Introduction to POL 217

Brad Jones[1]

[1] Department of Political Science
University of California, Davis

January 9, 2007

Course Topics
Course Requirements
Preliminary Material

Topics of Course

- Models for Categorical Data.

Topics of Course

- Models for Categorical Data.
- Models for Events Data.

Topics of Course

- Models for Categorical Data.
- Models for Events Data.
- There is a close connection.

Course Requirements

- Several Applied Problem Sets (50 percent).

Course Requirements

- Several Applied Problem Sets (50 percent).
- One Take-Home Exam . . .

Course Requirements

- Several Applied Problem Sets (50 percent).
- One Take-Home Exam . . .
- or Paper and Presentation (40 percent).

Course Requirements

- Several Applied Problem Sets (50 percent).
- One Take-Home Exam . . .
- or Paper and Presentation (40 percent).
- Participation beyond Breathing (10 percent).

Pros and Cons

- Exams may help prepare for comprehensives.

Pros and Cons

- Exams may help prepare for comprehensives.
- Exams don't (usually) result in articles.

Pros and Cons

- Exams may help prepare for comprehensives.
- Exams don't (usually) result in articles.
- ...though all of you will need to discuss quantitative work.

Categorical Response Variables

- Binary (logit or probit).

Categorical Response Variables

- Binary (logit or probit).
- Ordinal (proportional odds/cumulative probit).

Categorical Response Variables

- Binary (logit or probit).
- Ordinal (proportional odds/cumulative probit).
- Nominal (baseline category logit).

Categorical Response Variables

- Binary (logit or probit).
- Ordinal (proportional odds/cumulative probit).
- Nominal (baseline category logit).
- (We'll worry about events data in a few weeks).

Why Logit or Probit?

- Suppose y is binary.

Why Logit or Probit?

- Suppose y is binary.
- Regression assumes y is unbounded and continuous.

Why Logit or Probit?

- Suppose y is binary.
- Regression assumes y is unbounded and continuous.
- Hence, $y = \beta_0 + \beta_X + \epsilon$.

Why Logit or Probit?

- Suppose y is binary.
- Regression assumes y is unbounded and continuous.
- Hence, $y = \beta_0 + \beta_X + \epsilon$.
- $\therefore$ $\hat{y}$ must be unbounded.

Why Logit or Probit?

- Suppose y is binary.
- Regression assumes y is unbounded and continuous.
- Hence, $y = \beta_0 + \beta_X + \epsilon$.
- $\therefore$ $\hat{y}$ must be unbounded.
- Linear change in $E(y) \mid x$ not natural.

Why Logit or Probit?

- Suppose y is binary.
- Regression assumes y is unbounded and continuous.
- Hence, $y = \beta_0 + \beta_X + \epsilon$.
- $\therefore$ $\hat{y}$ must be unbounded.
- Linear change in $E(y) \mid x$ not natural.
-

$$\begin{aligned} P(y=1) &= P_i \\ P(y=0) &= (1-P_i) = Q_i \\ E(y) &= P_i(1) + Q_i(0) \\ E(y) &= P_i(1) \end{aligned}$$

Why Logit or Probit?

- More Trouble.

Why Logit or Probit?

- More Trouble.
- $\hat{\epsilon} = y_i - \sum \hat{\beta}_k x_i$.

Why Logit or Probit?

- More Trouble.
- $\hat{\epsilon} = y_i - \sum \hat{\beta}_k x_i$.
- $$\begin{aligned} y &= 1: \quad 1 - \sum \hat{\beta}_k x_i \\ y &= 0: \quad 0 - \sum \hat{\beta}_k x_i \end{aligned}$$

Why Logit or Probit?

- More Trouble.
- $\hat{\epsilon} = y_i - \sum \hat{\beta}_k x_i$.
-

$$
\begin{aligned}
y &= 1: \quad 1 - \sum \hat{\beta}_k x_i \\
y &= 0: \quad 0 - \sum \hat{\beta}_k x_i
\end{aligned}
$$

- $\forall\ x_i$, ϵ assumes two values.

Why Logit or Probit?

- Heteroskedasticity

Why Logit or Probit?

- Heteroskedasticity
- $E(y) = \sum \hat{\beta}_k x_i = P_i$ and $1 - \sum \hat{\beta}_k x_i = Q_i$

Why Logit or Probit?

- Heteroskedasticity
- $E(y) = \sum \hat{\beta}_k x_i = P_i$ and $1 - \sum \hat{\beta}_k x_i = Q_i$
- Noting (without proof) that $\mathrm{var}(\epsilon) = (1 - \sum \hat{\beta}_k x_i)^2 P_i + (-\sum \hat{\beta}_k x_i)^2 Q_i$
- $\therefore$

$$
\begin{aligned}
\mathrm{var}(\epsilon) &= (Q_i)^2 P_i + (-P_i)^2 Q_i \\
&= Q_i P_i (Q_i P_i) \\
&= Q_i P_i ([1 - P_i] + P_i) \\
&= Q_i P_i \\
&= (1 - \sum \hat{\beta}_k x_i)(\sum \hat{\beta}_k x_i)
\end{aligned}
$$

Motivating Logit (or Probit)

- Suppose $E(y) = P(y = 1 \mid x) = \beta_k x_{ik}$, and y is binary.

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-\sum \beta_k x_{ik})}$$

Motivating Logit (or Probit)

- Suppose $E(y) = P(y = 1 \mid x) = \beta_k x_{ik}$, and y is binary.

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-\sum \beta_k x_{ik})}$$

- Let $Z = \sum \beta_k x_{ik}$, then

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-Z)} = \frac{\exp(Z)}{1 + \exp(Z)}$$

Motivating Logit (or Probit)

- Suppose $E(y) = P(y = 1 \mid x) = \beta_k x_{ik}$, and y is binary.

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-\sum \beta_k x_{ik})}$$

- Let $Z = \sum \beta_k x_{ik}$, then

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-Z)} = \frac{\exp(Z)}{1 + \exp(Z)}$$

- This is the c.d.f. for the logistic distribution.

Motivating Logit (or Probit)

- Suppose $E(y) = P(y = 1 \mid x) = \beta_k x_{ik}$, and y is binary.

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-\sum \beta_k x_{ik})}$$

- Let $Z = \sum \beta_k x_{ik}$, then

$$\Pr(y = 1 \mid x) = \frac{1}{1 + \exp(-Z)} = \frac{\exp(Z)}{1 + \exp(Z)}$$

- This is the c.d.f. for the logistic distribution.
- Problems Solved:
  Z is unbounded; P_i must stay in unit interval.
  P_i is nonlinearlly related to parameters (though logit is linear!)

Logit Model

- Odds Ratios are given by $P_i/(1 - P_i) = \exp(Z)$

Logit Model

- Odds Ratios are given by $P_i/(1-P_i) = \exp(Z)$
- Log-odds are then $\log[P_i/(1-P_i)] = Z$

Logit Model

- Odds Ratios are given by $P_i/(1 - P_i) = \exp(Z)$
- Log-odds are then $\log[P_i/(1 - P_i)] = Z$
- The Logit Model

$$\log \frac{P_i}{1 - P_i} = Z = \sum \beta_k x_{ik}$$

Logit Model

- Odds Ratios are given by $P_i/(1-P_i) = \exp(Z)$
- Log-odds are then $\log[P_i/(1-P_i)] = Z$
- The Logit Model

$$\log \frac{P_i}{1-P_i} = Z = \sum \beta_k x_{ik}$$

- This is the *logit transformation* and yields the logit model.

Logit Model

- Odds Ratios are given by $P_i/(1 - P_i) = \exp(Z)$
- Log-odds are then $\log[P_i/(1 - P_i)] = Z$
- The Logit Model

$$\log \frac{P_i}{1 - P_i} = Z = \sum \beta_k x_{ik}$$

- This is the *logit transformation* and yields the logit model.
- Again, Z unbounded; perfect prediction impossible.