Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Some Useful Results from Wooldridge

Simple linear regression
Assumes a linear conditional expectation function (CEF) with a causal interpretation. Estimating the popluation regression function (PRF) is thus estimating the (causal) CEF.

$$E(y|x) = \beta_0 + \beta_1 x$$
$$y = \beta_0 + \beta_1 x + u$$

where $E(u|x) = 0$ (mean independence, or the zero conditional mean assumption). Note this is a stronger assumption than $Corr(u, x) = 0$.

Estimation: use a random sample of (x_i, y_i) of size n to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$. Under the assumptions $E(u) = 0$ and $Cov(x, u) = E(xu) = 0$, the method of moments yields:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\hat{\beta}_1$ is the sample covariance between x and y divided by the variance of x. Two alternate ways of writing $\hat{\beta}_1$ are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_1 = \hat{\rho}_{xy} \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$$

where $\hat{\rho}_{xy}$ is the sample correlation between x_i and y_i and $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the sample standard deviations of x_i and y_i, respectively. The predicted value of y and residual u are defined by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$\hat{u}_i = (y_i - \hat{y}_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

OLS: the sum of squared residuals (SSR) is $\sum \hat{u}_i^2$. It can be shown $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the SSR (why they are called OLS estimators). Other useful algebraic properties:

- $\sum_{i=1}^n \hat{u}_i = 0$, so $\bar{y} = \bar{\hat{y}}$
- $\sum_{i=1}^n x_i \hat{u}_i = 0$ (sample covariance between x and u is zero)

- $\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$ (sample covariance between $\hat{y}$ and u is zero)
- $(\bar{x}, \bar{y})$ is on the regression line
- SST = SSE + SSR, or $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$. SSE is the "explained" sum of squares, and SST is the total sum of squares.
- $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$, used as a measure of goodness-of-fit

Gauss-Markov assumptions: used to derive the statistical properties of $\hat{\beta}_0$ and $\hat{\beta}_1$:

- SLR1: $y = \beta_0 + \beta_1 x + u$ is the population model.
- SLR2: Data represent a random sample of n draws of (x_i, y_i) from the population.
- SLR3: $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$ (there is variation in x_i).
- SLR4: $E(u|x) = 0$ (zero conditional mean)
- SLR5: $Var(u|x) = \sigma^2$ (homoskedasticity)

Small sample properties: under SLR1-SLR4, $\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased*. Adding SLR5, the sampling variance of $\hat{\beta}_1$ is:

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{SST_x}$$

where SST_x is the total variation in x, and the population error variance σ_u^2 can be estimated using residuals:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$$

Note $\sqrt{\hat{\sigma}_u^2}$ is the standard error of the regression, or RMSE. The estimated standard error of $\hat{\beta}_1$ is thus:

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Multiple linear regression
Assumes a linear conditional expectation function (CEF) with a causal interpretation. Estimating the popluation regression function (PRF) is thus estimating the (causal) CEF.

$$\begin{aligned} E(y|x_1, x_2, ..., x_k) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \end{aligned}$$

where $E(u|x) = 0$ (the zero conditional mean assumption).

Estimation: use a random sample of $(x_{1i}, x_{2i}, ..., x_{ki}, y_i)$ of size n to estimate $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$. Under the conditional mean assumption, $E(u) = 0$ and $Cov(x_j, u) = E(x_j u) = 0 \; \forall j$.
The predicted value of y and residual u are defined by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ...\hat{\beta}_k x_{ki}$$
$$\hat{u}_i = (y_i - \hat{y}_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_{2i} - ...\hat{\beta}_k x_{ki}$$

OLS: the sum of squared residuals (SSR) is $\sum \hat{u}_i^2$. It can be shown the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ minimize this. Other useful algebraic properties:

- $\sum_{i=1}^n \hat{u}_i = 0$, so $\bar{y} = \bar{\hat{y}}$
- $\sum_{i=1}^n x_{ji}\hat{u}_i = 0$ for all j (sample covariance between x_j and u is zero)
- $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k, \bar{y}$ is on the regression line

Without matrix algebra it is difficult to write a simple expression for the $\hat{\beta}_j$.

Gauss-Markov assumptions: used to derive statistical properties of $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$:

- MLR1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$ is the population model.
- MLR2: Data represent a random sample of n draws of $(x_{1i}, x_{2i}, ..., x_{ki}, y_i)$ from the population.
- MLR3: No perfect collinearity: fails if perfect collinearity between any two x_j or if $n < k + 1$
- MLR4: $E(u|x_1, x_2, ..., x_k) = 0$ (exogenous explanatory variables)
- MLR5: $Var(u|x_1, x_2, ..., x_k) = \sigma^2$ (homoskedasticity)

Small sample properties: under MLR1-MLR4, $\hat{\beta}_j$ is *unbiased*. Adding MLR5, the variance of $\hat{\beta}_j$ is:

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)}$$

where SST_j is the total variation in x_j, $\sum(x_{ij} - \bar{x}_j)^2$, R_j^2 is the R^2 from a regression of x_j on all other covariates, and the population error variance σ_u^2 can be estimated using residuals:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n - k - 1}$$

Factors that affect the sampling variation in $\hat{\beta}_j$ include: (1) population error variance (σ_u^2); (2) total variation in x_j; (3) sample size n; (4) degree of collinearity in the x (seen in R_j^2).

The estimated standard error of $\hat{\beta}_j$ is:

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1 - R_j^2)}}$$

The Gauss-Markov Theorem: under assumptions MLR1-MLR5, the OLS estimator for $\hat{\beta}_j$ is the *best linear unbiased estimator* (BLUE). That is, minimum variance among all linear estimators.

Sampling distributions, confidence intervals, hypothesis tests

The results above are informative about the *mean* and *variance* of the $\hat{\beta}_j$. That is, under certain assumptions we know that the estimator is *unbiased*, and we know and can estimate its standard error. Stronger assumptions are needed if we want to presume this distribution has a certain shape (e.g., normality), permitting us to construct confidence intervals and test hypotheses. Large samples allow us to relax these assumptions (see asymptotics below).

- MLR6: $u \sim N(0, \sigma^2)$ (normality). Adding this assumption implies:

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$
$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$$

In practice we estimate $sd(\hat{\beta}_j)$ with $se(\hat{\beta}_j)$ and thus use the t statistic and t distribution to test hypotheses and construct confidence intervals:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

F-test of multiple linear restrictions: MLR6 also allows us to test *joint* hypotheses. Let SSR_r be the sum of squared residuals from a *restricted model* (in which a null hypothesis is imposed) and let SSR_u be the sum of squared residuals from an *unrestricted model* (in which the null is not imposed). Then:

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$$

where q is the number of restrictions made under the null hypothesis. Alternatively written:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k - 1)}$$

The F-statistic for overall significance of the regression—where H_0 is that all slope coefficients are zero—is:

$$F = \frac{R^2/k}{(1 - R^2)(n - k - 1)}$$

Large samples (asymptotics): All of the above results apply in finite samples, assuming MLR6. As long as the key assumptions hold (e.g., normality), the properties are exact. Often the normality assumption is too strong, however. *Asympototics* rely on the fact that in large samples, the estimator's distribution converges to a known distribution, such as the normal distribution.

Under MLR1-MLR4, the OLS estimator $\hat{\beta}_j$ is *consistent*. That is, its sampling distribution collapses to a single point, β_j

Under MLR1-MLR5, $\sqrt{n}(\hat{\beta}_j - \beta_j)$ has an *asymptotic normal distribution* $N(0, \sigma_u^2/a_j^2)$, where $a_j^2 = \text{plim}\left(\sum \hat{r}_{ij}^2\right)$. $\hat{r}_{ij}^2$ is the residual from from regressing x_j on the other x's.