Quantitative Social Science Methods, I, Lecture Notes: Binary Outcome Models

Gary King[1]
Institute for Quantitative Social Science
Harvard University

August 17, 2020

[1] GaryKing.org

Linear Probability, Logit, Probit Models

Interpreting Functional Forms

Alternative Interpretations of Binary Models

General Rules for Presenting and Interpreting Statistical Results

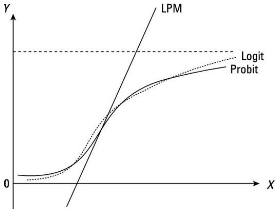
Linear Probability Model

The Model

1. Stochastic component for a binary outcome

$$Y_i \sim \text{Bernoulli}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y=1 \\ 1-\pi_i & \text{for } y=0 \end{cases}$$

2. Systematic component $\Pr(Y_i = 1|\beta) \equiv E(Y_i) \equiv \pi_i = x_i\beta$
3. Y_i and Y_j are independent $\forall\, i \neq j$, conditional on X



- Quiz: What's good? What's bad?
- for some x, $\Pr(Y) \notin [0, 1]$
- But models are approximations. Maybe ok for middling π?
- Unlikely to get uncertainties right

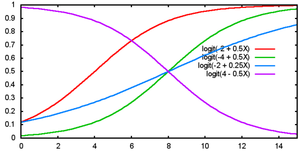
The Logistic Regression (Logit) model

The Model

1. Stochastic component

$$Y_i \sim \text{Bernoulli}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{cases}$$

2. Systematic Component: $\pi_i \equiv \Pr(Y_i = 1|\beta) = \frac{1}{1+e^{-x_i\beta}}$
3. Y_i and Y_j are independent $\forall\, i \neq j$, conditional on X



- Quiz: What's good? What's bad?
- $\Pr(y) \in [0,1]$ for any y
- One change for probit: $\pi_i = \Phi(X_i\beta)$
- Could be more flexible; OK for now

The Logit Log-Likelihood

Probability density of all the data

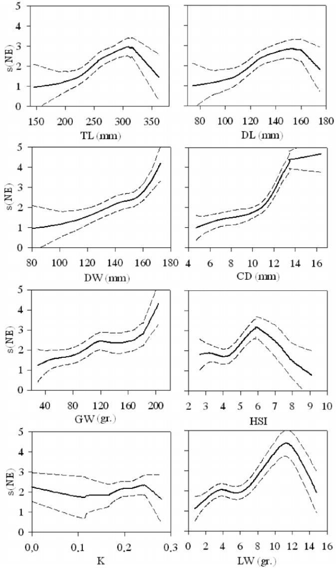
$$\mathrm{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \qquad \pi_i = \frac{1}{1+e^{-x_i\beta}}$$

Log-likelihood

$$\begin{aligned}
\ln L(\beta|y) &= \sum_{i=1}^{n} \{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \left\{-y_i \ln\left(1+e^{-x_i\beta}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-x_i\beta}}\right)\right\} \\
&= -\sum_{i=1}^{n} \ln\left(1+e^{(1-2y_i)x_i\beta}\right)
\end{aligned}$$

Quiz: What do we do with this?
How to interpret $\hat{\beta}$? What's the QOI?

Graphics to Interpret Functional Forms



- X horizontally; y vertically; uncertainty represented
- Use theoretical ranges, not observed X's
- Entire surface plot for a few X's
- Marginal effects: Hold some variables constant at their means, a typical value, or observed values
- Average effects: compute effects for every observation and average
- Be creative; choose graphs for impact

Fitted Values to Interpret Functional Forms

- Calculate fitted values given selected values of X, X_c for "typical" people, person types, regional representatives, stereotypes, etc.
- Compute $\hat{\theta}_c = g(X_c, \hat{\beta})$
- An example for logit: $\hat{\pi}_c = \frac{1}{1+e^{-X_c\hat{\beta}}}$

Sex	Age	Home	Income	Pr(vote)
Male	20	Chicago	\$33,000	0.20
Female	27	New York City	\$43,000	0.28
Male	50	Madison, WI	\$55,000	0.72
⋮				

- Include a measure of uncertainty (standard error, confidence interval, etc.) — for any quantity but a probability
- Easy to communicate
- Difficult to be comprehensive
- Better by simulation: point and uncertainty estimation

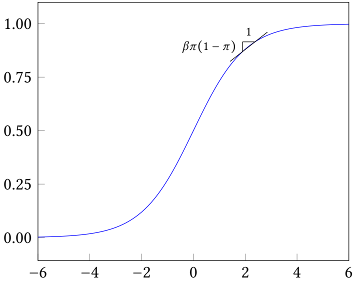
First Differences to Interpret Functional Forms

- aka Risk Differences (in epidemiology)
- Define X_s (starting point) and X_e (ending point) as $k \times 1$ vectors of values of X. Usually all values are the same but one.
- First difference
  - In general: $D = g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
  - Linear Model: $D = X_e\hat{\beta} - X_s\hat{\beta}$
  - Logit Model: $D = \frac{1}{1+e^{-X_e\hat{\beta}}} - \frac{1}{1+e^{-X_s\hat{\beta}}}$
- Example:

Variable	From		To	First Difference
Sex	Male	$\longrightarrow$	Female	.05
Age	65	$\longrightarrow$	75	−.10
Home	NYC	$\longrightarrow$	Madison, WI	.26
Income	\$35,000	$\longrightarrow$	\$75,000	.14

- Easier by simulation: point and uncertainty estimation

Derivative Rules of Thumb to Interpret Functional Forms



- Good for quick interpretation; probably not for presenting results
- Derivative rule: $\frac{\partial\theta}{\partial X_j} = \frac{\partial g(X,\beta)}{\partial X_j}$
- Linear: $\frac{\partial\mu}{\partial X_j} = \beta_j$ (unconditional)
- Logit: $\frac{\partial\pi}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j\hat{\pi}(1-\hat{\pi})$
- Max value of logit derivative: $\hat{\beta} \times 0.5(1-0.5) = \hat{\beta}/4$
- Max value for probit derivative: $\hat{\beta} \times 0.4$
- Presented so it's easy to remember; so remember!

Linear Probability, Logit, Probit Models

Interpreting Functional Forms

Alternative Interpretations of Binary Models

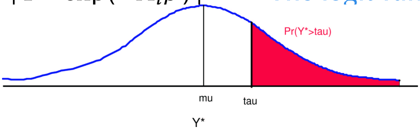
General Rules for Presenting and Interpreting Statistical Results

Logit Model Interpretation from Biology

- Continuous unobserved variable: Y^*. health, voting propensity
- A model: $Y_i^* \sim P(y_i^*|\mu_i), \quad \mu_i = x_i\beta, \quad Y_i \perp\!\!\!\perp Y_j|X$
- Quiz: what model has Y^* observed & $P(\cdot)$ normal?
- With observation mechanism: $y_i = \begin{cases} 1 & y_i^* \leq 0 \text{ if } i \text{ is alive} \\ 0 & y_i^* > 0 \text{ if } i \text{ is dead} \end{cases}$
- If only y_i is observed, and Y^* is standardized logistic,

$$P(y_i^*|\mu_i) = \text{STL}(y^*|\mu_i) = \frac{\exp(y_i^* - \mu_i)}{[1 + \exp(y_i^* - \mu_i)]^2}, \qquad \rightsquigarrow \text{logit model}$$

- Proof: $\Pr(Y_i = 1|\mu_i) = \Pr(Y_i^* \leq 0) = \int_{-\infty}^{0} \text{STL}(y_i^*|\mu_i)dy_i^* = F_{stl}(0|\mu_i) = [1 + \exp(-X_i\beta)]^{-1} \rightsquigarrow$ The logit functional form



Probit Model Interpretation from Biology

- Same setup as for logit, with one change
  - Stochastic component: $Y^* \sim \text{P}(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$
- Systematic component becomes

$$\Pr(Y_i = 1|\mu) = \int_{-\infty}^{0} N(y_i^*|\mu_i, 1)dy_i^* = \Phi(X_i\beta)$$

- Interpretation:
  - One unit of Y^*: one standard deviation
  - Interpret β: regression coefficients of Y^* on X
  - Interpret $\hat{\beta}_j$: what happens to Y^* on average (or μ_i exactly) when X_j goes up by one unit, holding constant the other covariates

Logit & Probit Interpretation from Economics

- Definitions:
  - Utility for the Democratic candidate: U_i^D
  - Utility for the Republican candidate: U_i^R
  - Utility difference, propensity to vote Dem: $Y^* \equiv U_i^D - U_i^R$
  - Same Observation mechanism: $y_i = \begin{cases} 1 & y_i^* \leq 0 \text{ if } i \text{ is Dem} \\ 0 & y_i^* > 0 \text{ if } i \text{ is Rep} \end{cases}$
- Assumptions:
  - $U_i^D \perp\!\!\!\perp U_i^R | X$
  - $U_i^k \sim \mathrm{P}(U_i^k | \eta_i^k)$ for $k = \{D, R\}$
- If $\mathrm{P}(\cdot)$ is normal: $\rightsquigarrow$ probit model
- If $\mathrm{P}(\cdot)$ is generalized extreme value: $\rightsquigarrow$ logit model
- Quiz: Of the three justifications for the same binary model, which do you prefer?
- Quiz: When would you choose LPM or logit or probit?

Linear Probability, Logit, Probit Models

Interpreting Functional Forms

Alternative Interpretations of Binary Models

General Rules for Presenting and Interpreting Statistical Results

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis		.609**
Age		.007**
Education		.289**
Family income		.151**
Race (non-White/White)		.695**
Gender (female/male)		.789**
Region (South/non-South)		.076
Network coverage		.000
Education × Time		–.003*
Time in months		.078**
Constant		–9.257**
Number	7,021	
–2 log-likelihood	7,215.231	
Goodness of fit	6,789.45	
Cox & Snell R^2		.212
Nagelkerke R^2		.295
Overall correct classification (%)	73.96	

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

- What do the each of the numbers mean?
- Why so much whitespace? Can you connect cols A and B?
- What's the star-gazing?
- Can any be interpreted as causal estimates?
- Can you compute a quantity of interest from these numbers?
- This is bad, not rare

The Goals of Interpretation and Presentation

- Statistical presentations should
  1. Convey precise estimates of quantities of interest
  2. Include measures of uncertainty
  3. Require little specialized knowledge to understand
  4. Exclude superfluous information (e.g., long lists of coefficients no one understands, star gazing, silly summary stats, too many decimal places)
- For example: Other things being equal, an additional year of education would increase your annual income by $1,500 on average, plus or minus about $500
- Try to satisfy someone like both me and my mom & dad
- Reading: King, Tomz, Wittenberg, "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" *American Journal of Political Science* (2000).

Simulating Quantities of Interest

- Quiz: What do quantity of interest simulations get us?
  - Summarize everything we know and don't know about the QOI
  - Complete flexibilty of presentation
  - A great test of whether we understand the model
- Goal: Simulate quantities of interest from the model

$$
\begin{aligned}
Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\
\theta_i &= g(x_i, \beta) && \text{systematic}
\end{aligned}
$$

- How to simulate QOIs
  1. Simulate β and α due to estimation uncertainty (because of inadequacies in your research design: $n < \infty$.)
  2. Simulate Y (given sims of α and β), representing fundamental uncertainty (due to the nature of nature!)
  3. Calculate quantity of interest (given sims of Y)

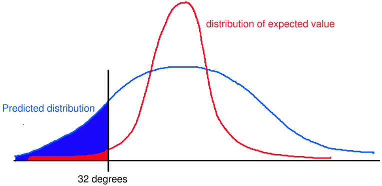
Simulate Parameters

- Goal: Random draws of parameters from sampling distribution (aka posterior with a flat prior)
- How to:
  1. Maximize likelihood function wrt $\gamma = \text{vec}(\beta, \alpha)$ (once)
  2. Record $\hat{\gamma}$ and $\hat{\text{V}}(\hat{\gamma})$
  3. Draw $\tilde{\gamma}$ from the multivariate normal (many times)

$$\tilde{\gamma} \sim \text{N}\left(\hat{\gamma}, \hat{\text{V}}(\hat{\gamma})\right)$$

Simulating Expected v. Predicted Values

- Definitions
  - Predicted: draws of Y that could in principle be observed
  - Expected: draws of distribution features, such as $E(Y)$
- Sources of variability
  - Predicted: estimation and fundamental uncertainty
  - Expected: estimation only (average over fundamental)
- Quiz: As $n \longrightarrow \infty$, does the variance go to zero?
  - Predicted: no
  - Expected: yes
- Example



  - Predicted: Pr(Temperature < 32°) tomorrow
  - Expected: Pr(Average Temperature < 32°) tomorrow

Simulating Predicted Values

- Predicted values can be for
  1. Forecasts: about the future
  2. Farcasts: about some area for which you have no y
  3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)
- Repeat once for each random draw of $\tilde{y}$
  1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha}) \sim \mathsf{N}\left(\hat{\gamma}, \hat{\mathsf{V}}(\hat{\gamma})\right)$
  2. Define vector X_c, which defines the predicted value to compute
  3. Extract simulated $\tilde{\beta}$ from $\tilde{\gamma}$; compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$
  4. Simulate outcome variable $\tilde{Y}_c \sim f(\tilde{\theta}_c, \tilde{\alpha})$
- Quiz: What can we do with many sims of $\tilde{y}$?
- E.g.: histogram, average, variance, percentile values, etc.

Simulating Expected Values: Algorithm

1. Draw *one* simulated expected value:
   (a) Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$ (estimation uncertainty)
   (b) Choose one value for each explanatory variable (X_c is a vector)
   (c) Compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$, given the one simulated $\tilde{\beta}$ from $\tilde{\gamma}$
   (d) Draw m sims of the outcome $\tilde{Y}_c^{(k)}$ ($k = 1, \dots, m$) (simulating fundamental uncertainty from stochastic component $f(\tilde{\theta}_c, \tilde{\alpha})$)
   (e) Average over fundamental uncertainty: average m simulations gives one simulated expected value $\tilde{E}(Y_c) = \sum_{k=1}^{m} \tilde{Y}_c^{(k)}/m$
2. Repeat algorithm M times leaving estimation uncertainty
3. Compute QOIs: histogram, average (point estimate), SE, CI

Interpretation

- When $m = 1$: same as predicted values.
- With large m: better fundamental uncertainty approximation
- When $E(Y_c) = \theta_c$: we *may* skip steps d–e. E.g., simulating π_i in logit model. If you're unsure, do it anyway!

Simulating First Differences

- Draw *one* simulated first difference
  1. Choose vectors X_s, the <u>s</u>tarting point, X_e, the <u>e</u>nding point.
  2. Apply the expected value algorithm twice, once for X_s and X_e
  3. Take the difference in the two estimated expected values
- Repeat M times
- Quiz: which QOIs do we want here?
- To save computation time, and improve approximation:
  Reuse the same simulated $\tilde{\beta}$ for both

Tricks for Simulating Parameters

- Simulate all parameters together (in γ), including ancillary parameters (unless you know they are orthogonal)
- Advantages of Reparameterization to unbounded scale
  - $\hat{\gamma}$ converges more quickly in n to multivariate normal. (MLEs don't change, but the posteriors and SEs do.)
  - maximization algorithm works faster without constraints
- How to reparameterize:
  - $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)
  - For a probability, $\pi = \left[1 + e^{-\eta}\right]^{-1}$ (logit transformation)
  - For $-1 \le \rho \le 1$, use $\rho = \left(e^{2\eta} - 1\right)/\left(e^{2\eta} + 1\right)$ (Fisher's Z trans)
  - In each case, η is unbounded: estimate it, simulate from it, and reparameterize back to the scale you care about.

Tricks for Simulating Quantities of Interest

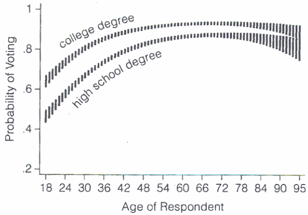
- Compute QOIs from sims of Y (unless you're sure)
- Simulating functions of Y
  - If analyzing $\ln(Y)$, simulate $\ln(Y)$ & apply inverse function $\exp(\ln(Y))$ to reveal Y
  - The wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate
  - Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
  - More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear
- Check approximation error: Run algorithm twice, check precision. If it's not enough for your tables, increase sims.
- Increase speed: Analytical calculations & other tricks
- Easily done in Clarify for Stata and Zelig for R

Replication of Rosenstone and Hansen

by King, Tomz and Wittenberg (2000)

- Logit of turnout on Age, Age^2, Education, Income, and Race
- QOI: effect of age on $\Pr(\text{vote}|X)$, given Income & Race
- Use $M = 1000$ and compute 99% CI

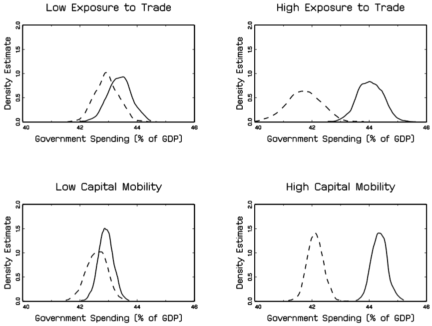
FIGURE 1 Probability of Voting by Age



Vertical bars indicate 99-percent confidence intervals

- Set age=24, education=high school, income=average, Race=white
- Run logistic regression
- Simulate 1000 $\tilde{\beta}$'s
- Compute 1000 $\tilde{\eta}_i = [1 + e^{-x_i \tilde{\beta}}]^{-1}$
- Sort in numerical order
- 99% CI: 5th and 995th values
- Plot vertical line at age=24 (the CI)
- Repeat for other ages and college

Replication of Garrett (King, Tomz and Wittenberg 2000)



- Dependent variable: Government Spending as % of GDP
- Key causal var: left-labor power (high = solid line; low = dashed)
- Garrett only reported the 8 point estimates.
- Quiz: What new information do we learn here?
- Left-labor power: only has effect with high exposure to trade or capital mobility
- Quiz: How can we summarize this with less real estate?