

The Gauss Markov Theorem

The usual statement is 'OLS is the best linear unbiased estimator' or 'OLS is BLUE'. Following Greene pp. 46-48, let $b_0 = Cy$ be another linear estimator; linear means 'linear in y '. Since this new estimator must be unbiased:

$$E(Cy|X) = E(CX\beta + C\varepsilon|X) = \beta,$$

from which we deduce that $CX = I$. The variance of b_0 is $V(C\varepsilon\varepsilon'C')$, i.e. $V(b_0|X) = V(C\varepsilon\varepsilon'C'|X) = \sigma^2CC'$. We need to show that this exceeds $\sigma^2(X'X)^{-1}$, the $V(b_{ols}|X)$, by a positive semidefinite matrix. (Recall a positive semidefinite matrix has nonnegative quadratic forms, i.e. $x'Qx \geq 0$ for all x).

The Gauss Markov Theorem (2)

We have $b_0 = Cy$, $CX = I$, $V(b_0|X) = \sigma^2 CC'$.

Define $D = C - (X'X)^{-1}X'$ or $C = D + (X'X)^{-1}X'$ so that $Dy = b_0 - b$.

Since

$$I = CX = (D + (X'X)^{-1}X')X = DX + (X'X)^{-1}X'X,$$

we have $DX = 0$. Calculate

$$\begin{aligned} V(b_0|X) &= \sigma^2 CC' = \sigma^2 [D + (X'X)^{-1}X'] [D' + X(X'X)^{-1}] \\ &= \sigma^2 [DD' + (X'X)^{-1}] \end{aligned}$$

So $V(b_0|X) > V(b|X)$ unless $D = 0$, in which case b_0 and b coincide.

Projections: The Matrices M and P

Following Greene pp. 24–26. The *residuals* e are defined as

$$e = y - Xb$$

$$e = y - X(X'X)^{-1}X'y$$

$$e = (I - X(X'X)^{-1}X')y$$

$$e = My$$

So M is called (well, at least by Greene) the residual ‘maker’. It is: (1) symmetric, $M' = M$; (2) idempotent, $M^2 = M$; (3) characterized by $MX = 0$. These properties can all be confirmed by direct calculation.

A related matrix is P :

$$\hat{y} = (y - e) = (I - M)y = Py, \quad \text{so}$$

$$P = X(X'X)^{-1}X'$$

M and P depend solely on the design X .

Partitioned/Partial Regression

We are following Greene pp. 26–30. Recall the normal equations

$$(X'X)b = X'y$$

You can partition X into $X_1 : X_2$ with $k_1 + k_2 = k$ columns and write:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

From these you can solve directly for

$$b_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2b_2 = (X_1'X_1)^{-1}X_1'(y - X_2b_2)$$

Interpret this as: b_1 can be obtained by regressing y on X_1 only and then making an adjustment (of size $-(X_1'X_1)^{-1}X_1'X_2b_2$). The adjustment is of size zero if: $X_1'X_2 = 0$ or $b_2 = 0$. The first case is that X_1 and X_2 are 'orthogonal'; the second is that the OLS coefficients of the X_2 variables were all zero anyway. The moral of this is that leaving out regressors changes the results, unless one of these 'unlikely' things is true. (Similar material/results.)

Improperly Included/Excluded Variables

(With reference to Greene pp. 148–151.) If we consider the model:

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

omitting variables that should be included corresponds to imposing, incorrectly, $\beta_2 = 0$. As the previous page shows, doing this changes the estimated coefficient on X_1 by $(X_1'X_1)^{-1}X_1'X_2b_2$, where b_2 is the coefficient that would be estimated in the full model (i.e. the one that includes both X_1 and X_2 and does not impose $\beta_2 = 0$.) Since it is the full model that is correct, b_2 is an unbiased estimate of β_2 and the estimate of β_1 in the incorrectly specified model is biased.

However, including X_2 incorrectly does not impose an incorrect restriction on the coefficients; the true coefficient of the incorrectly included variable X_2 is zero, which will be 'correctly' estimated (provided X_2 is exogenous.)

Multicollinearity

'Multicollinearity' refers to the situation in which there are columns of X that are nearly linearly dependent, which occurs e.g. when variables are highly correlated. A consequence of this (or rather, another way of saying the same thing) is that X approaches a condition in which it no longer has full rank and $(X'X)$ is no longer invertible (roughly, the inverse begins getting bigger.) There are various ways of formalizing this, by ratios of eigenvalues that define 'condition numbers'.

Multicollinearity weighs against the disposition to be sure to include every possible relevant variable. As we add variables, if they are successively more dependent, as is often the case, then $(X'X)^{-1}$ grows and the precision of our estimates suffers; the precise way in which this happens depends on the dependence structure of the sequence of variables.

Symptoms of Multicollinearity

Following the discussion of Greene pp. 56-59.

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high.
- Coefficients may have the 'wrong' sign or implausible magnitudes.

Speaking of R^2 : Goodness of fit and model assessment

It is natural that having estimated a model, we would like to know ‘how good an explanation does it give?’, and that we would want to be able to compare alternative models using a single standard. In elementary regression analysis, the simplest single measure is R^2 , which gets its name because under the standard conditions with an estimated intercept, it is the square of the correlation coefficient between $\hat{y} = Xb$ and y . It is also called the coefficient of determination.

Since residuals are orthogonal to X (by construction of the normal equations) the residuals are also orthogonal to $Xb = \hat{y}$. Consequently the total variation in y can be broken down into two separate sources: the variation of \hat{y} around its mean (which happens to also be the mean of y , see below) and the variation of the residuals around their mean (which is zero, also see below).

That is

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n u_i^2\end{aligned}$$

(This decomposition can also be confirmed by direct calculation, when there is an intercept in the model.) The usual interpretation is to call the first term, $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, the explained variation and the second term, rather more obviously, the unexplained variation. And R^2 is the ratio of the explained to the total variation, or equivalently, 1- the unexplained variation/total variation.

Is R^2 a good standard for comparing models? Its main advantage is intuitive; the models under comparison must have the same dependent variable for it to be meaningful. It coincides with, or is related to, some other measures with greater generality.

Properties of residuals, etc. (when there is an intercept).

When there is an intercept, the first of the normal equations $X'u = 0$ simply says $\sum_{i=1}^n 1 * u_i = 0$; and if this is so, $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$. It also follows that the 'fitted regression line passes through the point of means', i.e. that $\bar{y} = \bar{X}b$.

Cookbook F Tests

Another way of asking about the adequacy is to test the hypothesis that all the coefficients, except the intercept, are zero. We interpret acceptance of this hypothesis as saying that the regression in question does no better than the simple sample mean in predicting y .

This test is, in a way, a special case of the following strategy. To test a hypothesis, estimate the model *without* imposing the coefficient restrictions implied by the hypothesis, and compute the sum of squared residuals. Then estimate with the restriction imposed; this must increase the sum of squared residuals. Then compute

$$F = \frac{(SSR_r - SSR_u)/(\text{change in dimensionality of model})}{SSR_u/(\text{df of unrestricted model})}$$

and refer the test statistic to $F(df \text{ numerator}, df \text{ denominator})$.

Cookbook F Tests (2)

$$F = \frac{(SSR_r - SSR_u)/(\text{change in dimensionality of model})}{SSR_u/(\text{df of unrestricted model})}$$

The df of the unrestricted model is typically $n - k$, where k is the number of coefficients being estimated. The df for the numerator is the change of the dimensionality of the model; often this is just the change in the number of coefficients under estimate.

The intuition for the test is that if imposing the hypothesis does not cause the sum of squares to increase 'too much', (so that the explanatory value of the model is not decreased by too much), then the hypothesis is 'consistent with the data.'

When we get to maximum likelihood, we encounter the same idea in the likelihood ratio test (the F test here is a finite sample version of this.)

Examples of Cookbook F tests

One problem is that it is not always apparent how to write the restricted model so that it is easy to estimate. But it is surprising how much you can do.

Consider the Cobb-Douglas production function of the lecture 1:

$$\log(Y_i) = \log(\gamma) + \beta_K \log(K_i) + \beta_L \log(L_i) + \varepsilon_i$$

The hypothesis of constant returns to scale is the coefficient restriction $\beta_K + \beta_L = 1$. Thus

$$\log(Y_i) = \log(\gamma) + \beta_K \log(K_i) + (1 - \beta_K) \log(L_i) + \varepsilon_i,$$

$$\log(Y_i) = \log(\gamma) + \beta_K [\log(K_i) - \log(L_i)] + \log(L_i) + \varepsilon_i,$$

$$\log(Y_i) - \log(L_i) = \log(\gamma) + \beta_K [\log(K_i) - \log(L_i)] + \varepsilon_i,$$

$$\log(Y_i/L_i) = \log(\gamma) + \beta_K [\log(K_i/L_i)] + \varepsilon_i,$$

Examples of Cookbook F tests (2)

Let's use this to test whether the unrestricted model of $\log(\text{wage})$ for the NLS80 data that we had last time is 'significant.'

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.228176	0.138237	45.054	< 2e-16	***
ed10TRUE	-0.050020	0.141203	-0.354	0.723241	
ed11TRUE	0.115968	0.138660	0.836	0.403177	
ed12TRUE	0.178594	0.126922	1.407	0.159728	
ed13TRUE	0.310547	0.133813	2.321	0.020517	*
ed14TRUE	0.389017	0.135026	2.881	0.004055	**
ed15TRUE	0.478936	0.140341	3.413	0.000671	***
ed16TRUE	0.515503	0.131464	3.921	9.46e-05	***
ed17TRUE	0.600010	0.143027	4.195	2.99e-05	***
ed18TRUE	0.575781	0.138209	4.166	3.39e-05	***
exper	0.021231	0.003376	6.290	4.90e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3932 on 924 degrees of freedom

Multiple R-Squared: 0.1378, Adjusted R-squared: 0.1284

F-statistic: 14.76 on 10 and 924 DF, p-value: < 2.2e-16

Calculating: 165.6563 is SSR of log(wage) around its mean; 142.8355 is the SSR of the model (with all dummies). Thus

$$F = \frac{(165.6563 - 142.8355)/10}{142.8355/(935 - 11)} = 14.76 \text{ df}=(10,924)$$

What about restricting the marginal value of all years of education to be the same?

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.502710	0.112037	49.115	< 2e-16 ***
educ	0.077782	0.006577	11.827	< 2e-16 ***
exper	0.019777	0.003303	5.988	3.02e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.393 on 932 degrees of freedom

Multiple R-Squared: 0.1309, Adjusted R-squared: 0.129

F-statistic: 70.16 on 2 and 932 DF, p-value: < 2.2e-16

$$F = \frac{(143.9786 - 142.8355)/8}{142.8355/(935 - 11)} = .9243 \text{ df}=(8,924),$$

so we accept the restriction to a single coefficient.

By the way, $t^2(df) = F(1, df)$.

Some examples of coefficient restrictions (though addressed in a different framework for the F test solution) Greene p. 96.

Heteroscedasticity and Autocorrelation

(A few words on what these are. The failure of these assumptions affects the calculation of standard errors, but not unbiasedness of OLS.)

The same basic strategy underlies addressing both of these. First, use some method to estimate the 'irregularity', then use the estimate to transform the problem back to standard conditions.

The outline of a solution thus requires proving that the estimate of the irregularity in the first step is 'good enough.' The proof techniques for this, and a more general characterization of the remedies, is available in ML and QML estimation, so we postpone a full discussion until that time.

The Barest Outline of Instrumental Variables

Suppose one of the X 's is correlated with ε . This is one of the most difficult problems in econometrics. In the context of the wage equation, education is often thought to be correlated with (unobservable) 'ability', where ability is one of the most important constituents of the $\log(\text{wage})$ equation's disturbance. (Brief discussion.)

Suppose we have a variable that is (1) uncorrelated with the relevant disturbance and (2) correlated with the X in question. (So, stretching credulity a bit, this might be mother's education.) Let Z be the X matrix with the questionable variable's column replaced with the fitted value of the regression of the endogenous variable (education) on all the exogenous variables, including our 'instrument'.

Estimate b_{IV} by $(Z'X)^{-1}Z'y$. Then $E(b_{IV}) = (Z'X)^{-1}Z'(X\beta + \varepsilon) = (Z'X)^{-1}Z'X\beta + (Z'X)^{-1}Z'\varepsilon = \beta + E((Z'X)^{-1}Z'\varepsilon)$. What is $V(\beta_{IV})$?