Journal of Data Science **7**(2009), 255-266

Panel Regression of Arbitrarily Distributed Responses

Gordon G. Bechtel
University of Florida and Florida Research Institute

Abstract: The primary advantage of panel over cross-sectional regression stems from its control for the effects of omitted variables or "unobserved heterogeneity". However, panel regression is based on the strong assumptions that measurement errors are independently identically (i.i.d.) and normal. These assumptions are evaded by design-based regression, which dispenses with measurement errors altogether by regarding the response as a fixed real number.

The present paper establishes a middle ground between these extreme interpretations of longitudinal data. The individual is now represented as a panel of responses containing dependently non-identically distributed (d.n.d) measurement errors. Modeling the expectations of these responses preserves the Neyman randomization theory, rendering panel regression slopes approximately unbiased and normal in the presence of arbitrarily distributed measurement error. The generality of this reinterpretation is illustrated with German Socio-Economic Panel (GSOEP) responses that are discretely distributed on a 3-point scale.

Key words: Longitudinal weights, panel deviations, population of panels, single-stage panel sampling, stochastic measurement error, survey response expectation.

1. Random Individual-Wave Variables

Design-based sampling postulates the respondent in a fixed observable state that s(he) reports as a discrete rating, such as 0 1 2, or recalls on a continuous monetary scale. Thus, the value recorded on an opinion poll or economic survey is regarded as a real number in waiting. More realistically, however, the survey response may be viewed as a random variable containing measurement error (cf. Diggle, Liang, and Zeger, 1994). The present paper favors this more plausible interpretation and extends Bechtel's (2007) treatment of cross-sectional regressions to longitudinal regressions involving repeated measurements. These measurements make up a panel of random variables, which may be dependently non-identically distributed (d.n.d) within each respondent. A finite population of

these panels then gives rise to a finite population of realized random individual-wave variables. Each realization is a momentary numerical value governed by an individual-wave-specific mean and variance. This approach retains and enhances design-based regression, whose slopes are still normally distributed (over samples) for *any* stochastic distributions (over realizations) that prevail for individual-wave-specific responses.

Section 2 describes an unbalanced longitudinal population along with a single-stage sample of panels. Section 3 regresses response expectations over this population, defining new target parameters as functions of these expectations. Using the Horvitz-Thompson theorem, Sections 4, 5, and 6 show that these new parameters are estimated by the well-known design-based coefficients. Section 7 describes a user-friendly computation of these coefficients with STATA software. Section 8 uses this software to evaluate predictors of environmental concern in the German Socio-Economic Panel. The final section summarizes distribution-free panel regression and reemphasizes its applicability to *arbitrarily distributed survey responses.*

2. The Population and Sample of Panels

The term panel is used here to denote an intra-individual sequence of wave measures Y_{it} . This sequence is illustrated by a single row in Table 1, where $t = 1$ for individual i's first appearance. A population of panels is a finite set of panels exemplified by the seven rows in Table 1. This population is "unbalanced" because different individuals make different numbers of wave appearances. An unbalanced population of panels is also a series of incomplete censuses, such as the four columns in Table 1.

Table 1: An unbalanced longitudinal population of panels

Panel	Wave 1	Wave 2	Wave 3	Wave 4
Individual 1	$\mathbf{Y}_{11}$	$\mathbf{Y}_{12}$	$\mathbf{Y}_{13}$	$\mathbf{Y}_{14}$
Individual 2	Y_{21}	Y_{22}	Y_{23}	
Individual 3	$\mathbf{Y}_{31}$	$\mathbf{Y}_{32}$	$\mathbf{Y}_{33}$	
Individual 4	Y_{41}	Y_{42}		
Individual 5	$\mathbf{Y}_{51}$			
Individual 6	Y_{61}	Y_{62}		
Individual 7	Y_{71}	Y_{72}	Y_{73}	

The boldface rows in Table 1 exhibit an unbalanced sample of three panels drawn without replacement from our population of seven panels. Because every wave appearance in each sampled panel is measured, Table 1 illustrates single-stage cluster sampling (Lohr, 1999, pp. 136-145), which is called single-stage

panel sampling in the sequel. In this example a sample of eight individual-wave measures are drawn from a population of eighteen individual-wave values by single stage panel sampling.

In the sequel Y_{it} in Table 1 plays three roles: a *realizable* random variable, its *realized* value which is a real number, and an *observed* (i.e. sampled) realized value. A panel can be viewed, therefore, as a cluster of random variables or as a cluster of fixed realizations. The following sections emphasize the importance of stochastic measurement error in these distinctions.

3. Regressing Expectations in the Population of Panels

3.1 Stochastic measurement error

The present paper uses survey data that sharply departs from the (usually assumed) continuity, normality, and homoscedasticity of the panel response variable (cf. Baltagi, 2001; Hsiao, 2003). Here $Y_{it} = 0, 1, 2$ denotes a rating of environmental concern by German panelist i on wave t. The response options and coding for the GSOEP's three-point scale are:

Not concerned at all (0) Somewhat concerned (1) Very concerned (2)

This score is a discrete random variable that may be decomposed as

$$\begin{aligned} Y_{it} &= H_{it} + E_{it} \\ &= \alpha_i^* + \beta_1^* X_{1it} + \cdots + \beta_k^* X_{kit} + \gamma_{it}^* + E_{it}, \end{aligned} \tag{3.1}$$

where

$$E(Y_{it}) = H_{it} = \alpha_i^* + \beta_1^* X_{1it} + \cdots + \beta_k^* X_{kit} + \gamma_{it}^*,$$

α_i^* is a *fixed* individual intercept, $X_{1it}, \ldots, X_{kit}$ are *fixed* individual-wave-specific predictors, γ_{it}^* is a *fixed* individual-wave effect on H_{it} , and $E_{it} = Y_{it} - H_{it}$ is a measurement error for individual i on wave t, with $E(E_{it}) = 0$ and $\mathrm{Var}(E_{it}) = \sigma_{it}^2$.

In (3.1) our unit of interest, individual i on wave t, is represented by a pair of parameters; namely, a mean $_{it}$ and variance σ_{it}^2 that determine an *idiosyncratic, wave specific probability distribution* on the scale 0 1 2. The mean H_{it} is continuous in the interval $[0, 2]$ and is composed of an individual intercept, individual-wave-specific predictors, and an effect that is unique to individual i on wave t. This latter effect γ_{it}^* saturates the linear model for the H_{it} in (3.1), i.e. the structure

$$\alpha_i^* + \beta_1^* X_{1it} + \cdots + \beta_k^* X_{kit} + \gamma_{it}^*$$

fits the H_{it} exactly without constraining these expectations.

The E_{it} in (3.1) may be dependently non-identically distributed (d.n.d.) over waves within individuals. Fixing individual i and wave t, the random variable E_{it} can be displayed as follows:

$$\begin{array}{ccc} p_{0it} & p_{1it} & p_{2it} \\ 0 - H_{it} & 1 - H_{it} & 2 - H_{it} \end{array}$$

The (unknown) response probabilities p_{0it}, p_{1it} , and p_{2it} for *not concerned at all, somewhat concerned*, and *very concerned* are arbitrarily distributed over the points $0 - H_{it}$,$1 - H_{it}$, and $2 - H_{it}$. The standard deviation σ_{it} on this 3-point error scale denotes uncertainty in i's rating of environmental concern. A small σ_{it} represents a precisely reporting individual with a narrow error distribution. A broad error distribution has a large σ_{it} characterizing an individual with less consistent ratings over repeated realizations of the random variable Y_{it} .

3.2 New target parameters for design-based regression

In equation (3.1) the intercept α_i^*, the slopes $\beta_1^*, \ldots, \beta_k^*$, and the effects γ_{it}^* will be uniquely identified by the ordinary-least-squares (OLS) condition that $\sum_{it} \gamma_{it}^{*2}$ is minimal when the population of *expectations* H_{it} is regressed on the population of predictors $X_{1it}, \ldots, X_{kit}$. This OLS identification of $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_k^*)^T$ is given by the following function of these expectations:

$$\boldsymbol{\beta} = [\sum_{it} x_{it} x_{it}^T]^{-1} \sum_{it} x_{it} \eta_{it} \tag{3.2}$$

for $i = 1, \ldots, N; t = 1, \ldots, T_i$, where $x_{it} = X_{it} - X_{i\cdot}, \eta_{it} = H_{it\cdot} - H_{i\cdot}$.

In (3.2) $X_{it} = (X_{1it}, \ldots, X_{kit})$, and $X_{i\cdot\cdot}$ and $H_{i\cdot}$ are the means of X_{it} and H_{it} within panel i (StataCorp. 2001, p. 437 ; Hsiao 2003, pp. 30-33). Thus $\boldsymbol{\beta}$ in (3.2) is expressed in terms of the *deviations* of response expectations and predictors from their panel means. Equation (3.2) selects the unique parameterization α_i , $\beta_1, \ldots, \beta_k, \gamma_{it}$ from an infinite set $\{\alpha_i^*, \beta_1^*, \ldots, \beta_k^*, \gamma_{it}^*\}$ of exact characterizations of the H_{it}. This defines the new target parameters of design-based regression as $\beta_1, \ldots, \beta_k$.

4. Sampling from a Realized Population of Panels

4.1 Single-stage panel sampling

Our clustered population of individuals, each containing T_i survey waves for $i = 1, \ldots, N$, is anchored by $\sum_i T_i$ expectations H_{it}. In Table 1, for example, $i = 1, \ldots, 7$ panels and $\sum_i T_i = 18$ individual-waves. Now let the random variable Y_{it} be *realized* for every individual-wave in the population of panels. This population

realization occurs in a hypothetical (but possible) series of incomplete censuses. A single-stage cluster sample of n panels is then drawn without replacement from this population of N panels. The sample consists of $\sum_i T_i$ ratings Y_{it} for $i = 1, \ldots, n$. In Table 1 $\sum_i T_i = 8$ individual-waves are drawn from $i = 1, 2, 3$ sampled panels. This setup reinterprets conventional design-based sampling which treats the Y_{it} as constants rather than realizations of random variables.

4.2 Longitudinal weights

The sample inclusion probability for a panel is the cross-sectional inclusion probability of its initial wave multiplied by the retention probabilities for its subsequent waves. These retention probabilities are "the conditional probabilities of remaining in the panel" over these remaining waves (Haisken-DeNew and Frick 2005, p. 171). For example, the sample inclusion probability π_3 for individual 3 in Table 1 is her (his) cross-sectional inclusion probability in wave 2 multiplied by her (his) retention probabilities for waves 3 and 4. The sample inclusion probability π_5 for individual 5, however, is simply her (his) cross-sectional inclusion probability in wave 2. The final longitudinal weights for individuals 3 and 5 are the reciprocals of their inclusion probabilities, i.e. $w_3 = 1/\pi_3$ and $w_5 = 1/\pi_5$.

In the German Socio-Economic Panel each respondent is assigned a cross-sectional weight and a longitudinal weight for each wave. The cross-sectional weight for panel i's first participating year is multiplied by the longitudinal weights for her (his) subsequent participating years. Each longidudinal weight is the reciprocal of i's "staying" probability for that subsequent year, i.e. the conditional probability s(he) participates in that wave and in the previous waves of her (his) panel (Haisken-DeNew and Frick 2005, p. 180). The product of panel i's initial cross-sectional weight and subsequent longitudinal weights produces i's final longitudinal weight w_i . This weight w_i covers the sequence of years individual i is monitored within the time span 1999-2005.

5. Estimating Panel Regression Coefficients

5.1 The conventional moving target

Each element of the $k \times k$ matrix $\sum_{it} x_{it} x_{it}^T$ is a population sum of products, as is each element of the $k \times 1$ vector $\sum_{it} x_{it} y_{it}$ (Lohr, 1999, p. 360). Each sum of products contains panel deviation scores

$$\begin{aligned} x_{it} &= X_{it} - X_{i\cdot} \\ y_{it} &= Y_{it} - Y_{i\cdot} \end{aligned}$$

for $t = 1, \ldots, T_i$, where $X_{i\cdot}$ and $Y_{i\cdot}$ are the means of X_{it} and Y_{it} within panel

i for $i = 1, \ldots, N$. Due to Horvitz and Thompson (1952), unbiased estimates of the matrix $\sum_{it} x_{it}x_{it}^T$ and the vector $\sum_{it} x_{it}y_{it}$ are given by $\sum_{it} w_i x_{it}x_{it}^T$ and $\sum_{it} w_i x_{it}y_{it}$ for $i = 1, \ldots, n; t = 1, \ldots, T_i$. The weight w_i is individual i's final longitudinal weight described in Section 4.2. When the sample size n is large, the Horvitz-Thompson (HT) estimator

$$\mathbf{B} = [\sum_{it} w_i x_{it}x_{it}^T]^{-1} \sum_{it} w_i x_{it}y_{it} \tag{5.1}$$

for $i = 1, \ldots, n; t = 1, \ldots, T_i$, is consistent and almost unbiased for the conventional target parameter

$$\boldsymbol{\theta} = [\sum_{it} x_{it}x_{it}^T]^{-1} \sum_{it} x_{it}y_{it} \tag{5.2}$$

for $i = 1, \ldots, N; t = 1, \ldots, T_i$.

The unbiasedness of $\mathbf{B}$ is approximate because it is the *product* of matrix and vector estimators (Binder, 1983; Nathan, 1988, pp. 255-256; Thompson, 1997, pp. 106-107; Valliant, Dorfman, and Royall, 1999, pp. 40-41; Lohr, 1999, pp. 354-361; StataCorp. 2001, Volume 4, pp. 29-30; Chaudhuri and Stenger 2005, pp. 264-265.) The parameter $\boldsymbol{\theta}$ in (5.2) is called a moving target because it is a function of the transient deviations $y_{it} = Y_{it} - Y_{i\cdot}$.

5.2 The new stationary target

The important result here is that the conventional formula (5.1) also estimates the more profound and anchored target parameter (3.2), which is a function of constant expectations H_{it} rather than momentary realizations Y_{it}. To obtain this result we take the expected value of (5.2) *over realizations* of the stochastic ratings Y_{it}:

$$\begin{aligned} E(\theta) &= [\sum_{it} x_{it}x_{it}^T]^{-1} \sum_{it} x_{it}E(y_{it}) \\ &= [\sum_{it} x_{it}x_{it}^T]^{-1} \sum_{it} x_{it}\eta_{it} = \boldsymbol{\beta} \end{aligned}$$

for $i = 1, \ldots, N; t = 1, \ldots, T_i$. Because $E(\theta_j) = \beta_j$ and $\mathrm{Var}(\theta_j) \to 0$ as the number of panels $N \to \infty$, the differences $\theta_j - \beta_j$ for $j = 1, \ldots, k$ are infinitesimal for a *given* large population realization. Thus $\mathbf{B}$ in (5.1), which is almost unbiased for $\boldsymbol{\theta}$ in (5.2), is almost unbiased for $\boldsymbol{\beta}$ in (3.2) as well.

6. Normality and Variances of the Estimated Coefficients

Fixing the momentary population realizations Y_{it} for $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$, the resulting *reals* $\theta_1, \ldots, \theta_k$ in (5.2) become the classic target parameters of design-based regression. Therefore, a strict design-based argument using the θ_j can be given for the normality *over large samples* of each element B_j in $\mathbf{B}$. This provides a statistic for testing hypotheses about the new target parameter β_j against the conventional estimate B_j .

First, given the population realizations Y_{it} , the coefficient θ_j for $j = 1, \ldots, k$ can be written as a smooth function of cross-product totals in the population

$$\{y_{it}, x_{1it}, \ldots, x_{kit} : i = 1, \ldots, N; t = 1, \ldots, T_i\}$$

of deviation scores. Then, from the sample

$$\{y_{it}, x_{1it}, \ldots, x_{kit} : i = 1, \ldots, n; t = 1, \ldots, T_i\}$$

the estimate B_j can be written as the *same* function of HT estimators of these population totals. The HT estimators are corresponding sample totals of cross products with each term weighted by w_i . For example, the sample total

$$\sum_{it} w_i x_{1it} y_{it} \qquad \text{for } i = 1, \ldots, n; t = 1, \ldots, T_i$$

is an HT estimator of the population total $\sum_{it} x_{1it} y_{it}$ for $i = 1, \ldots, N; t = 1, \ldots, T_i$ (cf. Lohr, 1999, pp. 352-360; Thompson, 1997, pp. 106-108).

The asymptotic normality of HT estimators (Sen, 1988, pp. 313-328) may now be used to justify the asymptotic normality of B_j , which is a nonlinear function of these estimators. A "linearization" of the error $B_j - \theta_j$ is provided by the first-order approximation $B_j - \theta_j \approx \epsilon_j$, where ϵ_j is the linear term in a Taylor series expansion of this error. Asymptotic multivariate normality of the HT estimators then implies that $(B_j - \theta_j)/\sqrt{\text{Var}(\epsilon_j)}$ is asymptotically $N(0, 1)$ (Lehmann, 1999, pp. 253-269, 309-315; Lohr, 1999, pp. 290-293, 310, 352-360; Thompson, 1997, pp. 58-64, 106-111). The estimate $Var(\epsilon_j)$ of $\text{Var}(\epsilon_j)$ is given by the j-th diagonal element of the matrix $Var(\mathbf{B})$ in (6.2) and is computed by software described in Section 7. Due to the infinitesimal difference between θ_j and β_j , the statistic

$$t = \frac{B_j - \beta_{j0}}{\sqrt{Var(\epsilon_j)}} \tag{6.1}$$

may be used to test an hypothesis $H : \beta_j = \beta_{j0}$ about our target coefficient β_j . This test for $\beta_{j0} = 0$ is illustrated for the regression coefficients in Table 2 below.

Table 2: Regression coefficients for predicting environmental concern ($R^2 = .28$)

Concern predictor	Coefficient	Standard error	t-statistic
General economic development	.045	.011	4.28
Your health	.133	.011	12.26
Maintaining peace	.214	.012	17.55
Crime in Germany	.111	.011	10.13
Hostility toward foreigners or minorities in Germany	.072	.010	6.88
Age	-.007	.002	2.93

Finally, again using "linearization", an estimate of the entire covariance matrix of **B** is

$$Var(\mathbf{B}) = (\sum_{it} w_i x_{it} x_{it}^T)^{-1} Var[\sum_{it} w_i x_{it}^T (y_{it} - x_{it}^T \mathbf{B})](\sum_{it} w_i x_{it} x_{it}^T)^{-1}, \qquad (6.2)$$

where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$ (Lohr, 1999, pp. 359-361; StataCorp., 2001, Volume 4, pp. 29-30). As described in Section 4.2, the longitudinal weight w_i in (6.2) is fixed over the T_i waves in individual i's panel. The j-th diagonal of $Var(\mathbf{B})$ is the estimated variance of B_j in the denominator of (6.1). Again note that the covariance estimator in (6.2) is expressed in panel deviation scores.

7. Software

The estimated regression coefficients, along with their standard errors and test statistics, are easily calculated with the STATA commands:

svyset pweight longitudinal_weight (7.1)

svyset psu panel (7.2)

svyreg devY devX1 ... devXk , noconstant (7.3)

(StataCorp. 2001, Volume 4, pp. 18- 31). In (7.1) *longitudinal_weight* is the variable containing the longitidunal weights. In (7.2) *panel* is the variable containing the panel identifications. The definitions of the deviation variables in (7.3) are:

$$\begin{aligned} \text{dev}Y &= Y_{it} - Y_{i\cdot} = y_{it} \\ \text{dev}X_1 &= X_{1it} - X_{1i\cdot} = x_{1it} \\ &\cdots\cdots \\ \text{dev}X_k &= X_{kit} - X_{ki\cdot} = x_{kit} \end{aligned}$$

The option *noconstant* in (7.3) suppresses the intercept because the response variable and its predictors are deviations from their panel means.

For large samples these three STATA commands return (approximately) normal and unbiased estimates $B_1, \ldots, B_k$ in the presence *any* distributions of the measurement errors E_{it} in (3.1). The standard errors of the estimated coefficients reflect the effects of these measurement errors on coefficient variance.

8. Environmental Concern in Germany

8.1 The GSOEP for 1999-2005

Because Germany has been at the forefront of environmental protection, the present investigation of environmental concern relies upon the well-established German Socio-Economic Panel. The first wave of the GSOEP was carried out in 1984 in the Federal Republic of Germany. The panelists studied here are residents of the former Federal Republic living in private households whose head is not Turkish, Greek, Yugoslavian, Spanish, or Italian. These respondents are known as the "west German sample" of the GSOEP (Haisken-DeNew and Frick 2005, p. 19).

The GSOEP interviews are conducted face-to-face with all persons in a household aged 16 and over. Our west German sample consists of 6634 respondents measured within the seven years of the present study, i.e. 1999-2005. Further details on the English Language Public Use File of the GSOEP, including instructions for obtaining the data, have been given by Wagner, Burkhauser, and Behringer (1993).

The survey firm *Infratest Burke Sozialforschung* in Munich carries out the fieldwork for the GSOEP. In addition to demographic information, the GSOEP questionnaire contains "objective" measures such as income and unemployment, as well as "subjective" ratings of satisfactions, worries and fears of the German population.

8.2 The GSOEP items for rating worry

The GSOEP contains ratings of concern, or worry, about various living conditions in Germany, Europe, and the world. These items are prefaced with the question:

What is your attitude toward the following areas - are you concerned about them?

The areas of concern studied here are:

Environmental protection; General economic development; Your health; Maintaining peace; Crime in Germany; Hostility toward foreigners or minorities in Germany

The response scale and coding for these items were described in Section 3.1 as:

Not concerned at all (0) Somewhat concerned (1) Very concerned (2)

8.3 The weighted panel regression

Using the STATA commands (7.1), (7.2), and (7.3), equation (3.1) is estimated as

$$\hat{Y}_{it} = A_i + B_1 X_{1it} + \cdots + B_6 X_{6it}.$$

Five significant predictors of environmental concern, along with age, are exhibited in Table 2. This west German regression was run over 34269 individual-wave measures generated by 6634 panels from 1999 to 2005. These panels ranged from one to seven waves, with an average of 5.2 waves. The estimates $A_1, \ldots, A_{6634}$ of the individual effects are not included in this report.

The five concern coefficients in Table 2 are commensurate because they share the three-point rating scale in Section 6.2. The strongest predictor of *environmental concern* is worry about *maintaining peace*, followed by worries about *your health* and *crime in Germany*. The negative *age* coefficient reveals that younger Germans have greater *environmental concern*.

The predictors in Table 2, except for *your health*, suggest that German *environmental concern* has an altruistic societal, rather than selfish individualistic, orientation. This is supported by the finding that potential regressors, such as *personal dwelling satisfaction*, and *concern with your own economic situation*, failed to reach statistical significance in predicting *environmental concern*.

9. Summary

An unbalanced longitudinal population of real numbers is reinterpreted as a set of momentary realizations of random variables Y_{it} , each governed by the parameters H_{it} and σ^2_{it} for individual i on wave t. (See Table 1.) This reinterpretation better justifies the usual design-based regression estimates and their standard errors. It opens up panel regression to design-based theory, response weighting, and arbitrary stochastic responding without reference to an abstract superpopulation (cf. Skinner, Holt, and Smith 1989).

The primary advantage of panel over cross-sectional regression lies in the possibility of bringing variable intercepts α^*_i into the model. These individual effects, which reside in the error term of a cross-sectional model, bias regression coefficients if they are related to *both* the response and its predictors. This potential bias is removed by (3.1) which contains α^*_i as an estimable effect. However, this individual effect, and the non-estimable individual-wave effect γ^*_{it} in (3.1), are not needed in defining our target $\boldsymbol{\beta}$ in (3.2) and its estimate $\mathbf{B}$ in (5.1).

The present panel extension of Bechtel (2007) differs from model-based sampling, where the finite population of realizations is itself a sample from a "su-

perpopulation" with assumed distribution and covariance properties (cf. Binder 1983; Nathan 1988; Skinner, Holt, and Smith 1989; Thompson 1997; Valliant, Dorfman, and Royall 1999; Binder and Roberts 2003). Here this "superpopulation" is simply a finite set of *arbitrarily distributed* wave variables that are clustered by individuals. These random variables are realized as responses to a hypothetical (but possible) sequence of incomplete censuses. The targets of inference are population regression coefficients that are functions of the *expectations* of individual-wave realizations. This longitudinal population, and its limited target parameters, establish a plausible bridge between design- and model-based regression theory.

Finally, the estimate $\mathbf{B}$ in (5.1) of the target $\boldsymbol{\beta}$ in (3.2) is asymptotically normal and almost unbiased (over samples) *whatever* the distribution (over realizations) of Y_{it} in the panel population. Thus, the reinterpretation of Y_{it} as a stochastic response rather than a fixed real number is a step forward in the Neyman paradigm (Bellhouse, 1988). By allowing this response to be *arbitrarily* stochastic, formulas (3.2) and (5.1) also strip away the distribution assumptions thought to be necessary for panel regression (cf. Baltagi, 2001; Hsiao, 2003).

Acknowledgements

This work has been supported by the University of Florida's Warrington College of Business. The author thanks the JDS editor and referee whose responses to the first draft have greatly improved the paper. The German Socio-Economic Panel data were supplied to the author under a contract with the German Institute for Economic Research in Berlin. These data were obtained from the Cornell University English Language Public Use File of the GSOEP. None of the ideas or analyses in the present paper are attributable to the University of Florida, Cornell University, or the German Institute for Economic Research.

References

Baltagi, B. H. (2001). *Econometric Analysis of Panel Data*, 2nd edition. Wiley.

Bechtel, G. G. (2007). Distribution-free regression: Reinterpreting design-based sampling. *Journal of Data Science* **5**, 535-553.

Bellhouse, D. R. (1988). A brief history of random sampling methods. In *Handbook of Statistics, Volume 6 (Sampling)* (Edited by P.R. Krishnaiah and C.R. Rao), pp. 1-14, North Holland.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex samples. *International Statistical Review* **51**, 279-292.

Binder, D. A. and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data* (Edited by R. L. Chambers and C. J. Skinner), pp. 29-48, Wiley.

Chaudhuri, A. and Stenger, H. (2005). *Survey Sampling: Theory and Methods*, 2nd edition. Chapman and Hall.

Diggle, P. J., Liang K-Y and Zeger S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

Haisken-DeNew, J. P. and Frick, J. R. (2005). Desktop Companion to the German Socio- Economic Panel, Version 8.0 – December 2005. German Institute for Economic Research.

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edition. Cambridge University Press.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Nathan, G. (1988). Inference based on data from complex sample designs. In *Handbook of Statistics, Volume 6 (Sampling)* (Edited by P.R. Krishnaiah and C.R. Rao), pp. 247-266, North Holland.

Sen, P. K. (1988). Asymptotics in finite population sampling. In *Handbook of Statistics, Volume 6 (Sampling)* (Edited by P.R. Krishnaiah and C.R. Rao), pp. 291-331, North Holland.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Wiley.

StataCorp. (2001). Stata Statistical Software: Release 7.0. Stata Corporation.

Thompson, M. E. (1997). *Theory of Sample Surveys*. Chapman and Hall.

Valliant, R., Dorfman, A. H. and Royall, R. M. (1999). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.

Wagner, G. G., Burkhauser, R. V. and Behringer, F. (1993). New data bases: The English language public use file of the German Socio-Economic Panel. *The Journal of Human Resources* **28**, 429-433.

Received August 20, 2007; accepted November 29, 2007.

Gordon G. Bechtel
University of Florida and Florida Research Institute
P.O. Box 117155
Gainesville, Florida 32611-7155, USA
bechtel@ufl.edu