Survey Data in Macroeconomics

III. Econometric Methods for the Analysis of Survey Microdata

Prof. Dr. Lena Dräger

Johannes Gutenberg-University Mainz, GSEFM field course
Email: ldraeger@uni-mainz.de

Outline

1. Analysis of Survey Data in STATA
2. Pooled Cross-Sections
3. Panel Estimators
4. Analysis of Binary and Ordinal (Qualitative) Variables
5. Sample Selection and Attrition Bias
6. Sampling Weights

Textbooks: Wooldridge (2010) and Greene (2012)

Analysis of Survey Data in STATA

Analysis of Survey Microdata

In general:

- Control for socio-demographic characteristics (where possible) or individual fixed effects (only panel data)
- Possibly control for time trends and macroeconomic control variables
- Careful with causal statements regarding policy changes, generally not possible
- Probit/logit estimators: Define a "representative survey respondent" and evaluate marginal effects at this point $\Rightarrow$ otherwise, marginal effects will be evaluated at the mean which may differ across models

Analysis of Survey Microdata in STATA

- If the dataset is repeated cross-sections, declare the data to be (pooled) time series: *tsset timevar*
- If the dataset has a panel structure, declare the data to be a panel: *xtset panelvar timevar*
- For a single cross-section, you can declare the dataset to be survey data: *svyset*
- Think about meaningful truncation
- Get to know your dataset!

Implications of Truncation

- Micro survey data of macroeconomic expectations is typically truncated, at least for non-experts
- We might also think about truncating extremely high incomes etc.
- Should we adjust our estimator to the truncated nature of the data?
  - $\Rightarrow$ E.g., use estimators for censored data like a tobit estimator?
  - $\Rightarrow$ No, because truncated observations are simply dropped, not set to some fixed value (censored)
  - $\Rightarrow$ Truncation may bias OLS estimates towards 0, maximum likelihood remains consistent
  - $\Rightarrow$ In practice, estimations are typically not adjusted for truncation, probably because the truncation only becomes binding in very few cases

Pooled Cross-Sections

Pooled Cross-Sections

- Pool micro cross-sections of survey waves over time
- Individual and time component

$\Rightarrow$ If identical individuals are followed over time: Use panel estimators

$\Rightarrow$ If there are repeated (random) cross-sections: Use a simple OLS estimator on the pooled cross-sections

Pooled OLS

- Assumptions:
  - Random samples from a population at different points in time
  - Observations are independent, but not identically distributed
- Include either macro control variables or time dummies to control for aggregate changes over time
- Include individual characteristics to avoid omitted variables bias
- Standard errors may be corrected to account for heteroscedasticity (STATA: option *vce(robust)*) or clustering (STATA: option *vce(cluster clustvar)*)

Pooled OLS

Pooled cross-section model:

$$y_t = \mathbf{x}_t\beta + u_t, \ t = 1, 2, ..., T \tag{1}$$

Coefficients in the vector β can be consistently estimated by OLS if:

1. $E\left(\mathbf{x}_t' u_t\right) = 0, \ t = 1, 2, ..., T$
   $\Rightarrow$ No contemporaneous correlation between explanatory variables and the error term

2. $rank\left[\sum_{t=1}^{T} E\left(\mathbf{x}_t'\mathbf{x}_t\right)\right] = K$
   $\Rightarrow$ No perfect linear dependencies (collinearity) between the explanatory variables

Pooled OLS

For unbiasedness of standard errors of the OLS estimator, we additionally need:

3. $E\left(u_t^2 \mathbf{x}_t' \mathbf{x}_t\right) = \sigma^2 E\left(\mathbf{x}_t' \mathbf{x}_t\right)$, $t = 1, 2, ..., T$, where $\sigma^2 = E\left(u_t^2\right)$ for all t
   $\Rightarrow$ Homoscedasticity

4. $E\left(u_t u_s \mathbf{x}_t' \mathbf{x}_s\right) = \mathbf{0}$, $t \neq s$, $t, s = 1, 2, ..., T$
   $\Rightarrow$ No serial correlation

Diff-in-diff Estimation

- Popular in applied microeconometrics for policy analysis (e.g. unexpected change in unemployment insurance only in one of two neighboring states)
- Natural experiment: One control group (A) and one treatment group (B, dummy dB)
- Two time periods: One before the policy change, one afterwards (dummy $d2$)

Diff-in-diff Estimation

- Research question: What is the effect of the change in unemployment insurance on the individual labor market experience?

- Estimation equation:

$$y_{i,t} = \beta_0 + \beta_2 dB_{i,t} + \beta_3 d2_t + \delta_1 d2_t * dB_{i,t} + u_{i,t} \tag{2}$$

- OLS estimator of δ_1:

$$\hat{\delta}_1 = (\overline{y}_{B,1} - \overline{y}_{B,2}) - (\overline{y}_{A,1} - \overline{y}_{A,2}) \tag{3}$$

where $\overline{y}_{A,1}$, $\overline{y}_{A,2}$, $\overline{y}_{B,1}$ and $\overline{y}_{B,2}$ are the sample averages of y for the control (A) and the treatment (B) group before and after the policy change (time periods 1 and 2)

$\Rightarrow$ $\hat{\delta}_1$ measures the relative difference in individual labor market experience that can be attributed to the policy change only experienced by group B.

Diff-in-diff Estimation

- Estimator allows for time- and group-specific effects
- The estimator is unbiased if the policy change is not systematically related to other factors that affect $y_{i,t}$ (and are hidden in $u_{i,t}$) $\Rightarrow$ usually include further control variables for both time and cross-sectional variation

Robust and Clustered Standard Errors

- Problem: OLS assumes that observations are drawn from identical distributions, but this is often not the case in survey data
- Cross-sectional or time heteroscedasticity: E.g., variance of income is larger for high income respondents than for respondents with lower income
- Clustering: Data may differ across certain groups, either different time periods or different cross-section groups

$\Rightarrow$ Both heteroscedasticity and clustering lead to correlation between residuals and regressors $\Rightarrow$ standard errors are biased

- Robust and clustered standard errors are available with many estimators, including panel and probit/logit models

Robust and Clustered Standard Errors

OLS standard errors:

$$Var(\beta) = \left[\mathbf{x}'\mathbf{x}\right]^{-1}\left[\mathbf{x}'uu'\mathbf{x}\right]\left[\mathbf{x}'\mathbf{x}\right]^{-1} = \sigma^2\left[\mathbf{x}'\mathbf{x}\right]^{-1} \tag{4}$$

Robust standard errors:

- Account for (cross-sectional) heteroscedasticity:

$$Var(\beta) = \left[\mathbf{x}'\mathbf{x}\right]^{-1}\left[\sum_{i=1}^{N} u_i^2 x_i x_i'\right]\left[\mathbf{x}'\mathbf{x}\right]^{-1} \tag{5}$$

- In STATA: option *vce(robust)*

Robust and Clustered Standard Errors

Clustered standard errors:

- Similar to robust standard errors (correct for correlations between individual residuals e_i and regressors x_i) with corrections summed over each cluster $\Rightarrow$ assume independence across clusters g

$$Var(\beta) = \left[\mathbf{x}'\mathbf{x}\right]^{-1} \left[\sum_{i=1}^{G} x_g' u_g u_g' x_g\right] \left[\mathbf{x}'\mathbf{x}\right]^{-1} \tag{6}$$

- Appropriate cluster variables may emerge from the estimation context

- Clustering is possible along several dimensions (multi-way clustering), otherwise include fixed effects along one dimension and cluster along the other

- In STATA: option *vce(cluster clustvar)*, e.g. *vce(cluster income)* or *vce(cluster year)*

Panel Estimators

Panel Estimators

- Accounts for time-variation of individual units
- Survey data: Usually unbalanced panel
- In principal, we can test for fixed vs. random effects (Hausman test), but intuitively, individual fixed effects make more sense $\Rightarrow$ account for unobserved constant individual effects

Fixed Effects Estimator

Unobserved effects model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \ t = 1, 2, ..., T \tag{7}$$

- The model includes time-invariant individual effects c_i by estimating a dummy variable for each individual unit over time

1. Exogeneity Assumption:

$$E\left(u_{it}|\mathbf{x}_i, c_i\right) = 0, \ t = 1, 2, ..., T \tag{8}$$

$\Rightarrow$ Strict exogeneity of $\mathbf{x}_{it}$ conditional on c_i

- With fixed effects, $\mathbf{x}_{it}$ cannot include any time-invariant characteristics such as gender, race etc.! All variables must be time-varying at least for SOME part of the sample

Fixed Effects Estimator

2. Rank condition:

$$rank\left(\sum_{t=1}^{T} E\left((\mathbf{x}_{it}-\overline{\mathbf{x}}_i)'(\mathbf{x}_{it}-\overline{\mathbf{x}}_i)\right)\right)=K \tag{9}$$

$\Rightarrow$ Consistent estimation if exogeneity assumption and rank condition hold

3. Homoscedasticity and no serial correlation of residuals:

$$E\left(\mathbf{u}_i\mathbf{u}_i'|\mathbf{x}_i,c_i\right)=\sigma_u^2\mathbf{I}_T \tag{10}$$

$\Rightarrow$ Efficient and unbiased estimation

Fixed Effects: Within and Between Estimator

- Within estimator:

$$(y_{it} - \overline{y}_i) = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\boldsymbol{\beta} + (u_{it} - \overline{u}_i) \tag{11}$$

$\Rightarrow$ Eliminates the unobserved effect by subtracting time-averages for each cross-section, with $\overline{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ $\Rightarrow$ Evaluate time-variance **within** cross-sections

$\Rightarrow$ Estimated by pooled OLS

$\Rightarrow$ In STATA: ***xtreg*** *depvar indepvars [if] [in] [weight],* ***fe*** *[options]*

- Between estimator:

$$\overline{y}_i = \overline{\mathbf{x}}_i\boldsymbol{\beta} + c_i + \overline{u}_i \tag{12}$$

$\Rightarrow$ Eliminates the time effect by calculating time averages **between** the cross-sections

$\Rightarrow$ In STATA: ***xtreg*** *depvar indepvars [if] [in] [weight],* ***be*** *[options]*

First Differencing: FD Estimator

- We can also first-difference the data to eliminate the unobserved individual component in (7):

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \ t = 1, 2, ..., T \tag{13}$$

- Use pooled OLS to estimate (13)
- Loses one observation
- Same assumptions as for FE estimator need to hold
- We can use this version to construct a diff-in-diff estimator

Dynamic Panel Models

- Standard FE estimators are biased if the residuals are serially correlated (Nickell (1981) bias)
- Arellano-Bond estimator to account for a lagged dependent variable (Arellano and Bond (1991), in STATA: *xtabond*)
- Nickell bias not a severe problem if T is significantly larger than N $\Rightarrow$ often the case in aggregate country panels
- Use Least Squares Dummy Variable Corrected (LSDVC) estimator instead (Bruno (2005), in STATA: *xtlsdvc*)
- Generally: Serial correlation not a big issue with micro survey data, applies more to aggregated (macro) panel models

Accounting for Cross-Sectional Correlation in Panel Data

- Usual panel estimators assume that cross-sections are not correlated
- However, we may find contemporaneous correlation across either individuals or aggregate cross-sections
- Tests for cross-sectional correlation:
  - Breusch and Pagan (1980) statistic for cross-sectional independence of the residuals of a fixed effects model (see also Greene, 2012).
    In STATA: *xttest2*
  - Pesaran (2004) test for cross-sectional dependence, can also handle unbalanced panels. In STATA: *xtcsd, pesaran show*
  - *xtcsd* includes further tests by Frees (1995, 2004) and Friedman (1937)

Accounting for Cross-Sectional Correlation in Panel Data

- Solution:
  - Use OLS regression with clustered standard errors, robust to heteroscedasticity across cross-sections, and within cross-section correlation
  - Estimate with feasible generalized least squares (FGLS), estimator is robust to AR(1) autocorrelation within cross-sections, heteroscedasticity and cross-sectional correlation.
    Estimator in STATA: *xtgls*
  - Use OLS regression with panel-corrected standard errors that account for heteroscedasticity and contemporaneous correlation across cross-sections and, potentially, for autocorrelation.
    Estimator in STATA: *xtpcse*

Analysis of Binary and Ordinal (Qualitative) Variables

Binary Variables

- Dependent variable is a dummy, e.g. =1 if a person is employed and 0 otherwise
- We're interested in the response probability conditional on a set of explanatory variables $\mathbf{x}$:

$$p(\mathbf{x}) \equiv P(y = 1|\mathbf{x}) = P(y = 1|x_1, x_2, ..., x_K) \tag{14}$$

- Interpretation of coefficients not straightforward, except for the signs
- Estimate marginal effects post-regression

Probit or Logit Estimators

- Latent variable model:

$$y^* = \mathbf{x}\boldsymbol{\beta} + e, \quad y = 1[y^* > 0] \tag{15}$$

- e is a continuously distributed variable independent of $\mathbf{x}$ with a symmetric distribution around 0
- Let G be the cdf of e, then by the symmetry of the pdf of e around 0, we have $1 - G(-z) = G(z)$. Use this to get:

$$P(y = 1|\mathbf{x}) = P(y^* > 0|\mathbf{x}) = P(e > -\mathbf{x}\boldsymbol{\beta}|\mathbf{x}) = 1 - G(-\mathbf{x}\boldsymbol{\beta}) = G(\mathbf{x}\boldsymbol{\beta}) \tag{16}$$

- Estimation by maximum likelihood
- Robust or clustered standard errors to account for heteroscedasticity or clustered correlations

Probit or Logit Estimators

- Probit estimator: Assume a standard normal distribution for G:

$$G(\mathbf{x}\boldsymbol{\beta}) = \Phi(\mathbf{x}\boldsymbol{\beta}) \equiv \int_{-\infty}^{\mathbf{x}\boldsymbol{\beta}} \phi(\nu) d\nu \tag{17}$$

- In STATA: *probit depvar [indepvars] [if] [in] [weight] [, options]*

- Logit estimator: Assume a logistic distribution for G:

$$G(\mathbf{x}\boldsymbol{\beta}) = \Lambda(\mathbf{x}\boldsymbol{\beta}) \equiv \frac{exp(\mathbf{x}\boldsymbol{\beta})}{1 + exp(\mathbf{x}\boldsymbol{\beta})} \tag{18}$$

- In STATA: *logit depvar [indepvars] [if] [in] [weight] [, options]*

Marginal Effects

- In order to get an estimate of the marginal effect of a regressor on the probability that $y = 1$, we need to derive marginal effects after the estimation of probit/logit models
- For a continuous regressor x_j:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\mathbf{x}\boldsymbol{\beta})\beta_j, \quad \text{where } g(\mathbf{x}\boldsymbol{\beta}) \equiv \frac{dG}{d\mathbf{x}\boldsymbol{\beta}}(\mathbf{x}\boldsymbol{\beta}) \tag{19}$$

- For a discrete regressor x_K, the marginal effect of changing x_K from 0 to 1, holding all other variables fixed, is given by:

$$G(\beta_1 + \beta_2 x_2 + ... + \beta_{K-1}x_{K-1} + \beta_K) - G(\beta_1 + \beta_2 x_2 + ... + \beta_{K-1}x_{K-1}) \tag{20}$$

$\Rightarrow$ The marginal effect of a single regressor depends on all regressors in $\mathbf{x}$

Marginal Effects

- Default option: Marginal effects are evaluated at the mean of variables in **x**
- With survey micro data: Define a representative participant and evaluate marginal effects there, so that effects are comparable across different models
- In STATA: *margins, dydx(*) at(sex==1 age==49 inc==3 employ==3)*

Bi-Probit Estimators

- What if we are interested in the likelihood that two dummy variables equal 1 simultaneously?

$$y_1 = 1\left[\mathbf{x}_1\boldsymbol{\beta}_1 + e_1 > 0\right] \tag{21}$$
$$y_2 = 1\left[\mathbf{x}_2\boldsymbol{\beta}_2 + e_2 > 0\right] \tag{22}$$

- Error terms in $\mathbf{e} \equiv (e_1, e_2)$ are assumed to be bivariate normally distributed: $\mathbf{e}|\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Omega})$

Bi-Probit Estimators

- If e_1 and e_2 are correlated, estimating a joint maximum likelihood procedure is more efficient
- In STATA: *biprobit depvar1 depvar2 [indepvars] [if] [in] [weight] [, options]*
- Compute marginal effects for the likelihood of both y_1 and y_2 being 1: *margins, p11 dydx(*) (at sex==1 age==49 inc==3 employ==3)*
- *Margins* also computes conditional marginal effects or marginal effects for one variable being 1 and the other being 0

Ordered Probit or Logit Estimators

- Dependent variable is ordinal, e.g. has values from 1-5 where the ordering matters
- Often the case for qualitative survey data, e.g. Likert scale: (1 – like), (2 – like somewhat), (3 – neutral), (4 – dislike somewhat), (5 – dislike)
- Latent variable model:

$$
\begin{aligned}
y^* &= \mathbf{x}\beta + e, \quad e|\mathbf{x} \sim N(0,1) \qquad (23)\\
y &= 0 \quad \text{if } y^* \leq \alpha_1\\
y &= 1 \quad \text{if } \alpha_1 < y^* \leq \alpha_2\\
&\vdots\\
y &= J \quad \text{if } y^* > \alpha_J
\end{aligned}
$$

Ordered Probit or Logit Estimators

- Estimate parameters in α and β by maximum likelihood assuming either a normal or a logistic distribution for e
- Make sure the ordering is meaningful, i.e. qualitative expectations should assign higher numbers to higher expectations
- In STATA: *oprobit* or *ologit*
- Marginal effects have to be estimated separately for each ordinal realisation, e.g. in STATA: *margins, dydx(*) predict(outcome(5)) at(sex==1 age==49 inc==3 employ==3)*, gives the marginal effects of all regressors for the likelihood that the ordinal variable takes on the value of 5

Binary or Ordinal Dependent Variables in Panel Data

- All estimators for binary or ordinal dependent variables are also available for panel data
- With random cross-sections: Simply use the estimators on the pooled data, controlling for individual characteristics
- With repeated cross-sections:
  - Use within transformation or first-differencing to eliminate the unobserved effect c_i
  - Estimate random-effects or population-averaged probit or ordered probit models (in STATA: *xtprobit* and *xtoprobit*)
  - Population-averaged probit: Estimates at the average value of unobserved components c_i in the population, c
  - Random-effects probit: Assume a normal distribution for c_i, integrate out c_i when constructing the log-likelihood function, marginal effects can be estimated at $c = 0$

Sample Selection and Attrition Bias

Sample Selection and Attrition Bias

- Assumption: Survey samples are drawn randomly and are representative of the overall population

- What if there is non-random sample selection (= incidental truncation)? $\Rightarrow$ Survivorship or attrition bias
  - Respondents leave the survey permanently (e.g. professional forecasters get a new job at a firm not participating in the survey, a household drops out for private reasons or the respondent dies, a firm goes bankrupt)
  - Certain groups of participants are more likely to be selected into a panel dimension of the survey

- What if the decision to not answer a particular question is non-random? $\Rightarrow$ Non-response bias
  - Certain questions might be deemed more difficult than others

Heckman (1979) Estimator

- Two-step estimation procedure by Heckman (1979) to account for attrition
- Unobserved selection equation:

$$z_i^* = \mathbf{w}_i'\gamma + u_i,\ z_i = 1 \text{ if } z_i^* > 0 \text{ and 0 otherwise}$$
$$Prob(z_i = 1|\mathbf{w}_i) = \mathbf{\Phi}\left(\mathbf{w}_i'\gamma\right)$$
$$Prob(z_i = 0|\mathbf{w}_i) = 1 - \mathbf{\Phi}\left(\mathbf{w}_i'\gamma\right) \tag{24}$$

- Regression equation:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \text{ observed only if } z_i = 1 \tag{25}$$
$$(u_i, \varepsilon_i) \sim \text{bivariate normal}[0, 0, 1, \sigma_\varepsilon, \rho]$$

Heckman (1979) Estimator

- Need to assume that z_i and $\mathbf{w}_i$ are observed for a random sample
- Combining (24) and (40) gives:

$$E\left[y_i|z_i=1,\mathbf{x}_i,\mathbf{w}_i\right]=\mathbf{x}_i'\boldsymbol{\beta}+\rho\sigma_e\lambda\left(\mathbf{w}_i'\boldsymbol{\gamma}\right), \tag{26}$$
$$\text{with } \lambda\left(\mathbf{w}_i'\boldsymbol{\gamma}\right)=\phi\left(\mathbf{w}_i'\boldsymbol{\gamma}\right)/\boldsymbol{\Phi}\left(\mathbf{w}_i'\boldsymbol{\gamma}\right)$$

$\Rightarrow$ Not accounting for sample selection results in an omitted variable bias

$\Rightarrow$ OLS is not consistent!

Two-Step Heckman Correction

1. Estimate a probit model with maximum likelihood of (24) to obtain estimates of γ. For each observation, compute $\hat{\lambda}_i = \phi\left(\mathbf{w}_i'\hat{\gamma}\right) / \mathbf{\Phi}\left(\mathbf{w}_i'\hat{\gamma}\right)$ and $\hat{\delta}_i = \hat{\lambda}_i\left(\hat{\lambda}_i + \mathbf{w}_i'\hat{\gamma}\right)$
2. Estimate $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda\hat{\lambda} + \varepsilon_i$ by least squares to obtain estimates of $\boldsymbol{\beta}$ and β_λ

- The estimator of the correlation between the selection equation and the regression model, $\hat{\rho}$, provides a measure for the degree of attrition bias $\Rightarrow$ Test whether $\hat{\rho} = 0$

Two-Step Heckman Correction in STATA

- Cross-section data: *heckman depvar [indepvars], select([depvar_s =] varlist_s) twostep [options]*
- Binary data: *heckprobit depvar [indepvars] [if] [in] [weight], select([depvar_s =] varlist_s) [options]*
- Ordinal data: *heckoprobit depvar [indepvars] [if] [in] [weight], select([depvar_s =] varlist_s) [options]*
- So far no module to do a Heckman correction with panel data in STATA!

Sampling Weights

Sampling Weights

- Sampling weights are calculated by surveys in order to ensure the representative of the sample w.r.t. the overall population
- Check whether sample contains a weight variable (index). e.g. household head weight index in the Michigan Survey of Consumers
- Sampling weights (in STATA: *pweight*) denote the inverse of the probability that the observation is included due to the sampling design
- Use weights in the estimations, e.g.:
  oprobit cons_past infl_exp int_exp age sex income [pweight=weight], robust

Literature I

Arellano, M. and S. Bond (1991).

Some tests of specification for panel data: Monte carlo evidence and an application to employment equations.

Review of Economic Studies 58(2), 277–297.

Breusch, T. and A. R. Pagan (1980).

The Lagrange multiplier test and its applications to model specification in econometrics.

Review of Economic Studies 47(1), 239–253.

Bruno, G. S. (2005).

Approximating the bias of the LSDV estimator for dynamic unbalanced panel data models.

Economics Letters 87(3), 361–366.

Literature II



Frees, E. W. (1995).

Assessing cross-sectional correlations in panel data.

Journal of Econometrics 64, 393–414.



Frees, E. W. (2004).

Longitudinal and Panel Data: Analysis and Applications in the Social Sciences.

Cambridge University Press.



Friedman, M. (1937).

The use of ranks to avoid the assumption of normality implicit in the analysis of variance.

Journal of the American Statistical Association 32, 675–701.

Literature III


Greene, W. (2012).
Econometric Analysis (7th ed.).
Pearson Education.


Heckman, J. J. (1979).
Sample selection bias as a specification error.
Econometrica 47(1), 153–161.


Nickell, S. (1981).
Biases in dynamic models with fixed effects.
Econometrica 49(6), 1417–1426.


Pesaran, M. H. (2004).
General diagnostic tests for cross section dependence in panels.
Cambridge Working Paper in Economics 0435.

Literature IV

 Wooldridge, J. M. (2010).

Econometric Analysis of Cross Section and Panel Data (2nd ed.).

Cambridge, MA: MIT Press.