We'll have a population model

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u.$$

Assumption 1 (OLS 1)**.** $E[u|X] = 0$. This still implies via the LIE that $E[u] = 0$, $E[Xu] = 0$, and $Cov(u, X_j) = 0$. Under this assumption, $\hat{\beta} \xrightarrow{p} \beta_j$, even if there are omitted variables, measurement error, or simultaneity. (The interpretation of β_j will change if any of these hold, however.)

We can also write the system

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 X_{1,1} + \ldots + \beta_k X_{k,1} + u_1 \\ &\vdots \\ y_N &= \beta_0 + \beta_1 X_{1,N} + \ldots + \beta_k X_{k,N} + u_N, \end{aligned}$$

which can be condensed into the matrix equation

$$Y = X\boldsymbol{\beta} + U.$$

Note that the first entry of the vector $\boldsymbol{\beta}$ is 1 to capture the constant intercept.

Assumption 2 (OLS 2)**.** $E[X'X]$ *is positive definite, i.e. has rank k, i.e. is invertible.*

With OLS1 and OLS2, we can show that

$$\boldsymbol{\beta} = E[X'X]^{-1}E[X'Y],$$

and as an estimation,

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y = \left(\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i' y_i\right).$$

OLS1 and OLS2 do not necessarily imply unbiasedness.

When $E[X'X]$ is assumed to be nonsingular, we can write $E[X'X] = A$. By plugging in $\beta x_i + u_i$ for y_i, you can show that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ because $E[X'u] = 0$.

Measures of Fit

SST is the total sample variation in y,

$$SST = \sum_{i=1}^{N}(y_i - \bar{y})^2.$$

SSE is the variation between the fitted values in y,

$$SSE = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2.$$

SSR is the variation in the sample residuals or error,

$$SSR = \sum_{i=1}^{N}\hat{u}^2.$$

The explained variation plus the error (eg unexplained) variation equals the total variation, so $SST = SSE + SSR$. This can be seen because

$$y_i - \overline{y} = \hat{y}_i - \bar{y} + \hat{u}_i \quad \Longrightarrow \quad y_i = \hat{y}_i + \hat{u}_i.$$

If we have a good fit, then we'd expect the explained variation to be very close to the total variation, so the ratio SSE/SST should be close to 1. We call this ratio the ***R*-squared**. Note that we can also write

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{N}\hat{u}^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

Usually we will want to adjust for degrees of freedom. The total variation has only one parameter—the mean—so it has $N - 1$ degrees of freedom. The SSR is fitting K parameters, and so it has $N - K$ degrees of freedom. Thus the **adjusted *R*-squared** is given by

$$\overline{R}^2 = 1 - \frac{SSR/(N-K)}{SST/(N-1)} = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N}\hat{u}^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

The ***F*-statistic** utilizes the R^2 when testing joint hypotheses. Let q be the number of restrictions (variables dropped in the restricted model), and again K is the number of independent variables. Then

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}}\frac{N - K - 1}{q}$$

Note that the F-statistic is only valid if homoskedasticity holds, which is practically never.

Similar to the F-statistic is the **Lagrange multiplier statistic**. Suppose that for a partitioned model $y = x_1\beta_1 + x_2\beta_2 + u$, we want to test $H_0 : \beta_2 = 0$.

- Plug the null hypothesis into the model to get $y = x_1\beta_1 + \tilde{u}$, so that $\tilde{u} = y - x_1\beta_1$.
- Now regress $\tilde{u}$ on x_1 and x_2.
- Take the R^2 from the previous regression and multiply it by N. This is the LM statistic.

The LM statistic is distributed asymptotically according to $\mathcal{X}^2_{K_2}$, where K_2 is the number of restrictions being texted (i.e. the number of independent variables in x_2). This is also valid only under homoskedasticity. Without it, we will regress 1 on $\overline{u}\hat{r}$ without an intercept. Then $LM = N - SSR_0$.

Instrumental Variables (2SLS)

Variables correlated with the error term are called **endogenous** variables, whereas variables uncorrelated with the error term are called **exogenous** variables.

The model is $y = \beta_0 + \beta_1 y_2 + \beta_2 x + u$, where y_2 is endogenous (and thus problematic) and x is exogenous. Because y_2 and u are correlated, the OLS estimator is inconsistent. If there is a valid instrumental variable z, then the effect on y of a unit change in y_2 can be estimated by using the instrumental variables estimator.

A valid instrumental variable must satisfy

(a) $\text{Corr}(y_2, z) \neq 0$, *(instrument relevance)*

(b) $\text{Corr}(u, z) = 0$. *(instrument exogeneity)*

The **two-stage least squares** is used for instrumental variables. The first stage decomposes y_2 into a problematic component that might be correlated with u, and a problem-free component that is uncorrelated with the error. The second stage uses the problem-free component to estimate β_1.

- The **first-stage regression** is to regress the endogenous variable y_2 on the exogenous variables and the instrumental variable,

  $$y_2 = \pi_0 + \pi_1 x + \pi_2 z + v.$$

  This is called the **reduced form** for y_2. The problem-free component is $\pi_0 + \pi_1 z$, and v is the problematic component. Let

  $$\hat{y_2} = \hat{\pi}_0 + \hat{\pi_1} x + \hat{\pi_2} z,$$

  using the OLS estimates of each coefficient. If $\hat{\pi_2} \neq 0$, then we're good to go.
- The **second-stage regression** is

  $$y = \beta_0 + \beta_1 \hat{y_2} + \beta_2 x + u.$$

  In Stata, we could use the code

  ```
  ivregress y x (y_2 = z)
  ```

Okay, so we need A valid instrumental variable must satisfy $2SLS1$:

(a) $\text{Corr}(y_2, z) \neq 0$, *(instrument relevance)*

(b) $\text{Corr}(u, z) = 0$. *(instrument exogeneity)*

We also need $2SLS2$: $\text{rank}\, E[z'z] = L$ and $\text{rank}\, E(z'x) = K$, where $z = 1 \times L$ and $x = 1 \times K$. So it must be the case that $L \geq K$, meaning we have at least as many instrument as we have explanatory variables. Note that z includes any exogenous element of x, so z should have those plus whatever instruments are required for the endogenous variables in x.

Finally, we can also consider $2SLS3$, which is stupid homoskedasticity $E[u^2 z'z] = \sigma^2 E[z'z]$ where $\sigma^2 = E[u^2]$.

Bootstraps

Non-Parametric. You have i.i.d. data $\{y_i, x_i\}_{i=1}^N$. Draw with replacement from the sample to generate "new" collections $\{y_i^*, x_i^*\}_{i=1}^N$. Do this B times, so as to generate B samples of size N. Calculate $\hat{\beta}_b^*$ for each $b \in B$. This gives you a distribution of $\hat{\beta}$. If you draw 1000 times, then look between draws 25 and 975 for the 95% confidence interval.

Parametric. You have data $\{y_i, x_i\}_{i=1}^N$, not necessarily i.i.d. Consider the model $y_i = x_i \hat{\beta}_{OLS} + \hat{u}_i$, i.e. use OLS to obtain $\hat{\beta}$ and $\hat{u}_i$. If the model is well specified, then $\{\hat{u}_i\}_{i=1}^N$ is i.i.d. Pair $\{x_i, \hat{u}_i\}$ and draw with replacement B times of size N. Then we can generate $y_i^* = x_i^* \hat{\beta} + \hat{u}_i^*$. Use $\{y_i^*, x_i^*\}_{i=1}^N$ to calculate your desired statistics.