Econometrics II

Lecture 7
Discrete Response models

Marie Paul

University of Duisburg-Essen
Ruhr Graduate School in Economics

Summer Semester 2020

Outline of lecture

7. Discrete response models
   - Linear Probability Model
   - Logit and Probit Models
   - Non-linear Panel Data Models
   - Multinomial Models

Self-study plan for Lecture 7

- This lecture presents discrete response models with a focus on model choice and interpretation of results.
- For our next virtual meeting, please work through all slides of Lecture 7. Readings:
- On logit and probit: Cameron and Trivedi (2005) Chapter 14.1 to 14.4. as far as this is covered on the slides.
- On Non-linear Panel Data Models: Wooldridge (2010) Chapter 15.8.1, 15.8.2. and 15.8.4. as far as this is covered on the slides.
- On Multinomial Models: Cameron and Trivedi (2005) Chapter 15.1, 15.2, 15.4., 15.6.2., 15.8.1., 15.9.1. as far as this is covered on the slides.
- Please prepare answers to the "your own research" questions and collect questions and topics to discuss them in our virtual meeting.

On: Censoring, Truncation and Selection

- We will not cover "Censoring, Truncation and Selection".
- I suggest that you flip through Cameron and Trivedi (2005) Chapter 16.1 to 16.7 on Censored and Truncated Models, Sample Selection Models, and the Roy Model to make sure that you know what kinds of models exist should you ever need them.
- If you are interested please consider Angrist and Pischke (2009) on an argument against using Tobit and similar models in particular situations in the subchapter “Good COP, Bad COP: Conditional on Positive Effects” in Chapter 3.4.2.
- Details can be looked up when necessary, but one should know which censoring and selection problems exist to be able to detect them when they appear and have some idea which kind of solutions exist.
- For the examen the Chapters listed on this slide are not relevant.

Outline

7 Discrete response models
- Linear Probability Model
- Logit and Probit Models
- Non-linear Panel Data Models
- Multinomial Models

Definition

A **discrete response** means that the variable to be explained, y, takes on a **finite** number of outcomes. Because the outcome variables are dummies, x affects **probabilities**.

In practice, there are two cases:

1. The response is **binary**, $y = [0, 1]$. An individual is employed $(y = 1)$ or not $(y = 0)$.
2. **Multinomial** responses, $y = [1, 2..., m]$. Self-assessed health (Poor (1), Fair (2), Good (3), Very good (4), or Excellent (5)).

The econometric task is to model and estimate the **conditional response probability** of y

$$p(\mathbf{x}) = \Pr(Y = y|\mathbf{x}). \tag{7.1}$$

In the binary case we have only two outcomes and so

$$y = \begin{cases} 1 & \text{with probability } p(\mathbf{x}) \\ 0 & \text{with probability } 1 - p(\mathbf{x}), \end{cases} \tag{7.2}$$

where $p(\mathbf{x})$ is defined and estimated using regression.

The linear probability model

The **linear probability model (LPM)** for a **binary response** specifies

$$p(\mathbf{x}) = \mathrm{Pr}(y = 1|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}. \tag{7.3}$$

The interest lies in estimating the **marginal impact** of a regressor x_j on the conditional probability of "success"

$$\beta_j = \frac{\partial \mathrm{Pr}(y = 1|\mathbf{x})}{\partial x_j}. \tag{7.4}$$

Since y is **Bernoulli distributed**, its mean and variance are defined by

$$\begin{aligned} \mathrm{E}[y|\mathbf{x}] =& p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} \\ \mathrm{Var}[y|\mathbf{x}] =& p(\mathbf{x})(1 - p(\mathbf{x})) = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}). \end{aligned} \tag{7.5}$$

The implication is that, under (7.3), parameters can be correctly interpreted as marginal effects using OLS but that errors will generally be **heteroscedastic**.

The linear probability model

The LPM/OLS is strictly speaking not the correct model when the outcome is a **discrete qualitative variable** because it

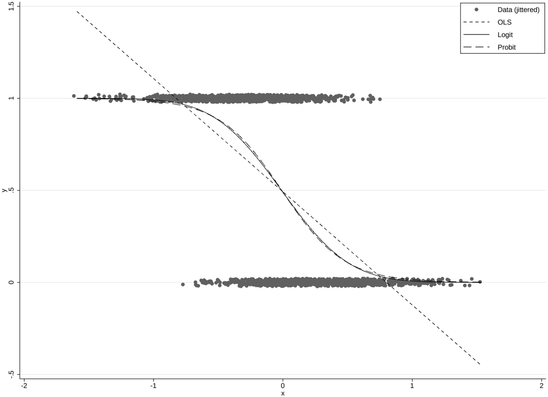
- **Ignores discreteness** and will treat the dependent variable as continuous.
- Does not **constrain predicted probabilities** to lie between zero and one.
- Does not take into account that outcomes might not be **naturally ordered**.

Linear regression does generally not fit the **CEF** $\mathrm{E}[y|\mathbf{x}]$ perfectly, but it is the **best linear predictor**. And **curve-fitting** might not be so important. A **misspecified non-linear model** might be more problematic than the LPM.

Specifically, the LPM may be preferable if

- Interest is in analyzing partial effects **averaged** over the distribution of **x**.
- The x_j take on only a **few values** and the model is **saturated**. Thus, the CEF is (close to) a linear function of the regressor.

Example: Predicted probabilities across models



NOTE.— Model: $y = \mathbf{1}[-1.2x + \mu]$ where $\mu = x = N(0, 1.5)$

Table 7.1. OLS, Logit and Probit estimates of $\boldsymbol{\beta}$

Regressor	**OLS**	**Logit**	**Probit**
Constant	0.49	-0.04	-0.02
	[56.00]	[-0.65]	[-0.77]
x	-0.62	-4.15	-2.41
	[-35.16]	[-22.31]	[-24.58]

LPM or Probit/Logit

For example, Pischke is a staunch defender of the LPM/OLS approach on the MHE blog: `http://www.mostlyharmlesseconometrics.com/2012/07/probit-better-than-lpm/`

Quote: *Structural parameters of a binary choice model, just like the probit index coefficients, are not of particular interest to us. We care about the marginal effects. The LPM will do a pretty good job estimating those. If the CEF is linear, as it is for a saturated model, regression gives the CEF – even for LPM. If the CEF is non-linear, regression approximates the CEF. Usually it does it pretty well. Obviously, the LPM won't give the true marginal effects from the right nonlinear model. But then, the same is true for the "wrong" nonlinear model! The fact that we have a probit, a logit, and the LPM is just a statement to the fact that we don't know what the "right" model is.*

Even if you plan to rely on the LPM you will often provide probit results as a **robustness check**. Furthermore, we need to understand binary discrete choice models to cover more **involved discrete choice models** in which case the LPM might not be an alternative....

LPM, Logit and Probit

Example 7.1 (Your own research)

In your research field: do people use LPM, probit or logit?

Outline

7 Discrete response models
- Linear Probability Model
- Logit and Probit Models
- Non-linear Panel Data Models
- Multinomial Models

The latent index function model

Consider the general **binary response model** of the form

$$p(\mathbf{x}) = \Pr(y = 1|\mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}). \tag{7.6}$$

It is also called an **single index model** due to that it restricts the way the **response probability** depends on the **index**, $\mathbf{x}'\boldsymbol{\beta}$.

The function $F(\cdot)$ **maps** the index to the response probability and is usually defined as a **cumulative distribution function** to ensure that $0 \leqslant p(x) \leqslant 1$.

It is useful to think of y as an indicator for whether a **latent continuous variable** y^* crosses a particular **threshold**. Consider the **index function model**:

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u, \quad y = \mathbf{1}[y^* > 0]. \tag{7.7}$$

Given this we have

$$\begin{aligned} \Pr(y = 1|\mathbf{x}) &= \Pr(y^* > 0|\mathbf{x}) \\ &= \Pr(-u < \mathbf{x}'\boldsymbol{\beta}) \\ &= F_u(\mathbf{x}'\boldsymbol{\beta}), \end{aligned} \tag{7.8}$$

where $F(\cdot)$ is the index mapping cdf of u if u is symmetric around zero.

The random utility model

The latent variable framework can have an **economic** interpretation based on **choice theory** via the **random utility model**.

Specify the **utilities** from an individual choosing between alternatives $y = [0, 1]$:

$$\begin{aligned} U_0 &= V_0 + \varepsilon_0 \\ U_1 &= V_1 + \varepsilon_1, \end{aligned} \tag{7.9}$$

where U_y constitutes the utility derived from choosing alternative $Y = y$, which depends on a **deterministic** (V_y) and a **random** (ε_y) component.

Here, $U_1 - U_0$ can be thought of as the **latent component** y^* in (7.7) from which we observe $y = \mathbf{1}[U_1 > U_0]$ for a utility-maximizing agent.

Hence, individuals maximize utility according to

$$\begin{aligned} \Pr[y = 1] &= \Pr[U_1 > U_0] \\ &= \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0] \\ &= F_u(\mathbf{x}'\boldsymbol{\beta}), \end{aligned} \tag{7.10}$$

where $u = \varepsilon_0 - \varepsilon_1$ and $\mathbf{x}'\boldsymbol{\beta} = V_1 - V_0$.

Parameterizations: LPM, Logit and Probit

The distributional assumptions on the errors in the latent index model will define the mapping function F and therefore also the model.

The most common specifications of F for index model z are:

1. The **identity** function used in the **linear probability model**

$$F(z) = z. \tag{7.11}$$

2. The **standard cumulative normal distribution** in the **probit model**

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v) dv. \tag{7.12}$$

   Or, when u in the latent formulation is **normally distributed**.

3. The **standard logistic distribution** in the **logit model**

$$F(z) = \Lambda(z) = \exp(z)/[1 + \exp(z)]. \tag{7.13}$$

   Or, when u in the latent formulation is **logistically distributed**.

Identification

- The maximum likelihood estimator is **consistent** if the CEF is correctly specified: $E[y_i|\mathbf{x}_i] = F(\mathbf{x}_i'\boldsymbol{\beta})$.
- **Identification** of the single-index model requires a **restriction of the variance** of u to secure uniqueness of β. In the probit model the error variance is set to one.
- Also the **mean** of the error distribution needs to be **normalized**: usually to zero.

Maximum Likelihood Estimation: Bernoulli Case

Consider a random sample $x_1, ..., x_N$ from a **parametric family of distributions** $f_0(\cdot) \in \{f(\cdot|\theta),\ \theta \in \Theta\}$ where $f_0 = f(\cdot|\theta_0)$.

The task is to find an estimator $\widehat{\theta}$ that is as close as θ_0 as possible.

Specify the **joint density function** for all iid N observations as

$$f(x_1, x_2, ..., x_N|\theta) = f(x_1|\theta) \times f(x_2|\theta)\times, ..., \times f(x_N|\theta). \tag{7.14}$$

Then the **log likelihood** that the x were generated by a specific θ is

$$\mathcal{L}(\theta; x_1, ..., x_N) = \ln f(x_1, x_2, ..., x_N|\theta) = \sum_{i=1}^{N} \ln f(x_i|\theta). \tag{7.15}$$

The **ML estimator** is the value of θ that maximizes the likelihood function:

$$\widehat{\theta}_{ML} = \underset{\theta \in \Theta}{\arg\max}(\mathcal{L}(\theta; x_1, ..., x_N)). \tag{7.16}$$

Maximum Likelihood Estimation: Bernoulli Case

The parameters of the discrete outcome models are estimated with ML.

In the binary case, given data $(y_i, \mathbf{x}_i)$ and cdf $p(\mathbf{x}_i) = F(\mathbf{x}'\boldsymbol{\beta})$, y is **Bernoulli distributed** with log density

$$\ln f(y_i|\mathbf{x}_i) = y_i \ln p(\mathbf{x}_i) + (1 - y_i)\ln(1 - p(\mathbf{x}_i)), \quad y_i = 0, 1 \tag{7.17}$$

and **log-likelihood function**

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i)\ln(1 - F(\mathbf{x}_i'\boldsymbol{\beta})) \right\}. \tag{7.18}$$

Differentiating the function wrt $\boldsymbol{\beta}$, the **ML estimator** $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ solves $(F_i \equiv F(\mathbf{x}_i'\boldsymbol{\beta}))$

$$\sum_{i=1}^{N} \left\{ \frac{y_i}{F_i} F_i' \mathbf{x}_i - \frac{1 - y_i}{1 - F_i} F_i' \mathbf{x}_i \right\} = \sum_{i=1}^{N} \frac{y_i - F_i}{F_i(1 - F_i)} F_i' \mathbf{x}_i = \mathbf{0}. \tag{7.19}$$

Maximum Likelihood Estimation: Bernoulli Case

No explicit solution exists for $\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}$ but the log-likelihood is **globally concave** $(F''(z) < 0$ for all $z)$ for both probit and logit models.

The MLE is **consistent** if the cdf of y given $\mathbf{x}$ is correctly specified which we know is Bernoulli distributed so consistency hinges on that $p(\mathbf{x}_i)$ is correctly specified.

For binary data, $\mathsf{E}[y_i|\mathbf{x}_i] = 1 \times p(\mathbf{x}_i) + 0 \times (1 - p(\mathbf{x}_i)) = p(\mathbf{x}_i) \equiv F(\mathbf{x}_i'\boldsymbol{\beta})$ implying that (7.19) has expected value zero and hence consistent.

Given correct density specification the MLE is distributed as

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ML}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, (-\mathsf{E}[\partial^2 \mathcal{L}_N / \partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'])^{-1}], \tag{7.20}$$

with asymptotic variance matrix

$$\widehat{\mathsf{V}}[\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}] = \left(\sum_{i=1}^{N} \frac{1}{F(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})(1 - F(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}))} F'(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}^2 \mathbf{x}_i\mathbf{x}_i') \right)^{-1}. \tag{7.21}$$

Marginal effects

In all binary response models, the main interest lies in estimating the **marginal effect** of a regressor on the conditional **probability** $\Pr[y = 1|\mathbf{x}]$.

For the **general** binary outcome model in (7.6), the marginal effect of change in a regressor j is equal to

$$\frac{\partial \Pr[y = 1|\mathbf{x}]}{\partial x_j} = F'(\mathbf{x}'\boldsymbol{\beta})\beta_j. \tag{7.22}$$

For **non-linear** models, where $F'(\mathbf{x}'\boldsymbol{\beta}) \neq c$ the marginal effect will vary with the **evaluation point x** and the **choice** of $F(\cdot)$.

Some common approaches are to evaluate at:

1. The sample average of the **marginal effects** (AME): $N^{-1}\sum_i F'(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$.
2. The sample average of the **regressors** (MEM): $F'(\bar{\mathbf{x}}'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$.
3. The marginal effect at a **representative point** $\mathbf{x}_1$ (MER): $F'(\mathbf{x}_1'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$.

Marginal effects and regression coefficients

The general marginal effect in the binary response model is

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = F'(\mathbf{x}'\boldsymbol{\beta})\beta_j, \quad F'(z) \equiv \frac{dF}{dz}(z). \tag{7.23}$$

i.e., the regression coefficient scaled by the index evaluated at a particular point.

Since $F(\cdot)$ is **strictly increasing**, $F'(z) > 0$ for all z and the **sign of the effect** is given by the sign of β_j.

Also, **relative effects** in single-index models does not depend on **x** since

$$\frac{\partial p(\mathbf{x})/\partial x_j}{\partial p(\mathbf{x})/\partial x_h} = \frac{F'(\mathbf{x}'\boldsymbol{\beta})\beta_j}{F'(\mathbf{x}'\boldsymbol{\beta})\beta_h} = \beta_j/\beta_h. \tag{7.24}$$

Table 7.2. Common binary outcome models

Model	**Probability** ($p = \Pr[y = 1\|\mathbf{x}_i]$)	**Marginal Effect** ($\partial p(\mathbf{x}_i)/\partial x_j$)
Linear Probability	$\mathbf{x}'\boldsymbol{\beta}$	β_j
Logit	$\Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$	$\Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_j$
Probit	$\Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)dz$	$\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$

The Logit model

The logit (logistic regression) model models the probability of an event as

$$p(\mathbf{x}) = \Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{1}{1+e^{-\mathbf{x}'\boldsymbol{\beta}}} \equiv \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}, \tag{7.25}$$

where $\Lambda(\cdot)$ is the **logistic cdf**.

The logit MLE FOC from (7.19) is simply (note: $\Lambda'(z) = \Lambda(z)[1-\Lambda(z)]$)

$$\sum_{i=1}^{N}(y_i - \Lambda(\mathbf{x}'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}. \tag{7.26}$$

The **marginal effects** are obtained by $\partial p(\mathbf{x}_i)/\partial x_{ij} = p(\mathbf{x}_i)(1-p(\mathbf{x}_i))\beta_j$. It is also common to interpret the coefficients as marginal effects on the **odds ratios**

$$\begin{aligned} p(\mathbf{x}) &= \exp(\mathbf{x}'\boldsymbol{\beta})/(1+\exp(\mathbf{x}'\boldsymbol{\beta})) \\ \Rightarrow \ln\frac{p(\mathbf{x})}{1-p(\mathbf{x})} &= \mathbf{x}'\boldsymbol{\beta}, \end{aligned} \tag{7.27}$$

where $p(\mathbf{x})/(1-p(\mathbf{x}))$ is the **relative risk** or **odds ratio**. So this gives the effect of a one unit change in x on the predicted odds ratio.

The Probit model

The probit model models the probability of an event as

$$p(\mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)dz, \tag{7.28}$$

where $\Phi(\cdot)$ is the normal cdf with density $\phi(z) = (1/\sqrt{2\pi})\exp(-z^2/2)$.

The probit MLE FOC is

$$\sum_{i=1}^{N} w_i(y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}, \tag{7.29}$$

where the **weights** $w_i = \phi(\mathbf{x}'\boldsymbol{\beta})/[\Phi(\mathbf{x}'\boldsymbol{\beta})(1-\Phi(\mathbf{x}'\boldsymbol{\beta}))]$ varies across observations (unlike the logit model).

The **marginal effects** are obtained by

$$\frac{\partial p(\mathbf{x}_i)}{x_{ij}} = \phi(\mathbf{x}'\boldsymbol{\beta})\beta_j, \tag{7.30}$$

which is more **complicated** than the logit model. It is still widely used since normal errors are often assumed in the **latent regression model**.

Model choice: Logit or probit

The choice of model depends on the unknown dgp. We know the **distribution** of the process (Bernoulli), but not the functional form of its **parameter**, p.

Hence, consistency of the MLE depends critically on the model's ability to correctly specify the conditional probability.

Empirically, the difference is typically very small when comparing **fitted log-likelihood** from the different specifications

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_i \{y_i \ln \widehat{p}(\mathbf{x}) + (1 - y_i) \ln(1 - \widehat{p}(\mathbf{x}))\}, \tag{7.31}$$

where $\widehat{p}(\mathbf{x})$ is either $\Lambda(\mathbf{x}'\widehat{\boldsymbol{\beta}}_{\text{Logit}})$ or $\Phi(\mathbf{x}'\widehat{\boldsymbol{\beta}}_{\text{Probit}})$.

Coefficients in the logit model are simpler to **interpret** while the **latent framework** make the probit model attractive in selection models.

Outline

7. Discrete response models
   - Linear Probability Model
   - Logit and Probit Models
   - Non-linear Panel Data Models
   - Multinomial Models

Motivation

The following slides are mainly based on Wooldridge (2010).

- It is not entirely clear how well the **LPM** works for estimating marginal effects for panel-data models. One sometimes sees **FE estimation of the LPM**, but e.g. Wooldridge notes in addition to the usual downsides this implies the unnatural restrictions $\mathbf{x}'\boldsymbol{\beta} \leqslant c_i \leqslant \mathbf{x}'\boldsymbol{\beta}$.
- The result for linear panel-data models that marginal effects can be consistently estimated without making assumptions about the **relationsship between $\mathbf{x}_i$ and c_i** does not hold for **non-linear models** → **incidental parameter problem**.
- Options like the **fixed-effects logit** estimator and "fixed-effects probit" estimators with **bias corrections** generally involve important downsides.
- A widely used non-linear panel data model is **(Chamberlains's correlated) (dynamic) random effects probit model** which we cover here.

Random Effects Probit

Assumptions:

1. **Strict exogeneity** conditional on the unoberved effect (in terms of conditional distributions):

$$D(y_{it}|\mathbf{x}_i, c_i) = D(y_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i) = D(y_{it}|\mathbf{x}_{it}, c_i), \quad t = 1, \ldots, T. \tag{7.32}$$

2. **Probit** specification:

$$P(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \quad t = 1, \ldots, T. \tag{7.33}$$

3. $y_{i1}, ..., y_{i1}$ are independent conditional on $(\mathbf{x}_i, c_i)$. (Analogous to assumption on serially uncorrelated errors in the linear case.)
4. Traditional **random effects probit assumption**: c_i and $\mathbf{x}_i$ independent and c_i has normal distribution:

$$c_i|\mathbf{x}_i \sim \mathcal{N}[0, \delta_c^2]. \tag{7.34}$$

Random Effects Probit

- Relative **importance** of unobserved effect is $\frac{\delta_c^2}{\delta_c^2+1}$.
- Estimate average **marginal effect**. Usually at $c = 0$.
- **Chamberlain-Mundlak approach**: relaxes random effects probit assumption allowing unobservables to be **correlated with some xses** by specifying:

$$c_i|\mathbf{x}_i \sim \mathcal{N}[\psi + \bar{\mathbf{x}}_i\zeta, \delta_a^2] \tag{7.35}$$

Easy to implement: Add **time averages** of selected xses to the regression. This holds time averages fixed → remember the linear correlated random effects model.

Dynamic Random Effects Probit

- The **dynamic model** adds lags of the dependent variable to model **state-depencence** and duration dependence.
- Hyslop (1999) has first shown that in labor supply models it is **crucial** for the random effects probit **assumption** to hold to account for state dependence.
- When building the joint distribution for the likelihood function we need (as usual) to integrate out c. This raises the problem how we treat y_{i0}: **initial condition problem**.
- If we treat it as a **nonstochastic starting point**, we assume that y_{i0} and c_i are independent. This is **undesirable** because if c_i influences y in later periods, it will also in this inital period (even if that is in some sense the start of a process, e.g. first year after school.).
- **Woodridge** suggests to solve the initial condition problem analogously to the Chamberlain-Mundlak approach and thus assuming a **linear dependence** structure. Thus, **add** y_{i0} among the regressors for each period.

This is finally **Chamberlains's correlated dynamic random effects probit model** which is often used e.g. in labor economics.

RE Probit with Endogenous Selection

Example 7.2 (Fitzenberger, Osikominu, Paul (2019): The Effects of Training Incidence and Planned Training Duration on Labor Market Transitions)

$$E_t^* = \psi_E(t, E^{t-1}, Q^{t-1}, P_{t-1}, X_t)'\gamma_E + \alpha_E + \varepsilon_{E,t} \quad \text{for } t = 1, \ldots, \bar{T} \quad (7.36)$$

$$Q_t^* = \psi_Q(t, E^t, Q^{t-1}, P_{t-1}, X_t)'\gamma_Q + \alpha_Q + \varepsilon_{Q,t} \quad (7.37)$$

$$\text{for } E_t = 0 \text{ and } \{S \geqslant t \text{ or } (S < t, Q_{t-1} = 1)\},\ t = 0, 1, \ldots, \bar{T}\,.$$

Outline

7. Discrete response models
   - Linear Probability Model
   - Logit and Probit Models
   - Non-linear Panel Data Models
   - Multinomial Models

Definition

The previous section covered the case where the outcome variable was **binary**.

We now consider the case where it can assume **more than two** (mutually exclusive) values: **the multinomial case**:

- Ways to commute: Bus, car, or walk.
- Health status: Poor, good, or excellent.
- Employment status: full-time, part-time, or none.

Just as the outcome was **binomially** distributed in the **binary** case, it is **multinomially** distributed in the **multinomial** case.

Maximum likelihood is used for **estimation** since we know the distribution of the outcome.

Estimation problems arise because of the **unknown** and potentially complex **nested** form of the probabilities of the multinomial distribution.

The general multinomial model

The **estimation problem** in multinomial models consists in obtaining a consistent estimate of the **probabilities** in the model.

Compared to the binary case we now have m **alternatives** denoted by $y = j$ if alternative j is taken so that

$$p_j = \text{Pr}[y = j], \quad j = 1, ..., m. \tag{7.38}$$

This is equivalent to a binary setting with m binary **variables** for each y

$$y_j = \begin{cases} 1 & \text{if } y = j \\ 0 & \text{if } y \neq j, \end{cases} \tag{7.39}$$

so that y_j equals one if alternative j is chosen and remaining $m - 1$ y_k are zero.

We can then write the **multinomial density** for one observation as

$$f(y) = p_1^{y_1} \times \cdots \times p_m^{y_m} = \prod_{j=1}^{m} p_j^{y_j}. \tag{7.40}$$

The general multinomial model

Introducing **heterogeneity**, we specify for individual i and regressors $\mathbf{x}_i$ a model for the probability that individual i chooses the jth **alternative**

$$p_{ij} = \Pr[y_i = j] = F_j(\mathbf{x}_i, \boldsymbol{\beta}), \quad j = 1, ..., m, \quad i = 1, .., N. \tag{7.41}$$

The F_j function should satisfy $p_{ij} = [0, 1]$ and $\sum_j^m p_{ij} = 1$ for all i.

As in the binary case, different **specifications** of F_j correspond to **specific models** which we will discuss in the following:

- **Multinomial probit/logit:** The simplest case further categorized into the **conditional**, **multinomial**, and **mixed** probit and logit models.
- **Nested logit:** Models the **sequential nature** of decision making when choices are sequentially **depending** on each other.
- **Ordered outcome models:** Uses the **natural ordering** of alternatives (if such exists) to model the probabilities.

The general multinomial model

Summing all N observations in (7.40) we can form the likelihood function $L_N = \prod_{i=1}^{N} \prod_{j=1}^{m} p_{ij}^{y_{ij}}$. Taking logs we obtain the **log-likelihood function**

$$\mathcal{L}_N(\boldsymbol{\beta}) = \ln L_n = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij}, \tag{7.42}$$

where $p_{ij}^{y_{ij}} = F_j(\mathbf{x}_i, \boldsymbol{\beta})$.

First order conditions for the MLE $\widehat{\boldsymbol{\beta}}$ solves

$$\frac{\partial \mathcal{L}_N}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \sum_{j=1}^{m} \frac{y_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{7.43}$$

If $F_j(\mathbf{x}_i, \boldsymbol{\beta})$ is correctly specified, then $\mathsf{E}[y_{ij}] = p_{ij}$, implying that (7.43) will hold since $\sum_{j=1}^{m} p_{ij} = 1$. The MLE is then distributed as

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ML}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, (-\mathsf{E}[\partial^2 \mathcal{L}_N / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'])^{-1}]. \tag{7.44}$$

Multinomial Models

Example 7.3 (Your own research)

How are / could multinomial models be applied in your research area?

Multinomial logit

The **multinomial logit model** is a class of models with the general form

$$p_{ij} = \frac{V_{ij}}{\sum_{l=1}^{m} V_{il}}, \tag{7.45}$$

where the V_{ij} are general functions of regressors $\mathbf{x}_i$ and parameters $\boldsymbol{\beta}$.

The simplest variations of this model distinguishes between **alternative-varying** (e.g., travel time) or **alternative-constant** (e.g., age) regressors:

- **Multinomial logit (MNL):** $V_{ik} = \exp(\mathbf{x}_i \boldsymbol{\beta}_k),$ for $k = (j, l)$.
- **Conditional logit (CL):** $V_{ik} = \exp(\mathbf{x}_{ik} \boldsymbol{\beta}),$ for $k = (j, l)$.
- **Mixed logit:** $V_{ik} = \exp(\mathbf{x}_{ik} \boldsymbol{\beta} + \mathbf{w}_i \gamma_k),$ for $k = (j, l)$.

From (7.43) the ML first order conditions for MNL and CL are

$$\frac{\partial \mathcal{L}_N}{\partial \boldsymbol{\beta}_k} = \begin{cases} \text{MNL:} & \sum_{i=1}^{N} (y_{ik} - p_{ik}) \mathbf{x}_i = 0 \\ \text{CL:} & \sum_{i=1}^{N} y_{ik} (\mathbf{x}_{ik} - \bar{\mathbf{x}}_i) = 0 \end{cases}, \tag{7.46}$$

for $k = 1, ..., m$ where $\bar{\mathbf{x}}_i = \sum_{l=1}^{m} p_{il} \mathbf{x}_{il}$.

Marginal effects in the multinomial logit

Care is needed in the **interpretation** of parameters in any non-linear model, but **in particular** in multinomial models.

For the CL model, consider the effect on the jth **probability** of changing by one unit the value of the **regressor** for the kth **alternative**.

$$\begin{aligned}\frac{\partial p_{ij}}{\partial \mathbf{x}_{ik}} &= \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}}{\sum_{l=1} e^{\mathbf{x}_{il}'\boldsymbol{\beta}}}\boldsymbol{\beta} - \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}}{\left(\sum_{l=1} e^{\mathbf{x}_{il}'\boldsymbol{\beta}}\right)^2} e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}\boldsymbol{\beta} \\ &= p_{ij}(\delta_{ijk} - p_{ik})\boldsymbol{\beta} \qquad (7.47)\end{aligned}$$

where $\delta_{ijk} = 1[j = k]$.

For the MNL model we have

$$\frac{\partial p_{ij}}{\partial \mathbf{x}_i} = p_{ij}(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_i) \qquad (7.48)$$

where $\bar{\boldsymbol{\beta}}_i = \sum_l p_{il}\boldsymbol{\beta}_l$ is a probability weighted average of the $\boldsymbol{\beta}_l$.

Marginal effects in the multinomial logit

Example 7.4 (Choice of Fishing Mode)

- Choose between four mutually exclusive modes of fishing: beach, pier, private boat, charter boat.
- Regressors price and catch rate vary by fishing mode and by individual → CL model (coefficient does not vary by mode but marginal effect does!) → marginal effect if x increases for one alternative and remains constant for the others: an increase of beach fishing by \$ 100 leads to a decrease of 0.272 in the probability beach fishing and an increase of 0.119 of pier fishing and so on....
- Regressor individual income does not vary with fishing mode → MNL model (coefficient and intercept for each mode, normalized to zero for one mode) → interpretation relative to base category (here: beach): a \$ 1000 increase in monthly income e.g. associated with an 0.033 increase in probability in fishing from private boat realative to beach fishing.
- The Mixed model combines both models.

Multinomial random utility models

The random utility model is useful in the multinomial case to have an **economic foundation** for the choice of model.

In the multinomial case with m choices, the random utility model specifies

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \quad j = 1, ..., m, \tag{7.49}$$

where the utility U_{ij} from choosing alternative j is decomposed into a **deterministic** (V_{ij}) and a **random** (ε_{ij}) part as in the binary case.

Individuals choose the option with the **highest utility** according to

$$\begin{aligned}
\Pr[y = j] &= \Pr[U_j \geqslant U_k, \text{ all } k \neq j] \\
&= \Pr[U_k - U_j \leqslant 0, \text{ all } k \neq j] \\
&= \Pr[\widetilde{\varepsilon}_{kj} \leqslant -\widetilde{V}_{kj}, \text{ all } k \neq j] \\
&= \int_{-\infty}^{-\widetilde{V}_{mj}} \int_{-\infty}^{-\widetilde{V}_{(m-1)j}} \cdots \int_{-\infty}^{-\widetilde{V}_{1j}} f(\widetilde{\varepsilon}_{1j}, \widetilde{\varepsilon}_{2j}, ..., \widetilde{\varepsilon}_{mj}) d\widetilde{\varepsilon}_{1j}, d\widetilde{\varepsilon}_{2j}, ..., d\widetilde{\varepsilon}_{mj}
\end{aligned} \tag{7.50}$$

where $\widetilde{q}_{kj} = q_k - q_j$ for $q = (V, \varepsilon)$ and the last equality is the $(m-1)$ integral over the joint density of the errors.

Multinomial random utility models

In general, the choice probabilities in (7.50) have no **closed-form solution** and therefore additional assumptions are imposed on the errors, ε.

The most natural assumption is to assume that $\varepsilon_1, \dots, \varepsilon_m$ are joint normal distributed, yielding the **multinomial probit model**.

Another common approach is to assume that the errors are distributed as **iid type 1 extreme** with density

$$f(\varepsilon_j) = e^{-\varepsilon_j} \exp(-e^{-\varepsilon_j}), \tag{7.51}$$

ε_j does not depend on the errors of the other alternatives. This specification simplifies the probabilities for choosing alternative j to

$$\Pr[y = j] = \frac{e^{V_j}}{e^{V_1} + e^{V_2} + \cdots + e^{V_m}}. \tag{7.52}$$

This is the **CL model** when $V_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$ or the **MNL model** when $V_{ij} = \mathbf{x}_i\boldsymbol{\beta}_j$.

While this is a convenient simplification, it hinges on the assumption of **uncorrelated errors across alternatives** which is potentially very restrictive (**independence of irrelevant alternatives (IIA)** e.g., "red bus-blue bus").

Multinomial Probit

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \quad j = 1, ..., m, \tag{7.53}$$

with errors following a **joint normal**. Usually regressors vary across individuals (sometimes coefficients).

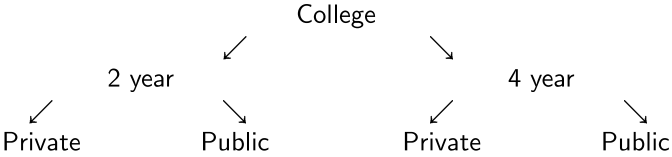
- In general this allows for **correlations of errors** across all alternatives. But **restrictions** need to be put on the covariance matrix.
- If **off-diagonal elements** are restricted to be zero, the model still has no closed-form solution, but obviously does **not allow for errors to be correlated** across alternatives. This is what mprobit in Stata does! MNL is then a better alternative...
- One way to **acchieve identification** is to set the error of the first mode to zero and at least one additional covariance element to one.
- MNP estimates are often imprecise and **not robust** at all. It is therefore strongly suggested to work with **exclusion restrictions** (i.e. regressors that appear only for one mode).
- Estimation is difficult as in the most general case an (m-1)-fold integral needs to be solved. For three choice numerical methods may work, for higher choices use **simulation methods** (e.g. MCMC).

Nested logit

The MNL model is equivalent to a series of **pairwise comparisons** assumed to be independent of all other alternatives than those under consideration.

The **nested logit model** is a **analytically tractable** model that relaxes the assumption of iid errors implied by the IIA assumption.

The nested logit is used when choices are **nested sequentially** such as, for example, when choosing college education.



(7.54)

The errors in this model are allowed to be correlated for each option **within** groups but uncorrelated **across** groups.

Ordered outcome models

Outcomes are sometimes naturally **ordered** such as in the case of **self-rated health** which may be rated from poor to excellent.

While the previous models are still applicable, it makes more sense to take such ordering into account in an **ordered model**.

To do this, consider the latent variable model

$$y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + u_i, \tag{7.55}$$

where, as y_i^* crosses a sequence of **unobserved thresholds** α_j, shifts it up in its ordering (health goes from poor to fair when $y^* > \alpha_{\text{poor}}$ etc.).

In general for a m-alternative **ordered model**,

$$y_i = j \quad \text{if } \alpha_{j-1} < y_i^* \leqslant \alpha_j, \tag{7.56}$$

where $\alpha_0 = -\infty$ and $\alpha_m = \infty$ corresponds to the respective endpoints.

Ordered outcome models

The probability of choosing alternative j is then

$$\begin{aligned}\Pr[y_i = j] &= \Pr[\alpha_{j-1} < y_i^* \leqslant \alpha_j] \\ &= \Pr[\alpha_{j-1} < \mathbf{x}_i'\boldsymbol{\beta} + u_i \leqslant \alpha_j] \\ &= \Pr[\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} < u_i \leqslant \alpha_j - \mathbf{x}_i'\boldsymbol{\beta}] \\ &= F_u(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}) - F_u(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta})\end{aligned} \tag{7.57}$$

u is assumed to be normally distributed with in the **ordered probit model** and logistically distributed in the **ordered logit model**.

The model parameters $\boldsymbol{\beta}$ and the $(m-1)$ threshold parameters $a_1, \ldots, \alpha_{m-1}$ can then be estimated with ML using (7.40); i.e.,

$$\mathcal{L} = \ln L_N = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij} \tag{7.58}$$

with $p_{ij} = \Pr[y_i = j]$ as in (7.57).

Choice of multinomial model

As in the binary case we know the distribution of the **dgp** (multinomial) but we still have to assume a functional form for the **probabilities**.

The **multinomial logit model** is mostly useful for **data description** owing to that the **independence of irrelevant alternatives** assumption is unlikely to hold.

The **nested logit model** is preferable when there is a clear nesting structure, but such structures are often **not obvious**.

If outcomes are **ordered**, the **ordered logit or probit model** are preferred as they efficiently include the **ordering structure** in the estimation.

The **multinomial probit** model and alternative models not covered are more **challenging to estimate**.

Paul (2016, SJOE): Multinomial probit with four alternatives and random effects

Example 7.5 (Selected Regressors)

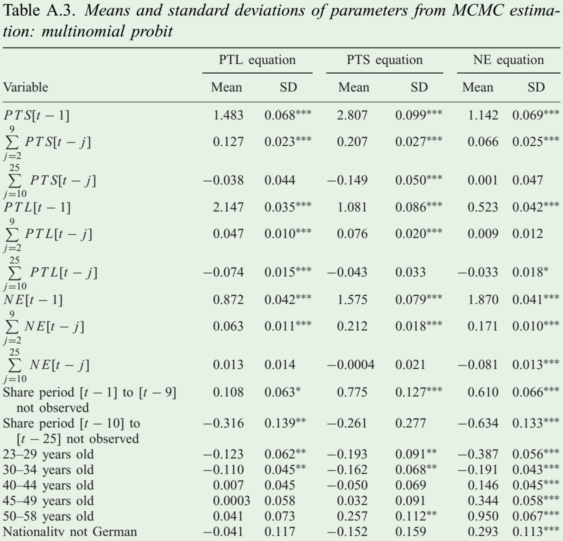
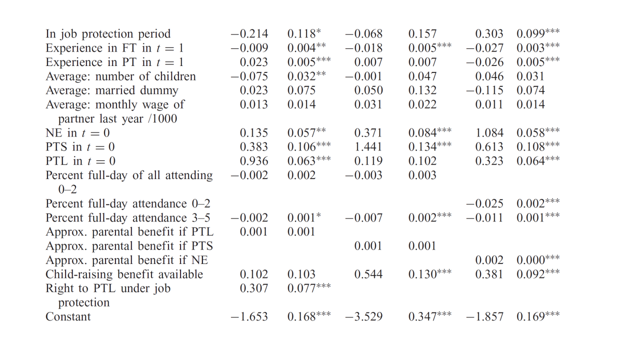


Table A.3. *Means and standard deviations of parameters from MCMC estimation: multinomial probit*

	PTL equation		PTS equation		NE equation	
Variable	Mean	SD	Mean	SD	Mean	SD
$PTS[t-1]$	1.483	0.068***	2.807	0.099***	1.142	0.069***
$\sum_{j=2}^{9} PTS[t-j]$	0.127	0.023***	0.207	0.027***	0.066	0.025***
$\sum_{j=10}^{25} PTS[t-j]$	−0.038	0.044	−0.149	0.050***	0.001	0.047
$PTL[t-1]$	2.147	0.035***	1.081	0.086***	0.523	0.042***
$\sum_{j=2}^{9} PTL[t-j]$	0.047	0.010***	0.076	0.020***	0.009	0.012
$\sum_{j=10}^{25} PTL[t-j]$	−0.074	0.015***	−0.043	0.033	−0.033	0.018*
$NE[t-1]$	0.872	0.042***	1.575	0.079***	1.870	0.041***
$\sum_{j=2}^{9} NE[t-j]$	0.063	0.011***	0.212	0.018***	0.171	0.010***
$\sum_{j=10}^{25} NE[t-j]$	0.013	0.014	−0.0004	0.021	−0.081	0.013***
Share period $[t-1]$ to $[t-9]$ not observed	0.108	0.063*	0.775	0.127***	0.610	0.066***
Share period $[t-10]$ to $[t-25]$ not observed	−0.316	0.139**	−0.261	0.277	−0.634	0.133***
23–29 years old	−0.123	0.062**	−0.193	0.091**	−0.387	0.056***
30–34 years old	−0.110	0.045**	−0.162	0.068**	−0.191	0.043***
40–44 years old	0.007	0.045	−0.050	0.069	0.146	0.045***
45–49 years old	0.0003	0.058	0.032	0.091	0.344	0.058***
50–58 years old	0.041	0.073	0.257	0.112**	0.950	0.067***
Nationality not German	−0.041	0.117	−0.152	0.159	0.293	0.113***

Paul (2016, SJOE): Multinomial probit with four alternatives and random effects

Example 7.6 (Selected Regressors: exclusions restrictions and ...)



In job protection period	−0.214	0.118*	−0.068	0.157	0.303	0.099***
Experience in FT in $t = 1$	−0.009	0.004**	−0.018	0.005***	−0.027	0.003***
Experience in PT in $t = 1$	0.023	0.005***	0.007	0.007	−0.026	0.005***
Average: number of children	−0.075	0.032**	−0.001	0.047	0.046	0.031
Average: married dummy	0.023	0.075	0.050	0.132	−0.115	0.074
Average: monthly wage of partner last year /1000	0.013	0.014	0.031	0.022	0.011	0.014
NE in $t = 0$	0.135	0.057**	0.371	0.084***	1.084	0.058***
PTS in $t = 0$	0.383	0.106***	1.441	0.134***	0.613	0.108***
PTL in $t = 0$	0.936	0.063***	0.119	0.102	0.323	0.064***
Percent full-day of all attending 0–2	−0.002	0.002	−0.003	0.003		
Percent full-day attendance 0–2					−0.025	0.002***
Percent full-day attendance 3–5	−0.002	0.001*	−0.007	0.002***	−0.011	0.001***
Approx. parental benefit if PTL	0.001	0.001				
Approx. parental benefit if PTS			0.001	0.001		
Approx. parental benefit if NE					0.002	0.000***
Child-raising benefit available	0.102	0.103	0.544	0.130***	0.381	0.092***
Right to PTL under job protection	0.307	0.077***				
Constant	−1.653	0.168***	−3.529	0.347***	−1.857	0.169***

Paul (2016, SJOE): Multinomial probit with four alternatives and random effects

Example 7.7 (Variance Parameters)

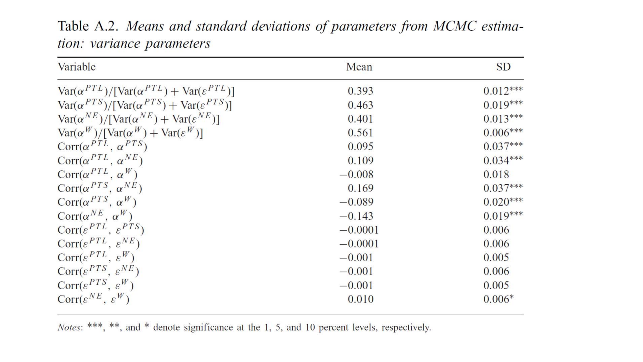


Table A.2. *Means and standard deviations of parameters from MCMC estimation: variance parameters*

Variable	Mean	SD
$\mathrm{Var}(\alpha^{PTL})/[\mathrm{Var}(\alpha^{PTL}) + \mathrm{Var}(\varepsilon^{PTL})]$	0.393	0.012***
$\mathrm{Var}(\alpha^{PTS})/[\mathrm{Var}(\alpha^{PTS}) + \mathrm{Var}(\varepsilon^{PTS})]$	0.463	0.019***
$\mathrm{Var}(\alpha^{NE})/[\mathrm{Var}(\alpha^{NE}) + \mathrm{Var}(\varepsilon^{NE})]$	0.401	0.013***
$\mathrm{Var}(\alpha^{W})/[\mathrm{Var}(\alpha^{W}) + \mathrm{Var}(\varepsilon^{W})]$	0.561	0.006***
$\mathrm{Corr}(\alpha^{PTL}, \alpha^{PTS})$	0.095	0.037***
$\mathrm{Corr}(\alpha^{PTL}, \alpha^{NE})$	0.109	0.034***
$\mathrm{Corr}(\alpha^{PTL}, \alpha^{W})$	−0.008	0.018
$\mathrm{Corr}(\alpha^{PTS}, \alpha^{NE})$	0.169	0.037***
$\mathrm{Corr}(\alpha^{PTS}, \alpha^{W})$	−0.089	0.020***
$\mathrm{Corr}(\alpha^{NE}, \alpha^{W})$	−0.143	0.019***
$\mathrm{Corr}(\varepsilon^{PTL}, \varepsilon^{PTS})$	−0.0001	0.006
$\mathrm{Corr}(\varepsilon^{PTL}, \varepsilon^{NE})$	−0.0001	0.006
$\mathrm{Corr}(\varepsilon^{PTL}, \varepsilon^{W})$	−0.001	0.005
$\mathrm{Corr}(\varepsilon^{PTS}, \varepsilon^{NE})$	−0.001	0.006
$\mathrm{Corr}(\varepsilon^{PTS}, \varepsilon^{W})$	−0.001	0.005
$\mathrm{Corr}(\varepsilon^{NE}, \varepsilon^{W})$	0.010	0.006*

Notes: ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.