Linear Models

Robert A. Miller

Structural Econometrics

November 2023

- The linear model is defined by the equation:

$$y_n = x_n' \beta_0 + \epsilon_n \tag{1}$$

  where $n \in \{1, 2, \ldots\}$ belongs to a population and:
  - y_n is a 1×1 observed dependent variable
  - x_n is a $k \times 1$ vector of observed explanatory variables
  - β_0 is a $k \times 1$ unknown parameter to be estimated
  - ϵ_n is a 1×1 unobserved idiosyncratic variable.
- The goal is to estimate β_0 from a sample $\{y_n, x_n\}_{n=1}^{N}$ of size N.
- There are essentially three reasons why the linear model has become the workhorse in econometrics:
  1. the model is easy to understand
  2. the estimator for the unknown coefficient is easy to compute
  3. the finite sample properties of the estimator are known
- To preface nonlinear estimation this lecture reviews the linear model.

- To illustrate one application of the linear model consider a *differences in differences* (DID) framework.
- Here the goal is to decontaminate the effects of a changing a regime, or more generally the effect of a particular factor of interest, from other extraneous factors, such as a time trend.
- We might write:

$$y_n = \beta_{00} + \beta_{01} t_n + \beta_{02} x_n + \beta_{03} x_n t_n + \epsilon_n$$

where $\beta_0 \equiv (\beta_{00}, \beta_{01}, \beta_{02}, \beta_{03})$ and $(x_n, t_n) \in \{0, 1\} \times \{0, 1\}$.
- Intuitively there are N observations, some of which are sampled in the first period, the others in the second, where a proportion are treated with a factor of interest (setting $x_n = 1$) and a proportion are left untreated (setting $x_n = 0$).
- This model is *saturated* because there are as many coefficients to be estimated as there are different combinations of (x, t).

- A second example is the *regression discontinuity design* (RDD) framework.
- Similar in some ways to DID, we seek to separate the effects of a changing a regime from other nonlinear effects that a particular explanatory variable might have on the dependent variable.
- For example let:

$$y_n = \beta_{00} + \beta_{0K} 1 \{x_n \leq c\} + \sum_{k=1}^{K-1} \beta_{0k} x_n^k + \epsilon_n$$

where $\beta_0 \equiv (\beta_{00}, \beta_{01}, \ldots, \beta_{0K})$ and $c \in \mathbb{R}$ is a cut-off value that might be crucial to determining how x affects y.

- This framework is used to flexibly model known discontinuities within an otherwise smooth nonlinear equation.

- Models of *fixed effects* (FE) arise when there are multiple observations on each individual $n \in \{1, 2, \ldots, N\}$, perhaps because they are sampled over time $t \in \{1, 2, \ldots, T\}$.
- Alternatively there might be several measurements of dependent variable, each of which is measured with error.
- We extend the notation for characterizing the data by writing:

$$y_{nt} = x'_{nt}\beta_0 + \gamma_n + \epsilon_{nt} \tag{2}$$

  where:

  - y_{nt} is a 1×1 observed dependent variable
  - x_{nt} is a $k \times 1$ vector of observed explanatory variables
  - β_0 is a $k \times 1$ unknown parameter to be estimated
  - γ_n is a $k \times 1$ unknown ancillary (or nuisance) parameter
  - ϵ_{nt} is a 1×1 unobserved idiosyncratic variable.
- We estimate β_0 from *panel data* $\{y_{nt}, x_{nt}\}_{n=1}^{NT}$ with NT observations.

- The *ordinary least squares* (OLS) estimator of β_0 is defined as:

$$\begin{aligned} \beta_{OLS}^{(N)} &\equiv \arg\min_{\beta} \left\{ \sum_{n=1}^{N} \left(y_n - x_n'\beta \right)^2 \right\} \\ &= \arg\min_{\beta} \sum_{n=1}^{N} \left[y_n^2 - 2\beta' x_n y_n + \left(x_n'\beta \right) \left(x_n'\beta \right) \right] \end{aligned}$$

- The $k \times 1$ *first order condition* (FOC) for this problem is:

$$0 = -2 \sum_{n=1}^{N} x_n y_n + 2 \sum_{n=1}^{N} x_n x_n' \beta_{OLS}^{(N)}$$

- If the $k \times k$ matrix $\frac{1}{N} \sum_{n=1}^{N} x_n x_n'$ has a nonzero determinant, then it is *invertible* and:

$$\beta_{OLS}^{(N)} = \left(\frac{1}{N} \sum_{n=1}^{N} x_n x_n' \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} x_n y_n \right) \quad (3)$$

- If $\frac{1}{N} \sum_{n=1}^{N} x_n x_n'$ is not invertible then the solution to this quadratic minimization problem is not unique.

Linear Projections

Metrics for approximations

- Let $F_{y,x}(y,x)$ denote the joint distribution function of (y,x) for the population, or data generating process.
- Also define an L_p space of real valued functions of (y,x), with elements $h(y,x) \in L_p$, by the condition:

$$\int |h(y,x)|^p \, dF_{y,x}(y,x) < \infty$$

equipped with norm:

$$\|h(y,x)\|_{L_p} \equiv \left[\int |h(y,x)|^p \, dF_{y,x}(y,x) \right]^{\frac{1}{p}}$$

- Given an L_p space define the *linear projection* of y on to x as:

$$\beta_{\|\cdot\|_{L_p}} \equiv \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\| y - x'\beta \right\|_{L_p} = \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\{ E\left[\left| y - x'\beta \right|^p \right] \right\} \tag{4}$$

- Thus $\beta_{\|\cdot\|_{L_p}}$ defines how closely a linear function of x is to central tendencies of the conditional distribution $F_{y|x}(y|x)$.

- If $p = 2$, then $\beta_{\|\cdot\|_{L_p}}$ becomes:

$$\beta_{OLS} \equiv \arg\min E\left[\left(y - x'\beta\right)^2\right] = \arg\min E\left[-2\beta' xy + \left(x'\beta\right)^2\right]$$

  with the FOC reducing to:

$$E\left[yx'\right] = E\left[xx'\right]\beta_{OLS}$$

- If $E\left[x_n x_n'\right]$ is invertible then:

$$\beta_{OLS} = E\left[xx'\right]^{-1} E\left[xy\right]$$

- In this case $\beta_{OLS}^{(N)}$ is the *sample analogue* of $\widehat{\beta}$, is found by replacing:

$$E\left[xx'\right] \text{ with } \left(\frac{1}{N}\sum\nolimits_{n=1}^{N} x_n x_n'\right) \text{ and } E\left[xy\right] \text{ with } \left(\frac{1}{N}\sum\nolimits_{n=1}^{N} x_n y_n\right)$$

- Using a different norm changes the solution to the linear projection.
- For example if $\|z\| \equiv E\left[|z|\right]$ then (4) reduces to:

$$\beta_{LAD} \equiv \arg\min_{\beta} E\left[\left|y - x'\beta\right|\right] = \arg\min_{\beta} E\left[\max\left\{y - x'\beta, x'\beta - y\right\}\right]$$

- The sample analogue of β_{LAD}, called the *least absolute deviations* (LAD) estimator, minimizes:

$$\frac{1}{N}\sum\nolimits_{n=1}^{N}\left|y_n - x_n'\beta\right| \tag{5}$$

- Note (5) is not differentiable with respect to β wherever $y_n = x_n'\beta$.
- Nevertheless $\widehat{\beta}_{LAD}^{(N)}$ is the solution to the linear program:

$$\begin{aligned} \widehat{\beta}_{LAD}^{(N)} &\equiv \arg\min_{\beta, u_1, \ldots, u_N} \frac{1}{N}\sum\nolimits_{n=1}^{N} u_n \\ \text{such that } u_n &\geq y_n - x_n'\beta \text{ and } u_n \geq x_n'\beta - y_n \end{aligned}$$

Quantile Estimators

Rationale and definition

- The LAD estimator is an example of a *quantile estimator*.
- For any $\tau \in (0,1)$ choose β to minimize:

$$E\left[(\tau-1)\int_{-\infty}^{x'\beta}(y-x'\beta)\,dF(y\,|x\,)+\tau\int_{x'\beta}^{\infty}(y-x'\beta)\,dF(y\,|x\,)\right]$$

(Note $y \leq x'\beta$ in the first integral and $y \geq x'\beta$ in the second.)

- The FOC for the solution β_τ is:

$$E\left[(1-\tau)\int_{-\infty}^{x'\beta_\tau}dF(y\,|x\,)=\tau\int_{x'\beta_\tau}^{\infty}dF(y\,|x\,)\right]$$

and a sample analogue, $\beta_\tau^{(N)}$, minimizes:

$$\frac{1}{N}\sum\nolimits_{n=1}^{N}\left[(\tau-1)\,I\left\{y_n \leq x_n'\beta\right\}+\tau I\left\{y_n > x_n'\beta\right\}\right](y_n - x_n'\beta)$$

- Setting $\tau = 0.5$, the median, defines the LAD estimator.

- Substituting (1) into (3) yields:

$$\begin{aligned}
\beta_{OLS}^{(N)} &= \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\left(x_n'\beta_0+\epsilon_n\right)\right] \\
&= \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)\beta_0 \\
&\quad + \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} x_n \epsilon_n\right) \\
&= \beta_0 + \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} x_n \epsilon_n\right)
\end{aligned}$$

- Thus the estimation error is:

$$\delta_{OLS}^{(N)} \equiv \beta_{OLS}^{(N)} - \beta_0 = \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} x_n \epsilon_n\right) \qquad (6)$$

Ordinary Least Squares

An orthogonality condition assumption

- Denote $x^{(N)} \equiv (x_1, \ldots, x_N)$ and assume $E\left[\epsilon_n \middle| x^{(N)}\right] = 0$.
- Then $E\left[\delta_{OLS}^{(N)} \middle| x^{(N)}\right] = 0$, and $\beta_{OLS}^{(N)}$ is *unbiased*, meaning:

$$E\left[\beta_{OLS}^{(N)} \middle| x^{(N)}\right] = \beta_0$$

- When $E\left[\epsilon_n \middle| x^{(N)}\right] = 0$ the variance of $\beta_{OLS}^{(N)}$ is:

$$E\left\{ \begin{array}{c} \left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} x_n \epsilon_n\right)\right] \\ \times\left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} x_n \epsilon_n\right)\right]' \end{array} \middle| x^{(N)} \right\} \tag{7}$$

$$= E\left[\begin{array}{l} \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1} \times \\ \left(\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} x_n \epsilon_n \epsilon_m x_m'\right)\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1} \end{array} \middle| x^{(N)} \right]$$

- Suppose it is also true that:

$$E\left[\epsilon_n \epsilon_m \left| x^{(N)} \right.\right] = \begin{cases} \sigma^2 \text{ if } m = n \\ 0 \text{ if } m \neq n \end{cases} \tag{8}$$

- Then (7) simplifies to:

$$\begin{aligned} & E\left[\delta_{OLS}^{(N)} \delta_{OLS}^{(N)\prime} \left| x^{(N)} \right.\right] \\ = & \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1} \times \\ & \left(\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} x_n E\left[\epsilon_n \epsilon_m \left| x_n, x_m \right.\right] x_m' \left| x^{(N)} \right.\right)\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1} \\ = & \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\left(\frac{\sigma^2}{N^2}\sum_{n=1}^{N} x_n x_n'\right)\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1} \\ = & \frac{\sigma^2}{N}\left\{\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n'\right)^{-1}\right\} \end{aligned}$$

Generalized Least Squares

A transformation

- Assume $E\left[\epsilon_n \middle| x^{(N)}\right] = 0$ for all $n \in \{1, \ldots, N\}$.
- Let $\epsilon^{(N)} \equiv (\epsilon_1, \ldots, \epsilon_N)'$ denote the vector of unobserved variables.
- Denote their covariance matrix by $\Psi \equiv E\left[\epsilon^{(N)}\epsilon^{(N)\prime} \middle| x^{(N)}\right]$.
- Since Ψ is positive definite, $\Psi^{-1/2}$ exists and satisfies:

$$\Psi^{-1} = \Psi^{-1/2}\Psi^{-1/2}$$

- Stack the individual equations and premultiply the resulting matrix equation by $\Psi^{-1/2}$ to obtain a transformation of (1):

$$y_n^* = x_n^{*\prime}\beta + \epsilon_n^* \tag{9}$$

where:

$$\begin{aligned} (y_1^*, \ldots, y_N^*)' &\equiv \Psi^{-1/2}(y_1, \ldots, y_N)' \\ (\epsilon_1^*, \ldots, \epsilon_N^*)' &\equiv \Psi^{-1/2}(\epsilon_1, \ldots, \epsilon_N)' \\ (x_1^*, \ldots, x_N^*) &\equiv (x_1, \ldots, x_N)\,\Psi^{-1/2} \end{aligned}$$

Generalized Least Squares

Definition

- We define the *generalized least squares* (GLS) estimator by:

$$\begin{aligned} \beta_{GLS}^{(N)} &\equiv \arg\min_{\beta} \left\{ \sum_{n=1}^{N} \left(y_n^* - x_n^{*\prime}\beta \right)^2 \right\} \\ &= \left(\frac{1}{N} \sum_{n=1}^{N} x_n^* x_n^{*\prime} \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} x_n^* y_n^* \right) \end{aligned}$$

- The assumptions in the previous slide imply:

$$\begin{aligned} E\left[\epsilon_n^* \middle| x^{(N)} \right] &= 0 \\ E\left[\epsilon_n^* \epsilon_m^* \middle| x^{(N)} \right] &= \begin{cases} 1 \text{ if } m = n \\ 0 \text{ if } m \neq n \end{cases} \end{aligned}$$

- Thus $\beta_{GLS}^{(N)}$ is unbiased.

- Returning to the model of panel data $\{y_{nt}, x_{nt}\}_{n=1}^{NT}$ with specification (2) we briefly consider the following two estimators.
- The first defines:

$$\widehat{\epsilon}_{nt} \equiv \gamma_n + \epsilon_{nt}$$

  and treats the equation be estimated as:

$$y_{nt} = x_{nt}'\beta_0 + \widehat{\epsilon}_{nt} \tag{10}$$

- A *random effects* estimator (RE) is to conduct OLS or GLS on (10).
- Without loss of generality $E\left[\epsilon_{nt}\,|\gamma_n\right] = 0$. The RE estimator is unbiased if:

$$E\left[\epsilon_{nt}\left|x^{(N)}\right.\right] = E\left[\gamma_n\left|x^{(N)}\right.\right] = 0$$

- Alternatively apply the *difference operator* to (2) and obtain:

$$\Delta y_{nt} = \Delta x'_{nt}\beta_0 + \Delta\epsilon_{nt} \tag{11}$$

where:

$$\begin{aligned} \Delta y_{nt} &\equiv y_{n,t+1} - y_{nt} \\ \Delta x_{nt} &\equiv x_{n,t+1} - x_{nt} \\ \Delta\epsilon_{nt} &\equiv \epsilon_{n,t+1} - \epsilon_{nt} \end{aligned}$$

- Then using (11) estimate β_0 from $\{y_{nt}, x_{nt}\}_{n=1}^{N,T-1}$ with OLS or GLS.
- The FD estimator is unbiased if:

$$E\left[\epsilon_{nt}\,\middle|\,x^{(N)}\right] = 0$$

but correlations between x_{nt} and γ_n do not affect the properties of this estimator.

Generalized Least Squares

Constructing the covariance matrices for these two GLS estimators

- Without loss of generality $E\left[\epsilon_{nt}\left|\gamma_{n}\right.\right]=0$ and hence:

$$E\left[\epsilon_{nt}\left|\gamma_{n}\right.\right]=0\Rightarrow E\left[\epsilon_{nt}\gamma_{n}\right]=0$$

- For now we also assume:
  - $E\left[\epsilon_{nt}\epsilon_{ms}\right]=0$ for all $m\neq n$ and all (s,t)
  - $E\left[\epsilon_{nt}\epsilon_{ns}\right]=0$ for all $s\neq t$
  - $E\left[\epsilon_{nt}^{2}\right]=\sigma_{\epsilon}^{2}$
- If $E\left[\gamma_{n}^{2}\right]=\sigma_{\gamma}^{2}$ and $E\left[\gamma_{n}\gamma_{m}\right]=0$ for all $m\neq n$, then the nonzero elements of Ψ_{RE} are:

$$E\left[\widehat{\epsilon}_{nt}\widehat{\epsilon}_{ns}\right]=\begin{cases}\sigma^{2}+\sigma_{\gamma}^{2} \text{ if } s=t\\ \sigma_{\gamma}^{2} \text{ if } s\neq t\end{cases}$$

- By way of contrast the only nonzero elements of Ψ_{FD} are:

$$E\left[\Delta\epsilon_{nt}\Delta\epsilon_{ns}\right]=\begin{cases}2\sigma_{\epsilon}^{2} \text{ if } s=t\\ -\sigma_{\epsilon}^{2} \text{ if } s=t+1\end{cases}$$

- Rearranging the FOC for the quadratic defining OLS gives:

$$0 = \sum_{n=1}^{N} x_n \left(y_n - x_n' \beta_{OLS}^{(N)} \right)$$

- As a matter of computation $\beta_{OLS}^{(N)}$ is obtained by:
    - premultiplying $\left(y_n - x_n' \beta_{OLS}^{(N)} \right)$ by x_n
    - solving the resulting k equations in k unknowns.
- Moreover its unbiasedness stems from the assumption that:

$$0 = E\left[\epsilon_n \middle| x^{(N)} \right] = E\left[y_n - x_n' \beta_0 \middle| x^{(N)} \right]$$

- Instead of premultiplying $\left(y_n - x_n' \beta_{OLS}^{(N)} \right)$ by x_n we could premultiply $\left(y_n - x_n' \beta_{OLS}^{(N)} \right)$ by $z_n \equiv A w_n$ for some $k \times l$ matrix A and some $l \times 1$ instrument vector, where $l > k$, and base the estimator on a different set of equations.

Linear Instrumental Variables

Definition

- Accordingly define an instrumental variables (IV) estimator by:

$$0 = \sum_{n=1}^{N} z_n \left(y_n - x_n' \beta_{IV}^{(N)} \right)$$

- If $\frac{1}{N} \sum_{n=1}^{N} z_n x_n'$ is invertible (has a nonzero determinant), then similar to above:

$$\beta_{IV}^{(N)} = \left(\frac{1}{N} \sum_{n=1}^{N} z_n x_n' \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} z_n y_n \right)$$

- To investigate the finite sample properties of $\beta_{IV}^{(N)}$ we follow the same reasoning we applied to $\beta_{OLS}^{(N)}$ by substituting for y_n to obtain:

$$\begin{aligned} \beta_{IV}^{(N)} &= \left(\frac{1}{N} \sum_{n=1}^{N} z_n x_n' \right)^{-1} \frac{1}{N} \sum_{n=1}^{N} z_n \left(x_n' \beta_0 + \epsilon_n \right) \\ &= \beta_0 + \left(\frac{1}{N} \sum_{n=1}^{N} z_n x_n' \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} z_n \epsilon_n \right) \end{aligned}$$

Linear Instrumental Variables

Conditions for the existence of an unbiased estimator

- In this case the estimation error is:

$$\delta_{IV}^{(N)} \equiv \beta_{IV}^{(N)} - \beta_0 = \left(\frac{1}{N} \sum_{n=1}^{N} z_n x_n' \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} z_n \epsilon_n \right) \quad (12)$$

- Let $v^{(N)} \equiv \left(x^{(N)}, w^{(N)} \right)$. If $E\left[\epsilon_n \middle| v^{(N)} \right] = 0$ then $E\left[\delta_{IV}^{(N)} \middle| v^{(N)} \right] = 0$ and $\beta_{IV}^{(N)}$ is unbiased, and (as we show on the next slides):

$$E\left[\delta_{IV}^{(N)} \delta_{IV}^{(N)\prime} \middle| v^{(N)} \right] = \frac{1}{N} \Upsilon^{(N)} E\left[\Omega^{(N)} \middle| v^{(N)} \right] \Upsilon^{(N)\prime}$$

where:

$$\Upsilon^{(N)} \equiv \left(\frac{1}{N} \sum_{n=1}^{N} z_n x_n' \right)^{-1}$$

$$\Omega^{(N)} \equiv \frac{1}{N} \sum_{n=1}^{N} z_n z_n' \epsilon_n^2 + \frac{1}{N} \sum_{s=2}^{N} \sum_{n=1}^{s-1} \left(z_n z_{n+s}' + z_{n+s} z_n' \right) \epsilon_n \epsilon_{n+s}$$

- From (12):

$$
\begin{aligned}
& E\left[\delta_{IV}^{(N)}\delta_{IV}^{(N)\prime}\,\middle|\,v^{(N)}\right] \\
= \quad & E\left\{\begin{array}{l} \left(\frac{1}{N}\sum_{n=1}^{N} z_n x_n'\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} z_n \epsilon_n\right) \\ \times\left(\frac{1}{N}\sum_{n=1}^{N} z_m \epsilon_m\right)'\left(\frac{1}{N}\sum_{n=1}^{N} z_n x_n'\right)^{-1\prime} \end{array}\,\middle|\,v^{(N)}\right\} \\
= \quad & \Upsilon^{(N)} E\left\{\left(\frac{1}{N}\sum_{n=1}^{N} z_n \epsilon_n\right)\left(\frac{1}{N}\sum_{m=1}^{N} z_m \epsilon_m\right)'\,\middle|\,v^{(N)}\right\}\Upsilon^{(N)\prime}
\end{aligned}
$$

- Focusing on the middle terms involving ϵ_n and ϵ_m:

$$
\begin{aligned}
& \left(\sum_{n=1}^{N} z_n \epsilon_n\right)\left(\sum_{n=1}^{N} z_m \epsilon_m\right)' \\
= \; & \sum_{n=1}^{N} \sum_{m=1}^{N} z_n \epsilon_n \epsilon_m z_m' \\
= \; & \sum_{n=1}^{N} z_n \epsilon_n^2 z_n' + \sum_{s=2}^{N} \sum_{n=1}^{s-1} \left(z_n z_{n+s}' + z_{n+s} z_n'\right) \epsilon_n \epsilon_{n+s}
\end{aligned}
$$

- The last line comes from visualizing the matrix of terms:

$$
\begin{bmatrix}
z_1 \epsilon_1^2 z_1' & \cdots & z_1 \epsilon_1 \epsilon_N z_N' \\
\vdots & \ddots & \vdots \\
z_N \epsilon_N \epsilon_1 z_1' & \cdots & z_N \epsilon_N^2 z_N'
\end{bmatrix}
$$

- Substituting the expression above back into the formula for the variance gives the result.

Constrained Least Squares

Definition and Solution

- Now suppose we have information about the unknown parameter vector β_0 that takes the form of a linear constraint, q equations in β_0:

$$Q\beta_0 = c \tag{13}$$

  where:
  - Q is a $q \times k$ matrix
  - c a $q \times 1$ vector
  - as before β_0 is $k \times 1$.
- The *constrained least squares* (CLS) estimator is defined by:

$$\beta_{CLS}^{(N)} \equiv \arg\min_{\beta} \left\{ \sum_{n=1}^{N} \left(y_n - x_n'\beta\right)^2 \text{ such that } Q\beta = c \right\}$$

- The next slides show $\beta_{CLS}^{(N)} - \beta_{OLS}^{(N)} =$

$$\left[\left(\frac{1}{N} \sum_{n=1}^{N} x_n x_n' \right)^{-1} Q' \right] \left[Q \left(\frac{1}{N} \sum_{n=1}^{N} x_n x_n' \right)^{-1} Q' \right]^{-1} \left(Q\beta_{OLS}^{(N)} - c \right)$$

- Define:

$$\eta \equiv \beta_{CLS}^{(N)} - \beta_{OLS}^{(N)} \qquad \gamma \equiv c - Q\beta_{OLS}^{(N)} \tag{14}$$

- The Lagrangian for the optimization problem can be written as:

$$\sum_{n=1}^{N} \left(y_n - x_n'\beta\right)^2 + \lambda\left(Q\beta - c\right)$$

and has FOC:

$$\begin{aligned}
0 &= -\left(\frac{2}{N}\sum_{n=1}^{N} x_n y_n\right) + \left(\frac{2}{N}\sum_{n=1}^{N} x_n x_n'\right)\beta_{CLS}^{(N)} + Q\lambda \\
&= \left(\frac{2}{N}\sum_{n=1}^{N} x_n x_n'\right)\left[\beta_{CLS}^{(N)} - \left(\sum_{n=1}^{N} x_n x_n'\right)^{-1}\sum_{n=1}^{N} x_n y_n\right] + Q\lambda \\
&= \left(\frac{2}{N}\sum_{n=1}^{N} x_n x_n'\right)\left(\beta_{CLS}^{(N)} - \beta_{OLS}^{(N)}\right) + Q\lambda \\
&= \left(\frac{2}{N}\sum_{n=1}^{N} x_n x_n'\right)\eta + Q\lambda
\end{aligned} \tag{15}$$

Constrained Least Squares

Proof of formula for CLS continued

- From (14) and (15):

$$\begin{aligned} 0 &= Q\beta_{CLS}^{(N)} - c = Q\left(\beta_{CLS}^{(N)} - \beta_{OLS}^{(N)}\right) - c + Q\beta_{OLS}^{(N)} = Q\eta - \gamma \\ \eta &= -\left(2N^{-1}\sum_{n=1}^{N} x_n x_n'\right)^{-1} Q'\lambda \qquad (16) \end{aligned}$$

- Solving for λ in terms of γ using (16):

$$\begin{aligned} \gamma &= Q\eta = -Q\left(2N^{-1}\sum_{n=1}^{N} x_n x_n'\right)^{-1} Q'\lambda \\ \Rightarrow \lambda &= -\left[Q\left(2N^{-1}\sum_{n=1}^{N} x_n x_n'\right)^{-1} Q'\right]^{-1}\gamma \\ \Rightarrow \eta &= \left[\left(\sum_{n=1}^{N} x_n x_n'\right)^{-1} Q'\right]\left[Q\left(\sum_{n=1}^{N} x_n x_n'\right)^{-1} Q'\right]^{-1}\gamma \end{aligned}$$

- Using the definitions of η and γ the formula now follows directly.

Specification Error versus Efficiency

Trading off efficiency with specification error

- Even if $E\left[\epsilon_n \,|x_n\right] \neq 0$ and hence $\beta_{OLS}^{(N)}$ is biased, an unbiased estimator $\beta_{IV}^{(N)}$ can be obtained if there exists some z_n satisfying:
    1. the invertibility assumption for $\frac{1}{N}\sum_{n=1}^{N} z_n x_n'$
    2. the orthogonality condition $E\left[\epsilon_n \,|z_n\right] = 0$.
- This raises the question of why OLS is ever used instead of IV, since the latter seems less restrictive.
- In Assignment 3 you are asked to show that:

$$E\left[\delta_{OLS}^{(N)}\delta_{OLS}^{(N)\prime}\right] \leq E\left[\delta_{IV}^{(N)}\delta_{IV}^{(N)\prime}\right]$$

- Similarly one can show that:

$$E\left[\delta_{CLS}^{(N)}\delta_{CLS}^{(N)\prime}\right] \leq E\left[\delta_{OLS}^{(N)}\delta_{OLS}^{(N)\prime}\right]$$

- Comparing the FE and the RE estimators raises similar issues. The former is based on $N\left(T-1\right)$ observations, but the latter requires $E\left[\gamma_n \,|x_{nt}\right] = 0$ for unbiasedness.

- The *mean square error* (MSE) is one way to evaluate the trade-off between bias and variance.
- Let $\theta \equiv \sum_{k=0}^{K-1} a_k \beta_k$ be a known linear combination of β defined by $a \equiv (a_0, \ldots, a_{K-1}) \in \mathbb{R}^K$.
- For any estimator $\theta^{(N)}$ of θ_0 we define the MSE as:

$$
\begin{aligned}
MSE\left(\theta^{(N)}\right) &\equiv E\left[\left(\theta^{(N)} - \theta_0\right)^2\right] \\
&= E\left[\left(\theta^{(N)} - E\left[\theta^{(N)}\right] + E\left[\theta^{(N)}\right] - \theta_0\right)^2\right] \\
&= E\left[\begin{array}{c} \left\{\theta^{(N)} - E\left[\theta^{(N)}\right]\right\}^2 + \left\{E\left[\theta^{(N)}\right] - \theta_0\right\}^2 \\ +2\left\{\theta^{(N)} - E\left[\theta^{(N)}\right]\right\}\left\{E\left[\theta^{(N)}\right] - \theta_0\right\} \end{array}\right] \\
&= E\left[\left\{\theta^{(N)} - E\left[\theta^{(N)}\right]\right\}^2\right] + \left\{E\left[\theta^{(N)}\right] - \theta_0\right\}^2
\end{aligned}
$$

Shrinkage Estimators

CLS as a response to overfitting

- Loosely speaking, the term overfitting means:
  - massaging the data with enough parameters and variables
  - in order to explain the sample very well
  - without reference to the underlying population.
- A fundamental limitation of this approach is that:
  - since the population does not exactly replicate the sample,
  - predicting out of sample is problematic.
- By imposing linear constraints on the model CLS:
  - reduces (or shrinks) the dimension of the basis defining the parameter space
  - and in this way increases the precision of the estimates,
  - that is if the constraints are (approximately) correct.
- One advantage of CLS, interpreted as a shrinkage estimator, is that the constraints are often easy to interpret, and may have some economic or institutional content.

- Another approach is to shrink the parameters by choosing β to minimize:

$$N^{-1} \sum_{n=1}^{N} \left(y_n - x_n'\beta\right)^2 \text{ subject to } \left(\sum_{k=1}^{K} |\beta_k|^p\right)^{1/p} \leq t \qquad (17)$$

  for some $p \in \mathbb{R}^+$ and $t \in \mathbb{R}^+$.
- The *lasso* (least absolute shrinkage and selection operator) estimator solves (17) for $p = 1$.
- The *ridge* (or Stein) estimator solves (17) for $p = 2$.
- A third variation, the *best subset selection*, is defined by requiring $t \in \{1, \ldots, K-1\}$ and replacing (17) with:

$$N^{-1} \sum_{n=1}^{N} \left(y_n - x_n'\beta\right)^2 \text{ subject to} \sum_{k=1}^{K} 1\{\beta_k \neq 0\} \leq t$$

- All three estimators (trivially) reduce overfitting, by constraining the objective function.
- Lasso and Ridge penalize all candidate values of β_k relative to their OLS counterparts.
- Lasso and best-subset-selection eliminate regressors with low explanatory power in OLS.
- Combining these estimators with machine learning could be useful in pointing to empirical patterns that guide the development of a structural model.
- However this class of estimators is not motivated by an economic theory that explains comovements within the population, so is not particularly useful for predictive purposes outside of the sample.