2, \ ..., \beta_q \text{ terms are } \ = 0.$$

Then we can use a subscript of '1' for sums of squares from the complete model above and '0' as a subscript for sums of squares for the reduced model when H_0 is true to give the F statistic for testing H_0 :

$$F_0 = \frac{(\text{RegSS}_1 - \text{RegSS}_0)/q}{\text{RSS}_1/(n - k - 1)},$$

which follows an F distribution with q and $n-k-1$ numerator and denominator degrees of freedom, respectively.

Bias due to underfitting with one or two covariates

The text makes some calculations to examine the bias in the slope estimator of a single variable (X_1) when there is another variable (X_2) that also explains variation in Y. They show that the expected value of the slope estimator is:

$$\frac{\sigma_{1Y}}{\sigma_1^2} - \beta_2 \frac{\sigma_{12}}{\sigma_1^2},$$

from which they conclude that not only must the additional variable X_2 truly explain variation in Y (so that $\beta_2 \neq 0$), but also X_1 must covary with X_2 in order for a bias to occur.

Measurement error in the explanatory variables

The text also makes some calculations to investigate the effect of having one of the X variables in a two-covariate model measured with error. The results indicate that the coefficient of the variable measured with error is downwardly biased (otherwise called attenuated), and the coefficient of the other variable, if the measurement variation is large enough, converges to its' regression coefficient value in a simple linear regression model (thus masking the effect of the covariate measured with error on the coefficient of the other covariate).