3
Linear models

Though modeling endogeneity may involve a variety of nonlinear or generalized linear, nonparametric or semiparametric models, and maximum likelihood or Bayesian estimation, much of the intuition is grounded in the basic linear model. This chapter provides a condensed overview of linear models and establishes connections with later discussions.

3.1 Standard linear model (*OLS*)

Consider the data generating process (*DGP*):

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim \left(0, \sigma^2 I\right)$, X is $n \times p$ (with rank p), and $E\left[X^T \varepsilon\right] = 0$, or more generally $E\left[\varepsilon \mid X\right] = 0$.

The *Gauss-Markov theorem* states that $b = \left(X^T X\right)^{-1} X^T Y$ is the minimum variance estimator of β amongst linear unbiased estimators. Gauss' insight follows from a simple idea. Construct b (or equivalently, the residuals or estimated errors, e) such that the residuals are orthogonal to every column of X (recall the objective is to extract all information in X useful for explaining Y — whatever is left over from Y should be unrelated to X).

$$X^T e = 0$$

where $e = Y - Xb$. Rewriting the orthogonality condition yields

$$X^T \left(Y - Xb\right) = 0$$

or the normal equations

$$X^T X b = X^T Y$$

Provided X is full column rank, this yields the usual *OLS* estimator

$$b = \left(X^T X\right)^{-1} X^T Y$$

It is straightforward to show that b is unbiased (conditional on the data X).

$$\begin{aligned}
E[b \mid X] &= E\left[\left(X^T X\right)^{-1} X^T Y \mid X\right] \\
&= E\left[\left(X^T X\right)^{-1} X^T (X\beta + \varepsilon) \mid X\right] \\
&= \beta + \left(X^T X\right)^{-1} X^T E[\varepsilon \mid X] = \beta + 0 = \beta
\end{aligned}$$

Iterated expectations yields $E[b] = E_X[E[b \mid X]] = E_X[\beta] = \beta$. Hence, unbiasedness applies unconditionally as well.

$$\begin{aligned}
Var[b \mid X] &= Var\left[\left(X^T X\right)^{-1} X^T Y \mid X\right] \\
&= Var\left[\left(X^T X\right)^{-1} X^T (X\beta + \varepsilon) \mid X\right] \\
&= E\left[\left\{\beta + \left(X^T X\right)^{-1} X^T \varepsilon - \beta\right\}\left\{\left(X^T X\right)^{-1} X^T \varepsilon\right\}^T \mid X\right] \\
&= \left(X^T X\right)^{-1} X^T E\left[\varepsilon\varepsilon^T\right] X \left(X^T X\right)^{-1} \\
&= \sigma^2 \left(X^T X\right)^{-1} X^T I X \left(X^T X\right)^{-1} \\
&= \sigma^2 \left(X^T X\right)^{-1}
\end{aligned}$$

Now, consider the stochastic regressors case,

$$Var[b] = Var_X[E[b \mid X]] + E_X[Var[b \mid X]]$$

The first term is zero since $E[b \mid X] = \beta$ for all X. Hence,

$$Var[b] = E_X[Var[b \mid X]] = \sigma^2 E\left[\left(X^T X\right)^{-1}\right]$$

the unconditional variance of b can only be described in terms of the average behavior of X.

To show that *OLS* yields the minimum variance linear unbiased estimator consider another linear unbiased estimator $b_0 = LY$ (L replaces $\left(X^T X\right)^{-1} X^T$). Since $E[LY] = E[LX\beta + L\varepsilon] = \beta$, $LX = I$.

Let $D = L - \left(X^T X\right)^{-1} X^T$ so that $DY = b_0 - b$.

$$\begin{aligned}
Var[b_0 \mid X] &= \sigma^2 \left[D + \left(X^T X\right)^{-1} X^T\right]\left[D + \left(X^T X\right)^{-1} X^T\right]^T \\
&= \sigma^2 \begin{pmatrix} DD^T + \left(X^T X\right)^{-1} X^T D^T + DX \left(X^T X\right)^{-1} \\ + \left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1} \end{pmatrix}
\end{aligned}$$

Since

$$LX = I = DX + \left(X^T X\right)^{-1} X^T X, DX = 0$$

and

$$Var\left[b_0 \mid X\right] = \sigma^2 \left(DD^T + \left(X^T X\right)^{-1}\right)$$

As DD^T is positive semidefinite, $Var\left[b\right]$ (and $Var\left[b \mid X\right]$) is at least as small as any other $Var\left[b_0\right]$ ($Var\left[b_0 \mid X\right]$). Hence, the Gauss-Markov theorem applies to both nonstochastic and stochastic regressors.

Theorem 3.1 *Rao-Blackwell theorem. If* $\varepsilon \sim N\left(0, \sigma^2 I\right)$ *for the above DGP,* b *has minimum variance of all unbiased estimators.*

Finite sample inferences typically derive from normally distributed errors and t (individual parameters) and F (joint parameters) statistics. Some asymptotic results related to the Rao-Blackwell theorem are as follows. For the Rao-Blackwell *DGP*, *OLS* is consistent and asymptotic normally (*CAN*) distributed. Since *MLE* yields b for the above *DGP* with normally distributed errors, *OLS* is asymptotically efficient amongst all *CAN* estimators. Asymptotic inferences allow relaxation of the error distribution and rely on variations of the laws of large numbers and central limit theorems.

3.2 Generalized least squares (*GLS*)

Suppose the *DGP* is $Y = X\beta + \varepsilon$ where $\varepsilon \sim (0, \Sigma)$ and $E\left[X^T \varepsilon\right] = 0$, or more generally, $E\left[\varepsilon \mid X\right] = 0$, X is $n \times p$ (with rank p). The *BLU* estimator is

$$b_{GLS} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} Y$$

$$E\left[b_{GLS}\right] = \beta$$

$$Var\left[b_{GLS} \mid X\right] = \left(X^T \Sigma^{-1} X\right)^{-1}$$

and

$$Var\left[b_{GLS}\right] = E\left[\left(X^T \Sigma^{-1} X\right)^{-1}\right] = \sigma^2 E\left[\left(X^T \Omega^{-1} X\right)^{-1}\right]$$

where scale is extracted to construct $\Omega^{-1} = \frac{1}{\sigma^2}\Sigma^{-1}$.

A straightforward estimation approach involves Cholesky decomposition of Σ.

$$\Sigma = \Gamma\Gamma^T = LD^{1/2}D^{1/2}L^T$$

where D is a matrix with pivots on the diagonal.

$$\Gamma^{-1}Y = \Gamma^{-1}\left(X\beta + \varepsilon\right)$$

and

$$\Gamma^{-1}\varepsilon \sim (0, I)$$

since $\Gamma^{-1}0 = 0$ and $\Gamma^{-1}\Sigma\left(\Gamma^{T}\right)^{-1} = \Gamma^{-1}\Gamma\Gamma^{T}\left(\Gamma^{T}\right)^{-1} = I$. Now, *OLS* applied to the regression of $\Gamma^{-1}Y$ (in place of Y) onto $\Gamma^{-1}X$ (in place of X) yields

$$\begin{aligned}
b_{GLS} &= \left(\left(\Gamma^{-1}X\right)^{T}\Gamma^{-1}X\right)^{-1}\left(\Gamma^{-1}X\right)^{T}\Gamma^{-1}Y \\
&= \left(X^{T}\left(\Gamma^{-1}\right)^{T}\Gamma^{-1}X\right)^{-1}X^{T}\left(\Gamma^{-1}\right)^{T}\Gamma^{-1}Y \\
b_{GLS} &= \left(X^{T}\Sigma^{-1}X\right)^{-1}X^{T}\Sigma^{-1}Y \text{ (Aitken estimator)}
\end{aligned}$$

Hence, *OLS* regression of suitably transformed variables is equivalent to *GLS* regression, the minimum variance linear unbiased estimator for the above *DGP*.

OLS is unbiased for the above *DGP* (but inefficient),

$$E\left[b\right] = \beta + E_X\left[\left(X^{T}X\right)^{-1}X^{T}E\left[\varepsilon \mid X\right]\right] = \beta$$

However, $Var\left[b \mid X\right]$ is not the standard one described above. Rather,

$$Var\left[b \mid X\right] = \left(X^{T}X\right)^{-1}X^{T}\Sigma^{-1}X\left(X^{T}X\right)^{-1}$$

which is typically estimated by Eicker-Huber-White asymptotic heteroskedasticity consistent estimator

$$\begin{aligned}
Est.Asy.Var\left[b\right] &= n\left(X^{T}X\right)^{-1}S_0\left(X^{T}X\right)^{-1} \\
&= n^{-1}\left(n^{-1}X^{T}X\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}e_i^2x_ix_i^T\right)\left(n^{-1}X^{T}X\right)^{-1}
\end{aligned}$$

where x_i is the ith row from X and $S_0 = 1/n\sum_{i=1}^{n}e_i^2x_ix_i^T$, or the Newey-West autocorrelation consistent covariance estimator where S_0 is replaced by $S_0 + n^{-1}\sum_{l=1}^{L}\sum_{t=l+1}^{n}w_le_ie_{t-l}\left(x_lx_{t-l}^T + x_{t-l}x_l^T\right)$, $w_l = 1 - \frac{l}{L+1}$, and the maximum lag L is set in advance.

3.3 Tests of restrictions and *FWL* (Frisch-Waugh-Lovell)

Causal effects are often the focus of accounting and economic analysis. That is, the question often boils down to what is the response to a change in one variable holding the others constant. *FWL* (partitioned regression or double residual regression) and tests of restrictions can help highlight causal effects in the context of linear models (and perhaps more broadly).

Consider the *DGP* for *OLS* where the matrix of regressors is partitioned $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ and X_1 represents the variables of prime interest and X_2 perhaps represents control variables.[1]

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

Of course, β can be estimated via *OLS* as b and b_1 (the estimate for β_1) can be extracted from b. However, it is instructive to remember that each β_k represents the response (of Y) to changes in X_k conditional on all other regressors X_{-k}. The *FWL* theorem indicates that b_1 can also be estimated in two steps. First, regress X_1 and Y onto X_2. Retain their residuals, e_1 and e_Y. Second, regress e_Y onto e_1 to estimate $b_1 = \left(e_1^T e_1\right)^{-1} e_1^T e_Y$ (a no intercept regression) and $Var\left[b_1 \mid X\right] = \sigma^2 \left(X^T X\right)_{11}^{-1}$, where $\left(X^T X\right)_{11}^{-1}$ refers to the upper left block of $\left(X^T X\right)^{-1}$.

FWL produces the following three results:
1.

$$\begin{aligned} b_1 &= \left(X_1^T \left(I - P_2\right) X_1\right)^{-1} X_1^T \left(I - P_2\right) Y \\ &= \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2 Y \end{aligned}$$

is the same as b_1 from the upper right partition of

$$b = \left(X^T X\right)^{-1} X^T Y$$

where $P_2 = X_2 \left(X_2^T X_2\right)^{-1} X_2^T$.
2.

$$\begin{aligned} Var\left[b_1\right] &= \sigma^2 \left(X_1^T \left(I - P_2\right) X_1\right)^{-1} \\ &= \sigma^2 \left(X_1^T M_2 X_1\right)^{-1} \end{aligned}$$

is the same as from the upper left partition of

$$Var\left[b\right] = \sigma^2 \left(X^T X\right)^{-1}$$

3. The regression or predicted values are

$$\begin{aligned} \widehat{Y} &= P_X Y = X \left(X^T X\right)^{-1} X^T Y = Xb \\ &= X_1 b_1 + X_2 b_2 = P_2 Y + \left(I - P_2\right) X_1 b_1 \\ &= P_2 Y + M_2 X_1 b_1 \end{aligned}$$

First, we demonstrate result 1. Since $e_1 = \left(I - P_2\right) X_1 = M_2 X_1$ and $e_Y = \left(I - P_2\right) Y = M_2 Y$,

$$\begin{aligned} b_1 &= \left(X_1^T \left(I - P_2\right) X_1\right)^{-1} X_1^T \left(I - P_2\right) Y \\ &= \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2 Y \end{aligned}$$

[1] When a linear specification of the control variables is questionable, we might employ partial linear or partial index regressions. For details see the discussion of these semi-parametric regression models in chapter 6. Also, a model specification test against a general nonparametric regression model is discussed in chapter 6.

To see that this is the same as from standard (one-step) multiple regression derive the normal equations from

$$X_1^T X_1 b_1 = X_1^T Y - X_1^T X_2 b_2$$

and

$$P_2 X_2 b_2 = X_2 b_2 = P_2 \left(Y - X_1 b_1 \right)$$

Substitute to yield

$$X_1^T X_1 b_1 = X_1^T Y - X_1^T P_2 \left(Y - X_1 b_1 \right)$$

Combine like terms in the normal equations.

$$\begin{aligned} X_1^T \left(I - P_2 \right) X_1 b_1 &= X_1^T \left(I - P_2 \right) Y \\ &= X_1^T M_2 Y \end{aligned}$$

Rewriting yields

$$b_1 = \left(X_1^T M_2 X_1 \right)^{-1} X_1^T M_2 Y$$

This demonstrates 1.[2]

[2] A more constructive demonstration of *FWL* result 1 is described below. From Gauss,

$$b = \left(\begin{bmatrix} X_2^T \\ X_1^T \end{bmatrix} \begin{bmatrix} X_2 & X_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_2^T \\ X_1^T \end{bmatrix} Y$$

(for convenience X is reordered as $\begin{bmatrix} X_2 & X_1 \end{bmatrix}$).

$$\left(\begin{bmatrix} X_2^T \\ X_1^T \end{bmatrix} \begin{bmatrix} X_2 & X_1 \end{bmatrix} \right)^{-1} = \begin{bmatrix} X_2^T X_2 & X_2^T X_1 \\ X_1^T X_2 & X_1^T X_1 \end{bmatrix}^{-1}$$

(by LDL^T block"rank-one" factorization)

$$\begin{aligned} &= \left(\begin{array}{c} \begin{bmatrix} I & 0 \\ X_1^T X_2 \left(X_2^T X_2 \right)^{-1} & I \end{bmatrix} \begin{bmatrix} X_2^T X_2 & 0 \\ 0 & X_1^T \left(I - P_2 \right) X_1 \end{bmatrix} \\ \times \begin{bmatrix} I & \left(X_2^T X_2 \right)^{-1} X_2^T X_1 \\ 0 & I \end{bmatrix} \end{array} \right)^{-1} \\ &= \begin{bmatrix} I & - \left(X_2^T X_2 \right)^{-1} X_2^T X_1 \\ 0 & I \end{bmatrix} \begin{bmatrix} \left(X_2^T X_2 \right)^{-1} & 0 \\ 0 & \left(X_1^T M_2 X_1 \right)^{-1} \end{bmatrix} \\ &\times \begin{bmatrix} I & 0 \\ -X_1^T X_2 \left(X_2^T X_2 \right)^{-1} & I \end{bmatrix} \end{aligned}$$

Multiply the first two terms and apply the latter inverse to $\begin{bmatrix} X_2 & X_1 \end{bmatrix}^T Y$

$$\begin{bmatrix} b_2 \\ b_1 \end{bmatrix} = \begin{bmatrix} \left(X_2^T X_2 \right)^{-1} & - \left(X_2^T X_2 \right)^{-1} X_2^T X_1 \left(X_1^T M_2 X_1 \right)^{-1} \\ 0 & \left(X_1^T M_2 X_1 \right)^{-1} \end{bmatrix} \begin{bmatrix} X_2^T Y \\ X_1^T \left(I - P_2 \right) Y \end{bmatrix}$$

$b_1 = (x^T M_2 x)^{-1} x^T M_2 Y$. This demonstrates *FWL* result 1.

FWL result 2 is as follows.

$$\begin{aligned}
Var\left[b\right] &= \sigma^2\left(X^TX\right)^{-1} = \sigma^2\begin{bmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{bmatrix}^{-1} \\
&= \sigma^2\begin{bmatrix} A_{11} & -\left(X_1^TX_1\right)^{-1}X_1^TX_2A_{22} \\ -\left(X_2^TX_2\right)^{-1}X_2^TX_1A_{11} & A_{22} \end{bmatrix}
\end{aligned}$$

where

$$A_{11} = \left(X_1^TX_1 - X_1^TX_2\left(X_2^TX_2\right)^{-1}X_2^TX_1\right)^{-1}$$

and

$$A_{22} = \left(X_2^TX_2 - X_2^TX_1\left(X_1^TX_1\right)^{-1}X_1^TX_2\right)^{-1}$$

Rewriting A_{11}, the upper left partition, and combining with σ^2 produces

$$\sigma^2\left(X_1^T\left(I-P_2\right)X_1\right)^{-1}$$

This demonstrates *FWL* result 2.

To demonstrate *FWL* result 3

$$X_1b_1 + X_2b_2 = P_2Y + \left(I-P_2\right)X_1b_1$$

refer to the estimated model

$$Y = X_1b_1 + X_2b_2 + e$$

where the residuals e, by construction, are orthogonal to X. Multiply both sides by P_2 and simplify

$$\begin{aligned}
P_2Y &= P_2X_1b_1 + P_2X_2b_2 + P_2e \\
&= P_2X_1b_1 + X_2b_2
\end{aligned}$$

Rearranging yields

$$X_2b_2 = P_2\left(Y - X_1b_1\right)$$

Now, add X_1b_1 to both sides

$$X_1b_1 + X_2b_2 = X_1b_1 + P_2\left(Y - X_1b_1\right)$$

Simplification yields

$$\begin{aligned}
X_1b_1 + X_2b_2 &= P_2Y + \left(I-P_2\right)X_1b_1 \\
&= P_2Y + M_2X_1b_1
\end{aligned}$$

This demonstrates *FWL* result 3.

3.4 Fixed and random effects

Often our data come in a combination of cross-sectional and time-series data, or panel data which can substantially increase sample size. A panel data regression then is

$$Y_{tj} = X_{tj}\beta + u_{tj}$$

where t refers to time and j refers to individuals (or firms). With panel data, one approach is multivariate regression (multiple dependent variables and hence multiple regressions as in, for example, seemingly unrelated regressions). Another common approach, and the focus here, is an error-components model. The idea is to model u_{tj} as consisting of three individual shocks, each independent of the others

$$u_{tj} = e_t + \nu_j + \varepsilon_{tj}$$

For simplicity, suppose the time error component e_t is independent across time $t = 1, \ldots, T$, the individual error component ν_j is independent across units $j = 1, \ldots, n$, and the error component ε_{tj} is independent across all time t and individuals j.

There are two standard regression strategies for addressing error components: (1) a fixed effects regression, and (2) a random effects regression. Fixed effects regressions model time effects e_t and/or individual effects ν_j conditionally. On the other hand, the random effects regressions are modeled unconditionally. That is, random effects regressions model time effects e_t and individual effects ν_j as part of the regression error. The trade-offs between the two involve the usual regression considerations. Since fixed effects regressions condition on e_t and ν_j, fixed effects strategies do not rely on independence between the regressors and the error components e_t and ν_j. On the other hand, when appropriate (when independence between the regressors and the error components e_t and ν_j is satisfied), the random effects model more efficiently utilizes the data. A Hausman test (Hausman [1978]) can be employed to test the consistency of the random effects model by reference to the fixed effects model.

For purposes of illustration, assume that there are no time-specific shocks, that is $e_t = 0$ for all t. Now the error components regression is

$$Y_{tj} = X_{tj}\beta + \nu_j + \varepsilon_{tj}$$

In matrix notation, the fixed effects version of the above regression is

$$Y = X\beta + D\nu + \varepsilon$$

where D represents n dummy variables corresponding to the n cross-sectional units in the sample. Provided ε_{tj} is iid, the model can be estimated via *OLS*. Or using *FWL*, the fixed effects estimator for β is

$$\widehat{\beta}^{WG} = \left(X^T M_D X\right)^{-1} X^T M_D Y$$

where $P_D = D\left(D^T D\right)^{-1} D^T$, projection into the columns of D, and $M_D = I - P_D$, the projection matrix that produces the deviations from cross-sectional group means. That is,

$$\left(M_D X\right)_{tj} = X_{tj} - \overline{X}_{\cdot j}$$

and

$$M_D Y_{tj} = Y_{tj} - \overline{Y}_{\cdot j}$$

where $\overline{X}_{\cdot j}$ and $\overline{Y}_{\cdot j}$ are the group (individual) j means for the regressors and regressand, respectively. Since this estimator only exploits the variation between the deviations of the regressand and the regressors from their respective group means, it is frequently referred to as a within-groups (*WG*) estimator.

Use of only the variation between deviations can be an advantage or a disadvantage. If the cross-sectional effects are correlated with the regressors, then the *OLS* estimator (without fixed effects) is inconsistent but the within-groups estimator is consistent. However, if the cross-sectional effects (i.e., the group means) are uncorrelated with the regressors then the within-groups (fixed effects) estimator is inefficient. In the extreme case in which there is an independent variable that has no variation between the deviations and only varies between group means, then the coefficient for this variable is not even identified by the within-groups estimator.

To see that *OLS* is inconsistent when the cross-sectional effects are correlated with the errors consider the complementary between-groups estimator. A between-groups estimator only utilizes the variation among group means.

$$\widehat{\beta}^{BG} = \left(X^T P_D X\right)^{-1} X^T P_D Y$$

The between-groups estimator is inconsistent if the (cross-sectional) group means are correlated with the regressors. Further, since the *OLS* estimator can be written as a matrix-weighted average of the within-groups and between-groups estimators, if the between-groups estimator is inconsistent, *OLS* (without fixed effects) is inconsistent as demonstrated below.

$$\widehat{\beta}^{OLS} = \left(X^T X\right)^{-1} X^T Y$$

Since $M_D + P_D = I$,

$$\widehat{\beta}^{OLS} = \left(X^T X\right)^{-1} \left(X^T M_D Y + X^T P_D Y\right)$$

Utilizing $\left(X^T X\right)^{-1} X^T X = \left(X^T X\right)^{-1} X^T \left(M_D + P_D\right) X = I$, we rewrite the *OLS* estimator as a matrix-weighted average of the within-groups and between-groups estimators

$$\begin{aligned}
\widehat{\beta}^{OLS} &= \left(X^T X\right)^{-1} X^T M_D X \widehat{\beta}^{WG} + \left(X^T X\right)^{-1} X^T P_D X \widehat{\beta}^{BG} \\
&= \left(X^T X\right)^{-1} X^T M_D X \left(X^T M_D X\right)^{-1} X^T M_D Y \\
&\quad + \left(X^T X\right)^{-1} X^T P_D X \left(X^T P_D X\right)^{-1} X^T P_D Y \\
&= \left(X^T X\right)^{-1} X^T M_D X \left(X^T M_D X\right)^{-1} X^T M_D \left(X\beta + u\right) \\
&\quad + \left(X^T X\right)^{-1} X^T P_D X \left(X^T P_D X\right)^{-1} X^T P_D \left(X\beta + u\right)
\end{aligned}$$

Now, if the group means are correlated with the regressors then

$$\begin{aligned}
p\lim \widehat{\beta}^{BG} &= p\lim \left(X^T P_D X\right)^{-1} X^T P_D \left(X\beta + u\right) \\
&= \beta + p\lim \left(X^T P_D X\right)^{-1} X^T P_D u \\
&= \beta + \alpha \quad \alpha \neq 0
\end{aligned}$$

and

$$\begin{aligned}
p\lim \widehat{\beta}^{OLS} &= \left(X^T X\right)^{-1} X^T X \beta + \left(X^T X\right)^{-1} X^T P_D X \alpha \\
&= \beta + \left(X^T X\right)^{-1} X^T P_D X \alpha \\
&\neq \beta \quad \text{if } \alpha \neq 0
\end{aligned}$$

Hence, *OLS* is inconsistent if the between-groups estimator is inconsistent, in other words, if the cross-sectional effects are correlated with the errors.

Random effects regressions are typically estimated via *GLS* or maximum likelihood (here we focus on *GLS* estimation of random effects models). If the individual error components are uncorrelated with the group means of the regressors, then *OLS* with fixed effects is consistent but inefficient. We may prefer to employ a random effects regression which is consistent and more efficient. *OLS* treats all observations equally but this is not an optimal usage of the data. On the other hand, a random effects regression treats ν_j as a component of the error rather than fixed. The variance of u_{jt} is σ_ν^2 + σ_ε^2. The covariance of u_{ti} with u_{tj} is zero for $i \neq j$, under the conditions described above. But the covariance of u_{tj} with u_{sj} is σ_ν^2 for $s \neq t$. Thus, the $T \times T$ variance-covariance matrix is

$$\Sigma = \sigma_\varepsilon^2 I + \sigma_\nu^2 \iota \iota^T$$

where ι is a T-length vector of ones and the data are ordered first by individual unit and then by time. And the covariance matrix for the u_{tj} is

$$Var\left[u\right] = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma \end{bmatrix}$$

GLS estimates can be computed directly or the data can be transformed and *OLS* applied. We'll briefly explore a transformation strategy. One transformation, derived via singular value decomposition (*SVD*), is

$$\Sigma^{-1/2} = \sigma_{\varepsilon}^{-1}\left(I - \alpha P_{\iota}\right)$$

where $P_{\iota} = \iota\left(\iota^{T}\iota\right)^{-1}\iota^{T} = \frac{1}{T}\iota\iota^{T}$ and α, between zero and one, is

$$\alpha = 1 - \sigma_{\varepsilon}\left(T\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2}\right)^{-\frac{1}{2}}$$

The transformation is developed as follows. Since Σ is symmetric, *SVD* combined with the spectral theorem implies we can write

$$\Sigma = Q\Lambda Q^{T}$$

where Q is an orthogonal matrix ($QQ^{T} = Q^{T}Q = I$) with eigenvectors in its columns and Λ is a diagonal matrix with eigenvalues along its diagonal; $T-1$ eigenvalues are equal to σ_{ε}^{2} and one eigenvalue equals $T\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2}$. To fix ideas, consider the $T = 2$ case,

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2} & \sigma_{\nu}^{2} \\ \sigma_{\nu}^{2} & \sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2} \end{bmatrix} \\ &= Q\Lambda Q^{T} \end{aligned}$$

where

$$\Lambda = \begin{bmatrix} \sigma_{\varepsilon}^{2} & 0 \\ 0 & 2\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2} \end{bmatrix}$$

and

$$Q = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

Since

$$\Sigma = Q\begin{bmatrix} \sigma_{\varepsilon}^{2} & 0 \\ 0 & 2\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2} \end{bmatrix}Q^{T}$$

and

$$\begin{aligned} \Sigma^{-1} &= Q\begin{bmatrix} \frac{1}{\sigma_{\varepsilon}^{2}} & 0 \\ 0 & \frac{1}{2\sigma_{\nu}^{2}+\sigma_{\varepsilon}^{2}} \end{bmatrix}Q^{T} \\ &= Q\left(\begin{bmatrix} \frac{1}{\sigma_{\varepsilon}^{2}} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2\sigma_{\nu}^{2}+\sigma_{\varepsilon}^{2}} \end{bmatrix}\right)Q^{T} \\ &= Q\begin{bmatrix} \frac{1}{\sigma_{\varepsilon}^{2}} & 0 \\ 0 & 0 \end{bmatrix}Q^{T} + Q\begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2\sigma_{\nu}^{2}+\sigma_{\varepsilon}^{2}} \end{bmatrix}Q^{T} \\ &= \frac{1}{\sigma_{\varepsilon}^{2}}Q\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}Q^{T} + \frac{1}{2\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2}}Q\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}Q^{T} \\ &= \frac{1}{\sigma_{\varepsilon}^{2}}\left(I - P_{\iota}\right) + \frac{1}{2\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2}}P_{\iota} \end{aligned}$$

Note, the key to the general case is to construct Q such that

$$Q\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} Q^T = P_\iota$$

Since $I - P_\iota$ and P_ι are orthogonal projection matrices, we can write

$$\begin{aligned} \Sigma^{-1} &= \Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}} \\ &= \left(\frac{1}{\sigma_\varepsilon}(I - P_\iota) + \left(\frac{1}{T\sigma_\nu^2 + \sigma_\varepsilon^2}\right)^{\frac{1}{2}} P_\iota\right) \\ &\quad \times \left(\frac{1}{\sigma_\varepsilon}(I - P_\iota) + \left(\frac{1}{T\sigma_\nu^2 + \sigma_\varepsilon^2}\right)^{\frac{1}{2}} P_\iota\right) \end{aligned}$$

and the above claim

$$\begin{aligned} \Sigma^{-1/2} &= \sigma_\varepsilon^{-1}(I - \alpha P_\iota) \\ &= \sigma_\varepsilon^{-1}\left(I - \left[1 - \sigma_\varepsilon\left(T\sigma_\nu^2 + \sigma_\varepsilon^2\right)^{-\frac{1}{2}}\right] P_\iota\right) \\ &= \frac{1}{\sigma_\varepsilon}(I - P_\iota) + \left(\frac{1}{T\sigma_\nu^2 + \sigma_\varepsilon^2}\right)^{\frac{1}{2}} P_\iota \end{aligned}$$

is demonstrated.

A typical element of

$$\Sigma^{-1/2}Y_{\cdot j} = \sigma_\varepsilon^{-1}\left(Y_{tj} - \alpha\overline{Y}_{\cdot j}\right)$$

and for

$$\Sigma^{-1/2}X_{\cdot j} = \sigma_\varepsilon^{-1}\left(X_{tj} - \alpha\overline{X}_{\cdot j}\right)$$

GLS estimates then can be derived from the following *OLS* regression

$$\left(Y_{tj} - \alpha\overline{Y}_{\cdot j}\right) = \left(X_{tj} - \alpha\overline{X}_{\cdot j}\right) + residuals$$

Written in matrix terms this is

$$(I - \alpha P_\iota) Y = (I - \alpha P_\iota) X + (I - \alpha P_\iota) u$$

It is instructive to connect the *GLS* estimator to the *OLS* (without fixed effects) estimator and to the within-groups (fixed effects) estimator. When $\alpha = 0$, the *GLS* estimator is the same as the *OLS* (without fixed effects) estimator. Note $\alpha = 0$ when $\sigma_\nu = 0$ (i.e., the error term has only one component ε). When $\alpha = 1$, the *GLS* estimator equals the within-groups estimator. This is because $\alpha = 1$ when $\sigma_\varepsilon = 0$, or the between groups variation is zero. Hence, in this case the within-groups (fixed effects) estimator is fully efficient. In all other cases, α is between zero and one and the *GLS* estimator exploits both within-groups and between-groups variation. Finally, recall consistency of random effects estimators relies on there being no correlation between the error components and the regressors.

3.5 Random coefficients

Random effects can be generalized by random slopes as well as random intercepts in a random coefficients model. Then, individual-specific or heterogeneous response is more fully accommodated. Hence, for individual i, we have

$$Y_i = X_i\beta_i + \varepsilon_i$$

3.5.1 Nonstochastic regressors

Wald [1947], Hildreth and Houck [1968], and Swamy [1970] proposed standard identification conditions and (*OLS* and *GLS*) estimators for random coefficients. To fix ideas, we summarize Swamy's conditions. Suppose there are T observations on each of n individuals with observable outcomes Y_i and regressors X_i and unobservables β_i and ε_i.

$$\underset{(T\times 1)}{Y_i} = \underset{(T\times K)}{X_i}\underset{(K\times 1)}{\beta_i} + \underset{(T\times 1)}{\varepsilon_i} \quad (i = 1, \ldots, n)$$

Condition 3.1 $E[\varepsilon_i] = 0 \quad E\left[\varepsilon_i\varepsilon_j^T\right] = \begin{matrix} \sigma_{ii}I & i = j \\ 0 & i \neq j \end{matrix}$

Condition 3.2 $E[\beta_i] = \beta$

Condition 3.3 $E\left[(\beta_i - \beta)(\beta_i - \beta)^T\right] = \begin{matrix} \Delta & i = j \\ 0 & i \neq j \end{matrix}$

Condition 3.4 *β_i and ε_i are independent*

Condition 3.5 *β_i and β_j are independent for $i \neq j$*

Condition 3.6 X_i $(i = 1, \ldots, n)$ *is a matrix of K nonstochastic regressors,* x_{itk} $(t = 1, \ldots, T; k = 1, \ldots, K)$

It's convenient to define $\beta_i = \beta + \delta_i$ $(i = 1, \ldots, n)$ where $E[\delta_i] = 0$ and

$$E\left[\delta_i\delta_i^T\right] = \begin{matrix} \Delta & i = j \\ 0 & i \neq j \end{matrix}$$

Now, we can write a stacked regression in error form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}\beta + \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_n \end{bmatrix}\begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or in compact error form

$$Y = X\beta + H\delta + \varepsilon$$

where H is the $nT \times nT$ block matrix of regressors and the $nT \times 1$ disturbance vector, $H\delta + \varepsilon$, has variance

$$\begin{aligned} V &\equiv Var\left[H\delta + \varepsilon\right] \\ &= \begin{bmatrix} X_1 \Delta X_1^T + \sigma_{11} I & 0 & \cdots & 0 \\ 0 & X_2 \Delta X_2^T + \sigma_{22} I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_n \Delta X_n^T + \sigma_{nn} I \end{bmatrix} \end{aligned}$$

Therefore, while the parameters, β or $\beta + \delta$, can be consistently estimated via *OLS*, *GLS* is more efficient. Swamy [1970] demonstrates that β can be estimated directly via

$$\begin{aligned} b^{GLS} &= \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y \\ &= \left[\sum_{j=1}^{n} X_j^T \left(X_j \Delta X_j^T + \sigma_{jj} I\right)^{-1} X_j\right]^{-1} \\ &\quad \times \sum_{i=1}^{n} X_i^T \left(X_i \Delta X_i^T + \sigma_{ii} I\right)^{-1} Y_i \end{aligned}$$

or equivalently by a weighted average of the estimates for $\beta + \delta$

$$b^{GLS} = \sum_{i=1}^{n} W_i b_i$$

where, applying the matrix inverse result in Rao [1973, (2.9)],[3]

$$W_i = \left[\sum_{j=1}^{n} \left(\Delta + \sigma_{jj} \left(X_j^T X_j\right)^{-1}\right)^{-1}\right]^{-1} \left(\Delta + \sigma_{ii} \left(X_i^T X_i\right)^{-1}\right)^{-1}$$

and $b_i = \left(X_i^T X_i\right)^{-1} X_i^T Y_i$ is an *OLS* estimate for $\beta + \delta_i$.

3.5.2 Correlated random coefficients

As with random effects, a key weakness of random coefficients is the condition that the effects (coefficients) are independent of the regressors. When this

[3] Rao's inverse result follows. Let A and D be nonsingular matrices of orders m and n and B be an $m \times n$ matrix. Then

$$\begin{aligned} \left(A + BDB^T\right)^{-1} &= A^{-1} - A^{-1} B \left(B^T A^{-1} B + D^{-1}\right)^{-1} B^T A^{-1} \\ &= A^{-1} - A^{-1} BEB^T A^{-1} + A^{-1} BE \left(E + D\right)^{-1} EB^T A^{-1} \end{aligned}$$

where $E = \left(B^T A^{-1} B\right)^{-1}$.

condition fails, *OLS* parameter estimation of β is likely inconsistent. However, Wooldridge [2002, ch. 18] suggests ignorability identification conditions. We briefly summarize a simple version of these conditions.[4] For a set of covariates W the following redundancy conditions apply:

Condition 3.7 $E[Y_i \mid X_i, \beta_i, W_i] = E[Y_i \mid X_i, \beta_i]$

Condition 3.8 $E[X_i \mid \beta_i, W_i] = E[X_i \mid W_i]$

Condition 3.9 $Var[X_i \mid \beta_i, W_i] = Var[X_i \mid W_i]$

and

Condition 3.10 $Var[X_i \mid W_i] > 0$ *for all* W_i

Then, β is identified as $\beta = E\left[\frac{Cov(X,Y|W)}{Var(X|W)}\right]$. Alternative ignorability conditions lead to a standard linear model.

Condition 3.11 $E[\beta_i \mid X_i, W_i] = E[\beta_i \mid W_i]$

Condition 3.12 *the regression of Y onto covariates W (as well as potentially correlated regressors, X) is linear*

Now, we can consistently estimate β via a linear panel data regression. For example, ignorable treatment allows identification of the average treatment effect[5] via the panel data regression

$$E[Y \mid D, W] = D\beta + H\delta + W\gamma_0 + D(W - E[W])\gamma_1$$

where D is (a vector of) treatments.

3.6 Ubiquity of the Gaussian distribution

Why is the Gaussian or normal distribution so ubiquitous? Jaynes [2003, ch. 7] argues probabilities are "states of knowledge" rather than long run frequencies. Further, probabilities as logic naturally draws attention to the Gaussian distribution. Before stating some general properties of this "central" distribution, we review it's development in Gauss [1809] as related by Jaynes [2003], p. 202. The Gaussian distribution is uniquely determined if we equate the error cancelling property of a maximum likelihood estimator (*MLE*; discussed in ch. 4) with the sample average. The argument proceeds as follows.

[4] Wooldridge [2002, ch. 18] discusses more general ignorable treatment (or conditional mean independence) conditions and also instrumental variables (*IV*) strategies. We defer *IV* approaches to chapter 10 when we consider average treatment effect identification strategies associated with continuous treatment.

[5] Average treatment effects for a continuum of treatments and their instrumental variable identification strategies are discussed in chapter 10.

Suppose we have a sample of $n+1$ observations, $x_0, x_1, \ldots, x_n$, and the density function factors $f(x_0, x_1, \ldots, x_n \mid \theta) = f(x_0 \mid \theta) \cdots f(x_n \mid \theta)$. The log-likelihood is

$$\sum_{i=0}^{n} \log f(x_i \mid \theta) = \sum_{i=0}^{n} g(x_i - \theta)$$

so the *MLE* $\widehat{\theta}$ satisfies

$$\sum_{i=0}^{n} \frac{\partial g\left(\widehat{\theta} - x_i\right)}{\partial \widehat{\theta}} = \sum_{i=0}^{n} g^{'}\left(\widehat{\theta} - x_i\right) = 0$$

Equating the *MLE* with the sample average we have

$$\widehat{\theta} = \overline{x} = \frac{1}{n+1} \sum_{i=0}^{n} x_i$$

In general, *MLE* and $\overline{x}$ are incompatible. However, consider a sample in which only x_0 is nonzero, that is, $x_1 = \cdots = x_n = 0$. Now, if we let $x_0 = (n+1)\,u$ and $\widehat{\theta} = u$ then

$$\widehat{\theta} - x_0 = u - (n+1)\,u = -nu$$

and

$$\sum_{i=0}^{n} g^{'}\left(\widehat{\theta} - x_i\right) = 0$$

becomes

$$\sum_{i=0}^{n} g^{'}(-nu) = 0$$

or since $u = \widehat{\theta} - 0$

$$g^{'}(-nu) + ng^{'}(u) = 0$$

The case $n = 1$ implies $g^{'}(u)$ must be anti-symmetric, $g^{'}(-u) = -g^{'}(u)$. With this in mind, $g^{'}(-nu) + ng^{'}(u) = 0$ reduces to

$$g^{'}(nu) = ng^{'}(u)$$

Apparently, (and naturally if we consider the close connection between the Gaussian distribution and linearity)

$$g^{'}(u) = au$$

that is, $g^{'}(u)$ is a linear function and

$$g(u) = \frac{1}{2}au^2 + b$$

For this to be a normalizable function, a must be negative and b determines the normalization. Hence, we have

$$f(x \mid \theta) = \sqrt{\tfrac{\alpha}{2\pi}} \exp\left[-\tfrac{1}{2}\alpha(x-\theta)^2\right] \quad 0 < \alpha < \infty$$

and the "natural" way to think of error cancellation is the Gaussian distribution with only the scale parameter α unspecified. Since the maximum of the Gaussian likelihood function always equals the sample average and in the special case above this is true only for the Gaussian likelihood, the Gaussian distribution is necessary and sufficient.

Then, the ubiquity of the Gaussian distribution follows from its error cancellation properties described above, the central limit theorem (discussed in the appendix), and the following general properties (Jaynes [2003], pp. 220-221).

A. When any smooth function with a single mode is raised to higher and higher powers, it approaches a Gaussian function.

B. The product of two Gaussian functions is another Gaussian function.

C. The convolution of two Gaussian functions is another Gaussian function (see discussion below).

D. The Fourier transform of a Gaussian function is another Gaussian function.

E. A Gaussian probability distribution has higher entropy than any other distribution with equal variance.

Properties A and E suggest why various operations result in convergence toward the Gaussian distribution. Properties B, C, and D suggest why, once attained, a Gaussian distribution is preserved.

3.6.1 Convolution of Gaussians

Property C is pivotal as repeated convolutions lead to the central limit theorem. First, we discuss discrete convolutions (see Strang [1986],pp. 294-5). The convolution of f and g is written $f * g$. It is the sum (integral) of two functions after one has been reversed and shifted. Let $f = (f_0, f_1, \ldots, f_{n-1})$ and $g = (g_0, g_1, \ldots, g_{n-1})$ then

$$f * g = \left(\begin{array}{c} f_0 g_0 + f_1 g_{n-1} + f_2 g_{n-2} + \cdots + f_{n-1} g_1, f_0 g_1 + f_1 g_0 + f_2 g_{n-1} \\ + \cdots + f_{n-1} g_1, \ldots, f_0 g_{n-1} + f_1 g_{n-2} + f_2 g_{n-3} + \cdots + f_{n-1} g_0 \end{array} \right)$$

For example, the convolution of $(1, 2, 3)$ and $(4, 5, 6)$ is $(1, 2, 3) * (4, 5, 6) = (1 \cdot 4 + 2 \cdot 6 + 3 \cdot 5, 1 \cdot 5 + 2 \cdot 4 + 3 \cdot 6, 1 \cdot 6 + 2 \cdot 5 + 3 \cdot 4) = (31, 31, 28)$.

Now, we discuss property C. The convolution property applied to Gaussians is

$$\int_{-\infty}^{\infty} \varphi\left(x - \mu_1 \mid \sigma_1\right) \varphi\left(y - x - \mu_2 \mid \sigma_2\right) dx = \varphi\left(y - \mu \mid \sigma\right)$$

where $\varphi\left(\cdot\right)$ is a Gaussian density function, $\mu = \mu_1 + \mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. That is, two Gaussians convolve to make another Gaussian distribution with additive means and variances. For convenience let $w_i = \frac{1}{\sigma_i^2}$ and write

$$\left(\frac{x - \mu_1}{\sigma_1}\right)^2 + \left(\frac{y - x - \mu_2}{\sigma_2}\right)^2 = \left(w_1 + w_2\right)\left(x - \widehat{x}\right)^2 + \frac{w_1 w_2}{w_1 + w_2}\left(y - \mu_1 - \mu_2\right)^2$$

where $\widehat{x} \equiv \frac{w_1 \mu_1 + w_2 y - w_2 \mu_2}{w_1 + w_2}$. Integrating out x produces the above result.

3.7 Interval estimation

Finite sampling distribution theory for the linear model $Y = X\beta + \varepsilon$ follows from assigning the errors independent, normal probability distributions with mean zero and constant variance σ^2.[6] Interval estimates for individual model parameters β_j are Student t distributed with $n - p$ degrees of freedom when σ^2 is unknown. This follows from

$$\frac{b_j - \beta_j}{Est.Var\left[b_j\right]} \sim t\left(n - p\right)$$

where $Est.Var\left[b_j\right] = s^2\left(X_j^T M_{-j} X_j\right)^{-1}$, X_j is column j of X, $M_{-j} = I - P_{-j}$ and P_{-j} is the projection matrix onto all columns of X except j, $s^2 = \frac{e^T e}{n-p}$, and e is a vector of residuals. By *FWL*, the numerator is

$$\left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j} Y$$

Rewriting yields

$$b_j = \left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j}\left(X\beta + \varepsilon\right)$$

As M_{-j} annihilates all columns of X except X_j, we have

$$\begin{aligned} b_j &= \left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j}\left(X_j\beta_j + \varepsilon\right) \\ &= \beta_j + \left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j}\varepsilon \end{aligned}$$

Now,

$$b_j - \beta_j = \left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j}\varepsilon$$

As this is a linear combination of independent, normal random variates, the transformed random variable also has a normal distribution with mean zero and variance $\sigma^2\left(X_j^T M_{-j} X_j\right)^{-1}$. The estimated variance of b_j is $s^2\left(X_j^T M_{-j} X_j\right)^{-1}$ and the t ratio is

$$\begin{aligned} \frac{b_j - \beta_j}{Est.Var\left[b_j\right]} &= \frac{b_j - \beta_j}{\sqrt{s^2\left(X_j^T M_{-j} X_j\right)^{-1}}} \\ &= \frac{b_j - \beta_j}{\sqrt{\frac{e^T e\left(X_j^T M_{-j} X_j\right)^{-1}}{n-p}}} \end{aligned}$$

[6] It's instructive to recall the discussion of the ubiquity of the Gaussian distribution adapted from Jaynes [2003].

This can be rewritten as the ratio of a standard normal random variable to the square root of a chi-square random variable divided by its degrees of freedom.

$$\begin{aligned}\frac{b_j-\beta_j}{Est.Var\left[b_j\right]} &= \frac{\left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j}\varepsilon}{\sqrt{\frac{\varepsilon^T M_X \varepsilon \left(X_j^T M_{-j} X_j\right)^{-1}}{n-p}}} \\ &= \frac{\left(X_j^T M_{-j} X_j\right)^{-1} X_j^T M_{-j}\left(\varepsilon/\sigma\right)}{\sqrt{\frac{\left(\varepsilon/\sigma\right)^T M_X \left(\varepsilon/\sigma\right) \left(X_j^T M_{-j} X_j\right)^{-1}}{n-p}}}\end{aligned}$$

In other words, a Student t distributed random variable with $n-p$ degrees of freedom which completes the demonstration.

Normal sampling distribution theory applied to joint parameter regions follow an F distribution. For example, the null hypothesis $H_0 : \beta_1 = ... = \beta_{p-1} = 0$ is tested via the F statistic $= \frac{MSR}{MSE} \sim F\left(p-1, n-p\right)$. As we observed above, the denominator is

$$\begin{aligned}MSE &= \frac{e^T e}{n-p} \\ &= \frac{\varepsilon^T M_X \varepsilon}{n-p} \sim \frac{\chi^2\left(n-p\right)}{n-p}\end{aligned}$$

The numerator is $\frac{\left(Xb-\overline{Y}\right)^T\left(Xb-\overline{Y}\right)}{p-1}$. *FWL* indicates $Xb = P_\iota Y + M_\iota X_{-\iota} b_{-\iota}$ where ι refers to a vector of ones (for the intercept) and the subscript $-\iota$ refers to everything but the intercept (i.e., everything except the vector of ones in X). Therefore, $Xb - \overline{Y} = P_\iota Y + M_\iota X_{-\iota} b_{-\iota} - P_\iota Y = M_\iota X_{-\iota} b_{-\iota}$. Now,

$$\begin{aligned}MSR &= \frac{b_{-\iota}^T X_{-\iota}^T M_\iota X_{-\iota} b_{-\iota}}{p-1} \\ &= \frac{Y^T P_{M_\iota X_{-\iota}} Y}{p-1} \\ &= \frac{\left(X\beta+\varepsilon\right)^T P_{M_\iota X_{-\iota}} \left(X\beta+\varepsilon\right)}{p-1}\end{aligned}$$

under the null $\beta_{-\iota} = 0$ and β_0 is negated by M_ι. Hence,

$$MSR = \frac{\varepsilon^T P_{M_\iota X_{-\iota}} \varepsilon}{p-1} \sim \frac{\chi^2\left(p-1\right)}{p-1}$$

which completes the demonstration.

When the our understanding of the errors is weak, we frequently appeal to asymptotic or approximate sampling distributions. Asymptotic tests of restrictions are discussed next (also see the appendix).

3.8 Asymptotic tests of restrictions: Wald, *LM*, *LR* statistics

Tests of restrictions based on *Wald*, *LM* (Lagrange multiplier), and *LR* (likelihood ratio) statistics have a similar heritage. Asymptotically they are the same; only in finite samples do differences emerge. A brief sketch of each follows.

Intuition for these tests come from the *finite sample F statistic* (see Davidson and MacKinnon [1993], p. 82-83 and 452-6). The F statistic is valid if the errors have a normal probability assignment.

For the restriction H_0: $R\beta - r = 0$

$$\begin{aligned} F &= \frac{\left(e_*^T e_* - e^T e\right)/J}{e^T e/\left(n-p\right)} \\ &= \frac{\left(Rb-r\right)^T \left(Rs^2\left(X^T X\right)^{-1} R^T\right)^{-1}\left(Rb-r\right)}{J} \sim F\left(J, n-p\right) \end{aligned}$$

where R is $J \times p$, $e_* = \left(I - P_{X_*}\right)Y = M_{X_*}Y$ and $e = \left(I - P_X\right)Y = M_X Y$ are the residuals from the restricted and unrestricted models, respectively, $P_x = X\left(X^T X\right)^{-1} X^T$, $P_{X_*} = X_*\left(X_*^T X_*\right)^{-1} X_*^T$ are the projection matrices, X_* is the restricted matrix of regressors, and $s^2 = e^T e/\left(n-p\right)$ is the sample variance. Recall the numerator and denominator of F are divided by σ^2 to yield the ratio of two chi-squared random variables. Since s^2 converges to σ^2 we have $p\lim\left(\frac{s^2}{\sigma^2}\right) = 1$ in the denominator. Hence, we have J squared standard normal random variables summed in the numerator or W converges in distribution to $\chi^2\left(J\right)$.

FWL provides another way to see the connection between the F statistic and the Wald statistic W,

$$\begin{aligned} F &= \frac{\left(e_*^T e_* - e^T e\right)/J}{e^T e/\left(n-p\right)} \\ &= \frac{\left(Rb-r\right)^T \left(Rs^2\left(X^T X\right)^{-1} R^T\right)^{-1}\left(Rb-r\right)}{J} \\ &= \frac{W}{J} \end{aligned}$$

Consider the (partitioned) *DGP*:

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

and restriction $H_0 : \beta_2 = 0$. By *FWL*,[7]

$$e^T e = Y^T M_X Y = Y^T M_1 Y - Y^T M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y$$

and $e_*^T e_* = Y^T M_1 Y$. Hence, the numerator of F is

$$\left(e_*^T e_* - e^T e\right)/J = Y^T M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y/J$$

and the denominator is s^2. Since $R = \begin{bmatrix} 1 & 0 \end{bmatrix}$, rearrangement yields

$$W = \frac{b_2^T \left(\left(X^T X\right)^{-1}\right)_{22}^{-1} b_2/J}{s^2}$$

where

$$\left(\left(X^T X\right)^{-1}\right)_{22} = \left(X_2^T X_2 - X_2^T X_1 \left(X_1^T X_1\right)^{-1} X_1^T X_2\right)^{-1}$$

is the lower right hand block of $\left(X^T X\right)^{-1}$. Both F and W are divided by J in the numerator and have s^2 in the numerator. Now, we show that

$$Y^T M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y = b_2^T \left(\left(X^T X\right)^{-1}\right)_{22}^{-1} b_2$$

by rewriting the right hand side

$$\begin{aligned} b_2^T \left(\left(X^T X\right)^{-1}\right)_{22}^{-1} b_2 &= b_2^T \left(X_2^T X_2 - X_2^T X_1 \left(X_1^T X_1\right)^{-1} X_1^T X_2\right) b_2 \\ &= b_2^T \left(X_2^T M_1 X_2\right) b_2 \end{aligned}$$

by *FWL* for b_2

$$\begin{aligned} & b_2^T \left(X_2^T M_1 X_2\right) b_2 \\ = \; & Y^T M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} \left(X_2^T M_1 X_2\right) \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y \\ = \; & Y^T M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y \end{aligned}$$

[7] From *FWL*,

$$\begin{aligned} Xb &= P_1 Y + M_1 X_2 b_2 \\ &= P_1 Y + M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y \\ &= P_1 Y + M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y \\ &= P_1 Y + P_{M_1 X_2} Y \end{aligned}$$

Since $P_1 P_{M_1} = 0$ (by orthogonality),

$$\begin{aligned} e^T e &= Y^T \left(I - P_1 Y - P_{M_1 X_2}\right) Y \\ &= Y^T M_1 Y - Y^T P_{M_1 X_2} Y \\ &= Y^T M_1 Y - Y^T M_1 X_2 \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y \end{aligned}$$

This completes the demonstration as the right hand side from W is the same as the left hand side from F.

If the errors do not have a normal probability assignment, the F statistic is invalid (even asymptotically) but the *Wald statistic* may be asymptotically valid

$$W = (Rb - r)^T \left(Rs^2 \left(X^T X \right)^{-1} R^T \right)^{-1} (Rb - r) \xrightarrow{d} \chi^2 (J)$$

To see this apply the multivariate Lindberg-Feller version of the central limit theorem (see appendix on asymptotic theory) and recall if $x \sim N(\mu, \Sigma)$, then $(x - \mu)^T \Sigma^{-1} (x - \mu) \sim \chi^2 (n)$. W is the quadratic form as under the null Rb has mean r and $Est.Var[Rb] = Rs^2 \left(X^T X \right)^{-1} R^T$.

The *LR statistic* is based on the log of the ratio of the restricted to unrestricted likelihoods

$$\begin{aligned} LR &= -2 \left[LnL_* - LnL \right] \\ &= nLn \left[e_*^T e_* / e^T e \right] \xrightarrow{d} \chi^2 (J) \end{aligned}$$

Asymptotically LR converges to W.

The *Lagrange Multiplier* (*LM*) *test* is based on the gradient of the log-likelihood. If the restrictions are valid then the derivatives of the log-likelihood evaluated at the restricted estimates should be close to zero.

Following manipulation of first order conditions we find

$$\lambda_* = \left(Rs_*^2 \left(X^T X \right)^{-1} R^T \right)^{-1} (Rb - r)$$

A Wald test of $\lambda_* = 0$ yields the statistic $LM = \lambda_*^T \left\{ Est.Var[\lambda_*] \right\}^{-1} \lambda_*$ which simplifies to

$$LM = (Rb - r)^T \left(Rs_*^2 \left(X^T X \right)^{-1} R^T \right)^{-1} (Rb - r)$$

It is noteworthy that, unlike the *Wald* statistic above, the variance estimate is based on the restrictions.

In the classical regression model, the *LM* statistic can be simplified to an nR^2 test. Under the restrictions, $E\left[\frac{\partial Ln\, L}{\partial \beta} \right] = E\left[\frac{1}{\sigma^2} X^T \varepsilon \right] = 0$ and $Asy.Var\left[\frac{\partial Ln\, L}{\partial \beta} \right]$ $= \left[\frac{\partial^2 Ln\, L}{\partial \beta \partial \beta^T} \right]^{-1} = \sigma^2 \left(X^T X \right)^{-1}$. The *LM* statistic is

$$\frac{e_*^T X \left(X^T X \right)^{-1} X^T e_*}{e_*^T e_* / n} = nR_*^2 \xrightarrow{d} \chi^2 (p - J)$$

LM is n times R^2 from a regression of the (restricted) residuals e_* on the full set of regressors.

From the above, we have $W > LR > LM$ in finite samples.

3.8.1 Nonlinear restrictions

More generally, suppose the restriction is nonlinear in β

$$H_0 : f(\beta) = 0$$

The corresponding *Wald* statistic is

$$W = f(b)^T \left[G(b) s^2 \left(X^T X \right)^{-1} G(b)^T \right]^{-1} f(b) \xrightarrow{d} \chi^2(J)$$

where $G(b) = \left[\frac{\partial f(b)}{\partial b^T} \right]$. This is an application of the *Delta method* (see the appendix on asymptotic theory). If $f(b)$ involves continuous functions of b such that $\Gamma = \left[\frac{\partial f(\beta)}{\partial \beta^T} \right]$, by the central limit theorem

$$f(b) \xrightarrow{d} N\left(f(\beta), \Gamma \left(\frac{\sigma^2}{n} Q^{-1} \right) \Gamma^T \right)$$

where $p\lim \left(\frac{X^T X}{n} \right)^{-1} = Q^{-1}$.

3.9 Misspecification and *IV* estimation

Misspecification arises from violation of $E\left[X^T \varepsilon \right] = 0$, or $E[\varepsilon \mid X] = 0$, or asymptotically, $p\lim \left(\frac{1}{n} X^T \varepsilon \right) = 0$. Omitted correlated regressors, measurement error in regressors, and endogeneity (including simultaneity and self-selection) produce such misspecification when not addressed.

Consider the *DGP*:

$$Y = X_1 \beta + X_2 \gamma + \varepsilon$$

where

$$\varepsilon \sim \left(0, \sigma^2 I \right), E\left[\left[\begin{array}{cc} X_1 & X_2 \end{array} \right]^T \varepsilon \right] = 0$$

and

$$p\lim \left(\frac{1}{n} \left[\begin{array}{cc} X_1 & X_2 \end{array} \right]^T \varepsilon \right) = 0$$

If X_2 is omitted then it effectively becomes part of the error term, say $\eta = X_2\gamma + \varepsilon$. *OLS* yields

$$b = \left(X_1^T X_1 \right)^{-1} X_1^T \left(X_1 \beta + X_2 \gamma + \varepsilon \right) = \beta + \left(X_1^T X_1 \right)^{-1} X_1^T \left(X_2 \gamma + \varepsilon \right)$$

which is unbiased only if X_1 and X_2 are orthogonal (so the Gauss-Markov theorem likely doesn't apply). And, the estimator is asymptotically consistent only if $p\lim \left(\frac{1}{n} \left(X_1^T X_2 \right) \right) = 0$.

Instrumental variables (IV) estimation is a standard approach for addressing lack of independence between the regressors and the errors. A "good" set of instruments Z has two properties: (1) they are highly correlated with the (endogenous) regressors and (2) they are orthogonal to the errors (or $p\lim \left(\frac{1}{n} Z^T \varepsilon \right) = 0$).

Consider the *DGP*:

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim \left(0, \sigma^2 I\right)$, but $E\left[X^T \varepsilon\right] \neq 0$, and $p\lim\left(\frac{1}{n} X^T \varepsilon\right) \neq 0$.

IV estimation proceeds as follows. Regress X onto Z to yield $\widehat{X} = P_Z X = Z\left(Z^T Z\right)^{-1} Z^T X$. Estimate β via b_{IV} by regressing Y onto $\widehat{X}$.

$$\begin{aligned} b_{IV} &= \left(X^T P_Z P_Z X\right)^{-1} X^T P_Z Y \\ &= \left(X^T P_Z X\right)^{-1} X^T P_Z Y \end{aligned}$$

Asymptotic consistency[8] follows as

$$\begin{aligned} p\lim\left(b_{IV}\right) &= p\lim\left(\left(X^T P_Z X\right)^{-1} X^T P_Z Y\right) \\ &= p\lim\left(\left(X^T P_Z X\right)^{-1} X^T P_Z \left(X\beta + \varepsilon\right)\right) \\ &= \beta + p\lim\left(\left(X^T P_Z X\right)^{-1} X^T P_Z \varepsilon\right) \\ &= \beta + p\lim\left(\left(\frac{1}{n} X^T P_Z X\right)^{-1} 1/n X^T Z \left(\frac{1}{n} Z^T Z\right)^{-1} \frac{1}{n} Z^T \varepsilon\right) \\ &= \beta \end{aligned}$$

Note in the special case $Dim\left(Z\right) = Dim\left(X\right)$ (where Dim refers to the dimension or rank of the matrix), each regressor has one instrument associated with it, the instrumental variables estimator simplifies considerably as $\left(X^T Z\right)^{-1}$ and $\left(Z^T X\right)^{-1}$ exist. Hence,

$$\begin{aligned} b_{IV} &= \left(X^T P_Z X\right)^{-1} X^T P_z Y \\ &= \left(X^T Z \left(Z^T Z\right)^{-1} Z^T X\right)^{-1} X^T Z \left(Z^T Z\right)^{-1} Z^T Y \\ &= \left(Z^T X\right)^{-1} Z^T Y \end{aligned}$$

and

$$Asy.Var\left[b_{IV}\right] = \sigma^2 \left(Z^T X\right)^{-1} Z^T Z \left(X^T Z\right)^{-1}$$

There is a finite sample trade-off in choosing the number of instruments to employ. Asymptotic efficiency (inverse of variance) increases in the number of instruments but so does the finite-sample bias. Relatedly, if *OLS* is consistent the use of instruments inflates the variance of the estimates since $X^T P_Z X$ is smaller by a positive semidefinite matrix than $X^T X$ ($I = P_Z + (I - P_z)$, *IV* annihilates the left nullspace of Z).

[8] Slutsky's theorem is applied repeatedly below (see the appendix on asymptotic theory). The theorem indicates $plim\left(g(X)\right) = g(plim\left(X\right))$ and implies $plim\left(XY\right) = plim\left(X\right) plim\left(Y\right)$.

Importantly, if $Dim\left(Z\right) > Dim\left(X\right)$ then *over-identifying restrictions* can be used to test the instruments (Godfrey and Hutton, 1994). The procedure is regress the residuals from the second stage onto Z (all exogenous regressors). Provided there exists at least one exogenous regressor, then $nR^2 \sim \chi^2\left(K - L\right)$ where K is the number of exogenous regressors in the first stage and L is the number of endogenous regressors. Of course, under the null of exogenous instruments R^2 is near zero.

A *Hausman test* (based on a *Wald* statistic) can be applied to check the consistency of *OLS* (and is applied after the above exogeneity test and elimination of any offending instruments from the *IV* estimation).

$$W = \left(b - b_{IV}\right)^T \left[V_1 - V_0\right]^{-1} \left(b - b_{IV}\right) \sim \chi^2\left(p\right)$$

where V_1 is the estimated asymptotic covariance for the *IV* estimator and $V_0 = s^2\left(X^T X\right)^{-1}$ where s^2 is from the *IV* estimator (to ensure that $V_1 > V_0$).

3.10 Proxy variables

Frequently in accounting and business research we employ proxy variables as direct measures of constructs are not readily observable. Proxy variables can help to address potentially omitted, correlated variables. An important question is when do proxy variables aid the analysis and when is the cure worse than the disease.

Consider the *DGP*: $Y = \beta_0 + X\beta + Z\gamma + \varepsilon$. Let W be a set of proxy variables for Z (the omitted variables). Typically, there are two conditions to satisfy:
(1) $E\left[Y \mid X, Z, W\right] = E\left[Y \mid X, Z\right]$ This form of mean conditional independence is usually satisfied.
For example, suppose $W = Z + \nu$ and the variables are jointly normally distributed with ν independent of other variables. Then, the above condition is satisfied as follows. (For simplicity, we work with one-dimensional variables but the result

can be generalized to higher dimensions.[9])

$$\begin{aligned} E\left[Y \mid X,Z,W\right] &= \mu_Y + \left[\begin{array}{ccc} \sigma_{YX} & \sigma_{YZ} & \sigma_{YZ} \end{array}\right] \\ &\quad \times \left[\begin{array}{ccc} \sigma_{XX} & \sigma_{XZ} & \sigma_{XZ} \\ \sigma_{ZX} & \sigma_{ZZ} & \sigma_{ZZ} \\ \sigma_{ZX} & \sigma_{ZZ} & \sigma_{ZZ}+\sigma_{\nu\nu} \end{array}\right]^{-1} \left[\begin{array}{c} x-\mu_X \\ z-\mu_Z \\ w-\mu_W \end{array}\right] \\ &= \mu_Y + \frac{\sigma_{YX}\sigma_{ZZ}-\sigma_{XZ}\sigma_{YZ}}{\sigma_{XX}\sigma_{ZZ}-\sigma_{XZ}^2}\left(x-\mu_X\right) \\ &\quad + \frac{\sigma_{YZ}\sigma_{XX}-\sigma_{XZ}\sigma_{YX}}{\sigma_{XX}\sigma_{ZZ}-\sigma_{XZ}^2}\left(z-\mu_Z\right) + 0\left(w-\mu_W\right) \end{aligned}$$

(2) $Cov\left[X_j, Z \mid W\right] = 0$ for all j. This condition is more difficult to satisfy. Again, consider proxy variables like $W = Z+\nu$ where $E\left[\nu\right]=0$ and $Cov\left[Z,\nu\right] = 0$, then $Cov\left[X,Z \mid W\right] = \frac{\sigma_{XZ}\sigma_\nu^2}{\sigma_{ZZ}+\sigma_\nu^2}$. Hence, the smaller is σ_ν^2, the noise in the proxy variable, the better service provided by the proxy variable.

What is the impact of imperfect proxy variables on estimation? Consider proxy variables like $Z = \theta_0 + \theta_1 W + \nu$ where $E\left[\nu\right] = 0$ and $Cov\left[Z,\nu\right] = 0$. Let $Cov\left[X,\nu\right] = \rho \neq 0$, $Q = \left[\begin{array}{ccc} \iota & X & W \end{array}\right]$, and

$$\omega^T = \left[\begin{array}{ccc} \left(\beta_0+\gamma\theta_0\right) & \beta & \gamma\theta_1 \end{array}\right]$$

The estimable equation is

$$Y = Q\omega + \epsilon = \left(\beta_0+\gamma\theta_0\right) + \beta X + \gamma\theta_1 W + \left(\gamma\nu+\epsilon\right)$$

[9] A quick glimpse of the multivariate case can be found if we consider the simple case where the *DGP* omits X. If W doesn't contribute to $E[Y \mid Z,W]$, then it surely doesn't contribute to $E[Y \mid X,Z,W]$. It's readily apparent how the results generalize for the $E[Y \mid X,Z,W]$ case, though cumbersome. In block matrix form $E[Y \mid Z,W] =$

$$\begin{aligned} &\mu_Y + \left[\begin{array}{cc} \Sigma_{YZ} & \Sigma_{YZ} \end{array}\right] \left[\begin{array}{cc} \Sigma_{ZZ} & \Sigma_{ZZ} \\ \Sigma_{ZZ} & \Sigma_{ZZ}+\Sigma_{\nu\nu} \end{array}\right]^{-1} \left[\begin{array}{c} z-\mu_Z \\ w-\mu_W \end{array}\right] \\ =\ &\mu_Y + \left[\begin{array}{cc} \Sigma_{YZ} & \Sigma_{YZ} \end{array}\right] \left[\begin{array}{cc} \Sigma_{ZZ}^{-1}+\Sigma_{\nu\nu}^{-1} & -\Sigma_{\nu\nu}^{-1} \\ -\Sigma_{\nu\nu}^{-1} & \Sigma_{\nu\nu}^{-1} \end{array}\right] \left[\begin{array}{c} z-\mu_Z \\ w-\mu_W \end{array}\right] \\ =\ &\mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}\left(z-\mu_Z\right) + 0\left(w-\mu_W\right) \end{aligned}$$

The key is recognizing that the partitioned inverse (following some rewriting of the off-diagonal blocks) for

$$\begin{aligned} &\left[\begin{array}{cc} \Sigma_{ZZ} & \Sigma_{ZZ} \\ \Sigma_{ZZ} & \Sigma_{ZZ}+\Sigma_{\nu\nu} \end{array}\right]^{-1} \\ =\ &\left[\begin{array}{cc} \left[\Sigma_{ZZ}-\Sigma_{ZZ}\left(\Sigma_{ZZ}+\Sigma_{\nu\nu}\right)^{-1}\Sigma_{ZZ}\right]^{-1} & -\Sigma_{ZZ}^{-1}\Sigma_{ZZ}\Sigma_{\nu\nu}^{-1} \\ -\left(\Sigma_{ZZ}+\Sigma_{\nu\nu}\right)^{-1}\Sigma_{ZZ}\Sigma_{ZZ}^{-1}\left(\Sigma_{ZZ}+\Sigma_{\nu\nu}\right)\Sigma_{\nu\nu}^{-1} & \left[\Sigma_{ZZ}+\Sigma_{\nu\nu}-\Sigma_{ZZ}\Sigma_{ZZ}^{-1}\Sigma_{ZZ}\right]^{-1} \end{array}\right] \\ =\ &\left[\begin{array}{cc} \left[\Sigma_{ZZ}-\Sigma_{ZZ}\left(\Sigma_{ZZ}+\Sigma_{\nu\nu}\right)^{-1}\Sigma_{ZZ}\right]^{-1} & -\Sigma_{ZZ}^{-1}\Sigma_{ZZ}\Sigma_{\nu\nu}^{-1} \\ -\left(\Sigma_{ZZ}+\Sigma_{\nu\nu}\right)^{-1}\Sigma_{ZZ}\Sigma_{ZZ}^{-1}\left(\Sigma_{ZZ}+\Sigma_{\nu\nu}\right)\Sigma_{\nu\nu}^{-1} & \left[\Sigma_{ZZ}+\Sigma_{\nu\nu}-\Sigma_{ZZ}\Sigma_{ZZ}^{-1}\Sigma_{ZZ}\right]^{-1} \end{array}\right] \\ =\ &\left[\begin{array}{cc} \Sigma_{ZZ}^{-1}+\Sigma_{\nu\nu}^{-1} & -\Sigma_{\nu\nu}^{-1} \\ -\Sigma_{\nu\nu}^{-1} & \Sigma_{\nu\nu}^{-1} \end{array}\right] \end{aligned}$$

The *OLS* estimator of ω is $b = \left(Q^T Q\right)^{-1} Q^T Y$. Let $p\lim \left(1/nQ^T Q\right)^{-1} = \Omega$.

$$\begin{aligned} p\lim\, b &= \omega + \Omega\, p\lim\, 1/n \begin{bmatrix} \iota & X & W \end{bmatrix}^T (\gamma\nu + \epsilon) \\ &= \omega + \gamma\rho \begin{bmatrix} \Omega_{12} \\ \Omega_{22} \\ \Omega_{32} \end{bmatrix} = \begin{bmatrix} \beta_0 + \gamma\theta_0 + \Omega_{12}\gamma\rho \\ \beta + \Omega_{22}\gamma\rho \\ \gamma\theta_1 + \Omega_{32}\gamma\rho \end{bmatrix} \end{aligned}$$

Hence, b is asymptotically consistent when $\rho = 0$ and inconsistency ("bias") is increasing in the absolute value of $\rho = Cov\left[X, \nu\right]$.

3.10.1 Accounting and other information sources

Use of proxy variables in the study of information is even more delicate. Frequently we're interested in the information content of accounting in the midst of other information sources. As complementarity is the norm for information, we not only have the difficulty of identifying proxy variables for other information but also a functional form issue. Functional form is important as complementarity arises through joint information partitions. Failure to recognize these subtle interactions among information sources can yield spurious inferences regarding accounting information content.

A simple example (adapted from Antle, Demski, and Ryan[1994]) illustrates the idea. Suppose a nonaccounting information signal (x_1) precedes an accounting information signal (x_2). Both are informative of firm value (and possibly employ the language and algebra of valuation). The accounting signal however employs restricted recognition such that the nonaccounting signal is ignored by the accounting system. Table 3.1 identifies the joint probabilities associated with the information partitions and the firm's liquidating dividend (to be received at a future date and expressed in present value terms). Prior to any information reports, firm value (expected present value of the liquidating dividend) is 50. The change in firm value at the time of the accounting report (following the second signal) as well as the valuation-scaled signals (recall accounting, the second signal, ignores the first signal) are reported in table 3.2. Due to the strong complementarity in the information and restricted recognition employed by accounting, response to earnings is negative. That is, the change in value moves in the opposite direction of the accounting earnings report x_2.

Table 3.1: Multiple information sources case 1 setup

probabilities; payoffs		x_1		
		1	2	3
x_2	1	0.10;0	0.08;45	0.32;99
	2	0.32;1	0.08;55	0.10;100

As it is difficult to identify other information sources (and their information partitions), often a proxy variable for x_1 is employed. Suppose our proxy variable

Table 3.2: Multiple information sources case 1 valuation implications

change in firm value		x_1		
		-49.238	0	49.238
x_2	20.56	-0.762	-5	-0.238
	-20.56	0.238	5	0.762

is added as a control variable and a linear model of the change in firm value as a function of the information sources is estimated. Even if we stack things in favor of the linear model by choosing $w = x_1$ we find

$$\begin{gathered}\text{case 1: linear model with proxy variable}\\ E\left[Y \mid w, x_2\right] = 0. + 0.0153w - 0.070x_2\\ R^2 = 0.618\end{gathered}$$

While a saturated design matrix (an *ANOVA* with indicator variables associated with information partitions and interactions to capture potential complementarities between the signals) fully captures change in value

$$\begin{gathered}\text{case 1: saturated } \textit{ANOVA}\\ E\left[Y \mid D_{12}, D_{13}, D_{22}, D_{12}D_{22}, D_{13}D_{22}\right] = -0.762 - 4.238D_{12}\\ +0.524D_{13} + 1.0D_{22} + 0.524D_{13} + 1.0D_{22}\\ +9.0D_{12}D_{22} + 0.0D_{13}D_{22}\\ R^2 = 1.0\end{gathered}$$

where D_{ij} refers to information signal i and partition j, the linear model explains only slightly more than 60% of the variation in the response variable. Further, the linear model exaggerates responsiveness of firm value to earnings. This is a simple comparison of the estimated coefficient for γ (-0.070) compared with the mean effect scaled by reported earnings for the *ANOVA* design $(\frac{1.0}{-20.56} = -0.05)$. Even if w effectively partitions x_1, without accommodating potential informational complementarity (via interactions), the linear model is prone to misspecification.

$$\begin{gathered}\text{case 1: unsaturated } \textit{ANOVA}\\ E\left[Y \mid D_{12}, D_{13}, D_{22}\right] = -2.188 + 0.752D_{12} + 1.504D_{13} + 2.871D_{22}\\ R^2 = 0.618\end{gathered}$$

The estimated earnings response for the discretized linear proxy model is $\frac{2.871}{-20.56} = -0.14$. In this case (call it case 1), it is even more overstated.

Of course, the linear model doesn't always overstate earnings response, it can also understate (case 2, tables 3.3 and 3.4) or produce opposite earnings response to the *DGP* (case 3, tables 3.5 and 3.6). Also, utilizing the discretized or partitioned proxy may yield earnings response that is closer or departs more from the *DGP* than the valuation-scaled proxy for x_1. The estimated results for case 2 are

$$\begin{gathered}\text{case 2: linear model with proxy variable}\\ E\left[Y \mid w, x_2\right] = 0. + 0.453w + 3.837x_2\\ R^2 = 0.941\end{gathered}$$

Table 3.3: Multiple information sources case 2 setup

probabilities; payoffs		x_1		
		1	2	3
x_2	1	0.10;0	0.08;45	0.32;40
	2	0.32;60	0.08;55	0.10;100

Table 3.4: Multiple information sources case 2 valuation implications

change in firm value		x_1		
		-19.524	0.0	19.524
x_2	-4.400	-46.523	86.062	54.391
	4.400	61.000	-140.139	-94.107

case 2: saturated *ANOVA*

$$E\left[Y \mid D_{12}, D_{13}, D_{22}, D_{12}D_{22}, D_{13}D_{22}\right] = -30.476 + 25.476D_{12}$$
$$+20.952D_{13} + 40.0D_{22} - 30.0D_{12}D_{22} + 0.0D_{13}D_{22}$$
$$R^2 = 1.0$$

case 2: unsaturated *ANOVA*

$$E\left[Y \mid D_{12}, D_{13}, D_{22}\right] = -25.724 + 8.842D_{12} + 17.685D_{13} + 33.762D_{22}$$
$$R^2 = 0.941$$

Earnings response for the continuous proxy model is 3.837, for the partitioned proxy is $\frac{33.762}{4.4} = 7.673$, and for the *ANOVA* is $\frac{40.0}{4.4} = 9.091$. Hence, for case 2 the proxy variable models understate earnings response and the partitioned proxy is closer to the *DGP* earnings response than is the continuous proxy (unlike case 1).

For case 3,we have The estimated results for case 3 are

Table 3.5: Multiple information sources case 3 setup

probabilities; payoffs		x_1		
		1	2	3
x_2	1	0.10;4.802	0.08;105.927	0.32;50.299
	2	0.32;65.864	0.08;26.85	0.10;17.254

case 3: linear model with proxy variable

$$E\left[Y \mid w, x_2\right] = 0. + 0.063w + 1.766x_2$$
$$R^2 = 0.007$$

case 3: saturated *ANOVA*

$$E\left[Y \mid D_{12}, D_{13}, D_{22}, D_{12}D_{22}, D_{13}D_{22}\right] = -46.523 + 86.062D_{12}$$
$$+54.391D_{13} + 61.062D_{22} - 140.139D_{12}D_{22} - 94.107D_{13}D_{22}$$
$$R^2 = 1.0$$

Table 3.6: Multiple information sources case 3 valuation implications

change in firm value		x_1		
		1.326	16.389	-7.569
x_2	0.100	-46.523	86.062	54.391
	-0.100	61.000	-140.139	-94.107

case 3: unsaturated *ANOVA*

$$E\left[Y \mid D_{12}, D_{13}, D_{22}\right] = 4.073 - 1.400D_{12} - 2.800D_{13} - 5.346D_{22}$$

$$R^2 = 0.009$$

Earnings response for the continuous proxy model is 1.766, for the partitioned proxy is $\frac{-5.346}{-0.100} = 53.373$, and for the *ANOVA* is $\frac{61.062}{-0.100} = -609.645$. Hence, for case 3 the proxy variable models yield earnings response opposite the *DGP*.

The above variety of misspecifications suggests that econometric analysis of information calls for nonlinear models. Various options may provide adequate summaries of complementary information sources. These choices include at least saturated *ANOVA* designs (when partitions are identifiable), polynomial regressions, and nonparametric and semiparametric regressions. Of course, the proxy variable problem still lurks. Next, we return to the equilibrium earnings management example discussed in chapter 2 and explore the (perhaps linear) relation between firm value and accounting accruals.

3.11 Equilibrium earnings management

The earnings management example in Demski [2004] provides a straightforward illustration of the econometric challenges faced when management's reporting behavior is endogenous and also the utility of the propensity score as an instrument. Suppose the objective is to track the relation between a firm's value P_t and its accruals z_t. To keep things simple, firm value equals the present value of expected future dividends, the market interest rate is zero, current period cash flows are fully paid out in dividends, and dividends $\widetilde{d}$ are normal *iid* with mean zero and variance σ^2. Firm managers have private information $\widetilde{y}_t^p$ about next period's dividend $\widetilde{y}_t^p = \widetilde{d}_{t+1} + \widetilde{\varepsilon}_t$ where $\widetilde{\varepsilon}$ are normal *iid* with mean zero and variance σ^2.[10] If the private information is revealed, ex dividend firm value at time t is

$$\begin{aligned} P_t &\equiv E\left[\widetilde{d}_{t+1} \mid \widetilde{y}_t^p = y_t^p\right] \\ &= \frac{1}{2} y_t^p \end{aligned}$$

Suppose management reveals its private information through income I_t (cash flows plus change in accruals) where fair value accruals $z_t = E\left[\widetilde{d}_{t+1} \mid \widetilde{y}_t^p = y_t^p\right]$

[10] For simplicity, there is no other information.

$= \frac{1}{2}y_t^p$ are reported. Then, income is

$$\begin{aligned} I_t &= d_t + (z_t - z_{t-1}) \\ &= d_t + \frac{1}{2}\left(y_t^p - y_{t-1}^p\right) \end{aligned}$$

and

$$\begin{aligned} P_t &\equiv E\left[\widetilde{d}_{t+1} \mid \widetilde{d}_t = d_t, I_t = d_t + \frac{1}{2}\left(y_t^p - y_{t-1}^p\right)\right] \\ &= E\left[\widetilde{d}_{t+1} \mid \widetilde{z}_t = \frac{1}{2}y_t^p\right] \\ &= z_t \end{aligned}$$

There is a linear relation between price and fair value accruals.

Suppose the firm is owned and managed by an entrepreneur who, for intergenerational reasons, liquidates his holdings at the end of the period. The entrepreneur is able to misrepresent the fair value estimate by reporting, $z_t = \frac{1}{2}y_t^p + \theta$, where $\theta \geq 0$. Auditors are unable to detect any accrual overstatements below a threshold equal to $\frac{1}{2}\Delta$. Traders anticipate the firm reports $z_t = \frac{1}{2}y_t^p + \frac{1}{2}\Delta$ and the market price is

$$P_t = z_t - E[\theta] = z_t - \frac{1}{2}\Delta$$

Given this anticipated behavior, the entrepreneur's equilibrium behavior is to report as conjectured. Again, there is a linear relationship between firm value and reported "fair value" accruals.

Now, consider the case where the entrepreneur can misreport but with probability α. Investors process the entrepreneur's report with misreporting in mind. The probability of misreporting D, given an accrual report of z_t, is

$$\Pr(D \mid \widetilde{z}_t = z_t) = \frac{\alpha\phi\left(\frac{z_t - 0.5\Delta}{\sqrt{0.5}\sigma}\right)}{\alpha\phi\left(\frac{z_t - 0.5\Delta}{\sqrt{0.5}\sigma}\right) + (1-\alpha)\phi\left(\frac{z_t}{\sqrt{0.5}\sigma}\right)}$$

where $\phi(\cdot)$ is the standard normal density function. In turn, the equilibrium price for the firm following the report is

$$\begin{aligned} P_t &= E\left[\widetilde{d}_{t+1} \mid \widetilde{z}_t = z_t\right] \\ &= \frac{\alpha(z_t - 0.5\Delta)\phi\left(\frac{z_t - 0.5\Delta}{\sqrt{0.5}\sigma}\right) + (1-\alpha)z_t\phi\left(\frac{z_t}{\sqrt{0.5}\sigma}\right)}{\alpha\phi\left(\frac{z_t - 0.5\Delta}{\sqrt{0.5}\sigma}\right) + (1-\alpha)\phi\left(\frac{z_t}{\sqrt{0.5}\sigma}\right)} \end{aligned}$$

Again, the entrepreneur's equilibrium reporting strategy is to misreport the maximum whenever possible and the accruals balance is $\alpha\left(\frac{1}{2}\Delta\right)$, on average. Price is no longer a linear function of reported "fair value".

Consider the following simulation to illustrate. Let $\sigma^2 = 2, \Delta = 4$, and $\alpha = \frac{1}{4}$. For sample size $n = 5,000$ and $1,000$ simulated samples, the regression is

$$P_t = \beta_0 + \beta_1 x_t$$

where

$$x_t = D_t \left(z_t^p + \frac{1}{2}\Delta \right) + (1 - D_t)\, z_t^p$$

$$D_t \sim Bernoulli\,(\alpha)$$

$$P_t = \frac{\alpha\,(x_t - 0.5\Delta)\,\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5}\sigma}\right) + (1-\alpha)\,x_t\phi\left(\frac{x_t}{\sqrt{0.5}\sigma}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5}\sigma}\right) + (1-\alpha)\,\phi\left(\frac{x_t}{\sqrt{0.5}\sigma}\right)}$$

and $z_t^p = \frac{1}{2}y_t^p$. A typical plot of the sampled data, price versus reported accruals is depicted in figure 3.1. There is a distinctly nonlinear pattern in the data.[11]

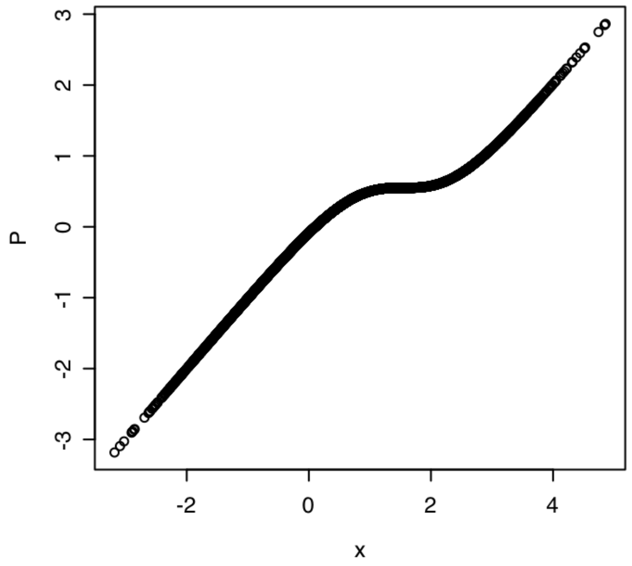

Figure 3.1: Price versus reported accruals

Sample statistics for the regression estimates are reported in table 3.7. The estimates of the slope are substantially biased downward. Recall the slope is one if there is no misreporting or if there is known misreporting. Suppose the analyst

[11] For larger (smaller) values of Δ, the nonlinearity is more (less) pronounced.

Table 3.7: Results for price on reported accruals regression

statistic	β_0	β_1
mean	-0.285	0.571
median	-0.285	0.571
standard deviation	0.00405	0.00379
minimum	-0.299	0.557
maximum	-0.269	0.584
$E[P_t \mid x_t] = \beta_0 + \beta_1 x_t$		

can ex post determine whether the firm misreported. Let $D_t = 1$ if the firm misreported in period t and 0 otherwise. Is price a linear function of reported accruals x_t conditional on D_t? Simulation results for the saturated regression

$$P_t = \beta_0 + \beta_1 x_t + \beta_2 D_t + \beta_3 x_t \times D_t$$

are reported in table 3.8. Perhaps surprisingly, the slope coefficient continues to be biased toward zero.

Table 3.8: Results for price on reported accruals saturated regression

statistic	β_0	β_1	β_2	β_3
mean	-0.244	0.701	0.117	-0.271
median	-0.244	0.701	0.117	-0.271
standard deviation	0.0032	0.0062	0.017	0.011
minimum	-0.255	0.680	0.061	-0.306
maximum	-0.233	0.720	0.170	-0.239
$E[P_t \mid x_t, D_t] = \beta_0 + \beta_1 x_t + \beta_2 D_t + \beta_3 x_t \times D_t$				

Before we abandon hope for our econometric experiment, it is important to remember investors do not observe D_t but rather are left to infer any manipulation from reported accruals x_t. So what then is the omitted, correlated variable in this earnings management setting? Rather than D_t it's the propensity for misreporting inferred from the accruals report, in other words $\Pr(D_t \mid \widetilde{x}_t = x_t) \equiv p(x_t)$. If the analyst knows what traders know, that is α, Δ, and σ, along with the observed report, then the regression for estimating the relation between price and fair value is

$$P_t = \beta_0 + \beta_1 x_t + \beta_2 p(x_t)$$

Simulation results are reported in table 3.9. Of course, this regression perfectly

Table 3.9: Results for price on reported accruals and propensity score regression

statistic	β_0	β_1	β_2
mean	0.000	1.000	-2.000
median	0.000	1.000	-2.000
standard deviation	0.000	0.000	0.000
minimum	0.000	1.000	-2.000
maximum	0.000	1.000	-2.000
$E\left[P_t \mid x_t, p\left(x_t\right)\right] = \beta_0 + \beta_1 x_t + \beta_2 p\left(x_t\right)$			

fits the data as a little manipulation confirms.

$$\begin{aligned}
P_t &= \frac{\alpha\left(x_t - 0.5\Delta\right)\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)x_t\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)} \\
&= \beta_0 + \beta_1 x_t + \beta_2 p\left(x_t\right) \\
&= \beta_0 + \beta_1 x_t + \beta_2 \frac{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)} \\
&= \frac{\left(\beta_0 + \beta_1 x_t\right)\left[\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)\right] + \beta_2\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)}
\end{aligned}$$

For $\beta_1 = 1, P_t = \frac{\beta_1 x_t\left[\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + (1-\alpha)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)\right] - \beta_1\alpha 0.5\Delta\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + (1-\alpha)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)}$. Hence, $\beta_0 = 0$ and the above expression simplifies

$$\begin{aligned}
&\frac{\left(\beta_0 + \beta_1 x_t\right)\left[\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)\right] + \beta_2\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)} \\
&= \frac{\beta_1\left[\alpha\left(x_t - 0.5\Delta\right)\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)x_t\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)\right]}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)} \\
&\quad + \frac{\left(\beta_1 0.5\Delta + \beta_2\right)\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right)}{\alpha\phi\left(\frac{x_t - 0.5\Delta}{\sqrt{0.5\sigma}}\right) + \left(1-\alpha\right)\phi\left(\frac{x_t}{\sqrt{0.5\sigma}}\right)}
\end{aligned}$$

Since the last term in the numerator must be zero and $\beta_1 = 1, \beta_2 = -\beta_1 0.5\Delta = -0.5\Delta$. In other words, reported accruals conditional on trader's perceptions of the propensity for misreporting map perfectly into price. The regression estimates the relation between price and fair value via β_1 and the magnitude of misreporting when the opportunity arises via β_2.

Of course, frequently the analyst (social scientist) suffers an informational disadvantage. Suppose the analyst ex post observes D_t (an information advantage relative to traders) but doesn't know α, Δ, and σ (an information disadvantage relative to traders). These parameters must be estimated from the data. An estimate of α is

$$\overline{D} = n^{-1} \sum_{t=1}^{n} D_t$$

An estimate of $\theta = \frac{1}{2}\Delta$ is

$$\widehat{\theta} = \frac{n^{-1} \sum_{t=1}^{n} x_t D_t}{\overline{D}} - \frac{n^{-1} \sum_{t=1}^{n} x_t \left(1 - D_t\right)}{\left(1 - \overline{D}\right)}$$

An estimate of $\nu^2 = \frac{1}{2}\sigma^2$ is

$$\widehat{\nu}^2 = \left(n-1\right)^{-1} \sum_{t=1}^{n} \left(x_t - \overline{x}\right)^2 - \widehat{\theta}^2 \overline{D} \left(1 - \overline{D}\right)$$

Combining the above estimates[12] produces an estimate of $p\left(x_t\right)$

$$\widehat{p}\left(x_t\right) = \frac{\overline{D}\phi\left(\frac{x_t - \widehat{\theta}}{\widehat{\nu}}\right)}{\overline{D}\phi\left(\frac{x_t - \widehat{\theta}}{\widehat{\nu}}\right) + \left(1 - \overline{D}\right)\phi\left(\frac{x_t}{\widehat{\nu}}\right)}$$

And the regression now is

$$P_t = \beta_0 + \beta_1 x_t + \beta_2 \widehat{p}\left(x_t\right)$$

Simulation results reported in table 3.10 support the estimated propensity score $\widehat{p}\left(x_t\right)$.

Table 3.10: Results for price on reported accruals and estimated propensity score regression

statistic	β_0	β_1	β_2
mean	0.0001	0.9999	-2.0006
median	-0.0000	0.9998	-2.0002
standard deviation	0.0083	0.0057	0.0314
minimum	-0.025	0.981	-2.104
maximum	0.030	1.019	-1.906
$E\left[P_t \mid x_t, \widehat{p}\left(x_t\right)\right] = \beta_0 + \beta_1 x_t + \beta_2 \widehat{p}\left(x_t\right)$			

[12] If D_t is unobservable to the analyst then some other means of estimating $p(x_t)$ is needed (perhaps initial guesses for α and Δ followed by nonlinear refinement).

Rather than $\widehat{p}(x_t)$, the propensity score can be estimated via logit, $\widetilde{p}(x_t)$, (discussed in chapter 5) where D_t is regressed on x_t.[13] As expected, simulation results reported in table 3.11 are nearly identical to those reported above (the correlation between the two propensity score metrics is 0.999).

Table 3.11: Results for price on reported accruals and logit-estimated propensity score regression

statistic	β_0	β_1	β_2
mean	-0.000	1.000	-1.999
median	-0.000	1.000	-1.997
standard deviation	0.012	0.008	0.049
minimum	-0.035	0.974	-2.154
maximum	0.040	1.028	-1.863
$E[P_t \mid x_t, \widetilde{p}(x_t)] = \beta_0 + \beta_1 x_t + \beta_2 \widetilde{p}(x_t)$			

This stylized equilibrium earnings management example illustrates two points. First, it provides a setting in which the intuition behind the propensity score, a common econometric instrument, is clear. Second, it reinforces our theme concerning the importance of the union of theory, data, and model specification. Consistent analysis requires all three be carefully attended and the manner in which each is considered depends on the others.

3.12 Additional reading

Linear models have been extensively studied and accordingly there are many nice econometrics references. Some favorites include Davidson and MacKinnon [1993, 2003], Wooldridge [2002], Cameron and Trivedi [2005], Greene [1997], Amemiya [1985], Theil [1971], Rao [1973], and Graybill [1976]. Angrist and Pischke [2009] provide a provocative justification for the linear conditional expectation function (see the end of chapter appendix). Davidson and MacKinnon in particular offer excellent discussions of *FWL*. Bound, Brown, and Mathiowetz [2001] and Hausman [2001] provide extensive review of classical and nonclassical measurement error and their implications for proxy variables. Christensen and Demski [2003, ch. 9-10] provide a wealth of examples of accounting as an information source and the subtleties of multiple information sources. Their discussion of the correspondence (or lack thereof) between accounting metrics and firm value suggests that association studies are no less prone to challenging specification issues than are information content studies. Discussions in this chapter

[13] The posterior probability of manipulation given a normally distributed signal has a logistic distribution (see Kiefer [1980]). Probit results are very similar although the logit intervals are somewhat narrower. Of course, if D_t is unobservable (by the analyst) then discrete choice methods like logit or probit are not directly accessible.

refer specifically to information content. Finally, we reiterate Jayne's [2003] discussion regarding the ubiquity of the Gaussian distribution is provocative.

3.13 Appendix

Angrist and Pischke [2009, ch. 3] layout a foundation justifying regression analysis of economic data and building linkages to causal effects. The arguments begin with the population-level conditional expectation function (*CEF*)

$$E\left[Y_i \mid X_i = x\right] = \int t f_y \left(t \mid X_i = x\right) dt$$

where $f_y \left(t \mid X_i = x\right)$ is the conditional density function evaluated at $Y_i = t$ and the law of iterated expectations

$$E\left[Y_i\right] = E_X \left[E\left[Y_i \mid X_i\right]\right]$$

The law of iterated expectations allows us to separate the response variable into two components: the *CEF* and a residual.

Theorem 3.2 *CEF decomposition theorem.*

$$Y_i = E\left[Y_i \mid X_i\right] + \varepsilon_i$$

where (i) ε_i is mean independent of X_i, $E\left[\varepsilon_i \mid X_i\right] = 0$, and (ii) ε_i is uncorrelated with any function of X_i.

Proof. (i)

$$\begin{aligned} E\left[\varepsilon_i \mid X_i\right] &= E\left[Y_i - E\left[Y_i \mid X_i\right] \mid X_i\right] \\ &= E\left[Y_i \mid X_i\right] - E\left[Y_i \mid X_i\right] = 0 \end{aligned}$$

(ii) let $h\left(X_i\right)$ be some function of X_i. By the law of iterated expectations,

$$E\left[h\left(X_i\right)\varepsilon_i\right] = E_X \left[h\left(X_i\right) E\left[\varepsilon_i \mid X_i\right]\right]$$

and by mean independence $E\left[\varepsilon_i \mid X_i\right] = 0$. Hence, $E\left[h\left(X_i\right)\varepsilon_i\right] = 0$. ■

The *CEF* optimally summarizes the relation between the response, Y_i, and explanatory variables, X_i, in a minimum mean square error (*MMSE*) sense.

Theorem 3.3 *CEF prediction theorem. Let $m\left(X_i\right)$ be any function of X_i. The CEF is the MMSE of Y_i given X_i in that it solves*

$$E\left[Y_i \mid X_i\right] = \underset{m(X_i)}{\arg\min} E\left[\left\{Y_i - m\left(X_i\right)\right\}^2\right]$$

Proof. Write

$$\begin{aligned}\{Y_i - m(X_i)\}^2 &= \{(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - m(X_i))\}^2 \\ &= (Y_i - E[Y_i \mid X_i])^2 + 2(Y_i - E[Y_i \mid X_i]) \\ &\quad \times (E[Y_i \mid X_i] - m(X_i)) + (E[Y_i \mid X_i] - m(X_i))^2\end{aligned}$$

The first term can be ignored as it does not involve $m(X_i)$. By the *CEF* decomposition property, the second term is zero since we can think of $h(X_i) \equiv 2(Y_i - E[Y_i \mid X_i])$. Finally, the third term is minimized when $m(X_i)$ is the *CEF*.
■

A closely related property involves decomposition of the variance. This property leads to the *ANOVA* table associated with many standard statistical analyses.

Theorem 3.4 *ANOVA theorem.*

$$Var[Y_i] = Var[E[Y_i \mid X_i]] + E_X[Var[Y_i \mid X_i]]$$

where $Var[\cdot]$ is the variance operator.

Proof. The *CEF* decomposition property implies the variance of Y_i equals the variance of the *CEF* plus the variance of the residual as the terms are uncorrelated.

$$Var[Y_i] = Var[E[Y_i \mid X_i]] + Var[\varepsilon_i \mid X_i]$$

Since $\varepsilon_i \equiv Y_i - E[Y_i \mid X_i]$ and $Var[\varepsilon_i \mid X_i] = Var[Y_i \mid X_i] = E[\varepsilon_i^2]$, by iterated expectations

$$\begin{aligned}E[\varepsilon_i^2] &= E_X[E[\varepsilon_i^2 \mid X_i]] \\ &= E_X[Var[Y_i \mid X_i]]\end{aligned}$$

■

This background sets the stage for three linear regression justifications. Regression justification *I* is the linear *CEF* theorem which applies, for instance, when the data are jointly normally distributed (Galton [1886]).

Theorem 3.5 *Linear CEF theorem (regression justification I). Suppose the CEF is linear.*

$$E[Y_i \mid X_i] = X_i^T \beta$$

where

$$\begin{aligned}\beta &= \arg\min_b E\left[\left(Y_i - X_i^T b\right)^2\right] \\ &= E\left[\left(X_i X_i^T\right)^{-1}\right] E[X_i Y_i]\end{aligned}$$

Then the population regression function is linear.

Proof. Suppose $E[Y_i \mid X_i] = X_i^T \beta^*$ for some parameter vector β^*. By the *CEF* decomposition theorem,

$$E[X_i (Y_i - E[Y_i \mid X_i]) \mid X_i] = 0$$

Substitution yields

$$E\left[X_i \left(Y_i - X_i^T \beta^*\right) \mid X_i\right] = 0$$

Iterated expectations implies

$$E\left[X_i \left(Y_i - X_i^T \beta^*\right)\right] = 0$$

Rearrangement gives

$$\beta^* = E\left[\left(X_i X_i^T\right)^{-1}\right] E[X_i Y_i] = \beta$$

■

Now, we explore approximate results associated with linear regression. First, we state the best linear predictor theorem (regression justification *II*). Then, we describe a linear approximation predictor result (regression justification *III*).

Theorem 3.6 *Best linear predictor theorem (regression justification II). The function $X_i^T \beta$ is the best linear predictor of Y_i given X_i in a MMSE sense.*

Proof. $\beta = E\left[\left(X_i X_i^T\right)^{-1}\right] E[X_i Y_i]$ is the solution to the population least squares problem as demonstrated in the proof to the linear CEF theorem. ■

Theorem 3.7 *Regression CEF theorem (regression justification III). The function $X_i^T \beta$ provides the MMSE linear approximation to $E[Y_i \mid X_i]$. That is,*

$$\beta = \arg\min_b E\left[\left(E[Y_i \mid X_i] - X_i^T b\right)^2\right]$$

Proof. Recall β solves $\arg\min_b E\left[\left(Y_i - X_i^T b\right)^2\right]$. Write

$$\begin{aligned}
\left(Y_i - X_i^T b\right)^2 &= \left\{(Y_i - E[Y_i \mid X_i]) + \left(E[Y_i \mid X_i] - X_i^T b\right)\right\}^2 \\
&= (Y_i - E[Y_i \mid X_i])^2 + \left(E[Y_i \mid X_i] - X_i^T b\right)^2 \\
&\quad +2(Y_i - E[Y_i \mid X_i])\left(E[Y_i \mid X_i] - X_i^T b\right)
\end{aligned}$$

The first term does not involve b and the last term has expected value equal to zero by the *CEF* decomposition theorem. Hence, the *CEF* approximation problem is the same as the population least squares problem (regression justification *II*). ■