Microeconometrics

Preliminaries

by Vanessa Berenguer-Rico

University of Oxford

Michaelmas Term 2016

Microeconometric Analysis

- **Microeconometric Analysis**: "The analysis of individual-level data on the economic behavior of individuals or firms," Cameron and Trivedi, 2005

- **Microeconomic Data**: cross-sections or panel data

- **Cross-sectional Data**: "consists of a sample of individuals taken at a given point in time," Wooldrige, 2013

- **Panel Data**: "A panel data (or longitudinal data) set consists of a time series for each cross-sectional member in the data set," Wooldrige, 2013

- **In this course**: Cross-sectional data

Cross-sectional Data

- Cross-sectional data. Examples:
- **California Test Score**: (Stock and Watson, 2012)

$$Data : tscore_i, str_i, \exp en_i, eng_i$$

- **Wage Equations**: (Wooldrige, 2013)

$$Data : w_i,\ educ_i,\ \exp er_i,\ female_i,\ married_i$$

- **Labor Force Participation**: (Wooldrige, 2013)

$$Data : inlf_i, nwifeinc_i, educ_i, \exp er_i, age_i, kidslt6_i, kidsge6_i$$

- **Crime**: (Wooldrige, 2013)

$$Data : crime_i, wage_i, othinc_i, freqarr_i, freqconv_i, avgsen_i, age_i$$

Cross-sectional Data: California Test Score

- **California Test Score**: Data: Stock and Watson (p. 51)
  - $tscore_i$: average of the math and science test scores for all fifth grades in 1999 in district i
  - str_i : average student-teacher ratio in district i
  - $\exp en_i$: average expenditure per pupil
  - eng_i: percentage of students still learning English

Cross-sectional Data: Wage Equations

- **Wage Equations**: Data: Wooldrige (p. 218)
  - w_i : hourly wage
  - $educ_i$: years of formal education
  - $\exp er_i$: years of workforce experience
  - $female_i$: 1 if person i is female, otherwise
  - $married_i$: 1 if person i is married, otherwise

Cross-sectional Data: Labor Force Participation

- **Labor Force Participation**: Data: Wooldrige (p. 239)
  - $inlf_i$: 1 if woman i reports working for a wage outside the home, 0 otherwise
  - $nwifeinc_i$: husband's earnings
  - $educ_i$: years of education
  - $\exp er_i$: past years of labor market experience
  - $kidslt6_i$: number of children less than six years old
  - $kidsge6_i$: number of kids between 6 and 18 years of age

Cross-sectional Data: Crime Data

- **Crime**: Data: Wooldrige (p. 4, 78,172, 295, 583)
  - $crime_i$: some measure of the frequency of criminal activity
  - Ex: $narr86_i$: number of times a man was arrested
  - $pcnv_i$: proportion of prior arrests leading to conviction
  - $tottime_i$: total time the man has spent in prison prior to 1986 since reaching the age of 18
  - $ptime86_i$: months spent in prison in 1986
  - $qemp86_i$: number of quarters in 1986 during which the man was legally employed

Databases

Some Sources of Microdata: Cameron and Trivedi (2005) p.58

- Panel Study in Income Dynamics (PSID)
- Current Population Survey (CPS)
- National Longitudinal Survey (NLS)
- National Longitudinal Surveys of Youth (NLSY)
- Survey of Income and Program Participation (SIPP)
- Health and Retirement (HRS)
- World Bank's Living Standards Measurement Study (LSMS)
- Data clearinghouses
- Journal data archives

Students Resources

Some Students Resources:

- Stock and Watson:
  - Stock: http://scholar.harvard.edu/stock/home
  - Watson: http://www.princeton.edu/~mwatson/
- Wooldrige: http://econ.msu.edu/faculty/wooldridge/
- Greene: http://people.stern.nyu.edu/wgreene/
- Cameron and Trivedi:
  - Cameron: http://cameron.econ.ucdavis.edu/
  - Trivedi: http://pages.iu.edu/~trivedi/

Microeconometric Analysis: Regressions

What is it usually done with a dataset, $y_i, x_{1i}, x_{2i}, ..., x_{ki}$, in microeconometrics?

REGRESSIONS

"In modern microeconometrics the term regression refers to a bewildering range of procedures for studying the relationship between an outcome variable y and a set of regressors x." Cameron and Trivedi (2005) p.66

Motivating Regressions

Conditional Prediction of y given x. (CT, 2005 p.66)

- Loss function
$$L(e) = L(y - h(x))$$
where $h(x)$ denotes the predictor defined as a function of x, $e = y - h(x)$ is the prediction error and $L(e)$ is the loss associated with the error e.

- Expected Loss:
$$E[L(y - h(x))|x]$$

- Optimal Predictor
$$\min_{h(x)} E[L(y - h(x))|x]$$

Motivating Regressions: Mean-square error loss

- The choice of the loss function depends on the nature of the problem being studied

- The quadratic loss function is often used in econometrics:

$$E\left[L\left(y-h\left(x\right)\right)|x\right]=E\left[e^{2}|x\right]$$

- **Important**: For the mean-square error loss function the optimal predictor is the conditional expectation $E\left[y|x\right]$, i.e. if

$$\min_{h(x)} E\left[\left(y-h\left(x\right)\right)^{2}|x\right]$$

  then $h\left(x\right)=E\left[y|x\right]$

Motivating Regressions: Mean-square error loss

- Two approaches: Nonparametric or Parametric $E[y|x]$
- In this course, we will specify a parametric model for $E[y|x] = g(x, \beta)$ where β needs to be estimated

Motivating Regressions: Mean-square error loss

- Sample Analog

$$\frac{1}{n}\sum_{i=1}^{n} L\left(e_i\right)$$

- For the mean-square error loss function:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(e_i\right) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(x_i, \beta\right)\right)^2,$$

  and the β that minimizes it is known as least squares. If g is linear, then it known as ordinary least squares

Motivating Regressions: Absolute error loss

- Absolute error loss: $L(e) = |e|$
- Optimal predictor: $med[y|x]$
- If $med[y|x] = x\beta$, then

$$\sum_{i=1}^{n} L(e_i) = \sum_{i=1}^{n} |y_i - x_i\beta|$$

  and the β that minimizes it is known as the least absolute deviations estimator
- Robustness (outliers)

Motivating Regressions: Asymmetric absolute error loss

- Asymmetric absolute error loss: penalty of $(1-\alpha)|e|$ on overprediction and $\alpha|e|$ on underprediction
- $\alpha \in (0,1)$ and $\alpha = 0.5$ implies symmetry
- Optimal predictor: Conditional quantile: $q_\alpha[y|x]$
- Basis for Quantile Regressions:

$$\sum_{i=1}^{n} L(e_i) = \sum_{i:y_i \geq x_i\beta}^{n} \alpha |y_i - x_i\beta_\alpha| + \sum_{i:y_i < x_i\beta}^{n} (1-\alpha)|y_i - x_i\beta_\alpha|$$

and the β_α that minimizes it is known as the α^{th} quantile regression estimator. For $\alpha = 0.5$, we get the median regression estimator or least absolute deviations estimator described above.

Motivating Regressions: Conditional Expectation

- **Main focus of this course**: Conditional Expectation: $E\left[y|x\right]$

$$y = E\left[y|x\right] + u$$

- $E\left[y|x\right]$ linear: Example: Linear wage equation

$$E\left[wage|x\right] = \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 female$$

- $E\left[y|x\right]$ is nonlinear. Example: Poisson Regression for number of arrests

$$E\left[narr86|x\right] = \exp\left(\beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime\right)$$

Conditional Expectations Review

Definition: *Conditional Expectation (Bivariate case)*: Let Y and X be random variables with joint density function $f(x, y)$. Let the conditional density function of Y given $x \in B$ be $f(y|x \in B)$. Let $g(Y)$ be a real-valued function of Y. Then the conditional expectation of $g(Y)$ given $x \in B$, is defined as
(i) Discrete case

$$E[g(Y)|x \in B] = \sum_{y \in R(Y)} g(Y) f(y|x \in B)$$

(ii) Continuous case

$$E[g(Y)|x \in B] = \int_{-\infty}^{\infty} g(Y) f(y|x \in B)\, dy$$

(Mittelhammer (2013) p. 125)

Conditional Expectations Review

Definition: *Conditional Density Function (Bivariate case)*: Let Y and X be random variables with joint density function $f(x, y)$ and let $f_X(x)$ be the marginal density function of X. The conditional density of Y given $x \in B$ is

$$f(y|x \in B) = \frac{f(x \in B, y)}{f_X(x \in B)}$$

Definition: *Marginal Density Function (Bivariate case)*: Let Y and X be random variables with joint density function $f(x, y)$. The marginal density function of X is

$$f_X(x) = \begin{cases} \sum_{y \in R(Y)} f(x, y) & \text{discrete case} \\ \int_{-\infty}^{\infty} f(x, y)\, dy & \text{continuous case} \end{cases}$$

Conditional Expectations Review

Example: (Mittelhammer, p. 82)

- A company has two processing plants, plant 1 and plant 2. The proportion of processing capacity at which each of the plants operates on any given day is the outcome of a bivariate random variable

- Joint density function:

$$f\left(x_1, x_2\right) = \left(x_1 + x_2\right) I_{[0,1]}\left(x_1\right) I_{[0,1]}\left(x_2\right)$$

Conditional Expectations Review

- Marginal for X_1: Integrate out x_2 from $f(x_1, x_2)$ as

$$
\begin{aligned}
f_1(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \\
&= \int_{-\infty}^{\infty} (x_1 + x_2)\, I_{[0,1]}(x_1)\, I_{[0,1]}(x_2)\, dx_2 \\
&= \int_0^1 (x_1 + x_2)\, I_{[0,1]}(x_1)\, dx_2 \\
&= \left(x_1 x_2 + \frac{x_2^2}{2} \right) I_{[0,1]}(x_1) \Big|_0^1 \\
&= \left(x_1 + \frac{1}{2} \right) I_{[0,1]}(x_1)
\end{aligned}
$$

Conditional Expectations Review

- Conditional density function of plan 1's capacity given that plant 2 operates at less than half of capacity

$$
\begin{aligned}
f\left(x_1 | x_2 \leq 0.5\right) &= \frac{\int_{-\infty}^{0.5} f\left(x_1, x_2\right) dx_2}{\int_{-\infty}^{0.5} f_2\left(x_2\right) dx_2} \\
&= \frac{\int_0^{0.5} \left(x_1 + x_2\right) I_{[0,1]}\left(x_1\right) dx_2}{\int_0^{0.5} \left(x_2 + \frac{1}{2}\right) dx_2} \\
&= \left(\frac{4}{3}x_1 + \frac{1}{3}\right) I_{[0,1]}\left(x_1\right)
\end{aligned}
$$

Conditional Expectations Review

- What about: Conditional density function for plant 1's capacity given that plant 2's capacity proportion is $x_2 = 0.75$?

$$f\left(x_1 | x_2 = 0.75\right) = \frac{\int_{0.75}^{0.75} f\left(x_1, x_2\right) dx_2}{\int_{0.75}^{0.75} f_2\left(x_2\right) dx_2} = \frac{0}{0}$$

- In that case, by an approximation argument (see Mittelhammer p.88),

$$\begin{aligned} f\left(x_1 | x_2 = 0.75\right) &= \frac{f\left(x_1, 0.75\right)}{f_2\left(0.75\right)} \\ &= \frac{\left(x_1 + 0.75\right) I_{[0,1]}\left(x_1\right)}{1.25} \\ &= \left(\frac{4}{5} x_1 + \frac{3}{5}\right) I_{[0,1]}\left(x_1\right) \end{aligned}$$

Conditional Expectations Review

- Conditional Expectation of X_1 given $x_2 = 0.75$?

$$
\begin{aligned}
E\left[X_1|x_2 = 0.75\right] &= \int_{-\infty}^{\infty} x_1 f\left(x_1|x_2 = 0.75\right) dx_1 \\
&= \int_0^1 x_1 \left(\frac{4}{5}x_1 + \frac{3}{5}\right) dx_1 \\
&= \frac{17}{30}
\end{aligned}
$$

Conditional Expectations Review

- Conditional Expectation of X_1 as a function of x_2:

$$\begin{aligned}
E\left[X_1|x_2\right] &= \int_{-\infty}^{\infty} x_1 f\left(x_1|x_2\right) dx_1 \\
&= \int_{-\infty}^{\infty} x_1 \frac{f\left(x_1, x_2\right)}{f_2\left(x_2\right)} dx_1 \\
&= \int_{0}^{1} x_1 \frac{\left(x_1 + x_2\right) I_{[0,1]}\left(x_2\right)}{\left(x_2 + \frac{1}{2}\right) I_{[0,1]}\left(x_2\right)} dx_1 \\
&= \left[\frac{\frac{1}{2}x_2 + \frac{1}{3}}{x_2 + \frac{1}{2}}\right] \quad \text{for} \ \ x_2 \in [0,1]
\end{aligned}$$

- When evaluated at $x_2 = 0.75$ same result as above

Motivating Regressions: Conditional Expectation

- **Main focus of this course**: Conditional Expectation: $E[y|x]$

$$y = E[y|x] + u$$

- First part of the course $E[y|x]$ linear: Example: Linear wage equation

$$E[wage|x] = \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 female$$

- Second part of the course $E[y|x]$ is nonlinear. Example: Poisson Regression for number of arrests

$$E[narr86|x] = \exp(\beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime)$$

Conditional Expectations Review

Conditional Expectation, $E[y|x]$**, Properties**:
(Wooldrige 2010, p. 30)

1. The conditional expectation is a linear operator: Let x and y be two random scalars and $a(x)$ and $b(x)$ two scalar functions of x. Then,

$$E[a(x)y + b(x)|x] = E[a(x)y|x] + E[b(x)|x] = a(x)E[y|x] + b(x)$$

provided that $E(|y|) < \infty$, $E(|a(x)y|) < \infty$, and $E(|b(x)|) < \infty$.

Conditional Expectations Review

1. (General) The conditional expectation is a linear operator: Let $a_1(x), \ldots, a_G(x)$ and $b(x)$ be scalar functions of x, and let $y_1, \ldots, y_G$ be random scalars. Then,

$$E\left[\sum_{j=1}^{G} a_j(x)\, y_j + b(x)\,|x\right] = \sum_{j=1}^{G} a_j(x)\, E\left[y_j|x\right] + b(x)$$

provided that $E\left(|y_j|\right) < \infty$, $E\left(|a_j(x)\, y_j|\right) < \infty$, and $E\left(|b(x)|\right) < \infty$.

Conditional Expectations Review

2. Law of Iterated Expectations: (Simplest case):

$$E(y) = E\left[E(y|x)\right]$$

3. Law of Iterated Expectations: (General case):

$$E[y|x] = E\left[E(y|w)|x\right]$$

where x and w are vectors with $x = f(w)$ for some nonstochastic function $f(.)$.

Conditional Expectations Review

4. If $f(x) \in \mathbb{R}^J$ is a function of x such that $E[y|x] = g(f(x))$ for some scalar function $g(.)$, then

$$E[y|f(x)] = E[y|x]$$

5. If the vector (u, v) is independent of the vector x, then

$$E[u|x, v] = E[u|v]$$

Conditional Expectations Review

6. If $u \equiv y - E\left[y|x\right]$, then

$$E\left[g\left(x\right)u\right] = 0$$

for any function $g\left(x\right)$, provided that $E\left(\left|g_j\left(x\right)u\right|\right) < \infty$, $j = 1, ..., J$, and $E\left(|u|\right) < \infty$. In particular, $E\left(u\right) = 0$ and $Cov\left(x_j, u\right) = 0, j = 1, ..., K$.

7. *Conditional Jensen's Inequality*: If $c : \mathbb{R} \rightarrow \mathbb{R}$ is a convex function defined on $\mathbb{R}$ and $E\left(|y|\right) < \infty$, then

$$c\left(E\left[y|x\right]\right) \leq E\left[c\left(y\right)|x\right]$$

Conditional Expectations Review

8. If $E\left(y^2\right) < \infty$ and $\mu(x) \equiv E[y|x]$, then μ is a solution to

$$\min_{m \in M} E\left[\left(y - m(x)\right)^2\right]$$

where M is the set of functions $m : \mathbb{R}^K \to \mathbb{R}$ such that $E\left[m(x)^2\right] < \infty$. (That is $E[y|x]$ is the best mean square predictor of y given x)

Motivating Regressions: Conditional Expectation

- **Main focus**: Conditional Expectation: $E\left[y|x\right]$

$$y = E\left[y|x\right] + u$$

- $E\left[y|x\right]$ linear: Example: Linear wage equation

$$E\left[wage|x\right] = \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 female$$

- $E\left[y|x\right]$ is nonlinear. Example: Poisson Regression for number of arrests

$$E\left[narr86|x\right] = \exp\left(\beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime\right)$$