Additional Notes on Estimators

OLS Estimators

Remember that a regression coefficient based on sample data is an estimate of a true regression parameter for the population the sample is drawn from. There are several desirable properties that we might want our coefficients to have.

An estimator is **unbiased** if its expected value (mean) is the same as the true parameter value in the population. In other words, an unbiased estimator

An estimator of a parameter is **consistent** if the estimate converges to the true value of the parameter as the sample size increases. I.e. its accuracy tends to improve as the sample size grows larger.

Efficiency refers to the accuracy of the estimates produced by the estimator.

An estimator may be referred to as **efficient** if it is the most accurate (i.e. its variance is the smallest) of all unbiased estimators for the given parameter.

We also want to know the sampling distribution of our statistics so we can test hypotheses, construct confidence intervals, and so forth.

Assumptions for Unbiased Coefficients

To show that the coefficients in a multiple regression are unbiased, we need four assumptions, which are extension of the assumptions we learned for simple regression.

1. Linear in Parameters
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean

The Gauss-Markov Assumptions

There are five Gauss-Markov assumptions in Wooldridge Chapter 3. The first four establish unbiasedness of OLS, whereas the fifth was added to derive the usual variance formulas and to conclude that OLS is best linear unbiased.

Assumption MLR.1 (Linear in Parameters)
The model in the population can be written as $y = \beta_0+\beta_1x_i+\ldots+\beta_kx_k+u$, where β_0, β_1,…, β_k are the unknown parameters (constants) of interest and u is an unobserved random error term.

Assumption MLR.2 (Random Sampling)
We have a random sample of n observations, $\{(x_{i1}, x_{i2},\ldots, x_{ik}, y_i): i = 1, 2,\ldots, n\}$, following the population model in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)
In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

Additional Notes on Estimators

Assumption MLR.4 (Zero Conditional Mean)
The error u has an expected value of zero given any values of the independent variables.
In other words, $E(u|x_1,x_2,\ldots,x_k) = 0$.

Assumption MLR.5 (Homoskedasticity)
The error u has the same variance given any value of the explanatory variables.
In other words, $Var(u|x_1,x_2,\ldots,x_k) = 0$.

Associative versus Causal models

What is the difference between the following two conditions?

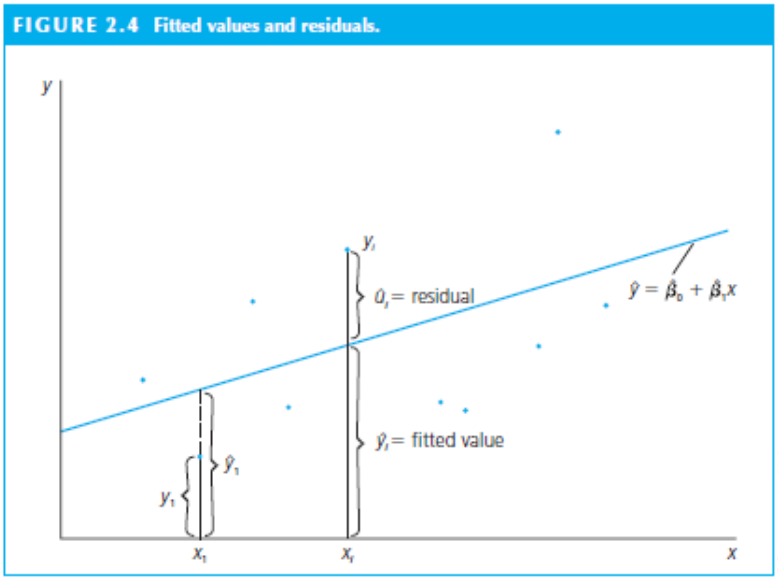
(1) $cov(x, \hat{u}) = 0$
(2) $cov\ (x, u) = 0$

Recall that given a random sample of size n from the population, $\{(x_i, y_i): i = 1,\ldots,n\}$, for each i we can write $y_i = \beta_0+\beta_1 x_i+u_i$, where u_i is the error term for observation i because it contains all factors affecting y_i other than x_i.

In the population, u is uncorrelated with x. Hence, u has zero expected value and the covariance between x and u is zero. In other words, E(u) = 0 and cov(x,u) = E(xu) = 0.

Recall that for any OLS estimates of β_0 and β_1, i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$, we can define a fitted value for y when $x=x_i$ as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The residual for observation i is the difference between the actual y_i and its fitted value, i.e. $\hat{u}_i = y_i - \hat{y}_i$.

By the OLS first order conditions, $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, which implies that each independent variable has zero sample covariance with $\hat{u}_i$.



Algebraic Properties of OLS Statistics

(1) By the OLS first order condition, the sum, and therefore the sample average of the OLS residuals, is zero, i.e. $\sum_{i=1}^{n} \hat{u}_i = 0$.
(2) The sample covariance between the regressors and the OLS residuals is zero. $\sum_{i=1}^{n} x_i \hat{u}_i = 0$. The sample average of the OLS residuals is zero, so the left-hand side of the equation is proportional to the sample covariance between x_i and $\hat{u}_i$.
(3) The point $(\bar{x}_i, \bar{y}_i)$ is always on the OLS regression line.

Residual Plots

- An **outlier** is a data point whose response *y* does not follow the general trend of the rest of the data.
- A data point has **high leverage** if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values (e.g., with two predictors that are positively correlated, an unusual combination of predictor values might be a high value of one predictor paired with a low value of the other predictor).
- A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.

Outliers and high leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential.