1 Background

Suppose we are interested in comparing the OLS estimator

$$\widehat{\beta}_{OLS} = \left(X'X\right)^{-1} X'Y = \left(\sum_i X_i X_i'\right)^{-1} \sum_i X_i Y_i \tag{1}$$

with the WLS estimator

$$\widehat{\beta}_{WLS} = \left(X'WX\right)^{-1} X'WY = \left(\sum_i W_i X_i X_i'\right)^{-1} \sum_i W_i X_i Y_i \tag{2}$$

Some economists view the standard sufficient conditions for consistency of WLS as more restrictive than the standard sufficient conditions for consistency of OLS. The key standard sufficient condition for OLS is

$$0 = E\left[(Y_i - X_i'\beta_0)X_i\right] \tag{3}$$

This can be viewed as defining the parameter of interest as the probability limit of the OLS estimator. The key standard sufficient condition for WLS is instead

$$0 = E\left[(Y_i - X_i'\beta_0)X_i W_i\right] \tag{4}$$

which can also be viewed as defining the parameter of interest as the probability limit of the WLS estimator.

To relate these two conditions, consider the condition

$$0 \quad = \quad E\left[(Y_i - X_i'\beta_0)X_i W_i | W_i\right] \tag{5}$$

for almost every w in the support of W_i. It is straightforward to show using iterated expectations that this condition implies condition (3) and condition (4). The condition in (5) is thus stronger than either (3) or (4), neither of which implies the other.

Note that condition (5) may fail, and we may nonetheless be interested in the probability limit of WLS. In that case, the sufficient condition of interest becomes (4). Symmetrically, condition (5) may fail, but we may nonetheless be interested in the probability limit of OLS. In that case, the sufficient condition of interest is instead (3). Finally, we may be willing to assume the stronger condition (5).

So far I have stressed a view of regression that maintains there may be misspecification of the functional form. Sometimes a stronger set of conditions are used as sufficient conditions. These conditions pertain to the conditional mean of residuals being zero, rather than the covariance of the residuals and the covariates being zero. For this case, note that the key sufficient condition for OLS is

$$0 = E\left[(Y_i - X_i'\beta_0)X_i | X_i\right] \tag{6}$$

This condition is equivalent to assuming that the conditional expectation of Y_i given X_i is $X_i'\beta_0$. The key sufficient condition for WLS is instead

$$0 = E\left[(Y_i - X_i'\beta_0)X_i W_i | X_i W_i\right] \tag{7}$$

As with the case considered previous, neither of these conditions imply the other, but both are implied by the condition

$$0 = E\left[(Y_i - X_i'\beta_0)X_i W_i | X_i, W_i\right] \tag{8}$$

To summarize, a Hausman test contrasting OLS and WLS can be viewed in one of three ways:

1. As a test of the condition for WLS, assuming that the condition for OLS holds
2. As a test of the condition for OLS, assuming that the condition for WLS holds
3. As a test of a stronger condition which implies that the condition for OLS holds and that for WLS holds

Note that under 1, 2, or 3, WLS and OLS share a probability limit. Consequently, the most accurate way to state the null hypothesis being tested is as a test of the notion that WLS and OLS share a probability limit.

2 Testing

Note that WLS is a particular stripe of IV estimator. That is, define $Z_i = W_i X_i$ and observe that then

$$\widehat{\beta}_{IV} \equiv \left(Z'X\right)^{-1} Z'Y = \left(X'WX\right)^{-1} X'WY = \widehat{\beta}_{WLS} \tag{9}$$

Contrasting an IV estimate with an OLS estimate as a means of gauging the plausibility of covariate exogeneity (while maintaining the exogeneity of the instruments) is the textbook example of a Hausman test and features prominently in Hausman's original 1978 *Econometrica* article.

The Hausman test rejects the null hypothesis of covariate exogeneity when the quadratic form

$$(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS})' V \left[\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}\right]^{-1} (\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}) \tag{10}$$

is large. This is generally not easy to implement because the covariance terms in

$$V\left[\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}\right] = V\left[\widehat{\beta}_{IV}\right] + V\left[\widehat{\beta}_{OLS}\right] - C\left[\widehat{\beta}_{IV}, \widehat{\beta}_{OLS}\right] - C\left[\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}\right] \tag{11}$$

are not immediately available from software. Frequently one is taught that since OLS is efficient under the null hypothesis of covariate exogeneity, the variance of the IV-OLS difference is just the difference between the variance of the IV estimator and the OLS estimator. This is an attractive result, because it means that we can ignore the covariance terms. This is also an illusory result, because it requires that the OLS estimator actually be fully efficient, which only holds (by Gauss-Markov) when the variance matrix of the residuals is spherical. If any kind of heteroskedasticity or "cluster" variance approach is being used, then OLS is not fully efficient and one needs to compute the covariance terms to do a Hausman test correctly, as discussed in Hausman (1978) and some graduate textbooks (e.g., Davidson and MacKinnon).

Fortunately, it is easy to avoid this complexity by recasting the problem into an auxiliary regression framework. Note that

$$\widehat{\beta}_{IV} - \widehat{\beta}_{OLS} = \left(Z'X\right)^{-1} Z'Y - \left(X'X\right)^{-1} X'Y = \left(Z'X\right)^{-1} Z'Y - \left(Z'X\right)^{-1} \left(Z'X\right) \left(X'X\right)^{-1} X'Y \tag{12}$$

$$= \left(Z'X\right)^{-1} \left\{Z'Y - Z'P_X Y\right\} = \left(Z'X\right)^{-1} Z'M_X Y \tag{13}$$

Observe that $0 = \text{plim}\ \widehat{\beta}_{IV} - \widehat{\beta}_{OLS}$ if and only if $0 = \text{plim}\ Z'M_X Y$, which in turn holds if and only if $0 = \text{plim}\ \left(Z'M_X Z\right)^{-1} Z'M_X Y$. So an auxiliary regression approach simply regresses Y on both X and Z and tests the null hypothesis that the coefficient on Z is equal to zero. This auxiliary regression approach can accomodate any particular heteroskedasticity or "cluster" variance approach.

This leads to the following algorithm for testing WLS vs. OLS.

```
1. for every covariate X(k), generate Z(k)=W*X(k)
2. regress Y X1 X2 X3 Z1 Z2 Z3 W
3. test Z1=Z2=Z3=W=0
```

Note that we include the weight itself in the algorithm, too, because of the presence of a constant column within the vector `X`. Note that in this setting the vector `Z` does *not* have a constant column unless the weights are constant, in which case the entire issue of WLS vs. OLS is moot.

Note that in Stata, a weighted least squares regression would be implemented as `regress Y X [aw=W]`. Stata expects for the weight `W` to be proportional to the *inverse* of the variance of the observation. For example, if the data pertain to averages in a human population, `W` is usually taken to be the size of the subsample corresponding to the average (say, `W=N` where `N` is a variable giving the size of the subsample for the given observation). The transformation one performs to construct the auxiliary variables `Z` is then `Z=X*N`. The issue of how to transform the original observations to implement a Hausman test can be confusing. For example, it is different than how one transforms the data manually if one wants to perform WLS. There, one transforms the variables as `Xnew=X/sqrt(N)`, `Ynew=Y/sqrt(N)`, and `uno=1/sqrt(N)`, and then calls `regress Ynew Xnew uno`.