Predicted Values and Residuals

The OLS estimator is $\hat{\beta} = (X'X)^{-1}X'Y$. The predicted values are $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY$, and the residuals are $e = Y - \hat{Y} = Y - X\hat{\beta} = Y - PY = (I-P)Y = MY$. The matrix $P$ is the projection matrix for predicted values, and the matrix $M$ is the projection matrix for residuals.

Properties of the Projection Matrices

1. $PX = X(X'X)^{-1}X'X = X$

2. $MX = (I-P)X = X-PX = X-X = 0$

   (Note that since $M$ is symmetric, $MX=0$ implies $(MX)' = X'M = 0$.)

3. $PM = X(X'X)^{-1}X'M = 0$

   (Note that since $P$ is symmetric, $PM=0$ implies $P'M = 0$.)

4. The residuals and predicted values are orthogonal.

   $\hat{Y}'e = Y'P'MY = 0$ since $PM=0$

This result states that the OLS estimator decomposes $Y$ into the orthogonal components $\hat{Y}$ and $e$. 
**Partitioned Blocks of Regressors**

Partition $X$ into two blocks of regressors. That is, $X = [X_1 \ X_2]$. There are $K_1$ regressors in $X_1$ and $K_2$ regressors in $X_2$.

1. $MX_1 = 0$
   
   Note that $MX = 0$ implies that $M[X_1 \ X_2] = [MX_1 \ MX_2] = 0$ which implies that $MX_1 = 0$.

2. $PX_1 = X_1$
   
   Note that $MX_1 = 0$ implies that $(I-P)X_1 = 0$ which implies that $X_1-PX_1 = 0$.

Let $P_1$ denote $X_1(X_1'X_1)^{-1}X_1'$ and $M_1$ denote $I-P_1$.

3. $PP_1 = P_1$
   
   Note that $PP_1 = PX_1(X_1'X_1)^{-1}X_1' = X_1(X_1'X_1)^{-1}X_1' = P_1$

4. $MM_1 = M$
   
   Note that $MM_1 = (I-P)(I-P_1) = I-P-P_1+PP_1 = I-P-P_1+P_1 = I-P = M$
The Frisch-Waugh-Lovell Theorem

Consider the sample regression function

\[ Y = X\hat{\beta} = X_{i1}\hat{\beta}_1 + X_{i2}\hat{\beta}_2 + e \]

where we have used the partitions \( X = [X_1, X_2] \) and \( \hat{\beta}'=[\hat{\beta}_1', \hat{\beta}_2'] \). Note that \( \hat{\beta} = (X'X)^{-1}X'Y \), and that \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) are the submatrices corresponding to the partition on \( X \). (When using this notation, it is important to recognize that, in general, \( \hat{\beta}_1 \neq (X_1'X_1)^{-1}X_1'Y \) and \( \hat{\beta}_2 \neq (X_2'X_2)^{-1}X_2'Y \).)

Premultiplying the sample regression function by \( M_1 \) gives

\[ M_1Y = M_1X_1\hat{\beta}_1 + M_1X_2\hat{\beta}_2 + M_1e \]

\[ = M_1X_2\hat{\beta}_2 + M_1e \quad \text{since } M_1X_1 = 0 \]

\[ = M_1X_2\hat{\beta}_2 + M_1MY \quad \text{since } e = MY \]

\[ = M_1X_2\hat{\beta}_2 + e \quad \text{since } M_1M = M \text{ implies } M_1MY = MY = e. \]

Consequently, the sample regression functions

\[ Y = X_{i1}\hat{\beta}_1 + X_{i2}\hat{\beta}_2 + e \]

and

\[ (M_1Y) = (M_1X_2)\hat{\beta}_2 + e \]

provide identical estimates of \( \beta_2 \) and an identical residual vector \( e \).
Note that in the second regression function above, $M_1Y$ is the vector of residuals that result from an OLS regression of $Y$ on $X_1$. Likewise, the columns of $M_1X_2$ are the vectors of residuals that result from an OLS regression of each of the columns of $X_2$ on $X_1$. The implication is that we can estimate the model either directly, or as a sequence of regressions. To estimate the model directly, we just regress $Y$ on $X$. Alternatively, regress $Y$ on $X_1$ and compute the residuals $M_1Y$. Repeat the process for each column of $X_2$ to get the matrix of residuals $M_1X_2$. Once this is done, regress the residuals $M_1Y$ on $M_1X_2$. The estimate of $\hat{\beta}_2$ and the corresponding residual vector will be identical to those obtained by direct estimation.

So what are the implications of all of this? Perhaps a simple example will help. Consider a linear consumption function with trend. The model is:

$$C_t = \alpha + \beta I_t + \delta t + u_t$$

where $C$ denotes consumption, $I$ denotes income, and $t$ is the observation index. The most obvious approach to estimation is to regress $C$ on $I$ and $t$. Alternatively, one could “detrend” the data first and then estimate the model. Specifically, the residuals from a regression of $C$ on $t$ (with an intercept) are considered “detrended” values of consumption. Likewise, the residuals from a regression of $I$ on $t$ (again with an intercept) are detrended values of income. The FWL theorem states that a regression (no intercept) of the detrended values of $C$ on the detrended values of $I$ will give an identical estimate of $\hat{\beta}$, and the same residual vector (and consequently goodness of fit) as that obtained with direct estimation. (In the notation of the theorem, the dependent variable $Y$ corresponds to $[C_t]$, the submatrix $X_1$ correspond to $[1 \ t]$, and the submatrix $X_2$ corresponds to $[I \ t]$.) In the end, there is no difference between including a trend in the model, or estimating a model using detrended data. (Note that for this to be true, we must detrend not only the dependent variable, but all regressors as well.)
The previous example does not illustrate the true power of the theorem. There are times when the two step approach results in considerable simplification of the estimation process. One such example is the use of a panel data set (the same group of individuals are observed over some period of time) with fixed effects (a time-invariant individual-specific intercept). The fixed effects are reflected by a set of dummy (binary) variables that identify the individuals. In such data set, the number of individuals is often quite large (thousands or millions) and the number of time periods relatively small. Direct estimation is difficult because of the number of dummies that must be included. The residuals from a regression of each variable on the set of dummies have a simple form, however. For each individual, the residuals are just the difference between the observed value of the variable and the sample mean (over time) for that individual. The FWL theorem states that we can account for individual fixed effects by just estimating the model after differencing the variables for each individual from their sample means.