Lecture 22: Treatment Effects with Limited Dependent Variables

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

April 27, 2022

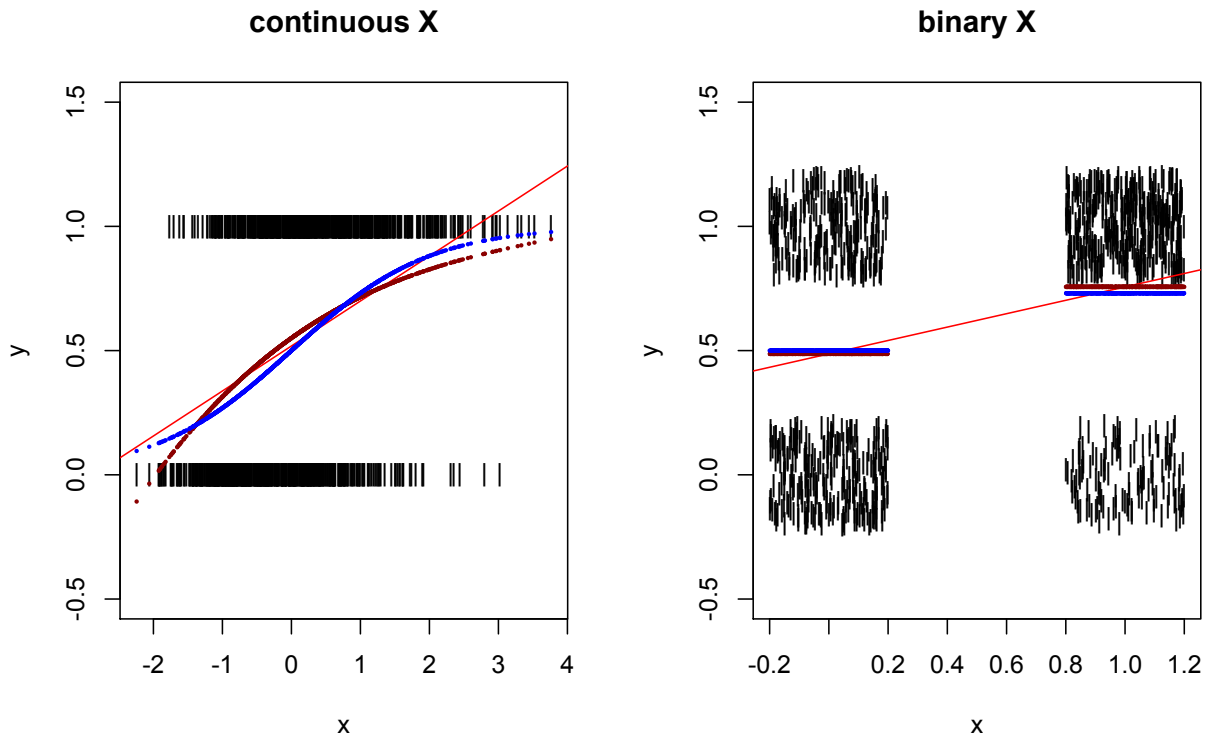
Motivation

- Limited dependent variables (LDVs): outcome variables with finite, truncated, or discrete support:
  - Binary.
  - Multichotomous (e.g., vote choice in 3-party race).
  - Censored or truncated (e.g., salaries, counts).

Motivation

- Limited dependent variables (LDVs): outcome variables with finite, truncated, or discrete support:
  - Binary.
  - Multichotomous (e.g., vote choice in 3-party race).
  - Censored or truncated (e.g., salaries, counts).
- Tradition is to use linear models (e.g. OLS) for continuous outcomes and "non-linear" models for LDVs:
  - Logits or probits for binary, multichotomous, or ordered.
  - Poisson or negative binomial for counts.
  - Cox or parametric duration models for censored durations.

Motivation



- Idea: "respect the support" of Y.
- Linear model can result in "non-sensical predictions" or failure to account for subtle functional form issues, like diminishing absolute effects (Long, 1997, p. 39)
- (Exception is dummy variable regression, which is inherently linear.)

Motivation

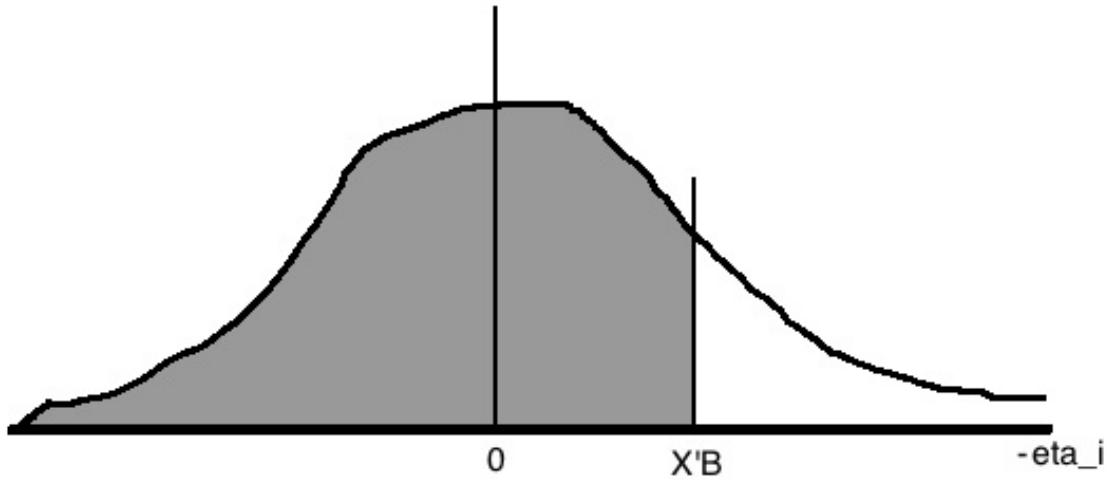
Goals of the lecture:

- Present basic ideas about how conventional LDV models work.
- Study their performance in estimating causal effects, using OLS as benchmark.
- Binary outcomes and logistic regression.
- Counts and Poisson regression (left-truncated at 0).
- Durations (left-truncated at 0 and often right-censored) and various non-parametric and semi-parametric approaches.

Binary Outcomes

- Put aside causal effect stuff for a moment and just think about an outcome, Y_i and regressors, X_i.
- Suppose $Y_i = 0, 1$.
- We would like a model for $\Pr[Y_i = 1|X_i]$ that is relatively easy to work with.
- Two ideas: "latent index" and "transformation to linearity."

Binary Outcomes



Latent index:

- Assume

$$Y_i^* = X_i'\beta + \eta_i \text{ where } Y_i = 1(X_i'\beta + \eta_i > 0)$$

- If CDF of $-\eta_i$ is given by $F(.)$, then $\Pr[Y_i = 1|X_i] = F(X_i'\beta)$.
- If $\eta_i \sim \mathrm{N}(0,1)$, then $F(X_i'\beta) = \Phi(X_i'\beta)$, "probit."
- If $\eta_i \sim \text{Logistic}(0, \frac{\pi^2}{3})$, "logit," $F(\mathbf{X}_i'\beta) = \Lambda(X_i'\beta) = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$.

Binary Outcomes

Transformation:

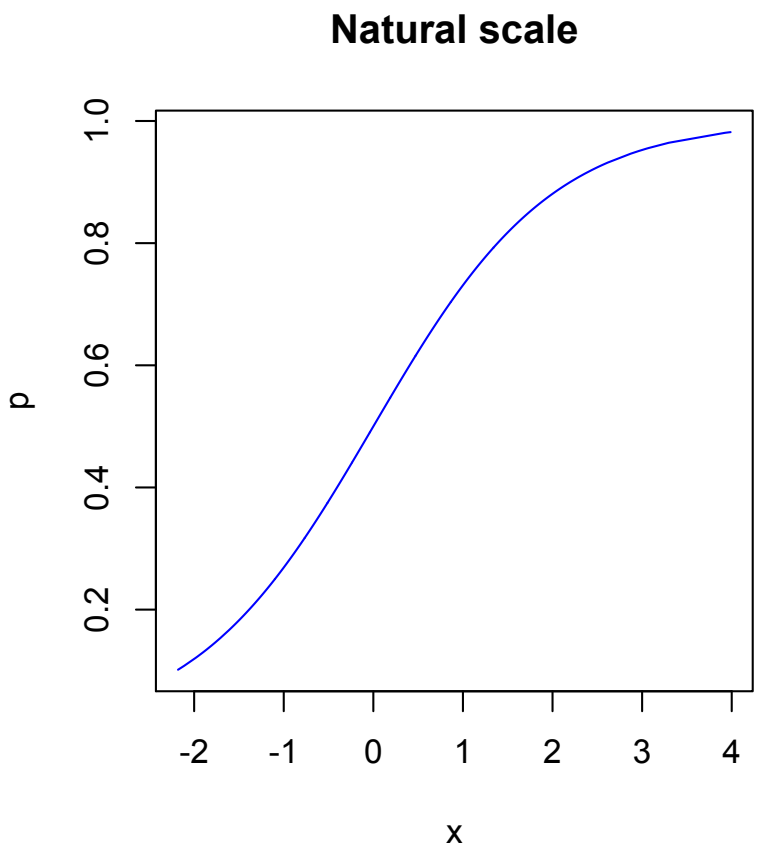
- Rescale $\mathrm{E}[Y_i|X_i] = \Pr[Y_i = 1|X_i]$ so that it can be expressed as linear: $g(\mathrm{E}[Y_i|X_i]) = X_i'\beta$.
- Yields "generalized linear models" (GLM): $g(.)$ is "link" function and $X_i'\beta$ "linear predictor."
- Logistic transformation is "canonical" link for binary outcomes (cf. Gill, 2000):

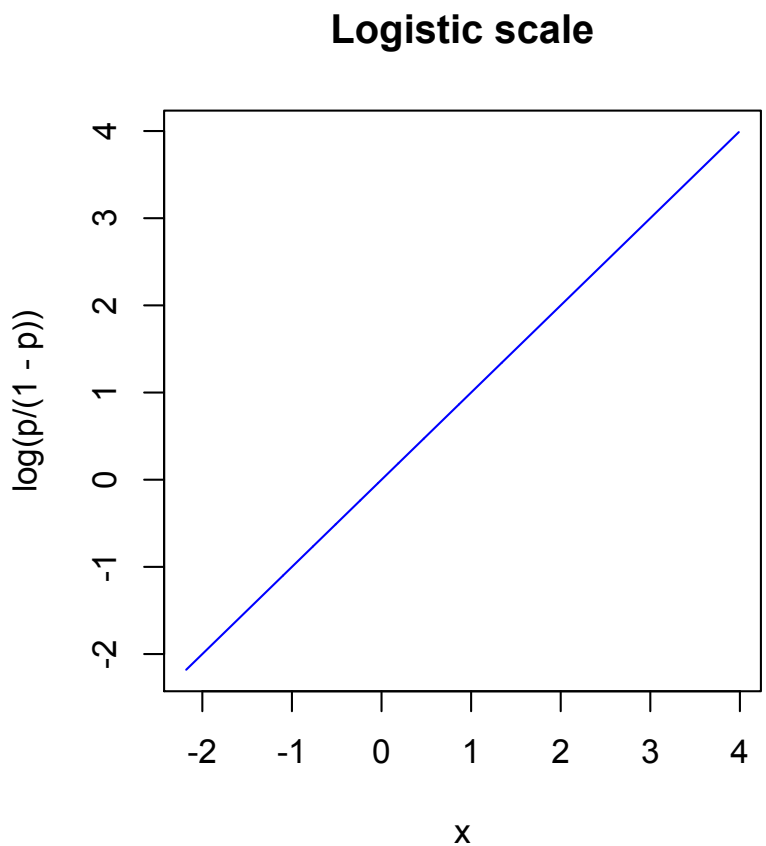
$$\log\left(\frac{\Pr[Y_i = 1|X_i]}{1 - \Pr[Y_i = 1|X_i]}\right) = X_i\beta$$

$$\Rightarrow \Pr[Y_i = 1|X_i] = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)} = \Lambda(X_i'\beta)$$

- Modeling "log-odds" rather than probabilities.

Binary Outcomes





Binary Outcomes

Estimation:

- Suppose data iid with $\Pr[Y_i = 1|X_i] = \Lambda(X_i'\beta)$.
- "Likelihood" defined by joint density of observed outcomes conditional on data and β:

$$\begin{aligned}\mathscr{L}(\beta|Y,\mathbf{X}) &\equiv p[Y_1,...,Y_N|X_1,...,X_n;\beta] \\ &= \prod_{i:Y_i=1} \Pr[Y_i = 1|X_i] \prod_{i:Y_i=0} (1 - \Pr[Y_i = 1|X_i]) \\ &= \prod_{i=1}^{N} F(X_i'\beta)^{Y_i}[1 - F(X_i'\beta)]^{(1-Y_i)}\end{aligned}$$

- "Maximum likelihood estimation" searches over β's to maximize $\mathscr{L}(\beta|Y,\mathbf{X})$.
- We can stabilize the computation by using $\log[\mathscr{L}(\beta|Y,\mathbf{X})]$.

Binary Outcomes

MLE properties:

- If distributional assumptions correct, MLE is fully efficient.
- Coefficient estimator is asymptotically normal.
- Under iid, standard errors are easy to compute from the second derivative matrix of the log-likelihood (cf., DeGroot and Schervish, 2002, pp. 435-444, "Fisher information").
- If iid is violated (e.g., clustering or heteroskedasticity), "robust" standard errors are still meaningful, although we lose efficiency and may be "biased" relative to the "true" parametric form.
- (NB: OLS *is* MLE solution for many linear models, including with homoskedastic normal errors.)
- Under misspecification, MLE minimizes Kullback-Leibler divergence between true conditional distribution and assumed distribution (White, 1982).

Binary Outcomes

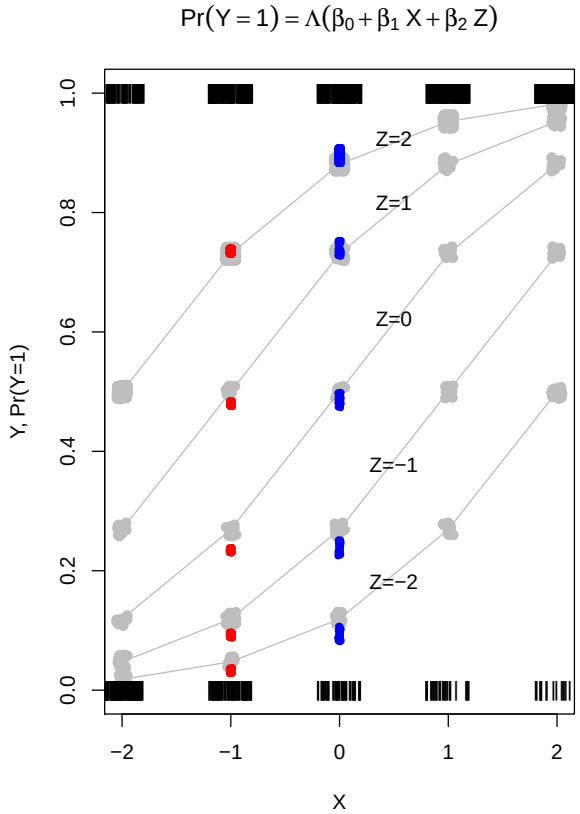
How do we interpret logistic regression estimates?

- When we fit a logistic regression, we get coefficients on the "log-odds" scale:

$$\log\left(\frac{\Pr[Y_i = 1|X_i]}{1 - \Pr[Y_i = 1|X_i]}\right) = X_i\beta.$$

  ("probits" scale for a probit regression).
- β's are only interpretable in signs ($+/-$) and significance (***).
- Convert back to probability scale to have substantive meaning.

Binary Outcomes



(Gray is true $\Pr[Y=1|X,Z]$, black is realized data, and red and blue are predictions from the logit fit.)

Derivative method:

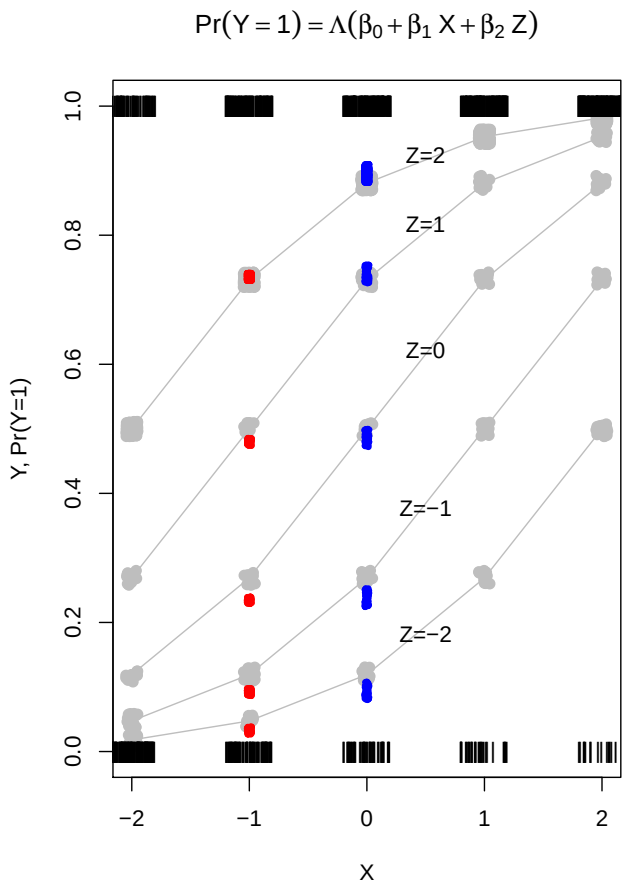
$$\frac{\partial \Pr[Y_i=1|X_i]}{\partial X_{ki}} = \frac{\partial}{\partial X_{ki}}\left[\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right] = \beta_k\left[\frac{\exp(X_i'\beta)}{[1+\exp(X_i'\beta)]^2}\right]$$

- "Avg. fitted partial derivative" wrt X_{ki}:

$$\bar{\Delta}_k = \beta_k \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\exp(X_i'\beta)}{[1+\exp(X_i'\beta)]^2}\right]$$

- On picture: $\approx$ avg. of derivatives at each gray dot.
- Stata's `margins, dydx(...)` and R's `margins` compute $\bar{\Delta}_k$.

Binary Outcomes



(Gray is true $\Pr[Y = 1|X, Z]$, black is realized data, and red and blue are predictions from the logit fit.)

Predicted probabilities method:

- Fix $X_{ik} = a$ for all i, compute predicted probabilities, leaving all other X_{i-k} at observed values. Store as $\hat{p}_a = p_1(a), ..., p_N(a)$.
- Same for $X_{ik} = b$ to get $\hat{p}_b = p_1(b), ..., p_N(b)$.
- Sample average effect of $X_i = b \rightarrow X_i = a$:

$$\bar{\Delta}_D = \frac{1}{N} \sum_{i=1}^{N} [p_i(a) - p_i(b)]$$

- On picture: $\approx$ avg. of changes from -1 to 0.
- Analogous to ATE (Hanmer & Kalkan, 2013).

Extensions

Extensions: "multinomial logit" model for multichotomous outcomes:

$$\Pr[Y_i = m|X_i] = \frac{\exp(X_i'\beta_m)}{\sum_{j=1}^{J}\exp(X_i'\beta_j)} \text{ with } \beta_1 = 0,$$

in which case

$$\Pr[Y_i = 1|X_i] = \frac{1}{1+\sum_{j=2}^{J}\exp(X_i'\beta_j)}$$

and

$$\Pr[Y_i = m|X_i, m > 1] = \frac{\exp(X_i'\beta_m)}{1+\sum_{j=2}^{J}\exp(X_i'\beta_j)},$$

with $m = 1$ the "base" category. Parametric identification requires that estimation be done relative to the base category, necessitating $\beta_1 = 0$.

Extensions

- Multinomial logit is rather demanding of the data.
- Consistent under so-called IIA assumption.
- Also "ordered logit" model for ordinal outcomes, "nested logit" for sequenced outcomes. Similar models for probit.
- OLS alternative: fit a regression on the long data,

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iJ} \end{pmatrix} = \begin{pmatrix} X_i\beta_1 + \varepsilon_{i1} \\ \vdots \\ X_i\beta_J + \varepsilon_{iJ} \end{pmatrix}$$

  where Y_{ij} is a dummy variable for whether i realizes outcome j.
- Use interactions with *outcome* dummies to estimate different slopes for each outcome.
- Cluster by i (or a higher level of aggregation), to account for $\sum_{j=1}^{J} Y_{ij} = 1$ (outcomes correlated by construction over i).

Treatment Effects with Binary Outcomes

Comparing MLE logistic to OLS for estimating treatment effects:

- Covariate, X_i, distributed as,

$$X_i = W_i + \alpha W_i^2 - \beta,$$

  where $W_i \sim \mathrm{N}(0,1)$ and α controls skew and β shift.
- Treatment, D_i, distributed as, $D_i \sim \text{Bernoulli}(\Lambda(X_i))$.
- Potential outcomes,

$$Y_{0i} \sim \text{Bernoulli}(p_0) \text{ and } Y_{1i} \sim \text{Bernoulli}(\Lambda(\gamma_0 + \gamma_1 X_i)).$$

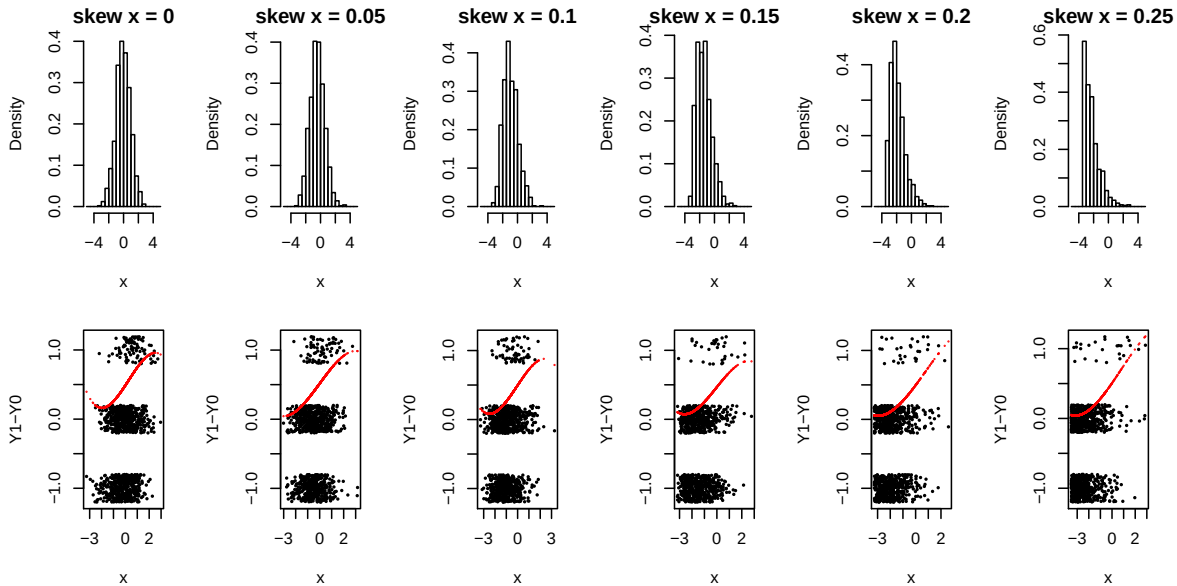
- Thus the true model *is* a logistic model. Very favorable to logit.
- We observe $Y_i = DY_1 + (1-D)Y_0$.
- Estimand: $\rho = \mathrm{E}[Y_1 - Y_0]$.

Treatment Effects with Binary Outcomes

Estimators that we will use (note: $\Lambda(.) = \text{logit}^{-1}(.)$):

Label	Specification	Fitting method	Estimator for ρ
`b.ols.d`	$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$	OLS	$\hat{\beta}_1$
`b.log.d`	$\Pr[Y_i = 1] = \text{logit}^{-1}(\beta_0 + \beta_1 D_i)$	MLE	$\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1) - \text{logit}^{-1}(\hat{\beta}_0)$
`b.ols.dx`	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i$	OLS	$\hat{\beta}_1$
`b.log.dx`	$\Pr[Y_i = 1] = \text{logit}^{-1}(\beta_0 + \beta_1 D_i + \beta_2 X_i)$	MLE	$\text{mean}[\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_1 X_i) - \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_i)]$
`b.ols.int`	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 (X_i - \bar{X}) + \beta_3 D_i (X_i - \bar{X}) + \varepsilon_i$	OLS	$\hat{\beta}_1$
`b.log.int`	$\Pr[Y_i = 1] = \text{logit}^{-1}(\beta_0 + \beta_1 D_i + \beta_2 X_i + \hat{\beta}_3 D_i X_i)$	MLE	$\text{mean}[\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_1 X_i + \hat{\beta}_3 D_i X_i) - \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_3 D_i X_i)]$

Treatment Effects with Binary Outcomes



(Top row shows histograms of X over the skew parameter. The bottom row plots unit-level treatment effects over X from a representative dataset for each of the simulation experiments. The red line overlaid on the bottom plots shows the expected value of D over values of X.)

- $\text{Cor}(X_i, D_i) > 0$ and $\text{Cor}(X_i, Y_i) > 0$. So, excluding X_i yields positive omitted variable bias. (Red line is mean of D_i over X_i.)
- $\text{Var}[D_i]$ is higher near $X_i = 0$ and effect of treatment is smaller near $X_i = 0$. So, aggregation bias that is negative in models including X_i but no interactions.

Treatment Effects with Binary Outcomes

Sim. #	1	2	3	4	5	6
skew	0.00	0.05	0.10	0.15	0.20	0.25
shift	0.00	0.50	1.00	1.50	2.00	2.50
baseline	0.50	0.50	0.50	0.50	0.50	0.50
b.true	-0.28	-0.34	-0.39	-0.42	-0.44	-0.46
b.ols.d	-0.22	-0.27	-0.32	-0.35	-0.37	-0.38
b.log.d	-0.22	-0.27	-0.32	-0.35	-0.37	-0.38
b.ols.dx	-0.29	-0.33	-0.37	-0.39	-0.41	-0.42
b.log.dx	-0.29	-0.33	-0.36	-0.38	-0.39	-0.40
b.ols.int	-0.29	-0.35	-0.40	-0.44	-0.47	-0.49
b.log.int	-0.28	-0.34	-0.39	-0.42	-0.44	-0.46
sd.baseline	0.02	0.02	0.02	0.02	0.02	0.02
sd.true	0.02	0.02	0.02	0.02	0.02	0.02
sd.ols.d	0.03	0.03	0.03	0.03	0.03	0.03
sd.log.d	0.03	0.03	0.03	0.03	0.03	0.03
sd.ols.dx	0.03	0.03	0.03	0.03	0.03	0.03
sd.log.dx	0.03	0.03	0.03	0.03	0.03	0.03
sd.ols.int	0.03	0.03	0.03	0.03	0.03	0.03
sd.log.int	0.03	0.03	0.03	0.03	0.02	0.02

Table: *Baseline rate of 0.50.*

Treatment Effects with Binary Outcomes

Sim. #	1	2	3	4	5	6
skew	0.00	0.05	0.10	0.15	0.20	0.25
shift	0.00	0.50	1.00	1.50	2.00	2.50
baseline	0.15	0.15	0.15	0.15	0.15	0.15
b.true	0.07	0.01	-0.03	-0.07	-0.09	-0.11
b.ols.d	0.13	0.08	0.04	0.00	-0.01	-0.03
b.log.d	0.13	0.08	0.04	0.00	-0.01	-0.03
b.ols.dx	0.06	0.02	-0.02	-0.04	-0.05	-0.07
b.log.dx	0.06	0.01	-0.02	-0.04	-0.05	-0.06
b.ols.int	0.06	-0.00	-0.05	-0.09	-0.12	-0.14
b.log.int	0.07	0.01	-0.03	-0.07	-0.09	-0.11
sd.baseline	0.01	0.01	0.01	0.01	0.01	0.01
sd.true	0.02	0.02	0.01	0.01	0.01	0.01
sd.ols.d	0.03	0.02	0.03	0.03	0.03	0.03
sd.log.d	0.03	0.02	0.03	0.03	0.03	0.03
sd.ols.dx	0.03	0.02	0.02	0.03	0.03	0.03
sd.log.dx	0.03	0.02	0.02	0.02	0.02	0.02
sd.ols.int	0.03	0.02	0.02	0.02	0.03	0.02
sd.log.int	0.02	0.02	0.02	0.02	0.02	0.02

Table: *Baseline rate of 0.15.*

Treatment Effects with Binary Outcomes

- Some differences between OLS and logit, but not huge.
- No differences in terms of efficiency.
- What matters in a first-order sense is the specification for Y in terms of D and X.
- The choice between OLS vs. logit is a second-order concern after this.
- If all regressors are dummy variable, the two models are equivalent.

Treatment Effects with Binary Outcomes and Fixed Effects

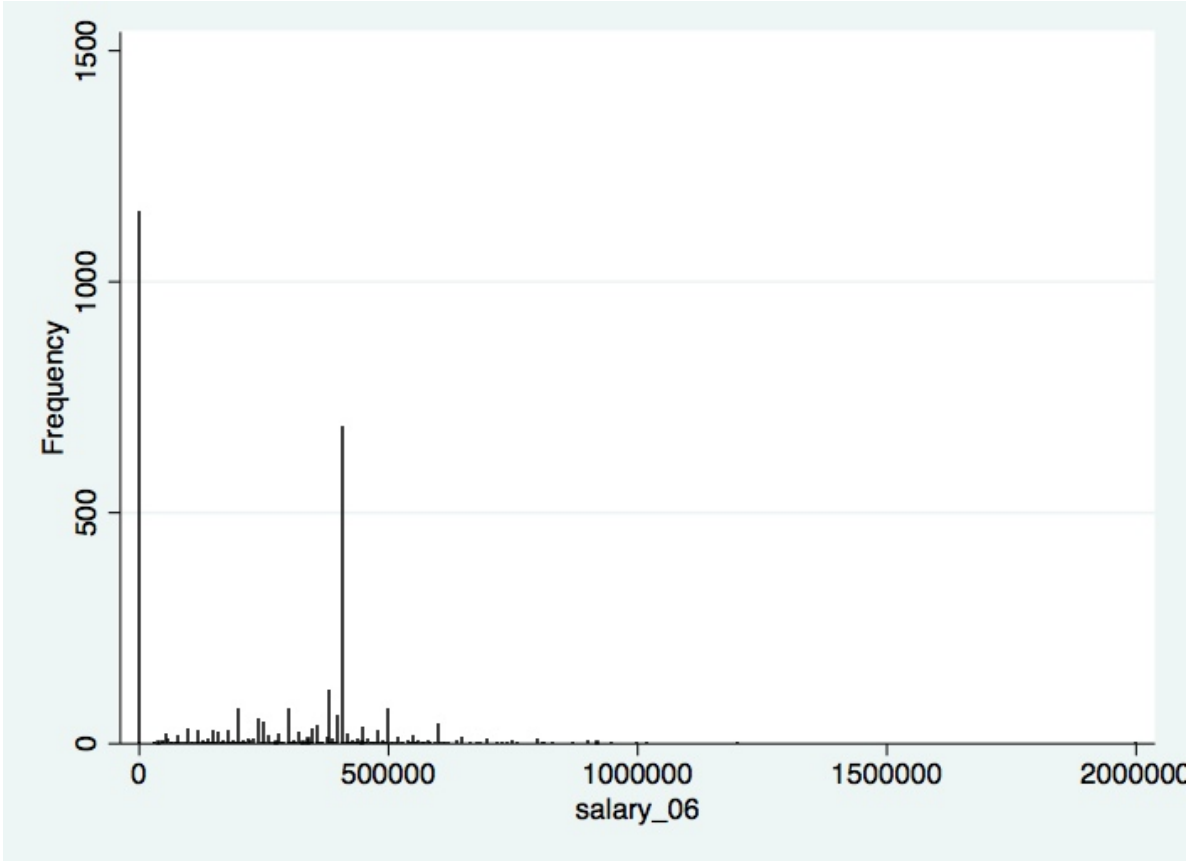
Fixed effects can create problems for MLE estimation, as we can see in the binary outcomes case:

- Suppose $t = 1, 2$.
- Consider the model,$\Pr[Y_{it} = 1 | X_{it}, \alpha_i] = \Lambda(\alpha_i + \rho D_{it})$. with $D_{it} = 0, 1$ and we use i-specific dummies to estimate the α_i's.
- For cases with $(Y_{i1}, Y_{i2}) = (0, 0)$ or $(Y_{i1}, Y_{i2}) = (1, 1)$, we have a problem, as the ML estimates for α_i in those situations are $-\infty$ and ∞, respectively. These problems propagate to estimates of ρ.
- OLS does not suffer these particular pathologies.

Treatment Effects with Binary Outcomes and Fixed Effects

- An MLE approach to this problem is "conditional logit" (Chamberlain estimator)
- Restricts estimation to units for which $(Y_{i1} \neq Y_{i2})$.
  - Problem: this changes the estimand: no longer estimating ρ for the population, but rather for this subgroup.
  - Will tend to produce estimates larger in absolute value than what would be the population coefficient.
  - Interpretation is problematic when conditioning on the outcome (rather than on covariates or treatment combinations).
- Other solutions include "regularized" estimators (e.g., Firth's correction, various multilevel models). Performance of these methods under misspecification is unclear (meaning more research needed).

Count/positive data



(Attanasio et al., 2011)

Count/positive data: Poisson regression

- Suppose Y_i is a count-type outcome such that $Y_i \in \mathbb{Z}^+$.
- $Y_i = X_i'\gamma + \varepsilon_i$ may fail to predict values within support.

Count/positive data: Poisson regression

- Suppose Y_i is a count-type outcome such that $Y_i \in \mathbb{Z}^+$.
- $Y_i = X_i'\gamma + \varepsilon_i$ may fail to predict values within support.
- GLM motivation: "linearizing" transformation,

$$\ln(\mathrm{E}[Y_i|X_i]) = X_i'\beta \Rightarrow \mathrm{E}[Y_i|X_i] = \exp(X_i'\beta).$$

- Ensures positive expected values based on linear predictor.

Count/positive data: Poisson regression

- Suppose Y_i is a count-type outcome such that $Y_i \in \mathbb{Z}^+$.
- $Y_i = X_i'\gamma + \varepsilon_i$ may fail to predict values within support.
- GLM motivation: "linearizing" transformation,

$$\ln(\mathrm{E}[Y_i|X_i]) = X_i'\beta \Rightarrow \mathrm{E}[Y_i|X_i] = \exp(X_i'\beta).$$

- Ensures positive expected values based on linear predictor.
- Under "working" assumption $\mathrm{Var}[Y_i|X_i] = \mathrm{E}[Y_i|X_i] = \exp(X_i'\beta)$, by GLM theory, MMSE estimate for β is,

$$\hat{\beta} \text{ s.t. } \sum_{i=1}^{N} [Y_i - \exp(X_i'\hat{\beta})]X_i = 0.$$

- Like OLS, but no closed form solution. Need to fit numerically.
- This is the Poisson regression model.

Count/positive data: Poisson regression

- Stochastic processes motivation: “rare events process” (cf. Cameron & Trivedi 1998, 1.1).

Count/positive data: Poisson regression

- Stochastic processes motivation: "rare events process" (cf. Cameron & Trivedi 1998, 1.1).
- M_i units performing Bernoulli trials each w/ π_i prob. of success.
- With (M_i, π_i) fixed, expected number of successes is $\lambda_i = M_i \pi_i$ and distribution of sum of successes, Y_i, is,

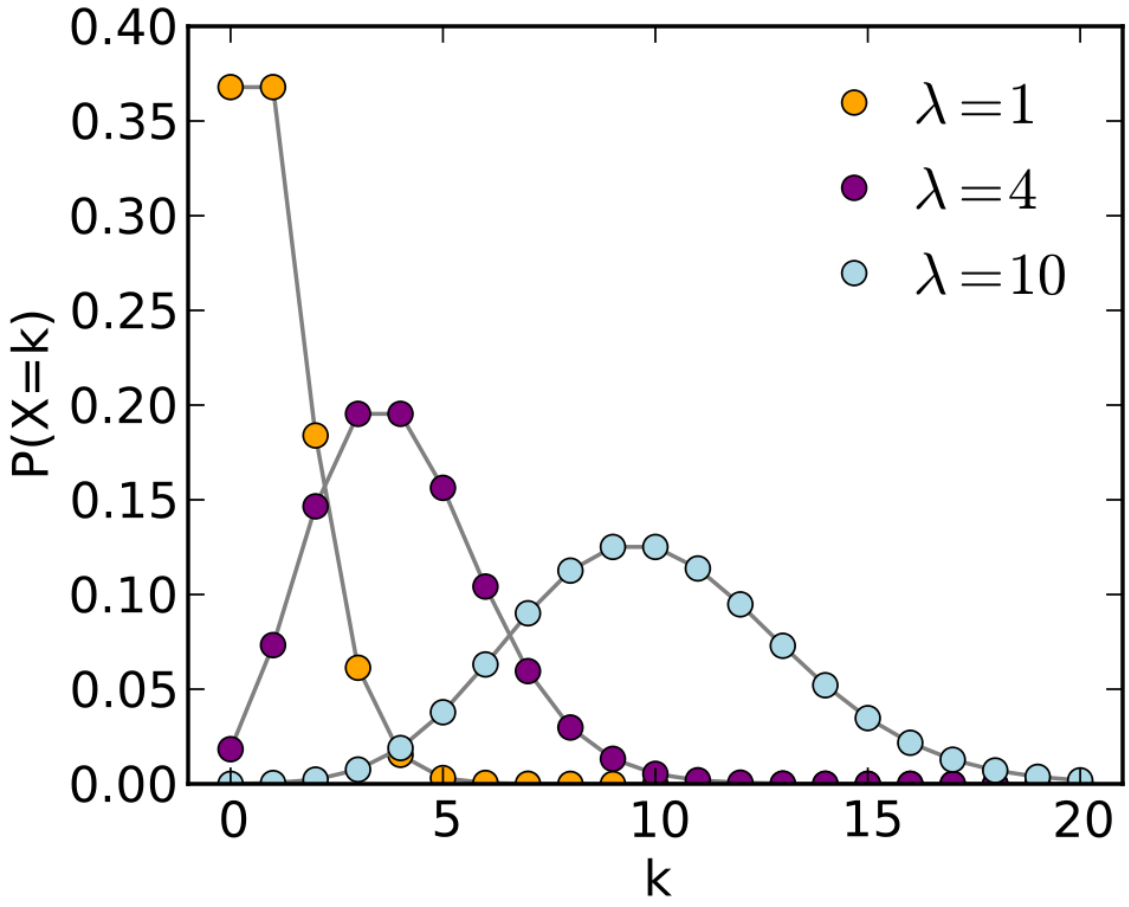
$$\Pr[Y_i = y] = \binom{M}{y} \pi_i^y (1 - \pi_i)^y$$

- Suppose for unit increase in M_i, reduce π_i by $M_i/(M_i + 1)$.
- Then,

$$\lim_{M_i \to \infty} \Pr[Y_i = y] = \frac{\lambda_i^y \exp(-\lambda_i)}{y!},$$

the Poisson distribution.

Count/positive data: Poisson regression



(Wikimedia Commons)

Count/positive data: Poisson regression

- Given

$$\Pr[Y_i = y] = \frac{\lambda_i^y \exp(-\lambda_i)}{y!},$$

  parameterize this with, $\lambda_i = \exp(X_i'\beta)$.
- Then $\mathrm{E}[Y_i|X_i] = \mathrm{Var}[Y_i|X_i] = \exp(X_i'\beta)$.

Count/positive data: Poisson regression

- Given

$$\Pr[Y_i = y] = \frac{\lambda_i^y \exp(-\lambda_i)}{y!},$$

  parameterize this with, $\lambda_i = \exp(X_i'\beta)$.
- Then $\mathrm{E}[Y_i|X_i] = \mathrm{Var}[Y_i|X_i] = \exp(X_i'\beta)$.
- With iid data, this yields the likelihood,

$$\mathscr{L}(\beta|Y,\mathbf{X}) = \prod_{i=1}^{N} \frac{[\exp(X_i'\beta)]^{Y_i} \exp[\exp(X_i'\beta)]}{Y_i!}$$

- Taking log and then maximizing yields MLE for β:

$$\hat{\beta} \text{ s.t. } \sum_{i=1}^{N} [Y_i - \exp(X_i'\hat{\beta})]X_i = 0.$$

  Cool!

Count/positive data: Remarks on Poisson regression

- The assumption $\mathrm{E}[Y_i|X_i] = \mathrm{Var}[Y_i|X_i]$ can be relaxed with few consequences. Solution for $\hat{\beta}$ is consistent for arbitrary assumptions on the variance.
- Run Poisson regression as usual, but just use robust standard errors (cf. Cameron & Trivedi 1998, 3.2).
- Violations of iid can also be handled using cluster-robust.

Count/positive data: Remarks on Poisson regression

- Current convention: use "negative binomial" and other "over dispersed" count models when presumed $\mathrm{E}[Y_i|X_i] < \mathrm{Var}[Y_i|X_i]$.
- Typically unnecessary and more sensitive to misspecification.

Count/positive data: Remarks on Poisson regression

- Current convention: use "negative binomial" and other "over dispersed" count models when presumed $\mathrm{E}[Y_i|X_i] < \mathrm{Var}[Y_i|X_i]$.
- Typically unnecessary and more sensitive to misspecification.
- Poisson handles FE without problems. So-called "conditional (FE) Poisson" is equivalent to dummy variable FE Poisson.
- Implementation in Stata (`poisson`, `xtpoisson`) and R (`glm()` with `family="poisson"`).

Count/positive data: Interpreting Poisson regression

- Derivative method:

$$\Delta_{\partial ki} = \frac{\partial}{\partial X_{ik}} \mathrm{E}\left[Y_i | X_i\right] = \exp(X_i' \hat{\beta}) \hat{\beta}_k,$$

  which is suggestive of Poisson as "multiplicative effects" model.

- Then, the sample average partial effect of X_{ik} is,

$$\bar{\Delta}_{\partial k} = \hat{\beta}_k \frac{1}{N} \sum_{i=1}^{N} \exp(X_i' \hat{\beta})$$

Count/positive data: Interpreting Poisson regression

- Derivative method:

$$\Delta_{\partial ki} = \frac{\partial}{\partial X_{ik}} \mathrm{E}\left[Y_i | X_i\right] = \exp(X_i'\hat{\beta})\hat{\beta}_k,$$

  which is suggestive of Poisson as "multiplicative effects" model.

- Then, the sample average partial effect of X_{ik} is,

$$\bar{\Delta}_{\partial k} = \hat{\beta}_k \frac{1}{N} \sum_{i=1}^{N} \exp(X_i'\hat{\beta})$$

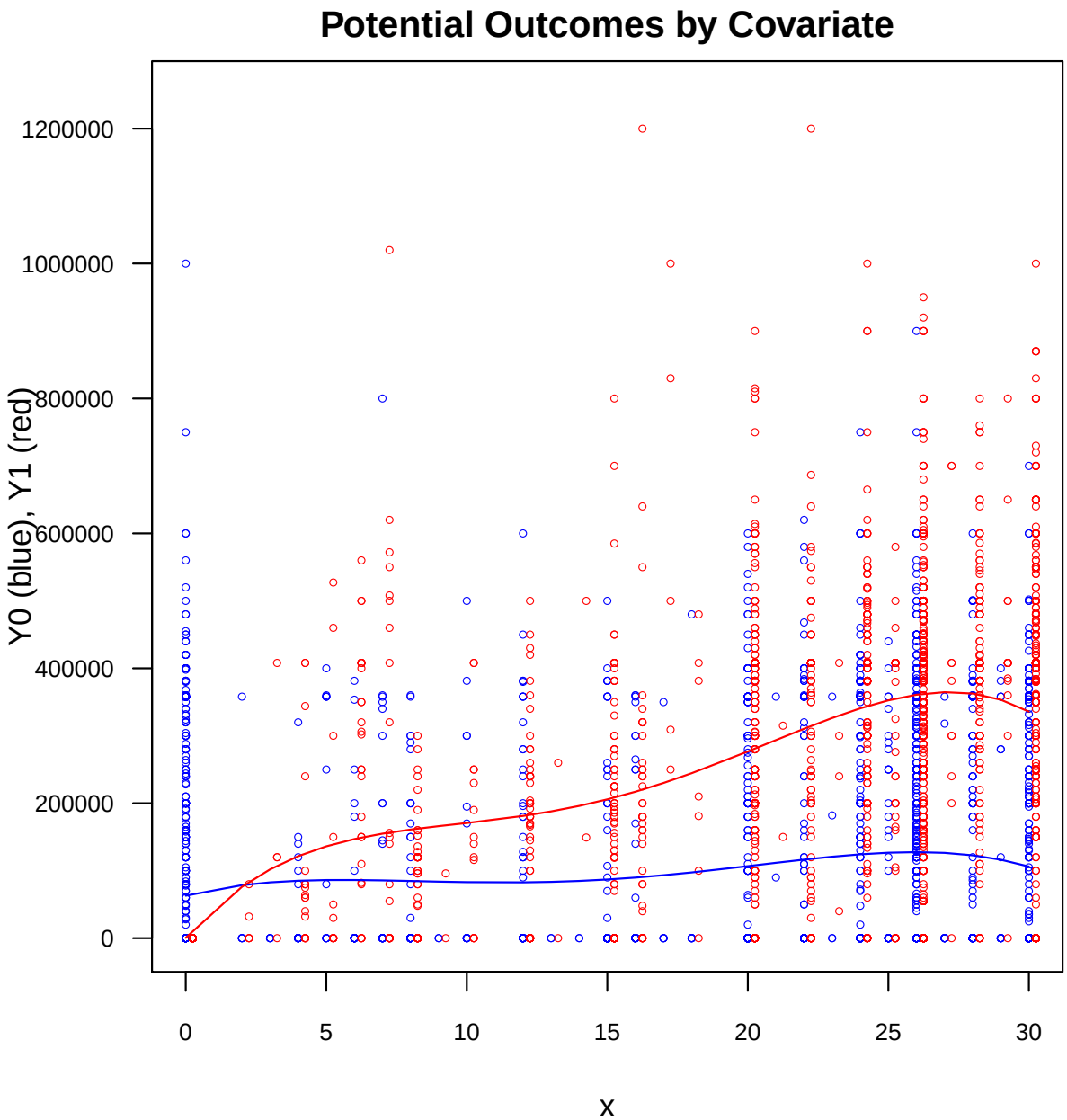
- Differences in predicted responses method: difference in predicted outcome values, like we saw for logit.
- E.g., for binary treatment D_i, $\widehat{ATE}$ given by,

$$\bar{\Delta}_D = \frac{1}{N} \sum_{i=1}^{N} \widehat{Y}_i(D_i = 1, X_{i,-D}) - \widehat{Y}_i(D_i = 0, X_{i,-D}).$$

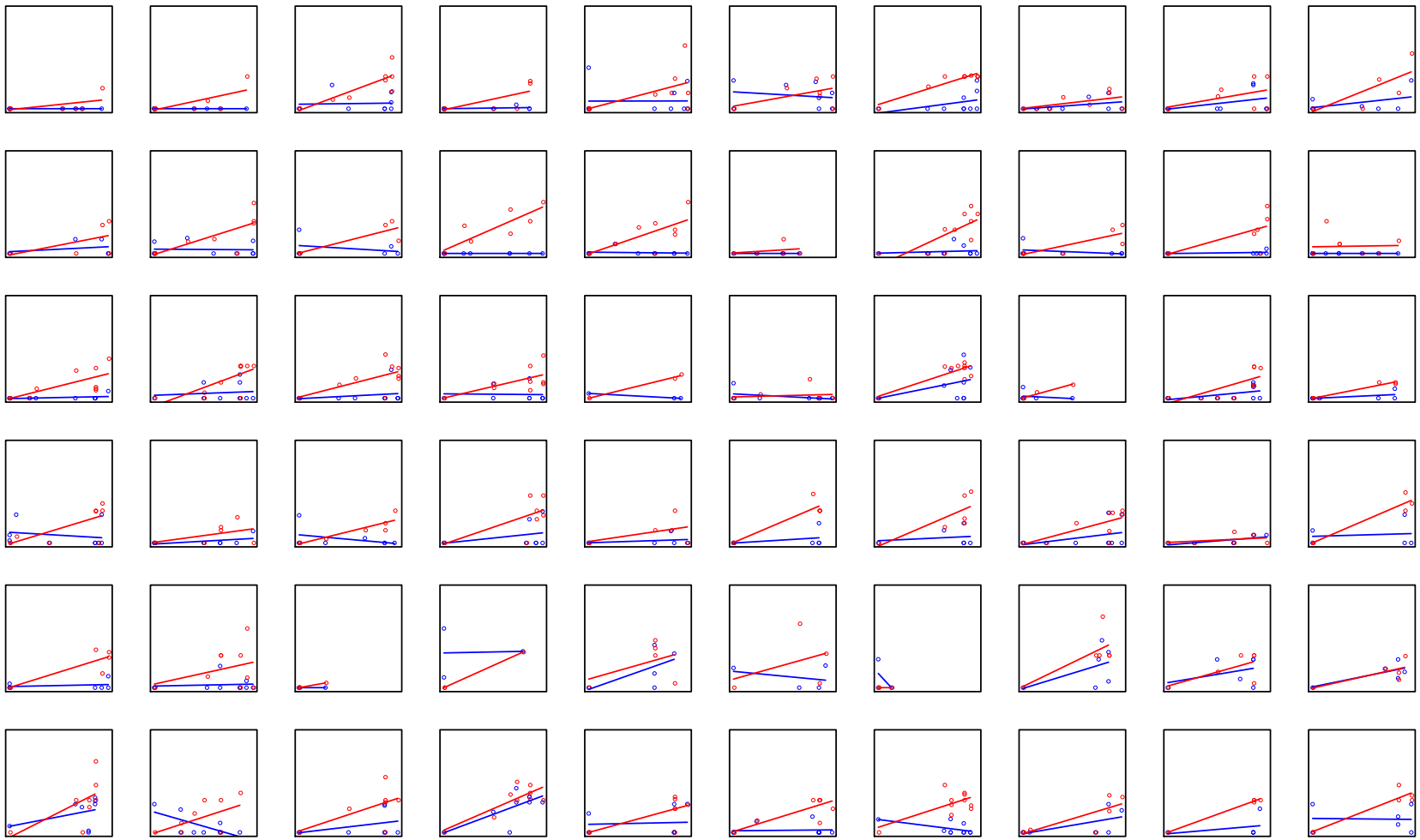
Count/positive data: Simulation

- Use data from Attanasio et al (2011) to construct simulation based on naturalistic data ($N = 3,237$).
- Simulate (Y_{0i}, Y_{1i})'s based on 2004 and 2006 salaries.
- Use avg. days worked per month as a covariate, X_i.
- Fixed effects for 441 strata defined by city and program type.
- Simulate two cases for a binary treatment variable, D_i:
  1. Random treatment (simple random assignment of D_i).
  2. Endogenous treatment, where stratum-specific Y_0 means are correlated with $\mathrm{E}[D_i]$, so FE needed.
- In all cases, $\mathrm{E}[Y_{1i} - Y_{0i}] = \mathrm{COP}135K/\text{month}$ ($\approx \$75/\text{month}$).

Count/positive data: Simulation



Count/positive data: Simulation

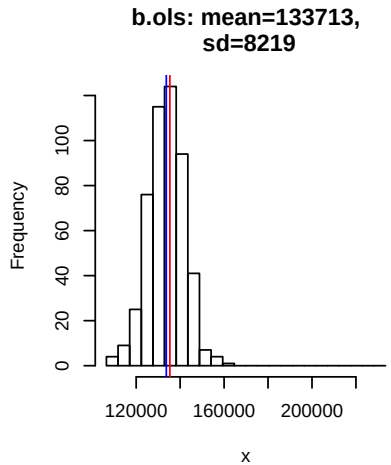


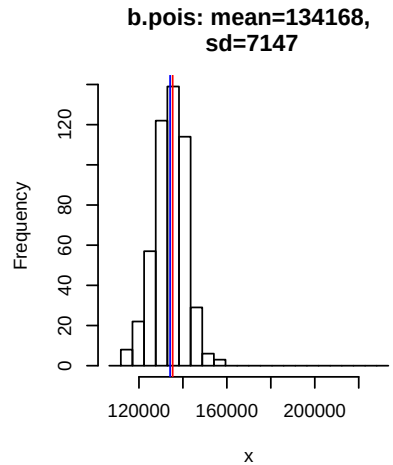
(60 out of 441 stratum-specific relationships.)

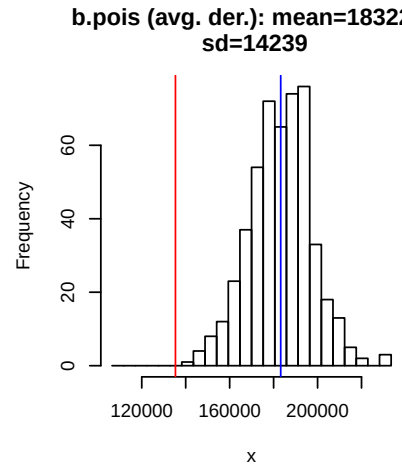
Count/positive data: Simulation

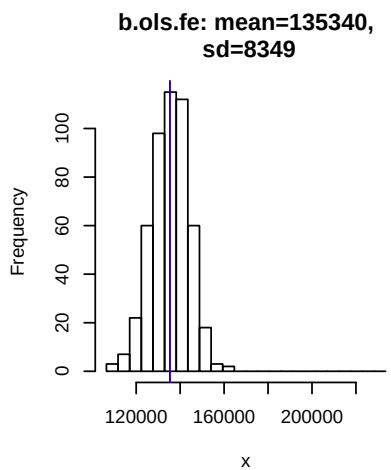
Estimator	Specification	Estimator
OLS linear	$Y_{is} = \beta_0 + \beta_1 D_{is} + \beta_2 X_{is} + \beta_3 X_{is}^2 + \varepsilon_{is}$	$\hat{\beta}_1$
OLS linear FE	$Y_i = \alpha_s + \beta_0 + \beta_1 D_{is} + \beta_2 X_{is} + \beta_3 X_{is}^2 + \varepsilon_{is}$	$\hat{\beta}_1$
MLE Poisson pred. prob.	$\mathrm{E}[Y_i] = \exp(\beta_0 + \beta_1 D_{is} + \beta_2 X_{is} + \beta_3 X_{is}^2)$	$\bar{\Delta}_D$
MLE Poisson avg. der.	"	$\bar{\Delta}_{\partial D}$
MLE Poisson FE pred. prob.	$\mathrm{E}[Y_i] = \exp(\alpha_s + \beta_0 + \beta_1 D_{is} + \beta_2 X_{is} + \beta_3 X_{is}^2)$	$\bar{\Delta}_D$
MLE Poisson FE avg. der.	"	$\bar{\Delta}_{\partial D}$

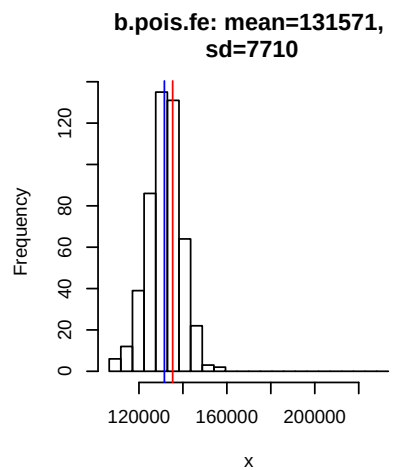
Count/positive data: Simulation result, random treatment

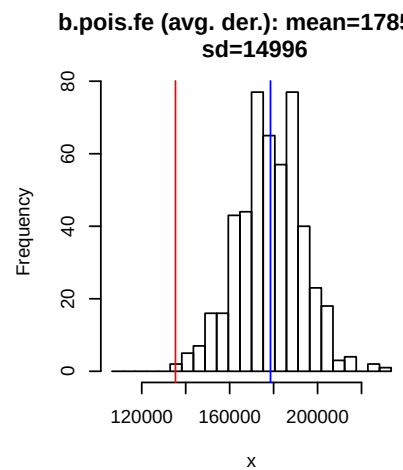




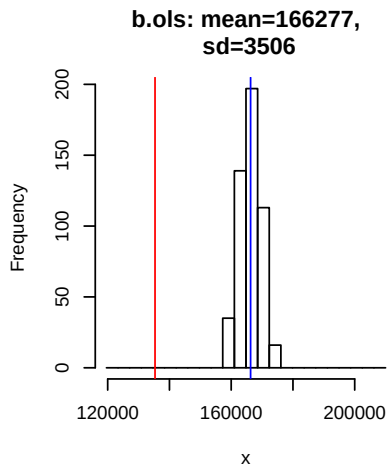


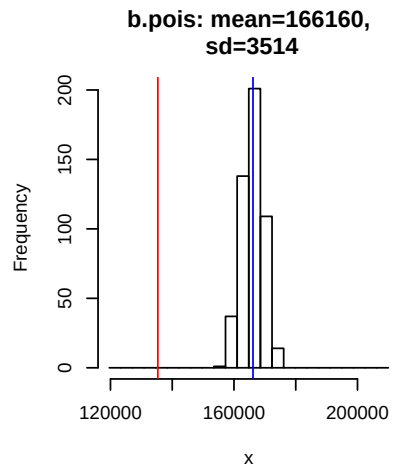


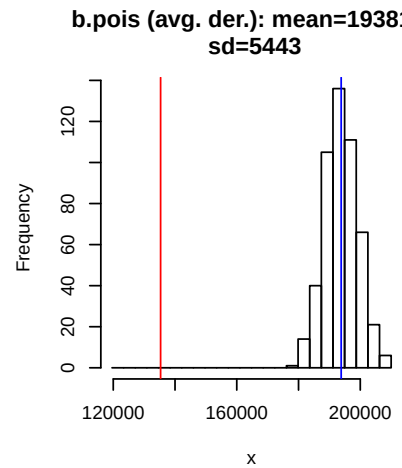


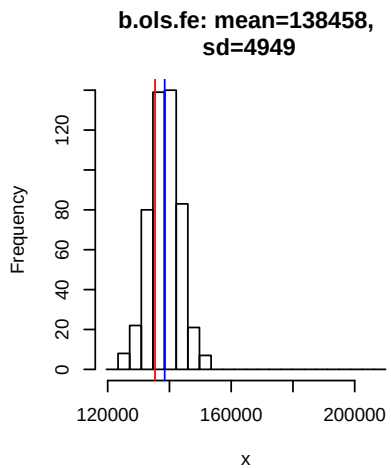


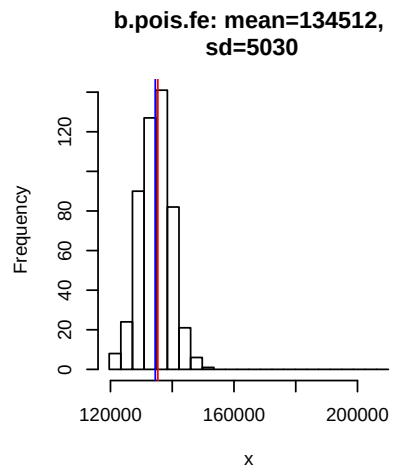
Count/positive data: Simulation result, endogenous treatment

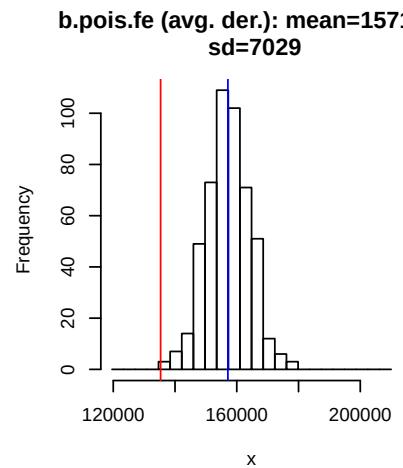








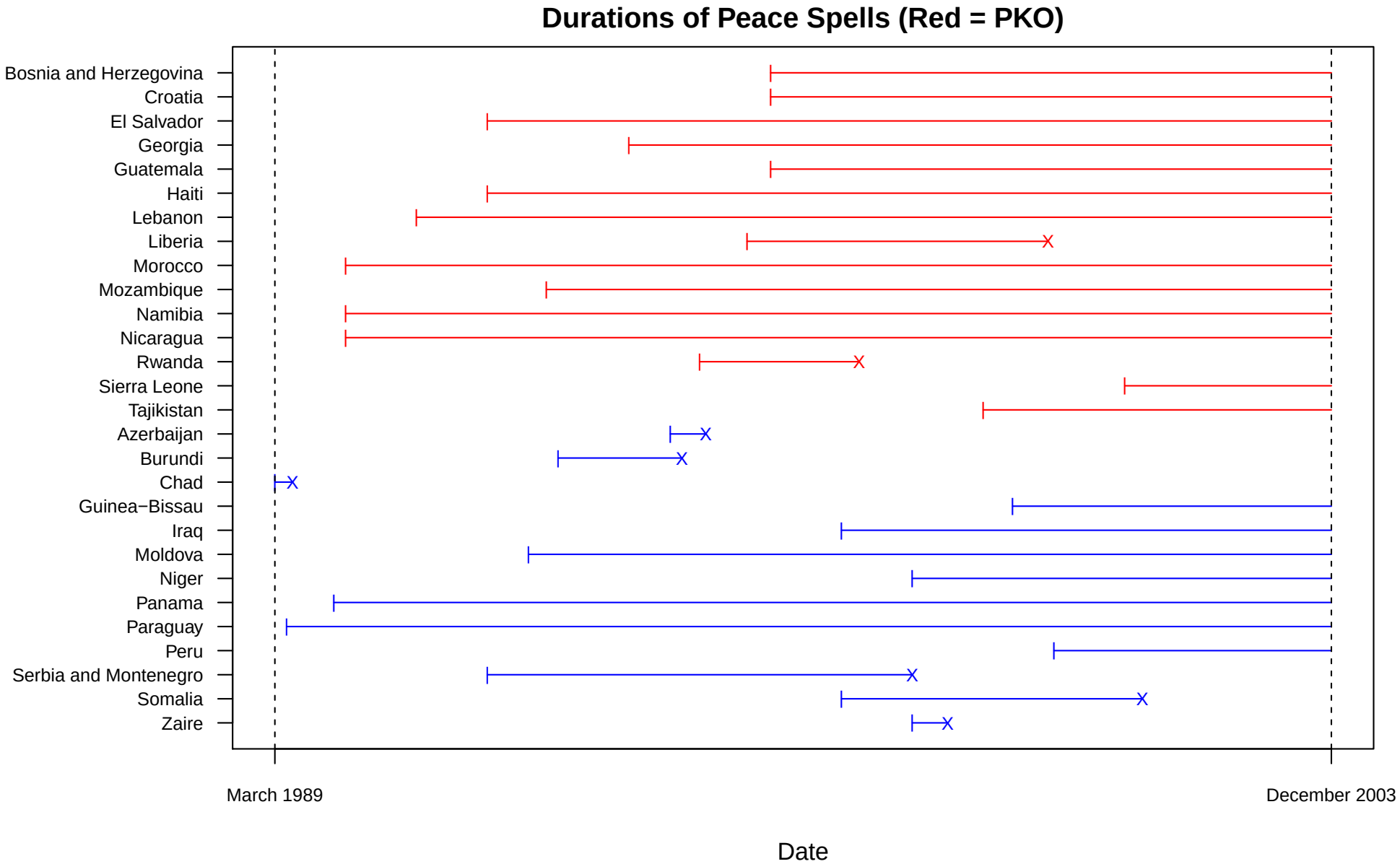




Count/positive data: Simulation

- Controlling for FE and having decent X specification really important.
- Accounting for effect heterogeneity really important.
  - Avg. derivative methods does not account for this properly.
- Whether you use OLS or Poisson not so important given the above.

Duration data



(Data from Gilligan & Sergenti, 2008)

Duration data

- Failing to account for right-censoring misrepresents differences in durations.
- Can't estimate $\mathrm{E}[Y_{1i} - Y_{0i}]$ directly if Y is defined as the duration.

Duration data

- Failing to account for right-censoring misrepresents differences in durations.
- Can't estimate $\mathrm{E}[Y_{1i} - Y_{0i}]$ directly if Y is defined as the duration.
- Common approach: shift attention to survival functions and hazard rates.
- Suppose T_i is time when failure occurs for i.
- Survival function: $S(t) \equiv \Pr[T_i > t] = 1 - \Pr[T_i \leq t] = 1 - F(t)$.
- Hazard function: If $F'(t) = f(t)$ is defined, hazard function is $h(t) = f(t)/S(t)$. "Instantaneous failure rate at time t among those surviving to t."

Duration data: Nonparametric methods

- In medicine, duration data common is outcomes in RCTs.
- Common approach for causal effects is non-parametric Kaplan-Meier estimation of survival function.

Duration data: Nonparametric methods

- In medicine, duration data common is outcomes in RCTs.
- Common approach for causal effects is non-parametric Kaplan-Meier estimation of survival function.
- Suppose we observe failure times for M of N subjects. Order these as $t_1 \leq \ldots \leq t_M$.
- Let n_k be the number of subjects surviving and uncensored up to just prior to t_k, and d_k be the number of failures at t_k.
- Then, we can estimate the survival function as,

$$\hat{S}(t) = \prod_{k:t_k<t} \frac{n_k - d_k}{n_k}$$

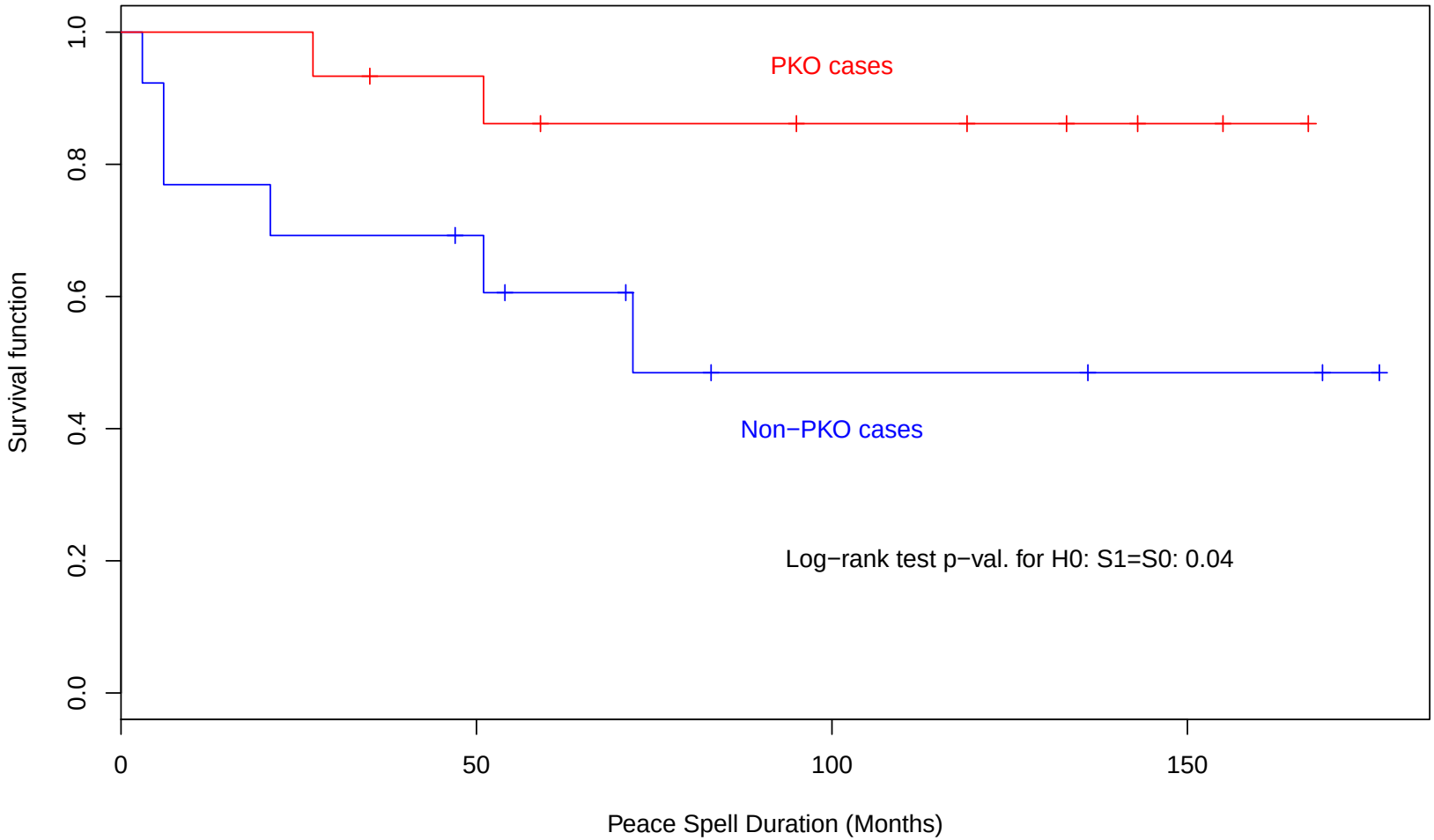
Duration data: Nonparametric methods

- In medicine, duration data common is outcomes in RCTs.
- Common approach for causal effects is non-parametric Kaplan-Meier estimation of survival function.
- Suppose we observe failure times for M of N subjects. Order these as $t_1 \leq \ldots \leq t_M$.
- Let n_k be the number of subjects surviving and uncensored up to just prior to t_k, and d_k be the number of failures at t_k.
- Then, we can estimate the survival function as,

$$\hat{S}(t) = \prod_{k:t_k<t} \frac{n_k - d_k}{n_k}$$

- For treatment effects, estimate survival functions for treated and control, then test for difference. Standard approach is to use "log-rank" test statistic, which is computed as normalized deviation from expected failures under the null. Can test against asymptotic normal distribution or do a permutation based test.

Duration data: Nonparametric methods



```
Call: survfit(formula = Surv(dur, .d) ~ UN)

                UN=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3     13       1    0.923  0.0739        0.789        1.000
    6     12       2    0.769  0.1169        0.571        1.000
   21     10       1    0.692  0.1280        0.482        0.995
   51      8       1    0.606  0.1382        0.387        0.947
   72      5       1    0.485  0.1548        0.259        0.906

                UN=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   27     15       1    0.933  0.0644        0.815            1
   51     13       1    0.862  0.0911        0.700            1
```

Duration data: Regression

- Kaplan-Meier does not allow for continuous covariate adjustment.
- You can compute separate KM estimates within strata, but this is tricky with either continuous covariates or when you have lots of dummy variables.

Duration data: Regression

- Kaplan-Meier does not allow for continuous covariate adjustment.
- You can compute separate KM estimates within strata, but this is tricky with either continuous covariates or when you have lots of dummy variables.
- A way to get beyond this is to suppose that a *hazard function* can be defined and then to model the *conditional* hazard function.
- The hazard function will tend to vary over time, so we need to specify how this occurs.

Duration data: Cox regression

- A somewhat agnostic approach is the Cox regression model:

$$h_i(t) = \exp[\alpha(t) + X_i'\beta] = \exp[\alpha(t)]\exp(X_i'\beta),$$

with $\exp[\alpha(t)] = h_0(t)$, the "baseline hazard," which, remarkably, we can leave unspecified. (Note: there is no constant in X_i.)

Duration data: Cox regression

- A somewhat agnostic approach is the Cox regression model:

$$h_i(t) = \exp[\alpha(t) + X_i'\beta] = \exp[\alpha(t)]\exp(X_i'\beta),$$

  with $\exp[\alpha(t)] = h_0(t)$, the "baseline hazard," which, remarkably, we can leave unspecified. (Note: there is no constant in X_i.)
- This model assumes that hazard rate ratios are constant over t for units with covariate profiles X_i and X_j:

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp[\alpha(t)]\exp(X_i'\beta)}{\exp[\alpha(t)]\exp(X_j'\beta)} = \frac{\exp(X_i'\beta)}{\exp(X_j'\beta)} = \exp[(X_i - X_j)\beta],$$

- Known as "proportional hazards" assumption.
- Fit via "partial likelihood."
- Robust and cluster-robust standard errors are available.

Duration data: Cox regression

- Signs and significance of treatment effect estimates can be read off of Cox model coefficients.
- But the coefficients are on the log-hazard rate scale.
- Hard to interpret substantively.
- Exponentiated coefficients indicate hazard multipliers, but this too is a bit esoteric.
- For interpretation, convert to survival functions (cf. Box-Steffensmeier & Jones, 2004, pp. 64-65):

Duration data: Cox regression

Retrieving the survival function:

- By definition (cf. Box-Steffensmeier & Jones, 2004, p. 14), for baseline group, $S_0(t) = \exp[-\int_0^t h_0(u)du]$.
- By our model,

$$\begin{aligned} S_i(t) &= \exp\left[-\int_0^t h_0(u)\exp(X_i'\beta)du\right] \\ &= \exp\left[-\int_0^t h_0(u)du\right]^{\exp(X_i'\beta)} \\ &= S_0(t)^{\exp(X_i'\beta)} \end{aligned}$$

- With this, we substitute $\hat{\beta}$ in, and then use one of a number of methods to back out $S_0(t)$ (e.g., Breslow estimator, Kalbfleisch/Prentice estimator).

Duration data: Cox regression

Effects of UN PKO on Hazard of War Recurrence (Cox regression)

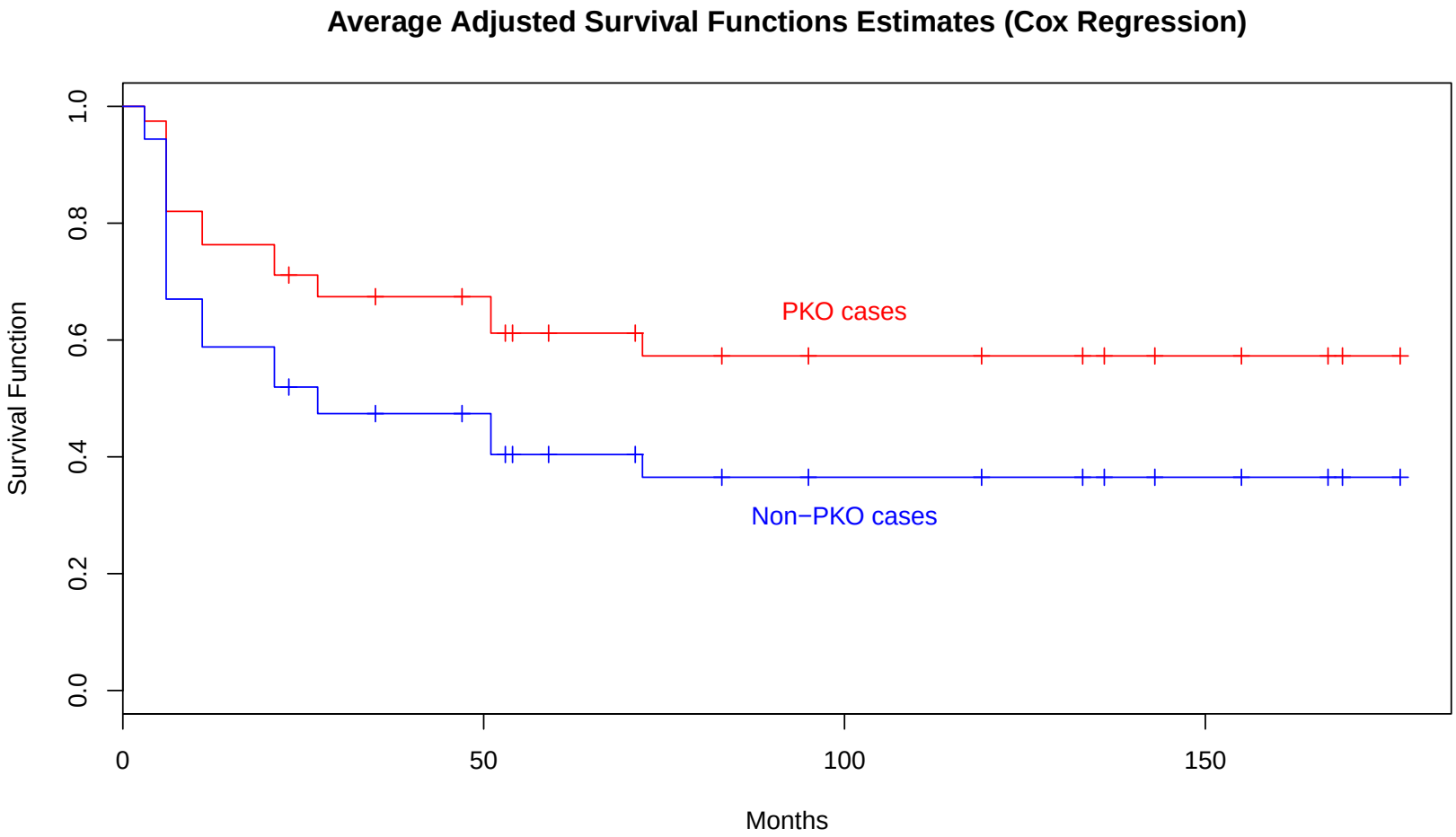
	Coef.*	exp(coef.)	Robust s.e.	z	Pr(>|z|)
UN	-2.45	0.09	1.66	-1.48	0.14
lwdeaths	0.46	1.58	0.44	1.05	0.30
lwdurat	-0.01	0.99	0.02	-0.78	0.43
ethfrac	0.01	1.01	0.05	0.26	0.80
pop	-0.36	0.70	0.78	-0.46	0.65
lmtnest	0.44	1.55	0.60	0.72	0.47
milper	0.27	1.31	0.68	0.40	0.69
bwgdp	-0.69	0.50	0.38	-1.82	0.07
bwplty2	-0.23	0.80	0.10	-2.24	0.03

$N = 38$, number of events $= 15$.

Likelihood ratio test stat. $= 27.3$ on 9 df, $p < 0.001$.

*Cox regression coefficients on the log-hazard rate scale.

Duration data: Cox regression



Duration data: Extensions

- Cox regression admits time-varying treatments and covariates. These should be used with caution, however, to avoid post-treatment bias.
- Cox and other duration models can be approximated by logistic regression (cf. Box-Steffensmeier & Jones, 2004, Ch. 4-5).

Duration data: Extensions

- Proportional hazards and other parametric assumptions may be wrong, leading to inconsistency.
- Unclear how Cox regression performs under misspecification (effect heterogeneity, higher order terms).

Duration data: Extensions

- Proportional hazards and other parametric assumptions may be wrong, leading to inconsistency.
- Unclear how Cox regression performs under misspecification (effect heterogeneity, higher order terms).
- Many elaborate duration models, but likely sensitive to misspecification and have convergence issues.
- Possible agnostic/semi-parametric alternatives:
  - Quantile regression and censored quantile regression (MHE, 7.1.1; Koenker, 2008).
  - OLS (or logit), using a panel data set-up.
- More research should be done with these alternatives!

Remarks

One perspective on what this all means:

> *[T]echnical challenges posed by LDV models come primarily from what I see as a counterproductive focus on structural parameters such a latent index coefficients or censored regression coefficients instead of directly interpretable causal effects. In my view, the problem of causal inference with LDV's is not fundamentally different from causal inference with continuous outcomes.*

(Angrist 2001, p. 3)

Angrist proposes that OLS and related methods (quantile regression, Abadie-type kappa weighting) produce causal effect estimates either identical or more reliable than those from non-linear models.

Remarks

But others disagree:

> *[T]he choice of the estimand is distinct from the statistical question of the specification of the model....The aim is to provide a flexible approximation. [F]or a binomial distribution the logistic regression model can be thought of as providing a linear approximation to the log odds ratio, this choice is...an appealing one....In cases with other limited dependent variables, alternative nonlinear models may be appropriate.*

(Imbens 2001, pp. 18-20)

Remarks

- My personal *belief*: fine to use MLE-glm estimators of treatment effects so long as they are interpreted correctly.
- My personal *experience*: in applied settings, MLE-glm estimators tend not to differ very much from linear OLS estimators.
- I have found that problems with fixed effects in MLE-glms tend to dominate over problems of non-sensical predictions from linear models (cf. Beck 2015).
- Moreover, these glms are hardly sophisticated: they are unlikely to be "right" either. Why not take this further and use *even more* flexible methods? (cf. Van der Laan & Rose book).
- These arguments are predicated on the idea that what interests us are average causal effects (e.g., ATEs or LATEs). If it is $\hat{Y}$ that you need, then, yes, it makes sense to work with models that produce predictions within the support of Y.