Quantitative Methods in Economics
Panel data and generalized least squares

Maximilian Kasy

Harvard University, fall 2016

Roadmap, Part I

1. Linear predictors and least squares regression
2. Conditional expectations
3. Some functional forms for linear regression
4. Regression with controls and residual regression
5. **Panel data and generalized least squares**

Takeaways for these slides

- Panel data: more than one outcome
- Estimator: GLS generalizes LS to multidimensional outcome
- Example 1: Siblings' earnings and education
- Example 2: Agricultural output over time

Panel data

- Data:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_M \end{pmatrix}, \quad X = \begin{pmatrix} X_1' \\ \vdots \\ X_M' \end{pmatrix} = \begin{pmatrix} X_{11} & \dots & X_{1K} \\ \vdots & & \vdots \\ X_{M1} & \dots & X_{MK} \end{pmatrix}.$$

- Example: Population of families with Y_t = log(earnings) and Z_t =(education, age) for family member t,
- $K \times 1$ X_t is constructed (polynomials, etc.) from Z_t ($t = 1, \dots, M$).

Generalized linear predictor

- Linear predictor with weighting:

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} = \arg \min_{c \in \mathscr{R}^K} E[(Y - Xc)' \Phi (Y - Xc)],$$

- Φ is a $M \times M$ symmetric, positive-definite matrix.
- Notation: $E^*_\Phi(Y \,|\, X) = X\beta$.

- Inner Product: If U and V are $M \times 1$ random vectors, and Φ is a (nonrandom) $M \times M$ positive-definite matrix,

$$\langle U, V \rangle_\Phi = E(U' \Phi V).$$

- Norm:

$$||V||_\Phi = \langle V, V \rangle_\Phi^{1/2}.$$

$$\beta = \arg\min_c ||Y - Xc||_\Phi^2.$$

Questions for you

Find the first order conditions for this minimization problem.

Solution:

- Let

$$X\beta = \begin{pmatrix} X^{(1)} & \dots & X^{(K)} \end{pmatrix} \beta = X^{(1)}\beta_1 + \cdots + X^{(K)}\beta_K,$$

- Orthogonal projection:

$$\langle Y - X\beta, X^{(k)} \rangle_\Phi = 0 \quad (k = 1, \dots, K) \quad \Rightarrow \quad E[(Y - X\beta)'\Phi X] = 0$$

- Solving for β:

$$E(Y'\Phi X) - \beta' E(X'\Phi X) = 0$$

$$E(X'\Phi X)\beta = E(X'\Phi Y),$$

and thus

$$\beta = [E(X'\Phi X)]^{-1} E(X'\Phi Y).$$

Sample version - generalized least squares (GLS)

- Data:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iM} \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

$$x_i = \begin{pmatrix} x_i^{(1)} & \dots & x_i^{(K)} \end{pmatrix} = \begin{pmatrix} x_{i11} & \dots & x_{i1K} \\ \vdots & & \vdots \\ x_{iM1} & \dots & x_{iMK} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

- y is $nM \times 1$ and x is $nM \times K$. Let A be a $nM \times nM$ symmetric, positive-definite matrix.
- GLS estimator:

$$b = \arg\min_c (y - xc)' A (y - xc).$$

- Inner Product: $u, v \in \mathscr{R}^{nM}$,

$$\langle u, v \rangle_A = u' A v.$$

- Norm:

$$||v||_A = \langle v, v \rangle_A^{1/2}.$$

$$b = \arg\min_c ||y - xc||_A.$$

- Orthogonal projection:

$$\langle y - xb, x^{(k)} \rangle_A = 0 \quad (k = 1, \ldots, K) \quad \Rightarrow \quad (y - xb)' A x = 0$$

- Solving for b:

$$y' A x - b' (x' A x) = 0$$

$$(x' A x) b = x' A y,$$

and thus

$$b = (x' A x)^{-1} x' A y.$$

Rewriting the GLS estimator

- Factorization of A :

$$A = C'C,$$

  where C is $nM \times nM$ and nonsingular.

- Define

$$\tilde{y} = Cy, \quad \tilde{x} = Cx.$$

- Note that with $\tilde{v} = Cv$,

$$||v||_A^2 = v'Av = v'C'Cv = \tilde{v}'\tilde{v} = ||\tilde{v}||_I^2,$$

- thus

$$b = \arg\min_c ||y - xc||_A = \arg\min_c ||\tilde{y} - \tilde{x}c||_I = (\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{y}.$$

- We can obtain a GLS fit using a least-squares program. In Matlab,

$$b = \tilde{x} \backslash \tilde{y}.$$

Panel data: Application to siblings

- Consider siblings in families with two children
- Y_{it} = log(earnings) of sibling t in family i
- Z_{it} = education of sibling t in family i ($t = 1, 2$)
- Data:

$$W_i = (Y_{i1}, Y_{i2}, Z_{i1}, Z_{i2}) \text{ i.i.d.} \qquad (i = 1, \ldots, n).$$

Latent variable model

- Assume that earnings are determined by

$$E^*(Y_{i1} \mid Z_{i1}, Z_{i2}, A_i) = \gamma_1 + \gamma_2 Z_{i1} + \gamma_3 A_i,$$
$$E^*(Y_{i2} \mid Z_{i1}, Z_{i2}, A_i) = \gamma_1 + \gamma_2 Z_{i2} + \gamma_3 A_i.$$

- Here A_i is a family background variable that is *not* observed.
- Note that, by assumption, sibling education does not show up given A_i!
- Now consider a regression of earnings on own- and sibling education:

$$E^*(Y_{i1} \mid 1, Z_{i1}, Z_{i2}) = \gamma_1 + \gamma_2 Z_{i1} + \gamma_3 E^*(A_i \mid 1, Z_{i1}, Z_{i2}),$$
$$E^*(Y_{i2} \mid 1, Z_{i1}, Z_{i2}) = \gamma_1 + \gamma_2 Z_{i2} + \gamma_3 E^*(A_i \mid 1, Z_{i1}, Z_{i2}).$$

- Next, consider a hypothetical regression of A_i on both siblings' education:

$$E^*(A_i \mid 1, Z_{i1}, Z_{i2}) = \lambda_0 + \lambda_1 Z_{i1} + \lambda_2 Z_{i2}.$$

Questions for you

Assume $\lambda_1 = \lambda_2$ and combine these equations to derive an expression for

$$E^*(Y_{it} \mid 1, Z_{it}, Z_{i1} + Z_{i2}).$$

- Solution:

$$E^*(Y_{i1} \mid 1, Z_{i1}, Z_{i1} + Z_{i2}) = (\gamma_1 + \gamma_3\lambda_0) + \gamma_2 Z_{i1} + \gamma_3\lambda_1(Z_{i1} + Z_{i2}),$$
$$E^*(Y_{i2} \mid 1, Z_{i2}, Z_{i1} + Z_{i2}) = (\gamma_1 + \gamma_3\lambda_0) + \gamma_2 Z_{i2} + \gamma_3\lambda_1(Z_{i1} + Z_{i2}).$$

- If the latent variable model is true, then the coefficient on own-education equals γ_2.
- Generalized Linear Predictor:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix}, \quad X_i = \begin{pmatrix} 1 & Z_{i1} & (Z_{i1} + Z_{i2}) \\ 1 & Z_{i2} & (Z_{i1} + Z_{i2}) \end{pmatrix},$$

$$E^*_\Phi(Y_i \mid X_i) = X_i\beta = \begin{pmatrix} \beta_1 + \beta_2 Z_{i1} + \beta_3(Z_{i1} + Z_{i2}) \\ \beta_1 + \beta_2 Z_{i2} + \beta_3(Z_{i1} + Z_{i2}) \end{pmatrix}.$$

Panel data: Application to production functions

- Suppose output is determined by

$$Q_{it} = L_{it}^{\gamma} F_i V_{it} \qquad (i = 1, \ldots, n;\ t = 1, \ldots, T),$$

- Q_{it} = output for farm i in year t,
  L_{it} = labor input, $0 < \gamma < 1$
  F_i = soil quality, V_{it} = rainfall.
- Profit maximization at time t, where V_{it} is uncertain:

$$\max_L E[P_t Q_{it} - W_t L \,|\, \text{Info}_{it}] = \max_L P_t [L^{\gamma} F_i E(V_{it} \,|\, \text{Info}_{it})] - W_t L$$

Questions for you

- Derive the first order condition for the firms' profit maximization problem.
- Solve for the implied demand for labor.

Solution:

- First order condition:

$$\gamma P_t L^{\gamma-1} F_i E(V_{it} \mid \text{Info}_{it}) = W_t.$$

- Derived demand for labor:

$$\log L_{it} = \frac{1}{1-\gamma}[\log \gamma - \log \frac{W_t}{P_t} + \log F_i + \log E(V_{it} \mid \text{Info}_{it})].$$

- We can write the production function as

$$\log Q_{it} = \gamma \log L_{it} + \log F_i + \log V_{it}.$$

- Equivalently

$$Y_{it} = \gamma Z_{it} + A_i + \log V_{it},$$

with $Y_{it} = \log Q_{it}$, $Z_{it} = \log L_{it}$, and $A_i = \log F_i$.

- Hypothetical regression of output at time t on labor input at all times and soil quality (unobserved):

$$E(Y_{it} \mid Z_{i1}, \ldots, Z_{iT}, A_i) = \gamma Z_{it} + A_i + E(\log V_{it} \mid Z_{i1}, \ldots, Z_{iT}, A_i)$$

- One more assumption: strict exogeneity of Z conditional on A: $E(\log V_{it} \,|\, Z_{i1}, \ldots, Z_{iT}, A_i) = \text{constant},$
- then

$$E(Y_{it} \,|\, Z_{i1}, \ldots, Z_{iT}, A_i) = \gamma Z_{it} + A_i + \text{constant}$$

Questions for you

Take "first differences":

- Calculate

$$E(Y_{it} - Y_{i,t-1} \,|\, Z_{i1}, \ldots, Z_{iT}, A_i).$$

- Then calculate

$$E(Y_{it} - Y_{i,t-1} \,|\, Z_{i1}, \ldots, Z_{iT}).$$

Solution:

- Strict exogeneity conditional on $A \Rightarrow$

$$\begin{aligned} E(Y_{it} - Y_{i,t-1} \,|\, Z_{i1}, \ldots, Z_{iT}, A_i) &= \gamma Z_{it} + A_i - (\gamma Z_{i,t-1} + A_i) \\ &= \gamma(Z_{it} - Z_{i,t-1}), \end{aligned}$$

- and thus

$$E(Y_{it} - Y_{i,t-1} \,|\, Z_{i1}, \ldots, Z_{iT}) = \gamma(Z_{it} - Z_{i,t-1}).$$

- Generalized Linear Predictor:

$$Ydif_i = \begin{pmatrix} Y_{i2} - Y_{i1} \\ \vdots \\ Y_{iT} - Y_{i,T-1} \end{pmatrix}, \quad X_i = \begin{pmatrix} Z_{i2} - Z_{i1} \\ \vdots \\ Z_{iT} - Z_{i,T-1} \end{pmatrix},$$

$$E^*_\Phi(Ydif_i \,|\, X_i) = X_i\beta = \begin{pmatrix} \beta(Z_{i2} - Z_{i1}) \\ \vdots \\ \beta(Z_{iT} - Z_{i,T-1}) \end{pmatrix}.$$

- If the latent variable model with the strict exogeneity assumption is true, then $\beta = \gamma$.

Within group estimator

- Strict exogeneity conditional on $A \Rightarrow$

$$E[\frac{1}{T}\sum_{t=1}^{T} Y_{it} \,|\, Z_{i1}, \ldots, Z_{iT}, A_i] = \frac{1}{T}\sum_{t=1}^{T}(\gamma Z_{it} + A_i + \text{constant})$$
$$= \gamma \bar{Z}_i + A_i + \text{constant},$$

- where

$$\bar{Z}_i = \frac{1}{T}\sum_{t=1}^{T} Z_{it}, \quad \bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}.$$

- Consider de-meaned regression:

$$E(Y_{it} - \bar{Y}_i \,|\, Z_{i1}, \ldots, Z_{iT}, A_i) = \gamma Z_{it} + A_i - (\gamma \bar{Z}_i + A_i) = \gamma(Z_{it} - \bar{Z}_i),$$

- and thus

$$E(Y_{it} - \bar{Y}_i \,|\, Z_{i1}, \ldots, Z_{iT}) = \gamma(Z_{it} - \bar{Z}_i).$$

Generalized Linear Predictor

- Define demeaned variables,

$$Ydev_i = \begin{pmatrix} Y_{i1} - \bar{Y}_i \\ \vdots \\ Y_{iT} - \bar{Y}_i \end{pmatrix}, \quad X_i = \begin{pmatrix} Z_{i1} - \bar{Z}_i \\ \vdots \\ Z_{iT} - \bar{Z}_i \end{pmatrix}.$$

- Generalized linear predictor:

$$E^*_\Phi(Ydev_i \mid X_i) = X_i\beta = \begin{pmatrix} \beta(Z_{i1} - \bar{Z}_i) \\ \vdots \\ \beta(Z_{iT} - \bar{Z}_i) \end{pmatrix}.$$

- If the latent variable model with the strict exogeneity assumption is true, then $\beta = \gamma$.

- Will return to panel data when talking about differences in differences!