# Lecture 22: Treatment Effects with Limited Dependent Variables II

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

April 24, 2018

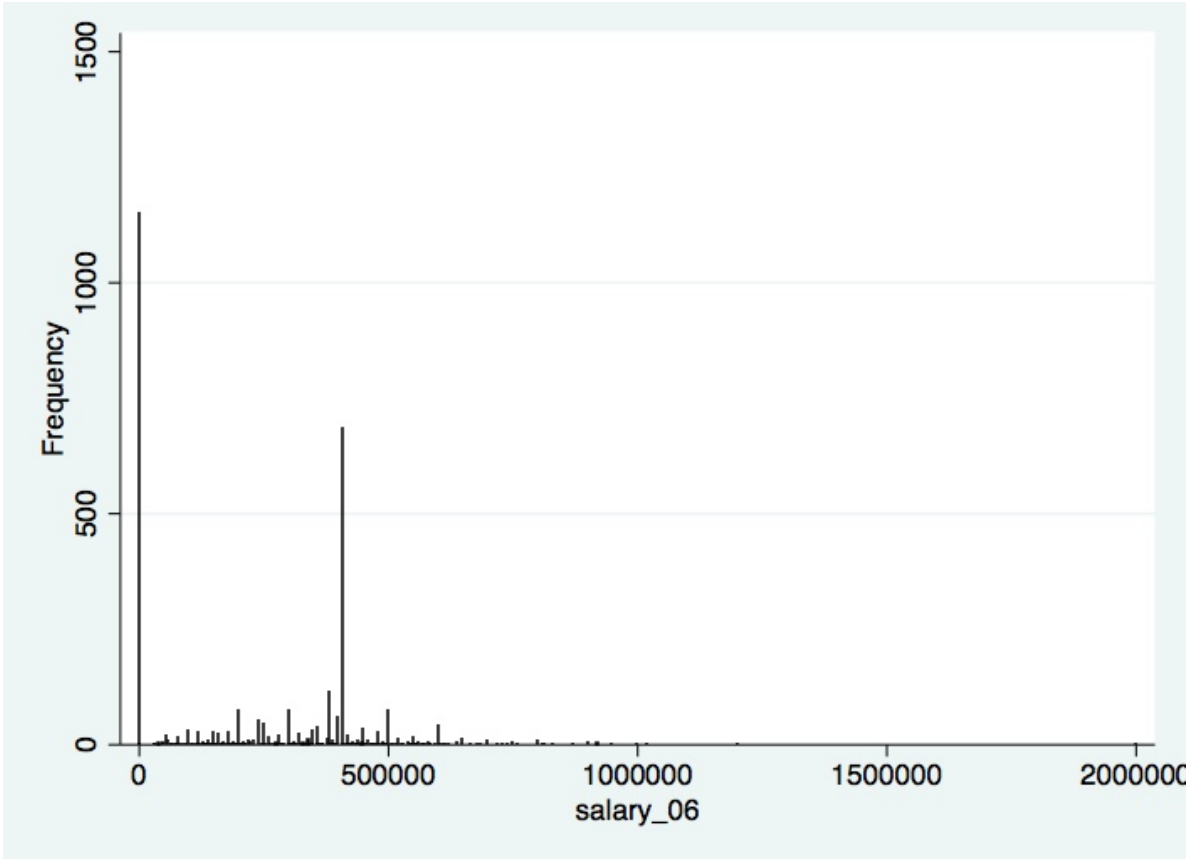
# Motivation

- Last time: treatment effects with discrete outcomes:
  - Binary.
  - Multichotomous, ordered, nested.
- Predicted values from linear model may extend beyond support of outcome.
- But in practice linear and non-linear models give similar results so long as specification accounted for confounds and effect heterogeneity.
- Linear models are well-behaved in situations that complicate nonlinear estimation (e.g., fixed effects).

# Motivation

- Today: truncated and censored outcomes:
  - Counts (left-truncated at 0).
  - Durations (left-truncated at 0 and often right-censored).

# Count data




(Attanasio et al., 2011)

# Count 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: "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 $\beta$ 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 data: Poisson regression

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

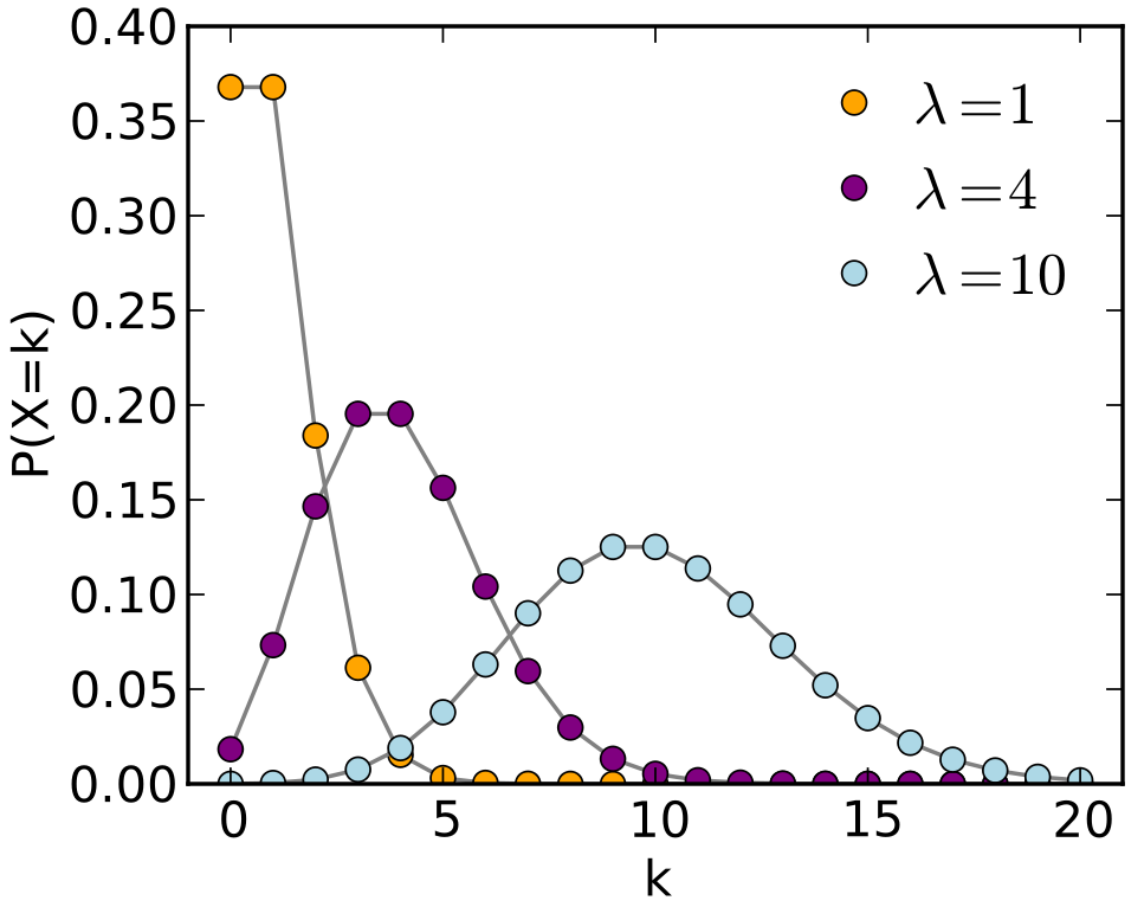
$$\Pr[Y_i = y] = \binom{M}{y} \pi_i^y (1 - \pi_i)^y$$

- Suppose for unit increase in $M_i$, reduce $\pi_i$ 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 data: Poisson regression




(Wikimedia Commons)

# Count 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 $\beta$:

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

  Cool!

# Count 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 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 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,

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

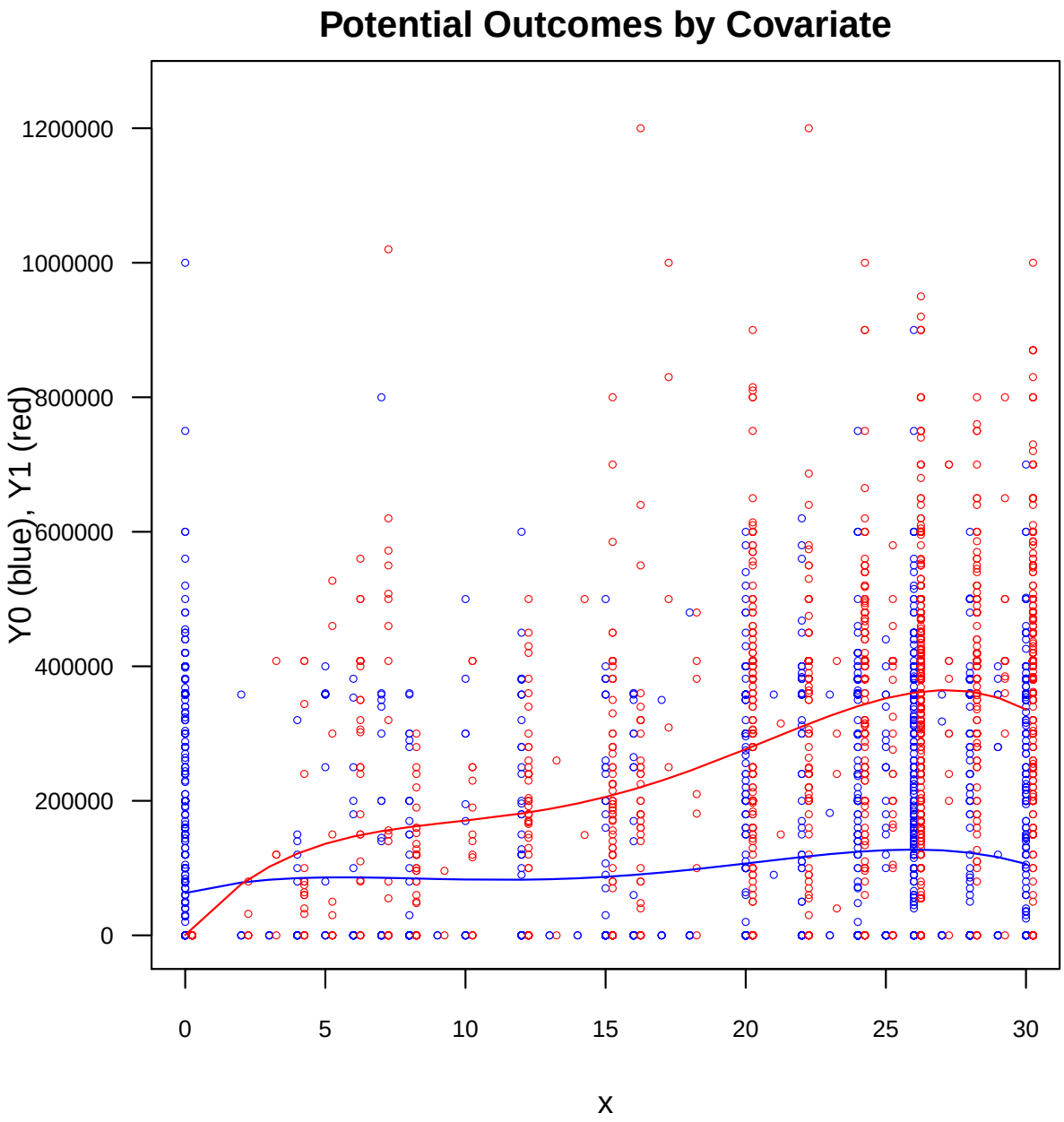
- Perhaps more interpretable is differences in predicted responses. E.g., for binary treatment $D_i$, $\widehat{ATE}$ given by,

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

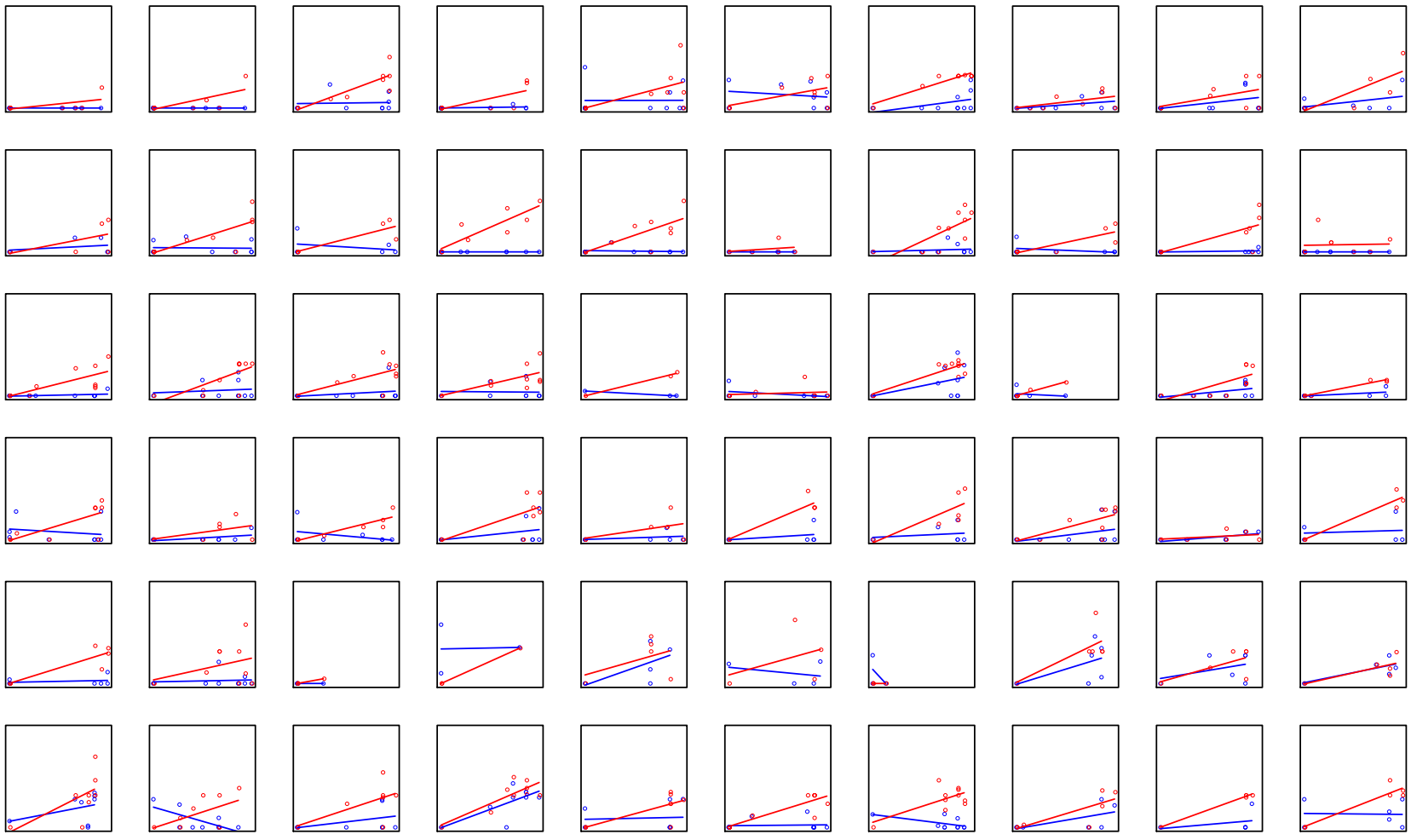
# Count data: Simulation

- 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 data: Simulation



# Count data: Simulation

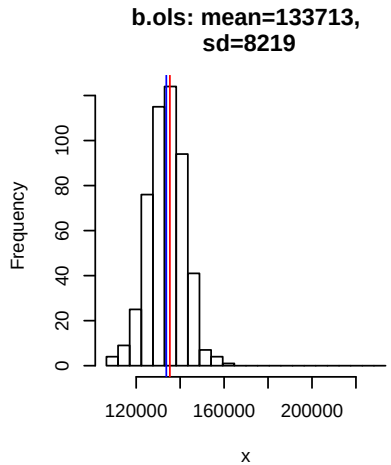


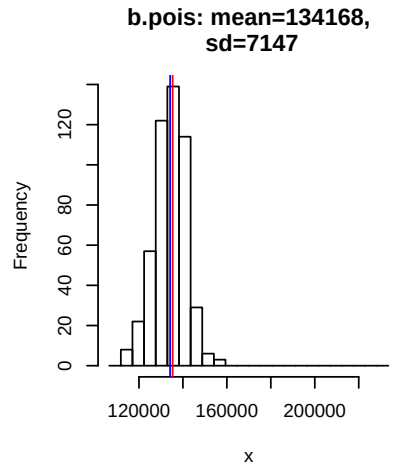
(60 out of 441 stratum-specific relationships.)

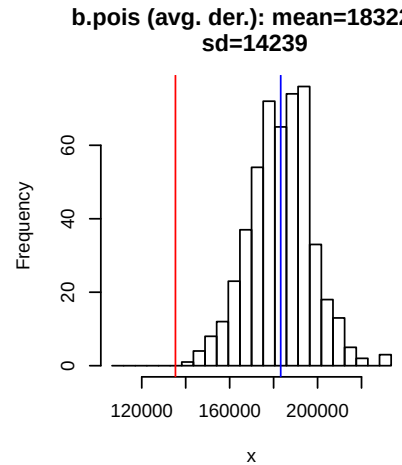
# Count data: Simulation

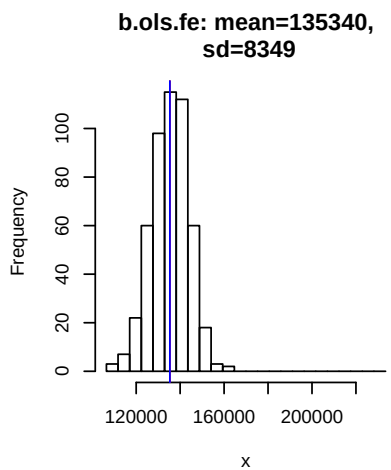
| Estimator | Specification | Estimator |
|---|---|---|
| OLS linear | $Y_{is} = \beta_0 + \beta_1 D_{is} + \beta_2 X_{is} + \beta_3 X_{is}^2 + \varepsilon_{is}$ | $\hat{\beta}_1$ |
| OLS linear FE | $Y_i = \alpha_s + \beta_0 + \beta_1 D_{is} + \beta_2 X_{is} + \beta_3 X_{is}^2 + \varepsilon_{is}$ | $\hat{\beta}_1$ |
| 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. | " | $\bar{\Delta}_{\partial D}$ |

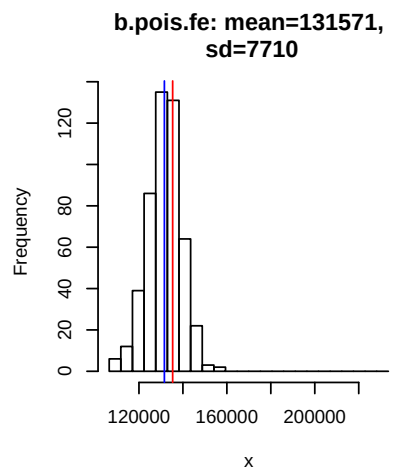
# Count data: Simulation result, random treatment

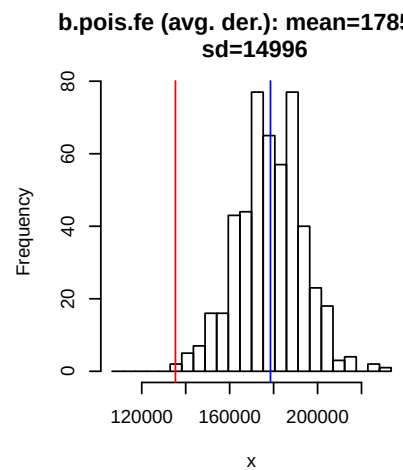


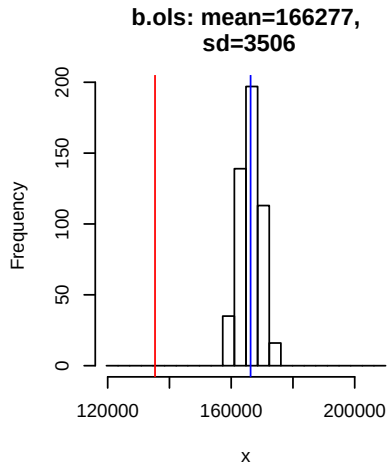



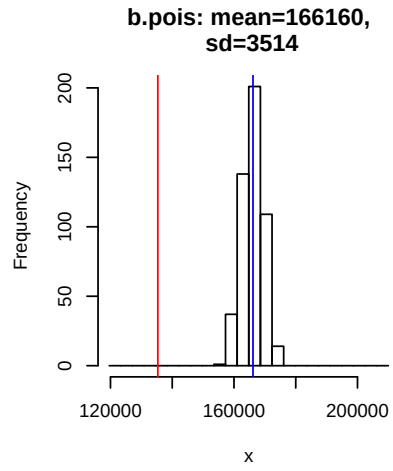



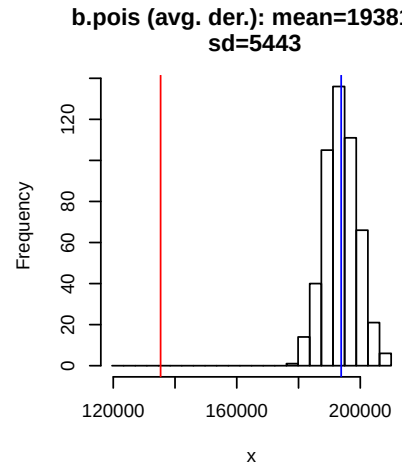



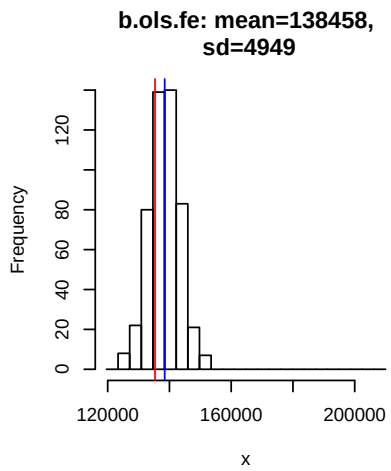



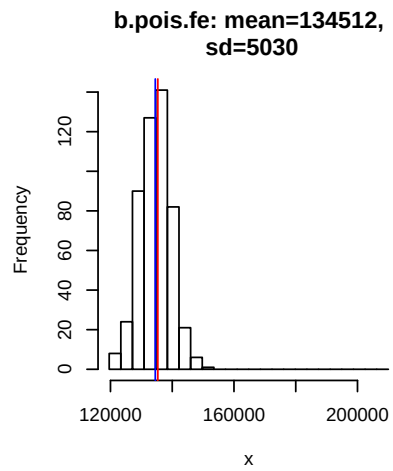
# Count data: Simulation result, endogenous treatment

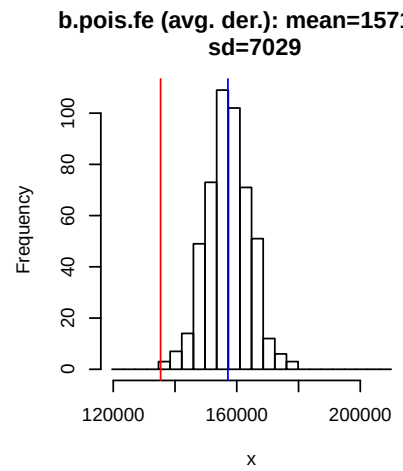


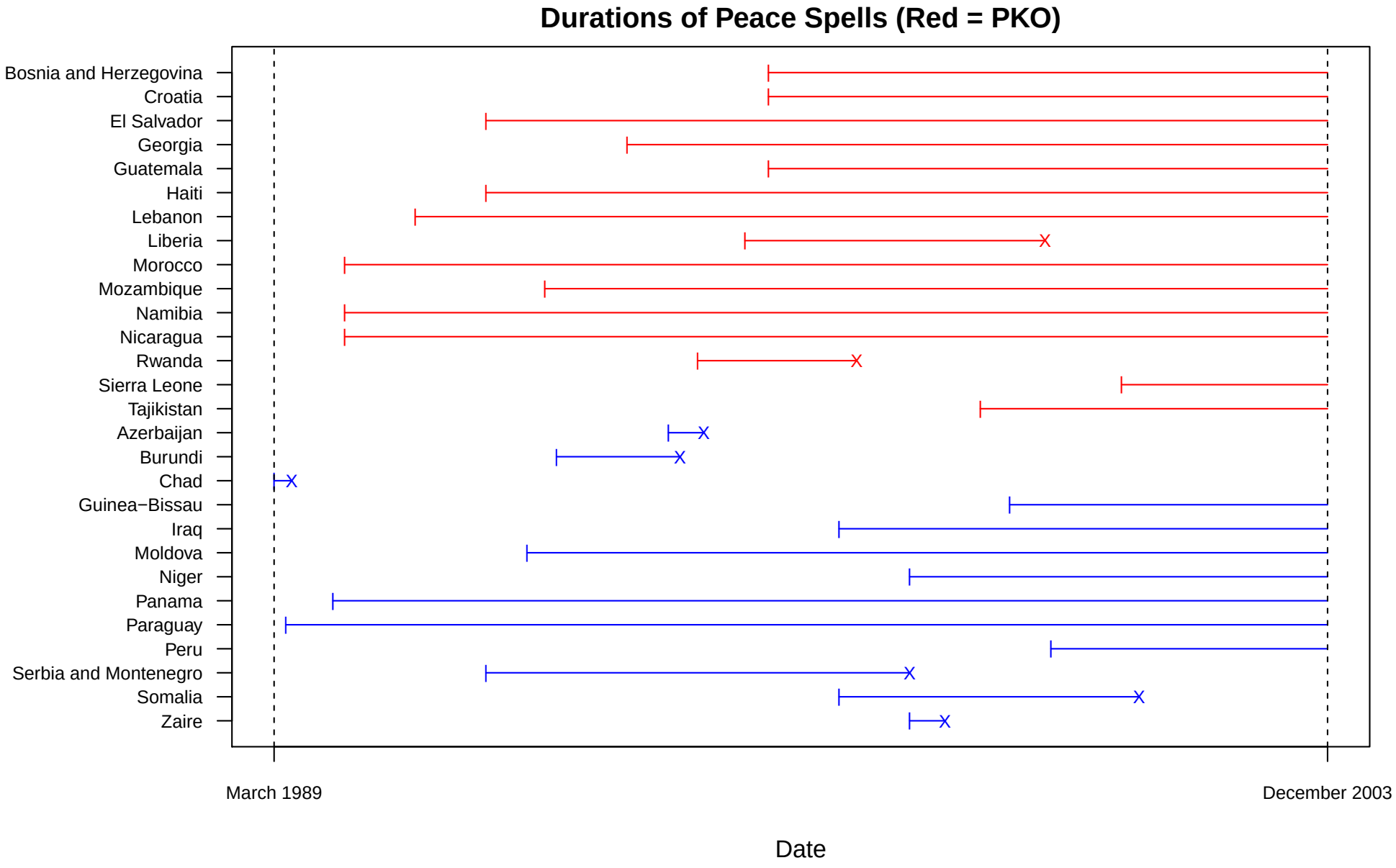



# Count data: Simulation

- 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.
- 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 $t$."

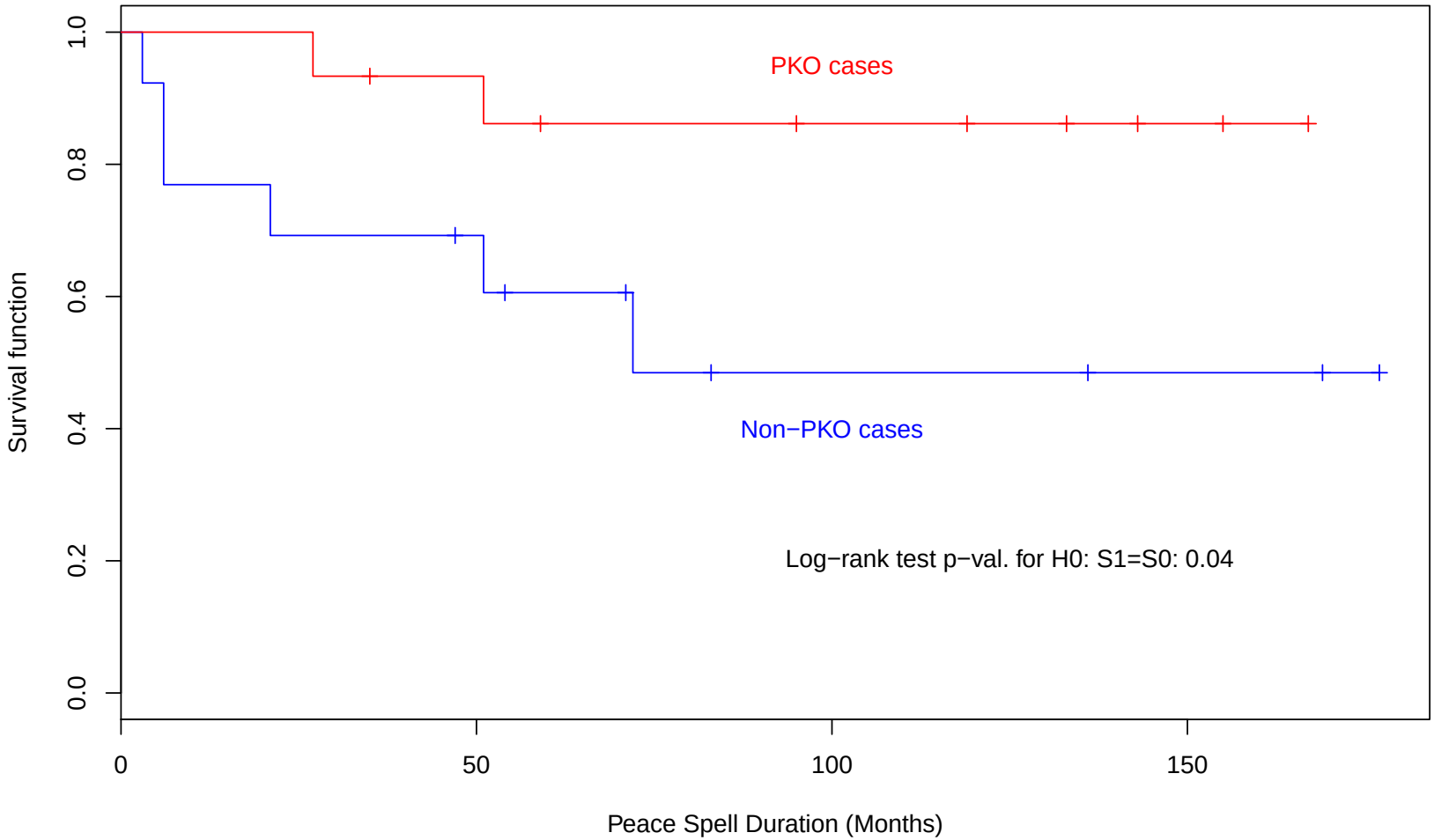
# Duration data: Nonparametric methods

- In medicine, duration data common is outcomes in RCTs.
- Common approach for causal effects is non-parametric Kaplan-Meier estimation of survival function.
- Suppose we observe failure times for $M$ of $N$ subjects. Order these as $t_1 \leq ... \leq t_M$.
- 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 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 $X_i$.)
- 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 presentation, you can convert back 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,

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

- 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)

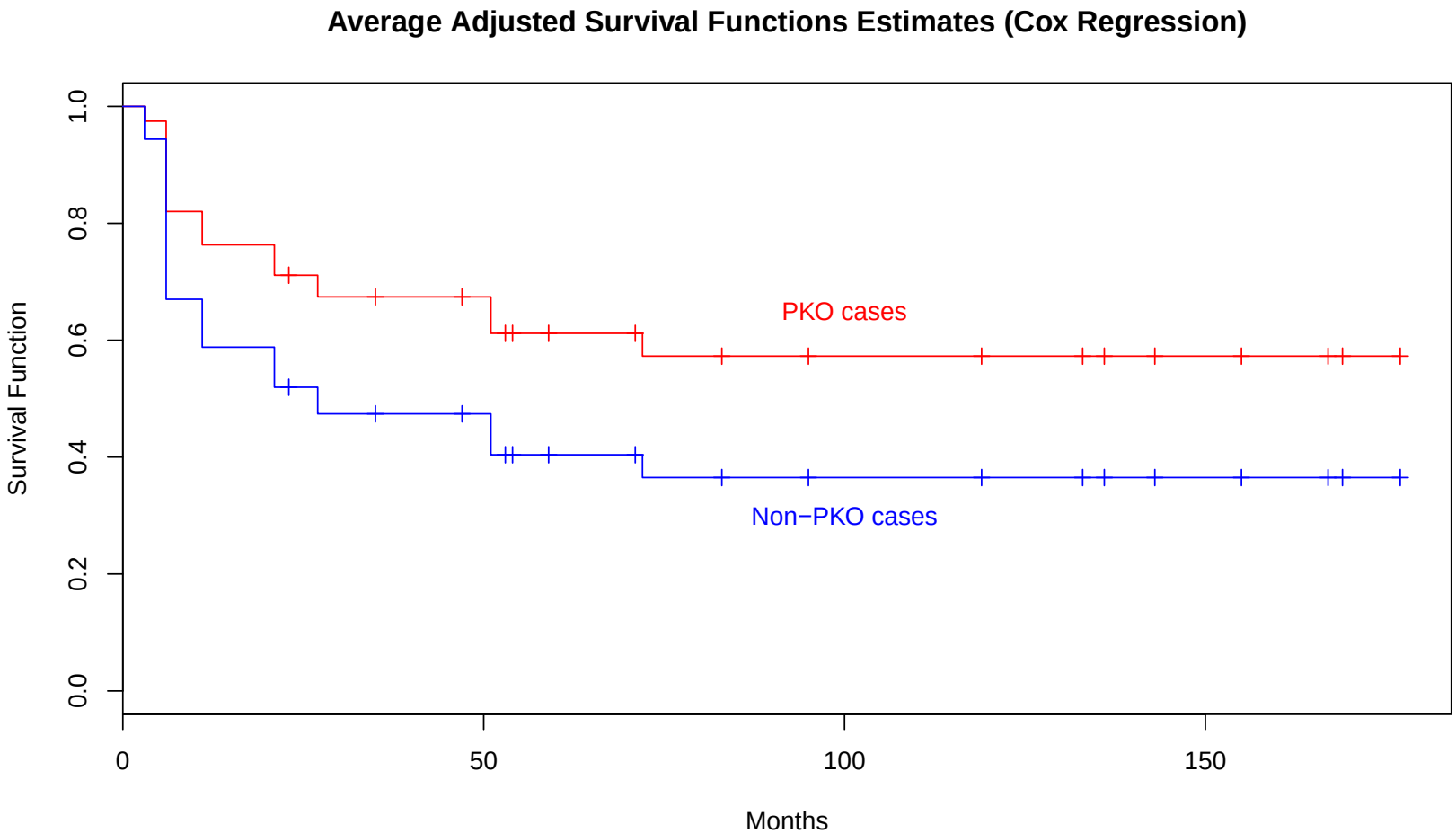
| | Coef.* | exp(coef.) | Robust s.e. | z | Pr(>\|z\|) |
|---|---|---|---|---|---|
| UN | -2.45 | 0.09 | 1.66 | -1.48 | 0.14 |
| lwdeaths | 0.46 | 1.58 | 0.44 | 1.05 | 0.30 |
| lwdurat | -0.01 | 0.99 | 0.02 | -0.78 | 0.43 |
| ethfrac | 0.01 | 1.01 | 0.05 | 0.26 | 0.80 |
| pop | -0.36 | 0.70 | 0.78 | -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.
- Unknown 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:
  - Censored quantile regression (MHE, 7.1.1; Koenker, 2008).
  - OLS (or logit), using a panel data set-up.
- More exploratory work should be done with these alternatives.

# Remarks

- Just the tip of the iceberg for LDVs.
- Work remains to determine how well they perform under misspecification in estimating *causal effects* on *natural scale* of outcome variable.