Multiple Regression

Goals

- To review/introduce the calculation and interpretation of the least-squares regression coefficients in multiple regression.
- To review/introduce the calculation and interpretation of the regression standard error and the multiple correlation coefficients.
- To introduce standardized regression coefficients.
- To introduce the standard statistical inference and assumptions for multiple linear regression.
- To describe properties of the least-squares coefficients as estimators of the parameters of the regression model.

Two Explanatory Variables

The linear multiple-regression equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

for two explanatory variables, X_1 and X_2, describes a plane in the three-dimensional X_1, X_2, Y space.

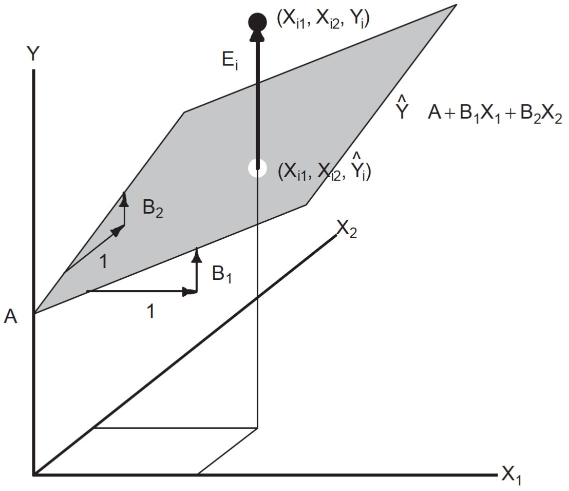

Figure 1: The multiple regression plane.

The residual is the signed vertical distance from the point to the plane:

$$\epsilon_i = Y_i - \widehat{Y}_i = Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$$

To make the plane come as close as possible to the points in the aggregate, we want the values of β_0, β_1, and β_2 that minimize the sum of squared residuals:

$$\sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$$

Differentiating the sum-of-squares function with respect to the regression coefficients, setting the partial derivatives to zero, and rearranging terms produces the normal equations.

This is a system of three linear equations in three unknowns, so it usually provides a unique solution for the least-squares regression coefficients β_0, β_1, and β_2.

The least-squares coefficients are uniquely defined unless X_1 and X_2 are perfectly correlated or unless one of the explanatory variables is invariant.

If X_1 and X_2 are perfectly correlated, then they are said to be collinear.

Interpretation for multiple regression

The slope coefficients for the explanatory variables in the multiple regression are partial coefficients, whereas the slope coefficient in simple regression gives the marginal relationship between the response variable and a single explanatory variable.

That is, each slope in multiple regression represents the 'effect' on the response variable of a one-unit increment in the corresponding explanatory variable holding constant the value of the other explanatory variable.

The simple regression slope effectively ignores the other explanatory variable.

This interpretation of the multiple regression slope is apparent in the figure showing the multiple regression plane. Because the regression plane is flat, its slope in the direction of X_1, holding X_2 constant, does not depend upon the specific value at which X_2 is fixed.

Algebraically, fix X_2 to the specific value x_2 and see how $\widehat{Y}$ changes as X_1 is increased by 1, from some specific value x_1 to $x_1 + 1$:

$$[\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2] = \beta_1$$

A similar result holds for X_2.

More than Two Explanatory Variables

For the general case of k explanatory variables, the multiple-regression equation is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

It is not possible to visualize the point cloud of the data directly when $k > 2$, but it is simple to find the values of the β's that minimize

$$\sum [Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik})]^2$$

Minimization of the sum-of-squares function produces the normal equations for general multiple regression.

Because the normal equations are linear, and because there are as many equations as unknown regression coefficients $(k + 1)$, there is usually a unique solution for the coefficients, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$.

Only when one explanatory variable is a perfect linear function of others, or when one or more explanatory variables are invariant, will the normal equations not have a unique solution.

The least-squares surface passes through the point of means $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k, \overline{Y})$.

Standard Error

As in simple regression, the standard error in multiple regression measures the 'average' size of the residuals.

As before, we divide by degrees of freedom, here $n - (k + 1) = n - k - 1$ to calculate the variance of the residuals; thus, the standard error is

$$S_\epsilon = \sqrt{\frac{\sum \epsilon_i^2}{n - k - 1}}$$

Heuristically, we 'lose' $k+1$ degrees of freedom by calculating the $k+1$ regression coefficients, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$.

Sums of Squares

The sums of squares in multiple regression are defined as in simple regression:

$$TSS = \sum (Y_i - \overline{Y})^2$$

$$RegSS = \sum (\widehat{Y}_i - \overline{Y})^2$$

$$RSS = \sum (Y_i - \widehat{Y}_i)^2$$

The fitted values $\widehat{Y}_i$ and residuals ϵ_i now come from the multiple regression equation.

We have a similar decomposition of variance: $TSS = RegSS + RSS$

The least-squares residuals are uncorrelated with the fitted values and with each of the X's.

The squared multiple correlation R^2 represents the proportion of variation in the response variable captured by the regression:

$$R^2 \equiv \frac{RegSS}{TSS}$$

The multiple correlation coefficient is the positive square root of R^2, and is interpretable as the simple correlation between the fitted and observed Y values.

Standardized Regression Coefficients

Researchers often wish to compare the coefficients of different explanatory variables in a regression analysis. When the explanatory variables are commensurable, comparison is straightforward. Standardized regression coefficients permit a limited assessment of the relative effects of incommensurable explanatory variables.

Imagine that the annual dollar income of wage workers is regressed on their years of education, years of labor force experience, producing the fitted regression equation

$$\widehat{Income} = \beta_0 + \beta_1 \times Education + \beta_2 \times Experience$$

If education and experience are measured in years, the coefficients β_1 and β_2 are both expressed in dollars/year, and can be directly compared.

More commonly, explanatory variables are measured in different units. Standardized regression coefficients rescale the β's using the standard deviations of the explanatory variables.

The usual practice standardizes the response variable as well, but this is not essential.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \epsilon_i$$

$$\overline{Y} = \beta_0 + \beta_1 \overline{X}_1 + \ldots + \beta_k \overline{X}_k$$

$$Y_i - \overline{Y} = \beta_1 (X_{i1} - \overline{X}_1) + \ldots + \beta_k (X_{ik} - \overline{X}_k) + \epsilon_i$$

$$\frac{(Y_i - \overline{Y})}{S_Y} = \left(\beta_1 \frac{S_1}{S_Y}\right) \frac{X_{i1} - \overline{X}_1}{S_1} + \ldots + \left(\beta_k \frac{S_k}{S_Y}\right) \frac{X_{ik} - \overline{X}_k}{S_k} + \frac{\epsilon_i}{S_Y}$$

$$Z_{iY} = \beta_1^* Z_{i1} + \ldots + \beta_k^* Z_{ik} + \epsilon_i^*$$

$Z_Y \equiv (Y - \overline{Y})/S_Y$ is the standardized response variable, linearly transformed to a mean of zero and a standard deviation of one.

$Z_1, \ldots, Z_k$ are the explanatory variables, similarly standardized.

$\epsilon^* \equiv \epsilon/S_Y$ is the transformed residual which, note, does not have a standard deviation of one.

$\beta_j^* \equiv \beta_j(S_j/S_Y)$ is the standardized partial regression coefficient for the jth explanatory variable.

The standardized coefficient is interpretable as the average change in Y, in standard deviation units, for a one standard deviation increase in X_j, holding constant the other explanatory variables.

A common misuse of standardized coefficients is to employ them to make comparisons of the effects of the same explanatory variable in two or more samples drawn from populations with different spreads.

Statistical Inference for Multiple Linear Regression

Assumptions

The assumptions underlying the model concern the errors, ϵ_i, and are identical to the assumptions in simple regression:

- Linearity: $E(\epsilon_i) = 0$
- Constant Variance: $V(\epsilon_i) = \sigma_\epsilon^2$
- Normality: $\epsilon_i \sim N(0, \sigma_\epsilon^2)$
- Independence: ϵ_i, ϵ_j are independent for $i \neq j$.
- Fixed X's or X's independent of ϵ

Under these assumptions (or particular subsets of them), the least-squares estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots \widehat{\beta}_k$ of $\beta_0, \beta_1, \ldots \beta_k$ are

- linear functions of the data, and hence relatively simple;
- unbiased;
- maximally efficient among unbiased estimators;
- maximum-likelihood estimators;
- normally distributed.

The slope coefficient $\widehat{\beta}_j$ in multiple regression has sampling variance

$$V(\widehat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (X_{ij} - \overline{X}_j)^2}$$

where R_j^2 is the squared multiple correlation from the regression of X_j on all of the other X's.

The second factor is essentially the sampling variance of the slope in simple regression, although the error variance σ_ϵ^2 will usually be smaller than before.

The first factor — called the *variance-inflation factor* — is large when the explanatory variable X_j is strongly correlated with other explanatory variables (the problem of collinearity).

Thus, the overall impact of additional predictors on $V(\widehat{\beta}_j)$ depends on the correlation of X_j with the additional predictors and how much additional variability in the outcome that those predictors explain.

Confidence Intervals and Hypothesis Tests

Individual Slope Coefficients

Confidence intervals and hypothesis tests for individual coefficients closely follow the pattern of simple regression analysis:

- The variance of the residuals provides an unbiased estimator of σ_ϵ^2:

$$\widehat{\sigma}_\epsilon^2 = \frac{\sum \widehat{\epsilon}_i^2}{(n - k - 1)}$$

- Using $\widehat{\sigma}_\epsilon^2$, we can calculate the standard error of $\widehat{\beta}_j$:

$$\widehat{SE(\beta_j)} = \frac{1}{\sqrt{1 - R_j^2}} \times \frac{\widehat{\sigma}_\epsilon}{\sqrt{\sum_{i=1}^n (X_{ij} - \overline{X}_j)^2}}$$

Confidence intervals and tests, based on the t-distribution with $n - k - 1$ degrees of freedom, follow straightforwardly.

With only two explanatory variables, $R_1^2 = R_2^2 = r_{12}^2$.

All Slopes

We can also test the global or 'omnibus' null hypothesis that all of the regression slopes are zero:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

which is not quite the same as testing the separate hypotheses

$$H_0^{(1)} : \beta_1 = 0; H_0^{(2)} : \beta_2 = 0; \ldots ; H_0^{(k)} : \beta_k = 0$$

An F-test for the omnibus null hypothesis is given by

$$F_0 = \frac{RegSS/k}{RSS/(n - k - 1)} = \frac{(n - k - 1)}{k} \times \frac{R^2}{1 - R^2}$$

Under the null hypothesis, this test statistic follows an F-distribution with k and $n - k - 1$ degrees of freedom.

The calculation of the test statistic can be organized in an analysis of variance table.

When the null hypothesis is true, regression mean squares, $RegMS$, and the residual mean squares, RMS, provide independent estimates of the error variance, so the ratio of the two mean squares should be close to one.

When the null hypothesis is false, $RegMS$ estimates the error variance plus a positive quantity that depends upon the β's:

$$E(F_0) \approx \frac{E(RegMS)}{E(RMS)} = \frac{\sigma_\epsilon^2 + \text{positive quantity}}{\sigma_\epsilon^2}$$

We consequently reject the omnibus null hypothesis for values of F_0 that are sufficiently larger than 1.

In the special case of $k = 1$, $t^2 = F$.

A Subset of Slopes

Consider the hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$ where $1 \leq q \leq k$.

The 'full' regression model, including all of the explanatory variables, may be written:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_q X_{iq} + \beta_{q+1} X_{i,q+1} + \ldots + \beta_k X_{ik} + \epsilon_i$$

If the null hypothesis is correct, then the first q of the β's are zero, yielding the 'null' model

$$Y_i = \beta_0 + \beta_{q+1} X_{i,q+1} + \ldots + \beta_k X_{ik} + \epsilon_i$$

The null model omits the first q explanatory variables, regressing Y on the remaining $k - q$ explanatory variables.

An F-test of the null hypothesis is based upon a comparison of these two models:

RSS_1 and $RegSS_1$ are the residual and regression sums of squares for the full model.

RSS_0 and $RegSS_0$ are the residual and regression sums of squares for the null model.

Because the null model is a special case of the full model, $RSS_0 \geq RSS_1$. Equivalently, $RegSS_0 \leq RegSS_1$.

If the null hypothesis is wrong and (some of) $\beta_1, \ldots, \beta_q$ are nonzero, then the incremental (or 'extra') sum of squares due to fitting the additional explanatory variables

$$RSS_0 - RSS_1 = RegSS_1 - RegSS_0$$

should be large.

The F-statistic for testing the null hypothesis is

$$F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-k-1)} = \frac{(n-k-1)}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2}$$

Under the null hypothesis, this test statistic has an F-distribution with q and $n - k - 1$ degrees of freedom.