Advanced Quantitative Research Methodology, Lecture Notes: Single Equation Models[1]

Gary King

GaryKing.org

March 3, 2013

Binary Variable Regression Models

The logistic regression (or "logit") model:

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

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

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

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

2. Systematic Component:

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

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

2. Systematic Component:

$$\Pr(Y_i = 1|\beta) \equiv E(Y_i) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$$

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

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

2. Systematic Component:

$$\Pr(Y_i = 1|\beta) \equiv E(Y_i) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$$

3. Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

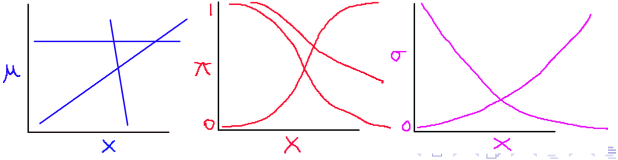
$$Y_i \sim Y_{\text{Bern}}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{cases}$$

2. Systematic Component:

$$\Pr(Y_i = 1|\beta) \equiv E(Y_i) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$$

3. Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X

(The graph in the middle:)



Binary Variable Regression Models

The probability density of all the data:

Binary Variable Regression Models

The probability density of all the data:

$$P(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Binary Variable Regression Models

The probability density of all the data:

$$\mathrm{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

Binary Variable Regression Models

The probability density of all the data:

$$P(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\}$$

Binary Variable Regression Models

The probability density of all the data:

$$P(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

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

Binary Variable Regression Models

The probability density of all the data:

$$P(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

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

Binary Variable Regression Models

The probability density of all the data:

$$P(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

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

What do we do with this?

Interpreting Functional Forms

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's
   (b) Can try entire surface plot for a small number of X's

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's
   (b) Can try entire surface plot for a small number of X's
   (c) Marginal effects: Can hold "other variables" constant at their means, a typical value, or at their observed values

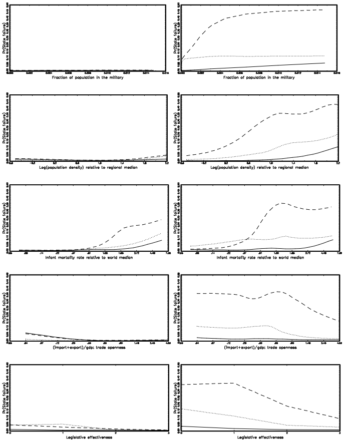
Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's
   (b) Can try entire surface plot for a small number of X's
   (c) Marginal effects: Can hold "other variables" constant at their means, a typical value, or at their observed values
   (d) Average effects: compute effects for every observation and average



Example Marginal Effect Plot of a neural network model (about which more later). Full democracies (dotted), partial democracies (dashed), and autocracies (solid). From Gary King and Langche Zeng. "Improving Forecasts of State Failure," *World Politics*, Vol. 53, No. 4 (July, 2001): 623-58.

Interpreting Functional Forms

2. Fitted Values for selected combinations of X's, or "typical" people or types:

2. Fitted Values for selected combinations of X's, or "typical" people or types:

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
⋮				

Interpreting Functional Forms

2. Fitted Values for selected combinations of X's, or "typical" people or types:

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
⋮				

For any quantity but a probability, always also include a measure of uncertainty (standard error, confidence interval, etc.)

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)

3. First Differences (called Risk Differences in epidemiology)
   (a) 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.

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) 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.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
    (a) 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.
    (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
    (c) $D = \frac{1}{1+e^{-X_e\hat{\beta}}} - \frac{1}{1+e^{-X_s\hat{\beta}}}$

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) 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.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
   (c) $D = \frac{1}{1+e^{-X_e\hat{\beta}}} - \frac{1}{1+e^{-X_s\hat{\beta}}}$
   (d) Better (and necessary to compute se's): do by simulation (we'll repeat the details soon)

3. First Differences (called Risk Differences in epidemiology)
   (a) 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.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
   (c) $D = \frac{1}{1+e^{-X_e\hat{\beta}}} - \frac{1}{1+e^{-X_s\hat{\beta}}}$
   (d) Better (and necessary to compute se's): do by simulation (we'll repeat the details soon)

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

Interpreting Functional Forms

4. Derivatives (some great rules of thumb)

Interpreting Functional Forms

4. Derivatives (some great rules of thumb)

$$\frac{\partial \pi_i}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

4. Derivatives (some great rules of thumb)

$$\frac{\partial \pi_i}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

(a) Max value of logit derivative: $\hat{\beta} \times 0.5(1 - 0.5) = \hat{\beta}/4$

4. Derivatives (some great rules of thumb)

$$\frac{\partial \pi_i}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

   (a) Max value of logit derivative: $\hat{\beta} \times 0.5(1 - 0.5) = \hat{\beta}/4$

   (b) Max value for probit $[\pi_i = \Phi(X_i\beta)]$ derivative: $\hat{\beta} \times 0.4$

Same Logit Model, Different Justification and Interpretation

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$Y_i^* \sim P(y_i^*|\mu_i)$$

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$Y_i^* \sim P(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$\begin{aligned} Y_i^* &\sim P(y_i^*|\mu_i) \\ \mu_i &= x_i\beta \end{aligned}$$

with observation mechanism:

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$Y_i^* \sim P(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

with observation mechanism:

$$y_i =$$

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$
\begin{aligned}
Y_i^* &\sim P(y_i^*|\mu_i) \\
\mu_i &= x_i\beta
\end{aligned}
$$

with observation mechanism:

$$
y_i = \begin{cases} 1 & y^* \leq \tau \text{ if } i \text{ is alive} \\ 0 & y^* > \tau \text{ if } i \text{ is dead} \end{cases}
$$

Same Logit Model, Different Justification and Interpretation

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$\begin{aligned} Y_i^* &\sim P(y_i^*|\mu_i) \\ \mu_i &= x_i\beta \end{aligned}$$

with observation mechanism:

$$y_i = \begin{cases} 1 & y^* \leq \tau \text{ if } i \text{ is alive} \\ 0 & y^* > \tau \text{ if } i \text{ is dead} \end{cases}$$

Since Y^* is unobserved anyway, define the threshold as $\tau = 0$. (Plus the same independence assumption, which from now on is assumed implicit.)

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.
2. If only y_i is observed, and Y^* is standardized logistic (which looks close to the normal),

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.
2. If only y_i is observed, and Y^* is standardized logistic (which looks close to the normal),

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

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.
2. If only y_i is observed, and Y^* is standardized logistic (which looks close to the normal),

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

   then we get a logit model.

Same Logit Model, Different Justification and Interpretation

3. The derivation:

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\Pr(Y_i = 1|\mu_i) = \Pr(Y_i^* \leq 0)$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$
\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i) dy_i^*
\end{aligned}
$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i) dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]}
\end{aligned}$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$
\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\
&= \int_{-\infty}^{0} \text{STL}(y_i^*|\mu_i)dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]} \\
&= [1 + \exp(-X_i\beta)]^{-1}
\end{aligned}
$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \le 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i) dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]} \\
&= [1 + \exp(-X_i\beta)]^{-1}
\end{aligned}$$

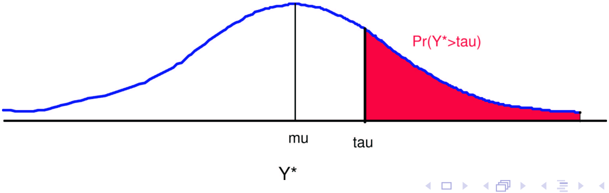
The same functional form!

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \le 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i)dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]} \\
&= [1 + \exp(-X_i\beta)]^{-1}
\end{aligned}$$

The same functional form!



Same Logit Model, Different Justification and Interpretation

4. For the Probit Model, we modify:

Same Logit Model, Different Justification and Interpretation

4. For the Probit Model, we modify:

$$P(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

Same Logit Model, Different Justification and Interpretation

4. For the Probit Model, we modify:

$$P(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

implying

Same Logit Model, Different Justification and Interpretation

4. For the Probit Model, we modify:

$$\text{P}(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

implying

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

Same Logit Model, Different Justification and Interpretation

4. For the Probit Model, we modify:

$$P(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

   implying

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

5. $\Longrightarrow$ interpret β as regression coefficients of Y^* on X: $\hat{\beta}_1$ is what happens to Y^* on average (or μ_i) when X_1 goes up by one unit, holding constant the other explanatory variables (and conditional on the model). In probit, one unit of Y^* is one standard deviation.

An Econometric Interpretation: Utility Maximization

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent
3. Assume $U_i^k \sim \mathrm{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent
3. Assume $U_i^k \sim \mathrm{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
4. Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent
3. Assume $U_i^k \sim \mathrm{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
4. Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.
5. If $\mathrm{P}(\cdot)$ is normal, we get a Probit model

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent
3. Assume $U_i^k \sim \mathrm{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
4. Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.
5. If $\mathrm{P}(\cdot)$ is normal, we get a Probit model
6. If $\mathrm{P}(\cdot)$ is generalized extreme value, we get logit.

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent
3. Assume $U_i^k \sim \mathrm{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
4. Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.
5. If $\mathrm{P}(\cdot)$ is normal, we get a Probit model
6. If $\mathrm{P}(\cdot)$ is generalized extreme value, we get logit.

- Of the three justifications for the same binary model, which do you prefer?

An Econometric Interpretation: Utility Maximization

1. Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
2. Assume U_i^D and U_i^R are independent
3. Assume $U_i^k \sim \mathrm{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
4. Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.
5. If $\mathrm{P}(\cdot)$ is normal, we get a Probit model
6. If $\mathrm{P}(\cdot)$ is generalized extreme value, we get logit.

- Of the three justifications for the same binary model, which do you prefer?
- Which would enable you to choose logit over probit?

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.

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.

1. This one is typical of current practice, not that unusual.

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.

1. This one is typical of current practice, not that unusual.
2. What do these numbers mean?

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.

1. This one is typical of current practice, not that unusual.
2. What do these numbers mean?
3. Why so much whitespace? Can you connect cols A and B?

How Not to Present Statistical Results: Continued

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.

How Not to Present Statistical Results: Continued

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.

4. What does the star-gazing add?

How Not to Present Statistical Results: Continued

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.

4. What does the star-gazing add?
5. Can any be interpreted as causal estimates?

How Not to Present Statistical Results: Continued

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.

4. What does the star-gazing add?
5. Can any be interpreted as causal estimates?
6. Can you compute a quantity of interest from these numbers?

Interpretation and Presentation

Interpretation and Presentation

1. Statistical presentations should

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, no long lists of coefficients no one understands, no star gazing, etc.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, no long lists of coefficients no one understands, no star gazing, etc.
2. 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.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, no long lists of coefficients no one understands, no star gazing, etc.
2. 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.
3. Your work should satisfy someone like me and someone like my grandmother.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, no long lists of coefficients no one understands, no star gazing, etc.
2. 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.
3. Your work should satisfy someone like me and someone like my grandmother.
4. Reading assignment: (at my web page) the handout on papers, and King, Tomz, Wittenberg, "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" American Journal of Political Science, Vol. 44, No. 2 (March, 2000): 341-355.

For any statistical model

For any statistical model

$$Y_i \sim f(y_i|\theta_i, \alpha) \qquad \text{stochatic}$$

For any statistical model

$$Y_i \sim f(y_i|\theta_i, \alpha)$$ stochatic

$$\theta_i = g(x_i, \beta)$$ systematic

For any statistical model

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

recall that we have

For any statistical model

$$Y_i \sim f(y_i|\theta_i, \alpha) \qquad \text{stochatic}$$
$$\theta_i = g(x_i, \beta) \qquad \text{systematic}$$

recall that we have

1. *Estimation uncertainty*: Lack of knowledge of β and α. Vanishes as n gets larger.

For any statistical model

$$Y_i \sim f(y_i|\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(x_i, \beta) \qquad \text{systematic}$$

recall that we have

1. *Estimation uncertainty*: Lack of knowledge of β and α. Vanishes as n gets larger.
2. *Fundamental uncertainty*: Represented by the stochastic component. Exists no matter what the researcher does.

Simulating Estimation Uncertainty: By Simulating the Parameters

To take one random draw of all the parameters $\gamma = \text{vec}(\beta, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

Simulating Estimation Uncertainty: By Simulating the Parameters

To take one random draw of all the parameters $\gamma = \text{vec}(\beta, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\text{V}}(\hat{\gamma})$.

Simulating Estimation Uncertainty: By Simulating the Parameters

To take one random draw of all the parameters $\gamma = \text{vec}(\beta, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\text{V}}(\hat{\gamma})$.
2. Draw the vector γ from the multivariate normal distribution:

Simulating Estimation Uncertainty: By Simulating the Parameters

To take one random draw of all the parameters $\gamma = \text{vec}(\beta, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\mathsf{V}}(\hat{\gamma})$.
2. Draw the vector γ from the multivariate normal distribution:

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

Simulating Estimation Uncertainty: By Simulating the Parameters

To take one random draw of all the parameters $\gamma = \text{vec}(\beta, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\text{V}}(\hat{\gamma})$.
2. Draw the vector γ from the multivariate normal distribution:

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

   Denote the draw $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$, which has k elements.

Simulating the Distribution of Predicted Values

Predicted values can be for:

Simulating the Distribution of Predicted Values

Predicted values can be for:

1. Forecasts: about the future

Simulating the Distribution of Predicted Values

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y

Simulating the Distribution of 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)

Simulating the Distribution of 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)

To simulate *one* predicted value, follow these steps:

Simulating the Distribution of 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)

To simulate *one* predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.

Simulating the Distribution of 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)

To simulate *one* predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Decide which kind of predicted value you wish to compute, and on that basis choose one value for each explanatory variable. Denote the vector of such values X_c.

Simulating the Distribution of 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)

To simulate *one* predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Decide which kind of predicted value you wish to compute, and on that basis choose one value for each explanatory variable. Denote the vector of such values X_c.
3. Taking the simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$, where $g(\cdot, \cdot)$ is the systematic component of the statistical model.

Simulating the Distribution of 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)

To simulate *one* predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Decide which kind of predicted value you wish to compute, and on that basis choose one value for each explanatory variable. Denote the vector of such values X_c.
3. Taking the simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$, where $g(\cdot, \cdot)$ is the systematic component of the statistical model.
4. Simulate the outcome variable $\tilde{Y}_c$ by taking a random draw from $f(\tilde{\theta}_c, \tilde{\alpha})$, the stochastic component of the statistical model.

Simulating the Distribution of Predicted Values

Predicted values can be for:

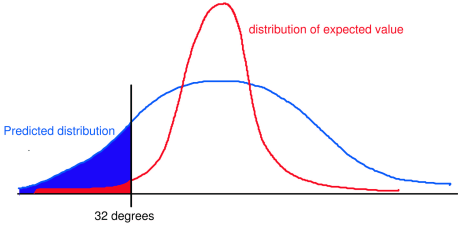
1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate *one* predicted value, follow these steps:

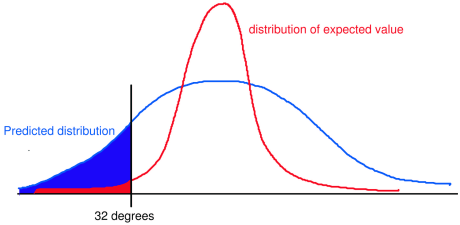
1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Decide which kind of predicted value you wish to compute, and on that basis choose one value for each explanatory variable. Denote the vector of such values X_c.
3. Taking the simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$, where $g(\cdot, \cdot)$ is the systematic component of the statistical model.
4. Simulate the outcome variable $\tilde{Y}_c$ by taking a random draw from $f(\tilde{\theta}_c, \tilde{\alpha})$, the stochastic component of the statistical model.

Repeat algorithm say $M = 1000$ times, to produce 1000 predicted values. Use these to compute a histogram for the full posterior, the average, variance, percentile values, or others.

How the Distribution of Expected and Predicted Values Differ

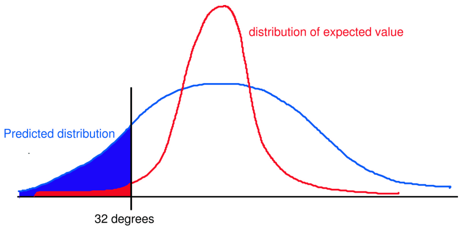


How the Distribution of Expected and Predicted Values Differ



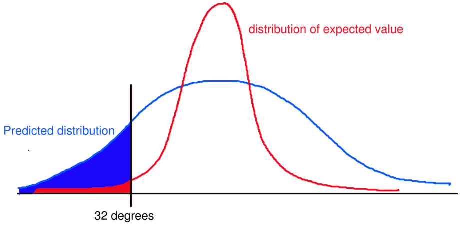
1. Predicted values are draws of Y that are or could be observed

How the Distribution of Expected and Predicted Values Differ



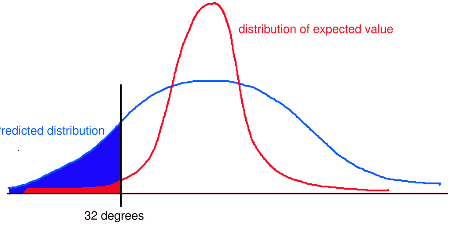
1. Predicted values are draws of Y that are or could be observed
2. Expected values are draws of fixed features of the distribution of Y, such as $E(Y)$.

How the Distribution of Expected and Predicted Values Differ

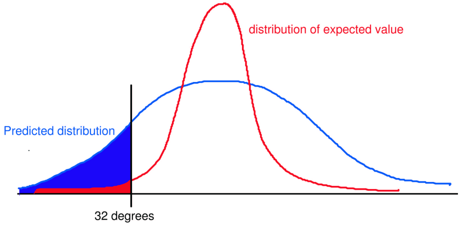


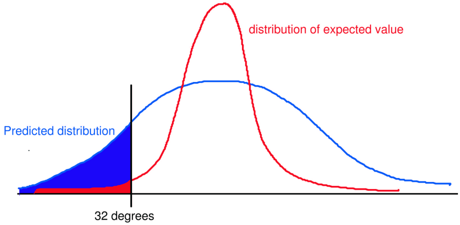
1. Predicted values are draws of Y that are or could be observed
2. Expected values are draws of fixed features of the distribution of Y, such as $E(Y)$.
3. Predicted values include estimation and fundamental uncertainty.

How the Distribution of Expected and Predicted Values Differ

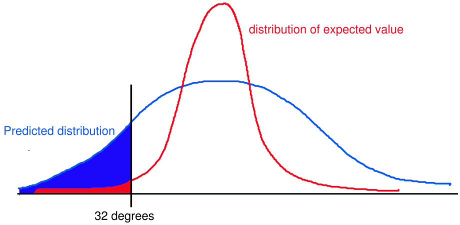


1. Predicted values are draws of Y that are or could be observed
2. Expected values are draws of fixed features of the distribution of Y, such as $E(Y)$.
3. Predicted values include estimation and fundamental uncertainty.
4. Expected values average over fundamental uncertainty

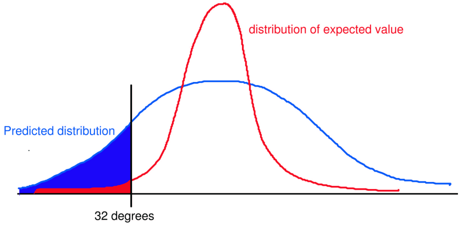

distribution of expected value
Predicted distribution
32 degrees



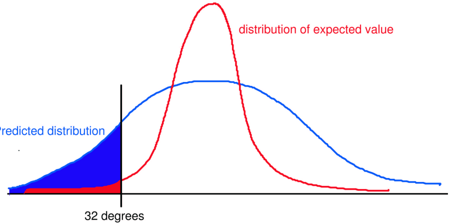
5. The variance of expected values (but not predicted values) go to 0 and n gets large.



5. The variance of expected values (but not predicted values) go to 0 and n gets large.
6. Example use of predicted value distribution: probability of temperature colder than 32° tomorrow. (Predicted temperature is uncertain because we have to estimate it and because of natural fluctuations.)



5. The variance of expected values (but not predicted values) go to 0 and n gets large.
6. Example use of predicted value distribution: probability of temperature colder than 32° tomorrow. (Predicted temperature is uncertain because we have to estimate it and because of natural fluctuations.)
7. Example use of expected value distribution: probability that the average temperature in this city on days like tomorrow will be colder than 32°. (Expected temperature is only uncertain because we have to estimate it; we don't care about natural fluctuations.)



5. The variance of expected values (but not predicted values) go to 0 and n gets large.
6. Example use of predicted value distribution: probability of temperature colder than 32° tomorrow. (Predicted temperature is uncertain because we have to estimate it and because of natural fluctuations.)
7. Example use of expected value distribution: probability that the average temperature in this city on days like tomorrow will be colder than 32°. (Expected temperature is only uncertain because we have to estimate it; we don't care about natural fluctuations.)
8. Which to use for causal effects & first differences?

Simulating the Distribution of Expected Values: An Algorithm

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)
3. Taking the one set of simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from the systematic component)

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)
3. Taking the one set of simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from the systematic component)
4. Draw *m* values of the outcome variable $\tilde{Y}_c^{(k)}$ ($k = 1, \ldots, m$) from the stochastic component $f(\tilde{\theta}_c, \tilde{\alpha})$. (This step simulates fundamental uncertainty.)

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)
3. Taking the one set of simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from the systematic component)
4. Draw m values of the outcome variable $\tilde{Y}_c^{(k)}$ $(k = 1, \dots, m)$ from the stochastic component $f(\tilde{\theta}_c, \tilde{\alpha})$. (This step simulates fundamental uncertainty.)
5. Average over the fundamental uncertainty by calculating the mean of the m simulations to yield one simulated expected value $\tilde{E}(Y_c) = \sum_{k=1}^{m} \tilde{Y}_c^{(k)}/m$.

Simulating Expected Values: Notes

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.
3. Repeat M times (say 1000), with results differing only due to estimation uncertainty

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.
3. Repeat M times (say 1000), with results differing only due to estimation uncertainty
4. Use to compute a histogram, average, standard error, confidence interval, etc.

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.
3. Repeat M times (say 1000), with results differing only due to estimation uncertainty
4. Use to compute a histogram, average, standard error, confidence interval, etc.
5. When $E(Y_c) = \theta_c$, we can skip the last two steps. E.g., in the logit model, once we simulate π_i, we don't need to draw Y and then average to get back to π_i.

Simulating First Differences

Simulating First Differences

To draw one simulated first difference:

Simulating First Differences

To draw one simulated first difference:

1. Choose vectors X_s, the <u>s</u>tarting point, X_e, the <u>e</u>nding point.

Simulating First Differences

To 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 (but reuse the random draws).

Simulating First Differences

To draw one simulated first difference:

1. Choose vectors X_s, the starting point, X_e, the ending point.
2. Apply the expected value algorithm twice, once for X_s and X_e (but reuse the random draws).
3. Take the difference in the two expected values.

Tricks for Simulating Parameters

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to (a) make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.) (b) make the maximization algorithm work faster and without explicit constraints.

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to (a) make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.) (b) make the maximization algorithm work faster and without explicit constraints.
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to (a) make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.) (b) make the maximization algorithm work faster and without explicit constraints.
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   (a) $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to (a) make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.) (b) make the maximization algorithm work faster and without explicit constraints.
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   (a) $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)
   (b) For a probability, $\pi = [1 + e^{-\eta}]^{-1}$ (a logit transformation).

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to (a) make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.) (b) make the maximization algorithm work faster and without explicit constraints.
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   (a) $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)
   (b) For a probability, $\pi = [1 + e^{-\eta}]^{-1}$ (a logit transformation).
   (c) For $-1 \leq \rho \leq 1$, use $\rho = (e^{2\eta} - 1)/(e^{2\eta} + 1)$ (Fisher's Z transformation)

Tricks for Simulating Parameters

1. Simulate all parameters, including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to (a) make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.) (b) make the maximization algorithm work faster and without explicit constraints.
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   (a) $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)
   (b) For a probability, $\pi = [1 + e^{-\eta}]^{-1}$ (a logit transformation).
   (c) For $-1 \leq \rho \leq 1$, use $\rho = (e^{2\eta} - 1)/(e^{2\eta} + 1)$ (Fisher's Z transformation)

   In all 3 cases, η is unbounded, so estimate it, simulate from it, and reparameterize back to the scale you care about.

Tricks for Simulating Quantities of Interest

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.
3. Check the *approximation error* of your simulation algorithm: Run it twice, check the number of digits of precision that don't change. If its not enough for your tables, increase M (or m) and try again.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.
3. Check the *approximation error* of your simulation algorithm: Run it twice, check the number of digits of precision that don't change. If its not enough for your tables, increase M (or m) and try again.
4. Analytical calculations and other tricks can speed simulation, or precision.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.
3. Check the *approximation error* of your simulation algorithm: Run it twice, check the number of digits of precision that don't change. If its not enough for your tables, increase M (or m) and try again.
4. Analytical calculations and other tricks can speed simulation, or precision.
5. Clarify does this all automatically in Stata. Zelig does the same and more in R.

Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

1. Logit of reported turnout on Age, Age^2, Education, Income, and Race

Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

1. Logit of reported turnout on Age, Age^2, Education, Income, and Race
2. Quantity of Interest: (nonlinear) effect of age on $\Pr(vote|X)$, holding constant Income and Race.

Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

1. Logit of reported turnout on Age, Age^2, Education, Income, and Race
2. Quantity of Interest: (nonlinear) effect of age on $\Pr(vote|X)$, holding constant Income and Race.
3. Use $M = 1000$ and compute 99% CI:

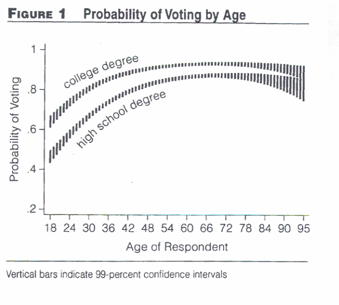
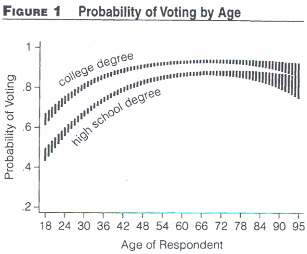
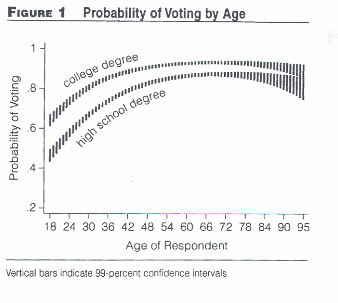


FIGURE 1 Probability of Voting by Age
college degree
high school degree
Probability of Voting
1
.8
.6
.4
.2
18 24 30 36 42 48 54 60 66 72 78 84 90 95
Age of Respondent
Vertical bars indicate 99-percent confidence intervals



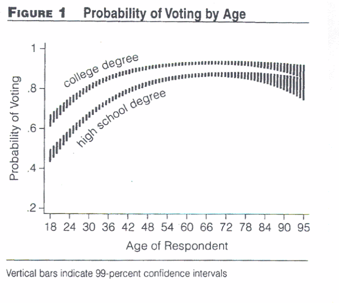
Vertical bars indicate 99-percent confidence intervals

To create this graph, simulate:



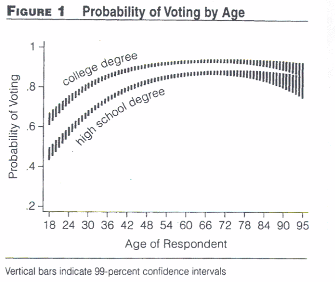
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white



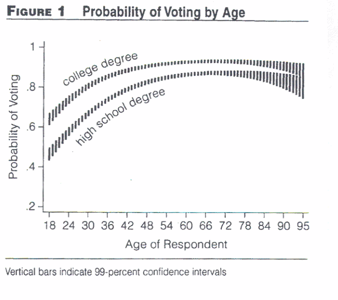
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression



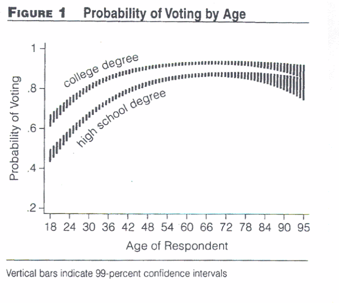
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s



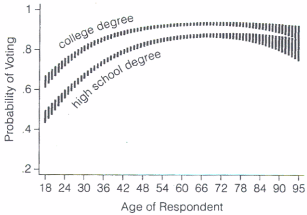
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$



To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort them in numerical order

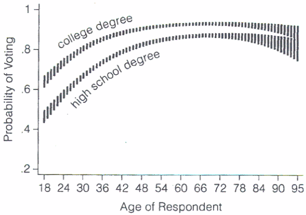


Vertical bars indicate 99-percent confidence intervals

To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort them in numerical order
6. Take 5th and 995th values as the 99% confidence interval

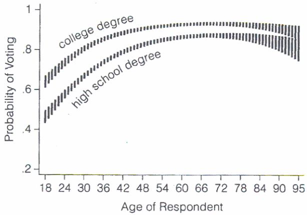
FIGURE 1 Probability of Voting by Age



Vertical bars indicate 99-percent confidence intervals

To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort them in numerical order
6. Take 5th and 995th values as the 99% confidence interval
7. Plot a vertical line on the graph at age=24 representing the CI.



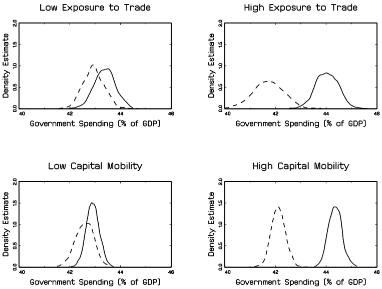
Vertical bars indicate 99-percent confidence intervals

To create this graph, simulate:

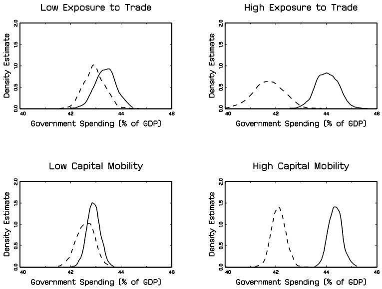
1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort them in numerical order
6. Take 5th and 995th values as the 99% confidence interval
7. Plot a vertical line on the graph at age=24 representing the CI.
8. Repeat for other ages and for college degree.

Replication of Garrett (King, Tomz and Wittenberg 2000)

Replication of Garrett (King, Tomz and Wittenberg 2000)

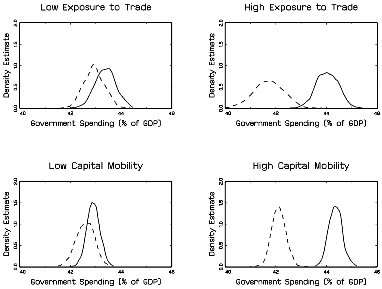


Replication of Garrett (King, Tomz and Wittenberg 2000)



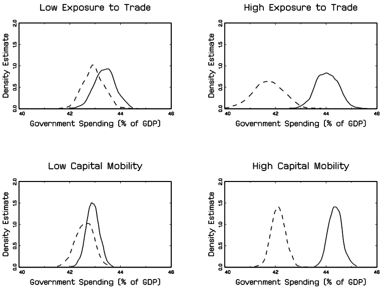
- Dependent variable: Government Spending as % of GDP

Replication of Garrett (King, Tomz and Wittenberg 2000)



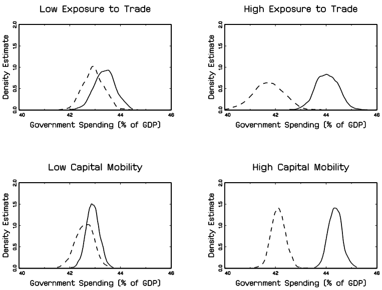
- Dependent variable: Government Spending as % of GDP
- Key explanatory variable: left-labor power (high = solid line; low = dashed)

Replication of Garrett (King, Tomz and Wittenberg 2000)



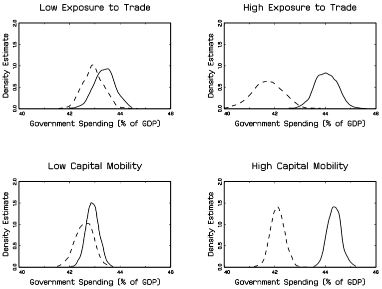
- Dependent variable: Government Spending as % of GDP
- Key explanatory variable: left-labor power (high = solid line; low = dashed)
- Garrett used only point estimates to distinguish the eight quantities represented above. What new information do we learn with this approach?

Replication of Garrett (King, Tomz and Wittenberg 2000)



- Dependent variable: Government Spending as % of GDP
- Key explanatory variable: left-labor power (high = solid line; low = dashed)
- Garrett used only point estimates to distinguish the eight quantities represented above. What new information do we learn with this approach?
- Left-labor power only has a clear effect when exposure to trade or capital mobility is high.

Replication of Garrett (King, Tomz and Wittenberg 2000)



- Dependent variable: Government Spending as % of GDP
- Key explanatory variable: left-labor power (high = solid line; low = dashed)
- Garrett used only point estimates to distinguish the eight quantities represented above. What new information do we learn with this approach?
- Left-labor power only has a clear effect when exposure to trade or capital mobility is high.
- How could we summarize this information with less real estate?

Ordered Dependent Variable Models

Ordered Dependent Variable Models

The model

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \mathrm{STN}(y_i^*|\mu_i)$$

The model

$$Y_i^* \sim \mathrm{STN}(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Ordered Dependent Variable Models

The model

$$
\begin{aligned}
Y_i^* &\sim \mathrm{STN}(y_i^* | \mu_i) \\
\mu_i &= x_i \beta
\end{aligned}
$$

Observation mechanism

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \text{STN}(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Observation mechanism

$$y_{ij} =$$

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \mathrm{STN}(y_i^* | \mu_i)$$
$$\mu_i = x_i \beta$$

Observation mechanism

$$y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1,i} \leq y_i^* \leq \tau_{j,i} \\ 0 & \text{otherwise} \end{cases}$$

Ordered Dependent Variable Models

The model

$$
\begin{aligned}
Y_i^* &\sim \mathrm{STN}(y_i^*|\mu_i) \\
\mu_i &= x_i\beta
\end{aligned}
$$

Observation mechanism

$$
y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1,i} \le y_i^* \le \tau_{j,i} \\ 0 & \text{otherwise} \end{cases}
$$

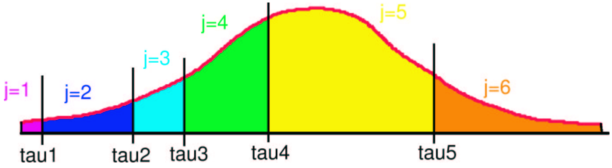
Ordered Dependent Variable Models

The model

$$Y_i^* \sim \text{STN}(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Observation mechanism

$$y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1,i} \le y_i^* \le \tau_{j,i} \\ 0 & \text{otherwise} \end{cases}$$



Ordinal Dependent Variable Models: Notes

Ordinal Dependent Variable Models: Notes

1. If we switch to $Y_i^* \sim \text{STL}(y_i^*|\mu_i)$, we get an ordinal logit model

Ordinal Dependent Variable Models: Notes

1. If we switch to $Y_i^* \sim \text{STL}(y_i^*|\mu_i)$, we get an ordinal logit model
2. One dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.

Ordinal Dependent Variable Models: Notes

1. If we switch to $Y_i^* \sim \text{STL}(y_i^*|\mu_i)$, we get an ordinal logit model
2. One dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, this is a linear-normal regression model

Ordinal Dependent Variable Models: Notes

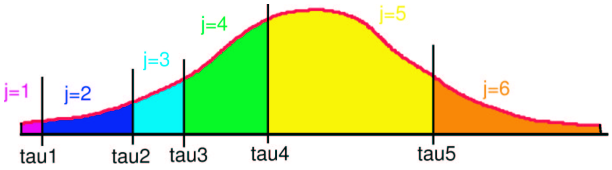
1. If we switch to $Y_i^* \sim \text{STL}(y_i^*|\mu_i)$, we get an ordinal logit model
2. One dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, this is a linear-normal regression model
4. If a dichotomous realization of Y^* is observed, its a probit model

Ordinal Dependent Variable Models: Notes

1. If we switch to $Y_i^* \sim \text{STL}(y_i^*|\mu_i)$, we get an ordinal logit model
2. One dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, this is a linear-normal regression model
4. If a dichotomous realization of Y^* is observed, its a probit model
5. This is the same model, and the same parameters are being estimated; only the observation mechanism differs.

Ordinal Dependent Variable Models: Notes

1. If we switch to $Y_i^* \sim \text{STL}(y_i^*|\mu_i)$, we get an ordinal logit model
2. One dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, this is a linear-normal regression model
4. If a dichotomous realization of Y^* is observed, its a probit model
5. This is the same model, and the same parameters are being estimated; only the observation mechanism differs.



Deriving the likelihood function

Deriving the likelihood function

First the probability of each observation, then the joint probability.

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\Pr(Y_{ji} = 1) = \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j)$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}\Pr(Y_{ji}=1) &= \Pr(\tau_{j-1} \le Y_i^* \le \tau_j)\\ &= \int_{\tau_{j-1}}^{\tau_j} \mathrm{STN}(y_i^*|\mu_i)dy_i^*\end{aligned}$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ji} = 1) &= \Pr(\tau_{j-1} \le Y_i^* \le \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^*|\mu_i) dy_i^* \\
&= F_{stn}(\tau_j|\mu_i) - F_{stn}(\tau_{j-1}|\mu_i)
\end{aligned}$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ji} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \mathrm{STN}(y_i^* | \mu_i) dy_i^* \\
&= F_{stn}(\tau_j | \mu_i) - F_{stn}(\tau_{j-1} | \mu_i) \\
&= F_{stn}(\tau_j | x_i\beta) - F_{stn}(\tau_{j-1} | x_i\beta)
\end{aligned}$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ji} = 1) &= \Pr(\tau_{j-1} \le Y_i^* \le \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \mathrm{STN}(y_i^*|\mu_i)dy_i^* \\
&= F_{stn}(\tau_j|\mu_i) - F_{stn}(\tau_{j-1}|\mu_i) \\
&= F_{stn}(\tau_j|x_i\beta) - F_{stn}(\tau_{j-1}|x_i\beta)
\end{aligned}$$

The joint probability is then:

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ji} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^* | \mu_i) dy_i^* \\
&= F_{stn}(\tau_j | \mu_i) - F_{stn}(\tau_{j-1} | \mu_i) \\
&= F_{stn}(\tau_j | x_i\beta) - F_{stn}(\tau_{j-1} | x_i\beta)
\end{aligned}$$

The joint probability is then:

$$P(Y) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \Pr(Y_{ji} = 1)^{y_{ij}} \right]$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ji} = 1) &= \Pr(\tau_{j-1} \le Y_i^* \le \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^* | \mu_i) dy_i^* \\
&= F_{stn}(\tau_j | \mu_i) - F_{stn}(\tau_{j-1} | \mu_i) \\
&= F_{stn}(\tau_j | x_i\beta) - F_{stn}(\tau_{j-1} | x_i\beta)
\end{aligned}$$

The joint probability is then:

$$P(Y) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \Pr(Y_{ji} = 1)^{y_{ij}} \right]$$

Bracketed portion has only one factor for each i.

Deriving the likelihood function

The Log-likelihood:

(Careful of constraints during optimization: $\tau_{j-1} < \tau_j$, $\forall j$)

Deriving the likelihood function

The Log-likelihood:

$$\ln L(\beta, \tau|y) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \Pr(Y_{ji} = 1)$$

(Careful of constraints during optimization: $\tau_{j-1} < \tau_j$, $\forall j$)

Deriving the likelihood function

The Log-likelihood:

$$\begin{aligned}\ln L(\beta,\tau|y) &= \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \ln \Pr(Y_{ji}=1) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \ln \left[F_{stn}(\tau_j|x_i\beta) - F_{stn}(\tau_{j-1}|x_i\beta)\right]\end{aligned}$$

(Careful of constraints during optimization: $\tau_{j-1} < \tau_j$, $\forall j$)

Interpretation: Ordinal Probit

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.
3. One first difference has an effect on all J probabilities.

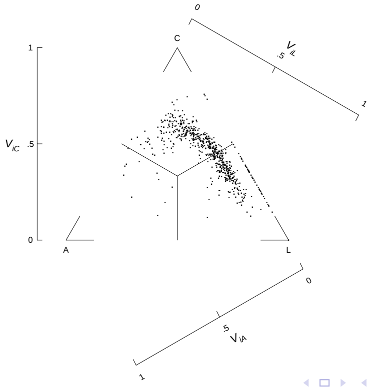
Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.
3. One first difference has an effect on all J probabilities.
4. When one probability goes up, at least one of the others must go down.

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.
3. One first difference has an effect on all J probabilities.
4. When one probability goes up, at least one of the others must go down.
5. Can use ternary diagrams if $J = 3$

Interpretation: Ordinal Probit



How do you know which model is better?

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set data and compare to the test set.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set data and compare to the test set.
   (e) Comparisons should be to the average prediction and also to the full distribution.

How do you know which model is better?

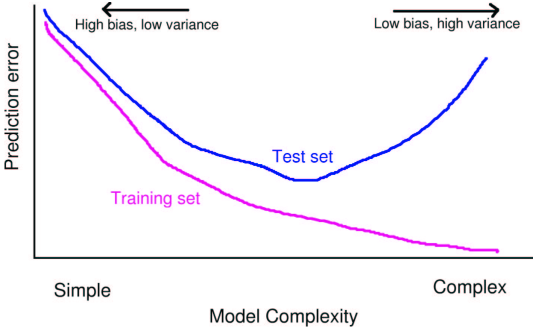
1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set data and compare to the test set.
   (e) Comparisons should be to the average prediction and also to the full distribution.
   (f) E.g., if a set of predictions have $\Pr(y = 1) = 0.2$, then 20% of these observations in the test set should be 1s.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set data and compare to the test set.
   (e) Comparisons should be to the average prediction and also to the full distribution.
   (f) E.g., if a set of predictions have $\Pr(y = 1) = 0.2$, then 20% of these observations in the test set should be 1s.
   (g) The best test sets are really out of sample, not even available yet.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job is to find the underlying structure that persists, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set data and compare to the test set.
   (e) Comparisons should be to the average prediction and also to the full distribution.
   (f) E.g., if a set of predictions have $\Pr(y = 1) = 0.2$, then 20% of these observations in the test set should be 1s.
   (g) The best test sets are really out of sample, not even available yet.
   (h) If the world changes between the time the training and test sets are observed, even an otherwise good model will fail. But this is still the right test.



See Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Chapter 7: Fig 7.1.

(i) Binary variable predictions require a normative decision.

(i) Binary variable predictions require a normative decision.
   i. Let C be number of times more costly misclassifying a 1 is than a 0.

(i) Binary variable predictions require a normative decision.

   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.

(i) Binary variable predictions require a normative decision.
   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.

(i) Binary variable predictions require a normative decision.

    i. Let C be number of times more costly misclassifying a 1 is than a 0.
    ii. C must be chosen independently of the data.
    iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
    iv. People often choose $C = 1$, but without justification.

(i) Binary variable predictions require a normative decision.
   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.

(i) Binary variable predictions require a normative decision.

   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$

(i) Binary variable predictions require a normative decision.

i. Let C be number of times more costly misclassifying a 1 is than a 0.
ii. C must be chosen independently of the data.
iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
iv. People often choose $C = 1$, but without justification.
v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
   - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
   - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$

(i) Binary variable predictions require a normative decision.

i. Let C be number of times more costly misclassifying a 1 is than a 0.
ii. C must be chosen independently of the data.
iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
iv. People often choose $C = 1$, but without justification.
v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
    - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
    - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(i) Binary variable predictions require a normative decision.

   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
      - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
   vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

(i) Binary variable predictions require a normative decision.

   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
      - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
   vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

   (i) Compute %1s and %0s correctly predicted for every possible value of C

(i) Binary variable predictions require a normative decision.

   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
      - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
   vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

   (i) Compute %1s and %0s correctly predicted for every possible value of C
   (ii) Plot %1s by %0s

(i) Binary variable predictions require a normative decision.
   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
      - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
   vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves
   (i) Compute %1s and %0s correctly predicted for every possible value of C
   (ii) Plot %1s by %0s
   (iii) Overlay curves for several models on the same graph.

(i) Binary variable predictions require a normative decision.
   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
      - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
   vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

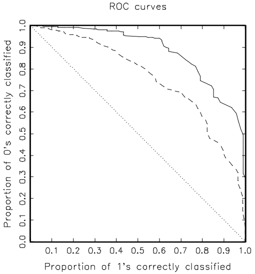
(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves
   (i) Compute %1s and %0s correctly predicted for every possible value of C
   (ii) Plot %1s by %0s
   (iii) Overlay curves for several models on the same graph.
   (iv) If one curve is above another the whole way, then that model *dominates* the other. It's better no matter your normative decision (about C)

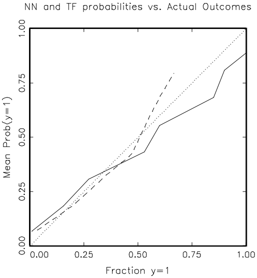
(i) Binary variable predictions require a normative decision.

   i. Let C be number of times more costly misclassifying a 1 is than a 0.
   ii. C must be chosen independently of the data.
   iii. C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
   iv. People often choose $C = 1$, but without justification.
   v. Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1+C)$ and 0 otherwise.
      - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
      - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
   vi. Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

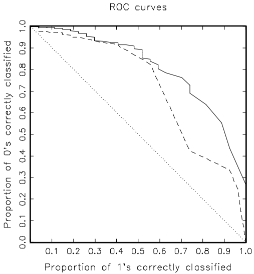
(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

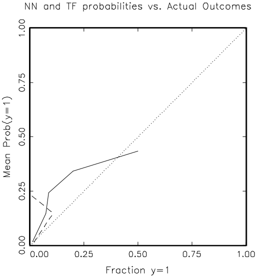
   (i) Compute %1s and %0s correctly predicted for every possible value of C
   (ii) Plot %1s by %0s
   (iii) Overlay curves for several models on the same graph.
   (iv) If one curve is above another the whole way, then that model *dominates* the other. It's better no matter your normative decision (about C)
   (v) Otherwise, one model is better than the other in only given specificed ranges of C (i.e., for only some normative perspectives).





In sample ROC, on left (from Gary King and Langche Zeng. "Improving Forecasts of State Failure," World Politics, Vol. 53, No. 4 (July, 2001): 623-58)





Out of sample ROC on left.

2. Cross-validation

2. Cross-validation

   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.

2. Cross-validation
   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
   (b) Useful for smaller data sets; real test sets are better.

2. Cross-validation
   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
   (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.

2. Cross-validation
    (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
    (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.
4. Fit: continuous variables

2. Cross-validation
   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
   (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.
4. Fit: continuous variables
   (a) The usual regression diagnostics

2. Cross-validation
   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
   (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.
4. Fit: continuous variables
   (a) The usual regression diagnostics
   (b) E.G., plots of $y - \hat{y}$ by X, Y or $\hat{y}$

2. Cross-validation
   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
   (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.
4. Fit: continuous variables
   (a) The usual regression diagnostics
   (b) E.G., plots of $y - \hat{y}$ by X, Y or $\hat{y}$
   (c) Be sure to check more than the means. E.g., plot $y - \hat{y}$ by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.

2. Cross-validation
    (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
    (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.
4. Fit: continuous variables
    (a) The usual regression diagnostics
    (b) E.G., plots of $y - \hat{y}$ by X, Y or $\hat{y}$
    (c) Be sure to check more than the means. E.g., plot $y - \hat{y}$ by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
    (d) Transform bounded variables

2. Cross-validation
   (a) The idea: set aside k observations as the test set, do comparisons, and then set aside another set of k observations. Repeat for all $\binom{n}{k}$ subsets, and average results.
   (b) Useful for smaller data sets; real test sets are better.
3. Fit, in general: Look for all possible observable implications of a model, and compare to observations.
4. Fit: continuous variables
   (a) The usual regression diagnostics
   (b) E.G., plots of $y - \hat{y}$ by X, Y or $\hat{y}$
   (c) Be sure to check more than the means. E.g., plot $y - \hat{y}$ by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
   (d) Transform bounded variables
   (e) Transform heteroskedastic results

5. Fit: dichotomous variables

5. Fit: dichotomous variables

   (a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,. . . , $[0.9, 1]$.

5. Fit: dichotomous variables
   (a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,. . . , $[0.9, 1]$.
   (b) From the observations in each bin, compute (a) the mean predictions (which will be near 0.05, 0.15, etc.) and (b) the average fraction of 1s.

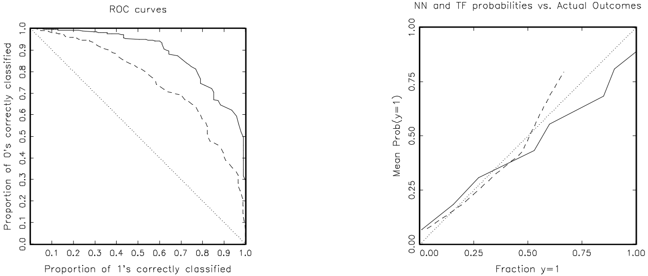
5. Fit: dichotomous variables

(a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,. . . , $[0.9, 1]$.

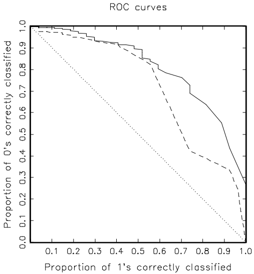
(b) From the observations in each bin, compute (a) the mean predictions (which will be near 0.05, 0.15, etc.) and (b) the average fraction of 1s.

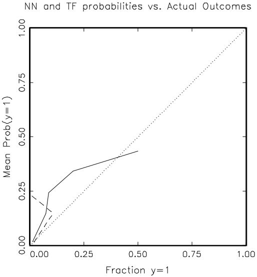
(c) Plot (a) by (b) and look for systematic deviation from 45° line.

5. Fit: dichotomous variables
    (a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,..., $[0.9, 1]$.
    (b) From the observations in each bin, compute (a) the mean predictions (which will be near 0.05, 0.15, etc.) and (b) the average fraction of 1s.
    (c) Plot (a) by (b) and look for systematic deviation from 45° line.



In sample calibration graph on right (from Gary King and Langche Zeng. "Improving Forecasts of State Failure," World Politics, Vol. 53, No. 4 (July, 2001): 623-58)





Out of sample calibration graph on right.

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{N_i - y_i}$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1+e^{-x_i\beta}]^{-1}$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{N_i - y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{N_i - y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

and a likelihood of

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{N_i - y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

and a likelihood of

$$L(\pi|y) \propto \prod_{i=1}^{n} \text{Binomial}(y_i|\pi_i)$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

and a likelihood of

$$\begin{aligned} L(\pi|y) &\propto \prod_{i=1}^{n} \text{Binomial}(y_i|\pi_i) \\ &= \prod_{i=1}^{n} \binom{N_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{N_i-y_i} \end{aligned}$$

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

and after substituting in the systematic component:

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

and after substituting in the systematic component:

$$\ln L(\beta|y) \doteq \sum_{i=1}^{n} \left\{ -y_i \ln[1 + e^{-x_i\beta}] + (N_i - y_i) \ln\left(1 - [1 + e^{-x_i\beta}]^{-1}\right) \right\}$$

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

and after substituting in the systematic component:

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

Grouped Uncorrelated Binary Variables

Notes:

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
    (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
    (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
    (c) Set X to your choice of values, X_c
    (d) Calculate simulations of the probability that any of the component binary variables is a one:

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:

$$\tilde{\pi}_c = [1 + e^{-x_c\tilde{\beta}}]^{-1}$$

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c\tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.
   (f) If simulations of y are needed, go one more step and draw $\tilde{y}$ from Binomial$(y_i|\pi_i)$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$f_{ebb}(y_i|\pi_i, \gamma) = \Pr(Y_i = y_i|\pi_i, \gamma, N)$$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$\begin{aligned} f_{ebb}(y_i|\pi_i,\gamma) &= \Pr(Y_i = y_i|\pi_i,\gamma,N) \\ &= \frac{N!}{y_i!(N-y_i)!} \prod_{j=0}^{y_i-1}(\pi_i+\gamma j) \prod_{j=0}^{N-y_i-1}(1-\pi_i+\gamma j)/\prod_{j=0}^{N-1}(1+\gamma j) \end{aligned}$$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$\begin{aligned} f_{ebb}(y_i|\pi_i,\gamma) &= \Pr(Y_i = y_i|\pi_i,\gamma,N) \\ &= \frac{N!}{y_i!(N-y_i)!}\prod_{j=0}^{y_i-1}(\pi_i+\gamma j)\prod_{j=0}^{N-y_i-1}(1-\pi_i+\gamma j)/\prod_{j=0}^{N-1}(1+\gamma j) \end{aligned}$$

and

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$\begin{aligned} f_{ebb}(y_i|\pi_i,\gamma) &= \Pr(Y_i = y_i|\pi_i,\gamma,N) \\ &= \frac{N!}{y_i!(N-y_i)!} \prod_{j=0}^{y_i-1}(\pi_i+\gamma j) \prod_{j=0}^{N-y_i-1}(1-\pi_i+\gamma j)/\prod_{j=0}^{N-1}(1+\gamma j) \end{aligned}$$

and

$$\pi_i = \frac{1}{1+e^{-x_i\beta}}$$

The probability model of all the data:

The probability model of all the data:

$$
\begin{aligned}
\Pr(Y = y|\beta, \gamma; N) = & \prod_{i=1}^{n} \left(\frac{N!}{y_i!(N-y_i)!} \right) \\
& \times \prod_{j=0}^{y_i-1} \left\{ [1+\exp(-x_i\beta)]^{-1} + \gamma j \right\} \\
& \times \prod_{j=0}^{N-y_i-1} \left\{ [1+\exp(x_i\beta)]^{-1} + \gamma j \right\} \Big/ \prod_{j=0}^{N-1} (1+\gamma j)
\end{aligned}
$$

The probability model of all the data:

$$
\begin{aligned}
\Pr(Y = y|\beta, \gamma; N) = & \prod_{i=1}^{n} \left(\frac{N!}{y_i!(N - y_i)!} \right) \\
& \times \prod_{j=0}^{y_i - 1} \left\{ [1 + \exp(-x_i\beta)]^{-1} + \gamma j \right\} \\
& \times \prod_{j=0}^{N - y_i - 1} \left\{ [1 + \exp(x_i\beta)]^{-1} + \gamma j \right\} / \prod_{j=0}^{N-1} (1 + \gamma j)
\end{aligned}
$$

$$
\begin{aligned}
\ln L(\beta, \gamma|y) = & \sum_{i=1}^{n} \left\{ \ln \left(\frac{N!}{y_i!(N - y_i)!} \right) \right. \\
& + \sum_{j=0}^{y_i - 1} \ln \left\{ [1 + \exp(-x_i\beta)]^{-1} + \gamma j \right\} \\
& \left. + \sum_{j=0}^{N - y_i - 1} \ln \left\{ [1 + \exp(x_i\beta)]^{-1} + \gamma j \right\} - \sum_{j=0}^{N-1} \ln(1 + \gamma j) \right\}
\end{aligned}
$$

The probability model of all the data:

$$
\begin{aligned}
\Pr(Y = y|\beta,\gamma;N) = &\prod_{i=1}^{n}\left(\frac{N!}{y_i!(N-y_i)!}\right)\\
&\times\prod_{j=0}^{y_i-1}\left\{[1+\exp(-x_i\beta)]^{-1}+\gamma j\right\}\\
&\times\prod_{j=0}^{N-y_i-1}\left\{[1+\exp(x_i\beta)]^{-1}+\gamma j\right\} \Big/ \prod_{j=0}^{N-1}(1+\gamma j)\\
\ln L(\beta,\gamma|y) = &\sum_{i=1}^{n}\left\{\ln\left(\frac{N!}{y_i!(N-y_i)!}\right)\right.\\
&+\sum_{j=0}^{y_i-1}\ln\left\{[1+\exp(-x_i\beta)]^{-1}+\gamma j\right\}\\
&\left.+\sum_{j=0}^{N-y_i-1}\ln\left\{[1+\exp(x_i\beta)]^{-1}+\gamma j\right\}-\sum_{j=0}^{N-1}\ln(1+\gamma j)\right\}\\
\doteq &\sum_{i=1}^{n}\left\{\sum_{j=0}^{y_i-1}\ln\left\{[1+\exp(-x_i\beta)]^{-1}+\gamma j\right\}\right.\\
&\left.+\sum_{j=0}^{N-y_i-1}\ln\left\{[1+\exp(x_i\beta)]^{-1}+\gamma j\right\}-\sum_{j=0}^{N-1}\ln(1+\gamma j)\right\}
\end{aligned}
$$

Notes:

Notes:

1. The math *looks* complicated.

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\hat{\gamma}$ from the multivariate normal with mean vector $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix that come from `optim`.

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
    (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
    (b) Draw many values of $\tilde{\beta}$ and $\hat{\gamma}$ from the multivariate normal with mean vector $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix that come from `optim`.
    (c) Set X to your choice of values, X_c

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
    (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
    (b) Draw many values of $\tilde{\beta}$ and $\hat{\gamma}$ from the multivariate normal with mean vector $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix that come from `optim`.
    (c) Set X to your choice of values, X_c
    (d) Calculate simulations of the probability that any of the component binary variables is a one:

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\hat{\gamma}$ from the multivariate normal with mean vector $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:

$$\tilde{\pi}_c = [1 + e^{-x_c\tilde{\beta}}]^{-1}$$

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\hat{\gamma}$ from the multivariate normal with mean vector $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c\tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.

Notes:

1. The math *looks* complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\hat{\gamma}$ from the multivariate normal with mean vector $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c\tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.
   (f) If simulations of y are needed, go one more step and draw $\tilde{y}$ from $f_{ebb}(y_i|\pi_i)$

Event Count Models: Poisson

Uses:

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,
2. The number of triplets born in Norway in each half-decade

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,
2. The number of triplets born in Norway in each half-decade
3. The annual number of presidential appointments to the Supreme Court

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,
2. The number of triplets born in Norway in each half-decade
3. The annual number of presidential appointments to the Supreme Court
4. The number of Coups d'Etat in black African states

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,
2. The number of triplets born in Norway in each half-decade
3. The annual number of presidential appointments to the Supreme Court
4. The number of Coups d'Etat in black African states
5. The number of medical consultations for each survey respondent

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,
2. The number of triplets born in Norway in each half-decade
3. The annual number of presidential appointments to the Supreme Court
4. The number of Coups d'Etat in black African states
5. The number of medical consultations for each survey respondent
6. For each of these examples, the upper limit on the number of observed events is theoretically infinite.

Event Count Models: Poisson

Uses:

1. The number of cooperative and conflictual international incidents,
2. The number of triplets born in Norway in each half-decade
3. The annual number of presidential appointments to the Supreme Court
4. The number of Coups d'Etat in black African states
5. The number of medical consultations for each survey respondent
6. For each of these examples, the upper limit on the number of observed events is theoretically infinite.
7. Some event count datasets go over time (count per year), some across areas (count per state), and some both.

Recall Poisson distribution's first principles:

Recall Poisson distribution's first principles:

1. Begin with an observation period:

Recall Poisson distribution's first principles:

1. Begin with an observation period:

Recall Poisson distribution's first principles:

1. Begin with an observation period:

2. All assumptions are about the events that occur between the start and when we observe the count. The process of event generation is assumed not observed.

Recall Poisson distribution's first principles:

1. Begin with an observation period:

2. All assumptions are about the events that occur between the start and when we observe the count. The process of event generation is assumed not observed.
3. 0 events occur at the start of the period

Recall Poisson distribution's first principles:

1. Begin with an observation period:

2. All assumptions are about the events that occur between the start and when we observe the count. The process of event generation is assumed not observed.
3. 0 events occur at the start of the period
4. Only observe number of events at the end of the period

Recall Poisson distribution's first principles:

1. Begin with an observation period:

2. All assumptions are about the events that occur between the start and when we observe the count. The process of event generation is assumed not observed.
3. 0 events occur at the start of the period
4. Only observe number of events at the end of the period
5. No 2 events can occur at the same time

Recall Poisson distribution's first principles:

1. Begin with an observation period:

2. All assumptions are about the events that occur between the start and when we observe the count. The process of event generation is assumed not observed.
3. 0 events occur at the start of the period
4. Only observe number of events at the end of the period
5. No 2 events can occur at the same time
6. $\Pr(\text{event at time } t \mid \text{all events up to time } t-1)$ is constant for all t.

The Poisson regression model:

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$

The Poisson regression model:

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(y_i|\lambda_i) \\
\lambda_i &= \exp(x_i\beta)
\end{aligned}
$$

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.

The Poisson regression model:

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(y_i|\lambda_i) \\
\lambda_i &= \exp(x_i\beta)
\end{aligned}
$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\text{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\text{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\text{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

$$\ln L(\beta|y) = \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\}$$

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$P(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

$$\begin{aligned}
\ln L(\beta|y) &= \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\} \\
&= \sum_{i=1}^{n} \{(x_i\beta)y_i - \exp(x_i\beta) - \ln y_i!\}
\end{aligned}$$

The Poisson regression model:

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(y_i|\lambda_i) \\
\lambda_i &= \exp(x_i\beta)
\end{aligned}
$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$
\text{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}
$$

The log-likelihood:

$$
\begin{aligned}
\ln L(\beta|y) &= \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\} \\
&= \sum_{i=1}^{n} \{(x_i\beta)y_i - \exp(x_i\beta) - \ln y_i!\} \\
&\doteq \sum_{i=1}^{n} \{(x_i\beta)y_i - \exp(x_i\beta)\}
\end{aligned}
$$

Poisson Model Interpretation

Poisson Model Interpretation

1. Derivative method:

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial\lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

so we could use $\bar{y}\beta$ for an approximate linearized effect.

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$
   (c) Compute $\tilde{\lambda}_c = \exp(X_c\tilde{\beta})$

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$
   (c) Compute $\tilde{\lambda}_c = \exp(X_c\tilde{\beta})$
   (d) Draw Y_c from Poisson$(y|\tilde{\lambda})$

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$
   (c) Compute $\tilde{\lambda}_c = \exp(X_c\tilde{\beta})$
   (d) Draw Y_c from Poisson$(y|\tilde{\lambda})$

3. Under the Poisson: $V(Y_i|X_i) = E(Y_i|X_i)$, which is heteroskedastic and fixed.

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$
   (c) Compute $\tilde{\lambda}_c = \exp(X_c\tilde{\beta})$
   (d) Draw Y_c from Poisson$(y|\tilde{\lambda})$
3. Under the Poisson: $V(Y_i|X_i) = E(Y_i|X_i)$, which is heteroskedastic and fixed.
   (a) Level of disperson is conditional on X, so it changes with the specification

Poisson Model Interpretation

1. Derivative method:

$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$
   (c) Compute $\tilde{\lambda}_c = \exp(X_c\tilde{\beta})$
   (d) Draw Y_c from Poisson$(y|\tilde{\lambda})$

3. Under the Poisson: $V(Y_i|X_i) = E(Y_i|X_i)$, which is heteroskedastic and fixed.
   (a) Level of disperson is conditional on X, so it changes with the specification
   (b) $V(Y_i|X_i) > E(Y_i|X_i)$ is overdispersion: standard errors will be too small (very common)

Poisson Model Interpretation

1. Derivative method:

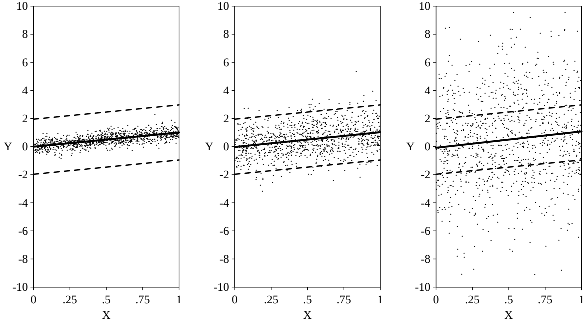
$$\frac{\partial \lambda_i}{\partial X_i^1} = \exp(x_i\beta)\beta = \lambda_i\beta$$

   so we could use $\bar{y}\beta$ for an approximate linearized effect.

2. To simulate:
   (a) Set X_c
   (b) Draw $\tilde{\beta}$ from $N\left(b|\hat{\beta}, \hat{V}(\hat{\beta})\right)$
   (c) Compute $\tilde{\lambda}_c = \exp(X_c\tilde{\beta})$
   (d) Draw Y_c from Poisson$(y|\tilde{\lambda})$

3. Under the Poisson: $V(Y_i|X_i) = E(Y_i|X_i)$, which is heteroskedastic and fixed.
   (a) Level of disperson is conditional on X, so it changes with the specification
   (b) $V(Y_i|X_i) > E(Y_i|X_i)$ is overdispersion: standard errors will be too small (very common)
   (c) $V(Y_i|X_i) < E(Y_i|X_i)$ is underdispersion: standard errors will be too big

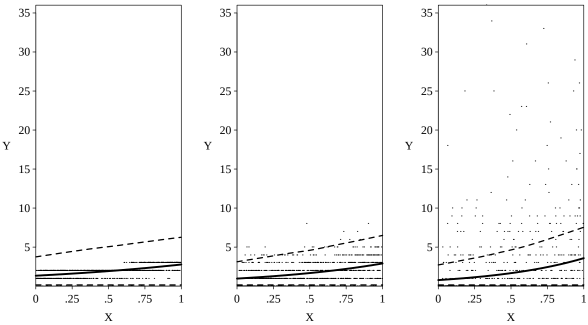
What happens without an extra parameter? Stylized Normal.



(a) (b) (c)

$E(Y|X)$ and 95% CI.

What happens without an extra parameter? Poisson



(a)

(b)

(c)

$E(Y|X)$ and 95% CI.

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$
$$E(Y_i) \equiv \phi = \exp(x_i\beta)$$

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$
$$E(Y_i) \equiv \phi = \exp(x_i\beta)$$

Interpretation:

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$
$$E(Y_i) \equiv \phi = \exp(x_i\beta)$$

Interpretation:

1. $V(Y|X) = \phi\sigma^2$, and $\sigma^2 > 1$

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$
$$E(Y_i) \equiv \phi = \exp(x_i\beta)$$

Interpretation:

1. $V(Y|X) = \phi\sigma^2$, and $\sigma^2 > 1$
2. Recall:

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$
$$E(Y_i) \equiv \phi = \exp(x_i\beta)$$

Interpretation:

1. $V(Y|X) = \phi\sigma^2$, and $\sigma^2 > 1$
2. Recall:
$$\lim_{\sigma^2 \to 1}\text{Negbin}(y_i|\phi_i, \sigma^2) = \text{Poisson}(y_i|\phi_i)$$

Negative Binomial Event Count Model

For overdispersed data (conditional on X), the model:

$$Y_i \sim \text{NegBin}(y_i|\phi, \sigma^2)$$
$$E(Y_i) \equiv \phi = \exp(x_i\beta)$$

Interpretation:

1. $V(Y|X) = \phi\sigma^2$, and $\sigma^2 > 1$
2. Recall:
$$\lim_{\sigma^2 \to 1}\text{Negbin}(y_i|\phi_i, \sigma^2) = \text{Poisson}(y_i|\phi_i)$$
3. So a test of Poisson vs NegBin can be based on σ^2, even though a likelihood ratio doesn't work since Poisson doesn't exactly nest within Negbin.

The Negative Binomial Likelihood

The probability density of all the data:

The Negative Binomial Likelihood

The probability density of all the data:

$$P(y|\phi, \sigma^2) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\phi}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\phi}{\sigma^2-1}\right)} \left(\frac{\sigma^2 - 1}{\sigma^2}\right)^{y_i} \left(\sigma^2\right)^{\frac{-\phi}{\sigma^2-1}}$$

The Negative Binomial Likelihood

The probability density of all the data:

$$P(y|\phi, \sigma^2) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\phi}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\phi}{\sigma^2-1}\right)} \left(\frac{\sigma^2 - 1}{\sigma^2}\right)^{y_i} \left(\sigma^2\right)^{\frac{-\phi}{\sigma^2-1}}$$

The log-likelihood:

The Negative Binomial Likelihood

The probability density of all the data:

$$P(y|\phi,\sigma^2) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\phi}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\phi}{\sigma^2-1}\right)} \left(\frac{\sigma^2-1}{\sigma^2}\right)^{y_i} \left(\sigma^2\right)^{\frac{-\phi}{\sigma^2-1}}$$

The log-likelihood:

$$\begin{aligned}\ln L(\phi,\sigma^2|y) = \sum_{i=1}^{n} \Bigg\{ & \ln\Gamma\left(\frac{\phi}{\sigma^2-1} + y_i\right) - \ln y! \\ & - \ln\Gamma\left(\frac{\phi}{\sigma^2-1}\right) + y_i \ln\left(\frac{\sigma^2-1}{\sigma^2}\right) - \left(\frac{\phi}{\sigma^2-1}\right)\ln(\sigma^2)\Bigg\}\end{aligned}$$

Computational Issues

Computational Issues

1. $\ln\Gamma(a)$ with large a is hard to compute in 2 steps (since $\Gamma(a) \approx a!$ is immense) but easy in one. In Gauss, a `lngamma` function exists; in R, see `lgamma`.

Computational Issues

1. $\ln \Gamma(a)$ with large a is hard to compute in 2 steps (since $\Gamma(a) \approx a!$ is immense) but easy in one. In Gauss, a `lngamma` function exists; in R, see `lgamma`.
2. β is unbounded as is, and so no need to reparameterize.

Computational Issues

1. $\ln \Gamma(a)$ with large a is hard to compute in 2 steps (since $\Gamma(a) \approx a!$ is immense) but easy in one. In Gauss, a `lngamma` function exists; in R, see `lgamma`.
2. β is unbounded as is, and so no need to reparameterize.
3. $\sigma^2 > 1$, and so we estimate γ, where $\sigma^2 = e^\gamma + 1$.

A Generalized Event Count (GEC) Model

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

$$Y_i \sim \text{GEC}(y_i|\lambda_i, \sigma^2) \equiv \text{P}(Y = y_i|\lambda_i, \sigma^2)$$

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

$$
\begin{aligned}
Y_i &\sim \text{GEC}(y_i|\lambda_i, \sigma^2) \equiv \text{P}(Y = y_i|\lambda_i, \sigma^2) \\
&= \text{GEC}\left(y_i|e^{X_i\beta}, \sigma^2\right)
\end{aligned}
$$

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

$$\begin{aligned}
Y_i \sim \mathrm{GEC}(y_i|\lambda_i, \sigma^2) &\equiv \mathrm{P}(Y = y_i|\lambda_i, \sigma^2) \\
&= \mathrm{GEC}\left(y_i|e^{X_i\beta}, \sigma^2\right) \\
&= \frac{1}{y_i!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(y_i, 1-\frac{1}{\sigma^2}\right)} \left[\sum_{j=0}^{y_i^{\max}} \frac{1}{j!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(j, 1-\frac{1}{\sigma^2}\right)}\right]^{-1},
\end{aligned}$$

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

$$
\begin{aligned}
Y_i \sim \text{GEC}(y_i|\lambda_i,\sigma^2) &\equiv \text{P}(Y=y_i|\lambda_i,\sigma^2)\\
&= \text{GEC}\left(y_i|e^{X_i\beta},\sigma^2\right)\\
&= \frac{1}{y_i!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(y_i,1-\frac{1}{\sigma^2}\right)}\left[\sum_{j=0}^{y_i^{\max}}\frac{1}{j!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(j,1-\frac{1}{\sigma^2}\right)}\right]^{-1},
\end{aligned}
$$

where $y_i^{\max}=\infty$ for $\sigma^2\geq 1$ and, letting $n_i=\lambda_i/(1-\sigma^2)$, $y_i^{\max}=[n_i+1)$ for $0<\sigma^2<1$, and where $[x)$ equals $x-1$ for integer x and floor(x) for non-integer x. Note also that:

$$
x^{(m,\delta)}=\begin{cases}\prod_{i=0}^{m-1}(x+\delta i)=x(x+\delta)(x+2\delta)\cdots[x+\delta(m-1)] & m\geq 1\\ 1 & m=0\end{cases}
$$

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

$$\begin{aligned}
Y_i \sim \text{GEC}(y_i|\lambda_i, \sigma^2) &\equiv \text{P}(Y = y_i|\lambda_i, \sigma^2) \\
&= \text{GEC}\left(y_i|e^{X_i\beta}, \sigma^2\right) \\
&= \frac{1}{y_i!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(y_i, 1-\frac{1}{\sigma^2}\right)} \left[\sum_{j=0}^{y_i^{\max}} \frac{1}{j!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(j, 1-\frac{1}{\sigma^2}\right)}\right]^{-1},
\end{aligned}$$

where $y_i^{\max} = \infty$ for $\sigma^2 \geq 1$ and, letting $n_i = \lambda_i/(1-\sigma^2)$, $y_i^{\max} = [n_i + 1)$ for $0 < \sigma^2 < 1$, and where $[x)$ equals $x - 1$ for integer x and floor(x) for non-integer x. Note also that:

$$x^{(m,\delta)} = \begin{cases} \prod_{i=0}^{m-1}(x + \delta i) = x(x+\delta)(x+2\delta)\cdots[x+\delta(m-1)] & m \geq 1 \\ 1 & m = 0 \end{cases}$$

And with systematic component

A Generalized Event Count (GEC) Model

An event count model that allows under-, Poisson, and over-dispersion:

$$
\begin{aligned}
Y_i \sim \text{GEC}(y_i|\lambda_i, \sigma^2) &\equiv \text{P}(Y = y_i|\lambda_i, \sigma^2) \\
&= \text{GEC}\left(y_i|e^{X_i\beta}, \sigma^2\right) \\
&= \frac{1}{y_i!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(y_i, 1-\frac{1}{\sigma^2}\right)} \left[\sum_{j=0}^{y_i^{\max}} \frac{1}{j!}\left(\frac{e^{X_i\beta}}{\sigma^2}\right)^{\left(j, 1-\frac{1}{\sigma^2}\right)}\right]^{-1},
\end{aligned}
$$

where $y_i^{\max} = \infty$ for $\sigma^2 \geq 1$ and, letting $n_i = \lambda_i/(1-\sigma^2)$, $y_i^{\max} = [n_i + 1)$ for $0 < \sigma^2 < 1$, and where $[x)$ equals $x - 1$ for integer x and floor(x) for non-integer x. Note also that:

$$
x^{(m,\delta)} = \begin{cases} \prod_{i=0}^{m-1}(x + \delta i) = x(x+\delta)(x+2\delta)\cdots[x+\delta(m-1)] & m \geq 1 \\ 1 & m = 0 \end{cases}
$$

And with systematic component

$$
E(Y_i|X_i) \equiv \lambda_i = \exp(X_i\beta)
$$

GEC Interpretation:

GEC Interpretation:

1. Special cases of the GEC

GEC Interpretation:

1. Special cases of the GEC
   (a) *Negative Binomial*, $\sigma^2 > 1$, the over-dispersed case.

GEC Interpretation:

1. Special cases of the GEC
   (a) *Negative Binomial*, $\sigma^2 > 1$, the over-dispersed case.
   (b) *Poisson*, $\sigma^2 = 1$

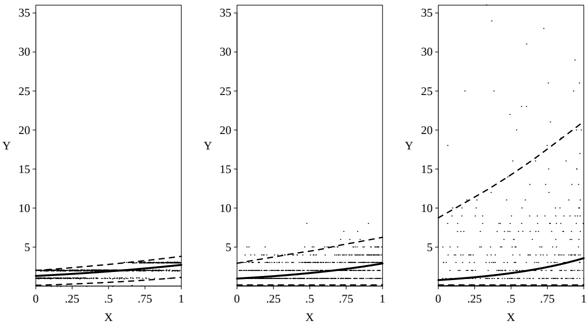
GEC Interpretation:

1. Special cases of the GEC
   (a) *Negative Binomial*, $\sigma^2 > 1$, the over-dispersed case.
   (b) *Poisson*, $\sigma^2 = 1$
   (c) *Continuous Parameter Binomial*, $0 < \sigma^2 < 1$, the underdispersed case. This special case itself reduces to an even more special case, the *Binomial*, when $\lambda_i/(1-\sigma^2)$ is an integer.

GEC Interpretation:

1. Special cases of the GEC
   (a) *Negative Binomial*, $\sigma^2 > 1$, the over-dispersed case.
   (b) *Poisson*, $\sigma^2 = 1$
   (c) *Continuous Parameter Binomial*, $0 < \sigma^2 < 1$, the underdispersed case. This special case itself reduces to an even more special case, the *Binomial*, when $\lambda_i/(1-\sigma^2)$ is an integer.
2. Can simulate in three parts, indexed by σ^2.

What happens with the extra parameter? GEC



(a) (b) (c)

$E(Y|X)$ and 95% CI.

King, Gary and Curtis S. Signorino. "The Generalization in the Generalized Event Count Model" in *Political Analysis*, 6 (1996): 225-252.

King, Gary. "Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator," *American Journal of Political Science*, 33, 3 (August, 1989): 762-784.

Duration Models and Censoring: The Exponential Model

King, Gary; James Alt; Nancy Burns; and Michael Laver. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies," American Journal of Political Science, Vol. 34, No. 3 (August, 1990): Pp. 846-871; Errata Vol. 34, No. 4 (November, 1990): P. 1168. (replication dataset: ICPSR s1115).

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

So the exponential duration model:

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

So the exponential duration model:

$$Y_i \sim \text{expon}(y_i|\lambda_i)$$

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

So the exponential duration model:

$$Y_i \sim \text{expon}(y_i|\lambda_i)$$
$$E(Y_i) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

So the exponential duration model:

$$Y_i \sim \text{expon}(y_i|\lambda_i)$$
$$E(Y_i) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

Log-likelihood:

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

So the exponential duration model:

$$Y_i \sim \text{expon}(y_i|\lambda_i)$$
$$E(Y_i) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

Log-likelihood:

$$\ln L(\beta|y) = \sum_{i=1}^{n} \{\ln \lambda_i - \lambda_i y_i\}$$

Duration Models and Censoring: The Exponential Model

A density with the same first principles as the Poisson, except that we observe Y, the duration between events. Then

$$Y_i \sim \text{expon}(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$$

So the exponential duration model:

$$Y_i \sim \text{expon}(y_i|\lambda_i)$$
$$E(Y_i) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

Log-likelihood:

$$\begin{aligned}\ln L(\beta|y) &= \sum_{i=1}^{n} \{\ln \lambda_i - \lambda_i y_i\} \\ &= \sum_{i=1}^{n} \left\{-X_i\beta - e^{-X_i\beta} y_i\right\}\end{aligned}$$

Censoring

Censoring

What do we do with observations that have not yet ended?

Censoring

What do we do with observations that have not yet ended?

1. For example

Censoring

What do we do with observations that have not yet ended?

1. For example
    (a) Parliamentary coalition duration; some still in office

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed
   (c) Duration in graduate school.

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed
   (c) Duration in graduate school. (What will we do with you?)

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed
   (c) Duration in graduate school. (What will we do with you?)
2. Could drop them. Selection bias.

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed
   (c) Duration in graduate school. (What will we do with you?)
2. Could drop them. Selection bias.
3. Could set duration to observed level: would underestimate in most cases

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed
   (c) Duration in graduate school. (What will we do with you?)
2. Could drop them. Selection bias.
3. Could set duration to observed level: would underestimate in most cases
4. Could guess. Even a good guess would produce standard errors that are too small.

Censoring

What do we do with observations that have not yet ended?

1. For example
   (a) Parliamentary coalition duration; some still in office
   (b) Duration of unemployment spells; some people still unemployed
   (c) Duration in graduate school. (What will we do with you?)
2. Could drop them. Selection bias.
3. Could set duration to observed level: would underestimate in most cases
4. Could guess. Even a good guess would produce standard errors that are too small.
5. We will model these observations explicitly as censored.

Censoring

$$Y_i^* \sim \text{expon}(y_i^*|\lambda_i)$$

Censoring

$$Y_i^* \sim \text{expon}(y_i^*|\lambda_i)$$

$$E(Y_i^*) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

Censoring

$$Y_i^* \sim \text{expon}(y_i^*|\lambda_i)$$
$$E(Y_i^*) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

An observation mechanism (with censoring at C):

Censoring

$$Y_i^* \sim \text{expon}(y_i^*|\lambda_i)$$
$$E(Y_i^*) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

An observation mechanism (with censoring at C):

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* < C \\ y_i^C & \text{if } y_i^* \geq C \end{cases}$$

Censoring

$$Y_i^* \sim \text{expon}(y_i^*|\lambda_i)$$

$$E(Y_i^*) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

An observation mechanism (with censoring at C):

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* < C \\ y_i^C & \text{if } y_i^* \geq C \end{cases}$$

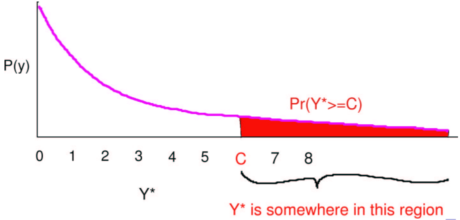
The likelihood function for censored observations. All we know is:

$$Y_i^* \sim \text{expon}(y_i^*|\lambda_i)$$
$$E(Y_i^*) \equiv \frac{1}{\lambda_i} = \frac{1}{e^{-x_i\beta}} = e^{x_i\beta}$$

An observation mechanism (with censoring at C):

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* < C \\ y_i^C & \text{if } y_i^* \geq C \end{cases}$$

The likelihood function for censored observations. All we know is:



Censoring

$$\Pr(Y_i = y_i^c) = \Pr(Y_i^* \geq y_i^c)$$

Censoring

$$\Pr(Y_i = y_i^c) = \Pr(Y_i^* \geq y_i^c)$$
$$= \int_{y_i^c}^{\infty} \text{expon}(y_i|\lambda_i)dy_i$$

Censoring

$$\begin{aligned}
\Pr(Y_i = y_i^c) &= \Pr(Y_i^* \geq y_i^c) \\
&= \int_{y_i^c}^{\infty} \text{expon}(y_i|\lambda_i) dy_i \\
&= \int_{y_i^c}^{\infty} \lambda_i e^{-\lambda_i y_i} dy_i
\end{aligned}$$

Censoring

$$\begin{aligned}
\Pr(Y_i = y_i^c) &= \Pr(Y_i^* \geq y_i^c) \\
&= \int_{y_i^c}^{\infty} \text{expon}(y_i|\lambda_i)dy_i \\
&= \int_{y_i^c}^{\infty} \lambda_i e^{-\lambda_i y_i} dy_i \\
&= e^{-\lambda_i y_i^c}
\end{aligned}$$

Censoring

$$\begin{aligned}
\Pr(Y_i = y_i^c) &= \Pr(Y_i^* \geq y_i^c) \\
&= \int_{y_i^c}^{\infty} \text{expon}(y_i|\lambda_i) dy_i \\
&= \int_{y_i^c}^{\infty} \lambda_i e^{-\lambda_i y_i} dy_i \\
&= e^{-\lambda_i y_i^c}
\end{aligned}$$

Thus, the full likelihood:

Censoring

$$
\begin{aligned}
\Pr(Y_i = y_i^c) &= \Pr(Y_i^* \geq y_i^c) \\
&= \int_{y_i^c}^{\infty} \text{expon}(y_i|\lambda_i) dy_i \\
&= \int_{y_i^c}^{\infty} \lambda_i e^{-\lambda_i y_i} dy_i \\
&= e^{-\lambda_i y_i^c}
\end{aligned}
$$

Thus, the full likelihood:

$$
L(\beta|y) = \left[\prod_{y_i^* < y_i^c} \text{expon}(y_i|\lambda_i)\right] \left[\prod_{y_i^* \geq y_i^c} \Pr(Y_i^* \geq y_i^c)\right]
$$