

Multiple Regression Analysis: Estimation

Suppose that what we are interested in the relationship between hourly wage and education, but we believe that wage is, in fact, determined not solely by education level, but also by experience. That is, the true model of hourly wage is given by:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \epsilon$$

Now, what we are interested in is still the coefficient, β_1 , holding fixed ALL OTHER FACTORS affecting wage. (That is, we want to know the effect of education on wage, ceteris parabus.)

We can write this model in a more general form where we have k independent variables as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

where β_0 is the intercept, β_1 is the parameter associated with x_1 and so forth. Note that since there are k independent variables and an intercept, this equations contains k+1 (unknown) population parameters.

NOTE: in multiple regression models, a model is considered “linear” if it is LINEAR IN THE PARAMETERS (those are the β 's). That means we can have the independent variables enter into the model however we would like – squared, cubed, log form, etc.

Key Assumption: A key assumption for multiple regression analysis is the relationship between the error term, ϵ , and the independent variables, $x_1 \dots x_k$. We will assume the following relationship, which can be described as a conditional expectation:

$$E(\epsilon | x_1, \dots, x_k) = 0$$

That is, GIVEN THE x's, the expected value of the error term is zero. The unobserved error term is uncorrelated with any of the independent variables.

OLS and Multiple Regression Estimation:

The estimated OLS equation is given by:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

where b_0 is the OLS estimate of β_0 , and so forth. The method of ordinary least squares is exactly the same as for the bivariate model. That is, the estimates are found by MINIMIZING the sum of squared errors:

$$\sum_{i=1}^N (y_i - b_0 - b_1 x_{i1} - \dots - b_k x_{ik})^2$$

Notation: note that there are TWO subscripts on the independent variables (x's). The first subscript refers to the observation number while the second refers to the independent variable.

Interpretation of the OLS Regression Equation:

Let's start with the model with two independent variables:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

The intercept, b_0 is the predicted value of y when x_1 and x_2 are both equal to zero.

The estimates b_1 and b_2 have partial effects, or ceteris paribus interpretations. To see this, consider how the estimated value of y changes when x_1 and x_2 both change:

$$\Delta \hat{y} = b_1 \Delta x_1 + b_2 \Delta x_2$$

If x_2 is held fixed, though, the change in the predicted value of y is then just:

$$\Delta \hat{y} = b_1 \Delta x_1$$

so, b_1 can be thought of as the effect that x_1 has on y when x_2 is held constant.

OLS Fitted Values and Residuals:

After we obtain the OLS estimators, we can obtain the fitted or predicted value for each observation (\hat{y}_i). The fitted or predicted value for observation i is given as:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}$$

Normally, \hat{y}_i will not equal y_i because OLS only minimizes the average squared prediction error. The RESIDUAL for observation i is defined as:

$$e_i = y_i - \hat{y}_i$$

The residuals and the OLS fitted values (the \hat{y}_i s) have some important properties:

1. The sample average of the residuals is zero.
2. The sample covariance between each independent variable and the residuals is zero. (So, the sample covariance between the residuals and the fitted values is also zero.)

3. The OLS fitted regression line ALWAYS passes through the mean values of the dependent and independent variables.

Statistical Properties of the OLS Estimators:

Assumptions:

1. The model in the population can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

where $\beta_0, \beta_1, \dots, \beta_k$ are the unknown fixed parameters of interest and ϵ is an unobservable random error term.

(Note that the model is LINEAR in the parameters.)

2. We have a random sample of N observations ($i = 1, \dots, N$) from the population model as described in (1), above.
3. $E(\epsilon | x_1, \dots, x_k) = 0$
4. No perfect collinearity between the X s. That is, there is NO exact linear relationship between the independent variables. If there is perfect collinearity between any of the x 's, the parameters CANNOT be estimated using OLS. (Eg. You can have income and income squared on the RHS but you can't have income and income/2 on the RHS. There would be no way to distinguish between those two variables when we try to minimize the sum of squared errors... This is a pretty obvious example, but things can get quite complicated when we introduce dummy variables.)

With assumptions 1-4, alone, we can establish that OLS leads to UNBIASED ESTIMATORS.

ASIDE: two further issues

- A. Inclusion of irrelevant variables in the regression model.

What happens if you include a variables on the RHS (an independent variable) that does NOT affect y ? That is, suppose you have a model with explanatory variables x_1, x_2 and x_3 but x_3 has no effect on y once x_1 and x_2 are taken into account. This would mean that the coefficient on x_3 (β_3) should be ZERO. In terms of conditional expectations, this would imply that:

$$E(y | x_1, x_2, x_3) = E(y | x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

If we didn't know that x_3 shouldn't be in there and INCLUDED it in our regression analysis, we would have included an irrelevant variable on the RHS of our regression. What is the effect of this?

NOTHING as the coefficient estimates on β_0 , β_1 , and β_2 would still be unbiased!

B. Omitted variables.

What happens if you OMIT a variable from the model that should be there? Now you get into trouble. Our OLS estimators will now be BIASED. The direction of the bias (if you over-estimate or under-estimate the parameter value) will depend on the relationship between the included variables and the OMITTED variable. In the SIMPLEST of cases (you estimated a bivariate model when you really have TWO RHS variables) you can summarize the bias in the following way:

Summary of Bias in b_1 when x_2 is OMITTED in the estimating equation.

	Corr (x_1, x_2) > 0	Corr (x_1, x_2) < 0
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

5. $\text{Var}(\epsilon | x_1, \dots, x_k) = \sigma^2$ (Homoskedasticity assumption)

Here, we assume that the variance of the error term, ϵ , conditional on the explanatory variables, is the SAME.

Assumptions 1-5 make up the Gauss-Markov assumptions and under these assumptions, OLS is BLUE (the best, linear, unbiased, estimator).

Estimating σ^2 : Standard Errors of the OLS Estimators

Recall that $\sigma^2 = E(\epsilon^2)$ since the $E(\epsilon | x) = 0$. So, an unbiased “estimator” for σ^2 could be:

$$n^{-1} \sum_i^N \epsilon_i^2$$

but, since we don’t observe any of the ϵ ’s we can’t do this. What we can do, however, is use the RESIDUALS, e (remember that $e = y - \hat{y}$). If we do this, we can get an unbiased estimator for σ^2 which is:

$$\hat{\sigma}^2 = \frac{\left(\sum_i^N e_i^2 \right)}{(N - k - 1)} = \text{SSR}/(N-k-1)$$

Note that $N-k-1$ is the DEGREES OF FREEDOM for the general OLS problem with N observations, K explanatory variables and an intercept term.

Multiple Regression Analysis: Inference

Now we will add a 6th assumption to the above 5:

6. The population error, ϵ is independent of the explanatory variables, x_1, \dots, x_k , and is normally distributed with a mean of zero and variance, σ^2 : $\epsilon \sim N(0, \sigma^2)$

What does this imply?

This means that $y|x \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \sigma^2)$. That is, conditional on the x 's, y follows a normal distribution.

Also, this means that:

$$b_j \sim N(\beta_j, \text{Var}(b_j))$$

Therefore,

$$\frac{b_j - \beta_j}{SD(b_j)} \sim N(0, 1)$$

Testing Hypotheses About a Single Population Parameter:

To conduct any hypothesis tests, we have to use our estimator for σ^2 , and then, we must make use of the fact that:

$$\frac{b_j - \beta_j}{SE(b_j)} \sim t_{N-k-1}$$

where $N - k - 1$ is the appropriate degrees of freedom and SE is the standard error of our OLS estimator.

Two sided test:

$$H_0: \beta_j = \beta \text{ (just some number)}$$

$$H_1: \beta_j \neq \beta$$

$$t\text{-statistic} = \frac{b_j - \beta}{SE(b_j)} \sim t_{N-k-1} \text{ under the null hypothesis.}$$

You REJECT the null hypothesis in this case if $|t\text{-statistic}| > t$ critical value.

One sided test:

$H_0: \beta_j < \beta$ (just some number)

$H_1: \beta_j \geq \beta$

$$t\text{-statistic} = \frac{b_j - \beta}{SE(b_j)} \sim t_{N-k-1} \text{ under the null hypothesis.}$$

You REJECT the null hypothesis in this case if $t\text{-statistic} < -t$ -critical value

NOTE: you should be able to construct confidence intervals around the population parameters.

Testing Hypotheses About a Single Linear Combination of Parameters:

Suppose what you wanted to do was something like the following:

$H_0: \beta_1 = \beta_3$ or $\beta_1 - \beta_3 = 0$

$H_1: \beta_1 \neq \beta_3$ or $\beta_1 - \beta_3 \neq 0$

Now, you are interested in knowing whether two parameters are the same or not. How would you construct your test now?

Your t-statistic would now be:

$$\frac{b_1 - b_3}{SE[b_1 - b_3]} \sim t_{N-k-1}$$

BUT remember that your denominator is now a bit more complicated because:

$\text{Var}(b_1 - b_3) = \text{var}(b_1) + \text{var}(b_3) - 2\text{cov}(b_1, b_3)$ and we don't necessarily KNOW the covariance between (b_1, b_3) . We can calculate this value (it's a little bit of a mess), or you can hope that your computer package will calculate this value for you.

Testing MULTIPLE Linear Restrictions: The F-test

Suppose that you wanted to have multiple hypotheses that were tested SIMULTANEOUSLY. For example:

$$H_0: \beta_1 = 0 \text{ AND } \beta_3 = 0$$

H_1 : H_0 is not true.

This is known as a multiple or joint hypothesis. Note that the alternative hypothesis can be that they both aren't equal to zero or that only one doesn't equal zero. How do you conduct this type of test?

FIRST: you cannot look at individual hypothesis tests and come to any conclusion about the joint hypothesis test.

You must construct a test statistic which, under the null hypothesis, will follow an F-distribution. The F-statistic is given by:

$$F = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{N - k - 1}} \sim F_{q, N-k-1} > 0$$

Where SSR = sum of squared residuals, r = restricted, ur = unrestricted, q = number of restrictions. The restricted SSR is the sum of squared residuals from the regression where the restrictions in the null hypothesis, H_0 , are imposed. The unrestricted SSR is the sum of squared residuals from the regression where the restrictions are NOT imposed. The F-statistic has q, and N-k-1 degrees of freedom. You REJECT the null hypothesis if the $F > F$ -critical value. (That means you fall in the tail of the distribution and under the null hypothesis, it would be highly unlikely that you'd observe the F-statistic that you get.)

SPECIAL CASE: Suppose you are testing the over-all significance of the regression. This is done by testing whether ALL of the partial regression coefficients are simultaneously equal to zero:

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

H_1 : H_0 is not true

Then, the F-test becomes:

$$\frac{\frac{ESS}{k}}{\frac{RSS}{N-k-1}} \sim F_{k, N-k-1}$$