Modeling Health Care Costs and Counts

Partha Deb

Hunter College and the Graduate Center, CUNY, and NBER

Willard G. Manning

University of Chicago

Edward C. Norton

University of Michigan and NBER

ASHEcon, Los Angeles 2014

Overview

Statistical issues - skewness and the zero mass

Studies with skewed outcomes but no zeroes

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Top Ten Urban Myths of Health Econometrics

What is the cost/use of interest?

1. **Costs in fixed period of time (e.g., stroke costs paid or visits in 2012)?**
2. **Per episode or per lifetime costs/use of stroke in incident cases?**

Our focus is on the former

Second question requires survival methods and consideration of right censoring in data (not covered here)

Characteristics of Health Care

Costs and Utilization

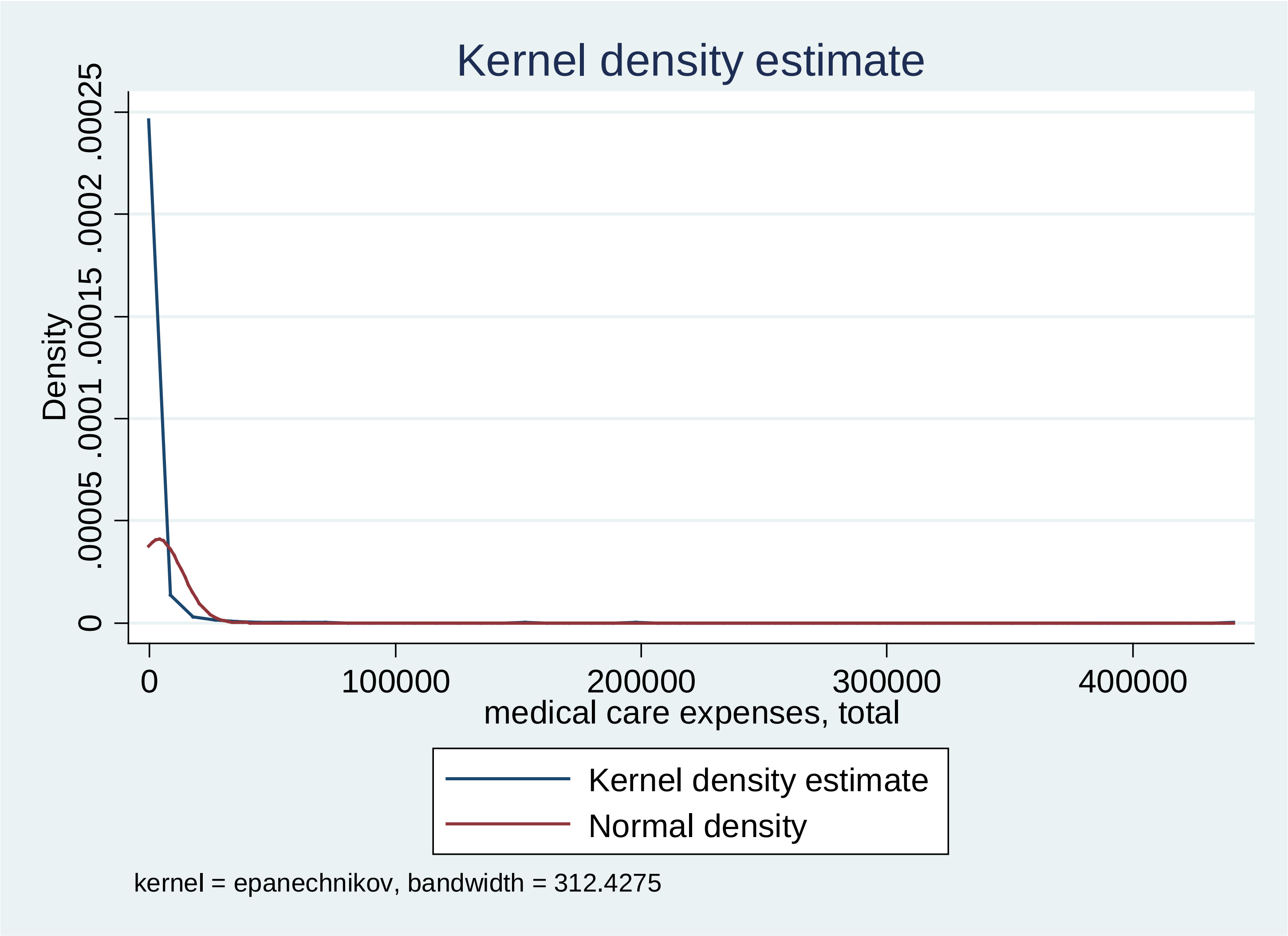
Large fraction of population without any care during period of observation

Consumption among those with any care is very skewed (visits, hospitalizations, costs)

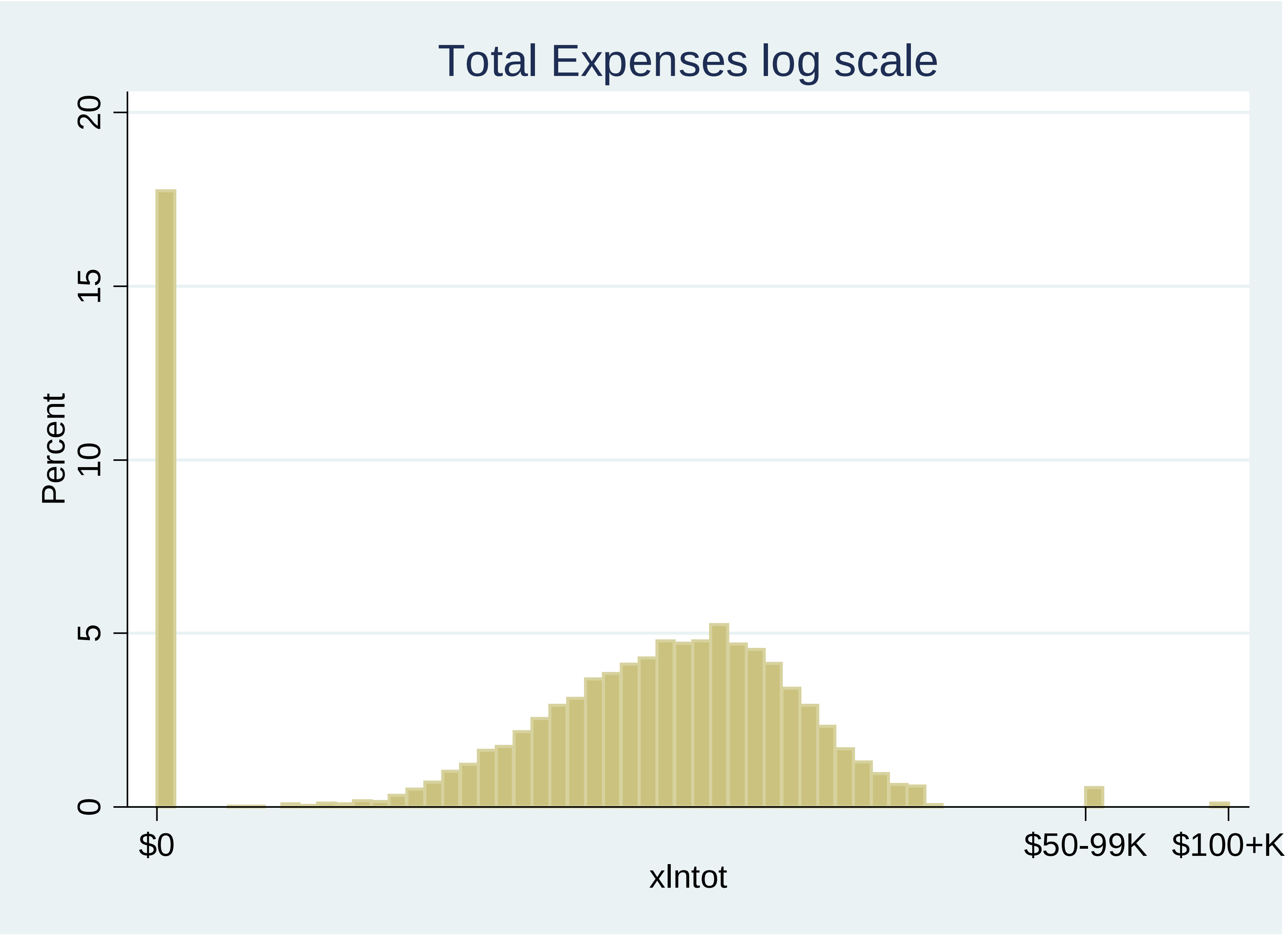
Nonlinearity in response to covariates

Cost response may change by level of consumption (e.g. outpatient versus inpatient, or low to high levels).

Density of Total Medical Expenditures, Adults, MEPS 2004

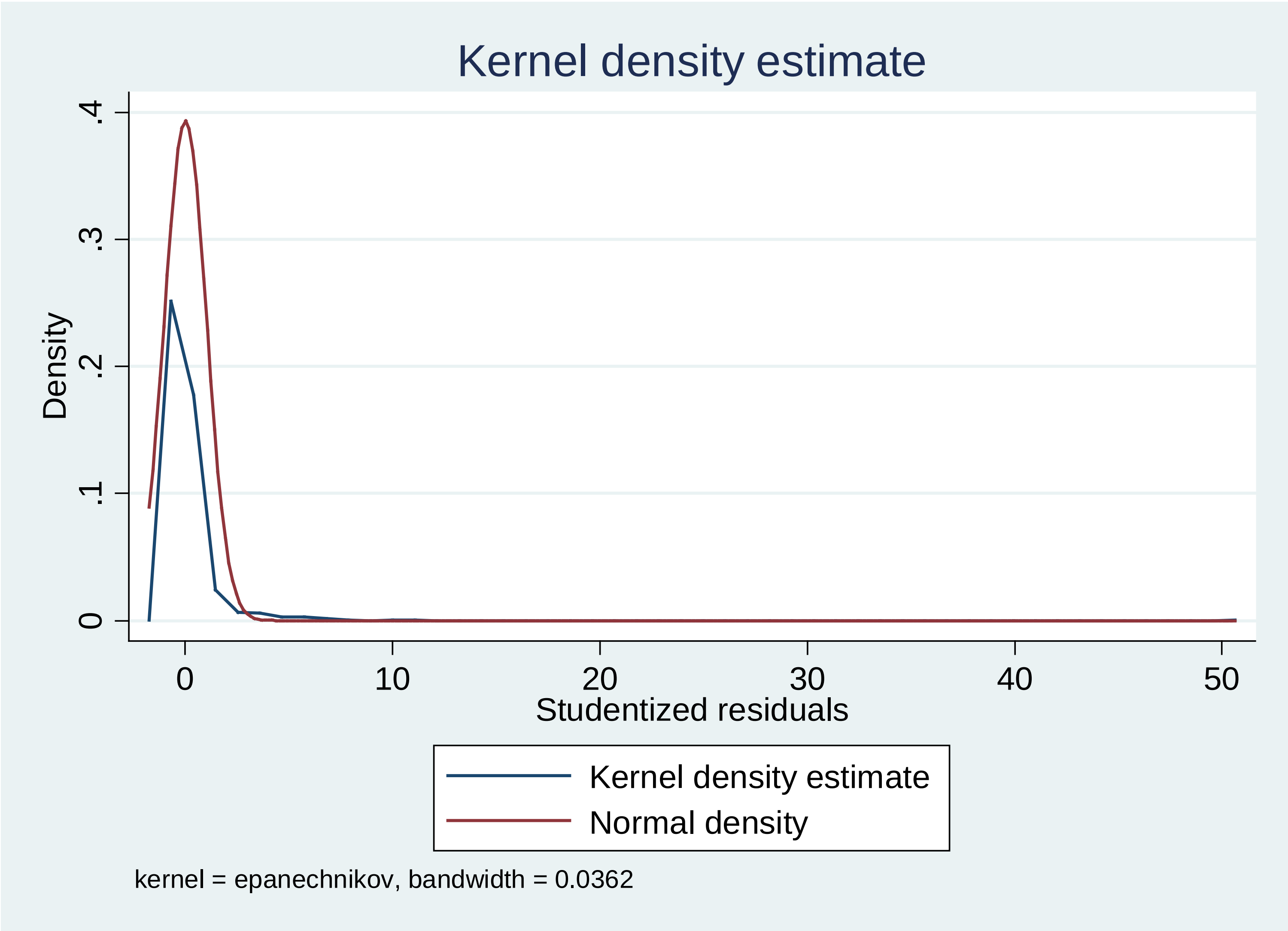


Discrete Version of Density for Total Medical Expenditures



“Bins” are $1000 wide, except upper two. Log-scaled to pull in right tail.

Density of Studentized Residuals for Total Medical Expenditures, Adults, MEPS 2004



Potential Problems from Ignoring Characteristics

Usual econometric (least squares) methods will yield less precise estimates of means and marginal effects

If failing to deal with inherently nonlinear response may lead to biased estimates for substantial subpopulations.

Results not robust to tail problems unless very large samples. Estimates from one subsample may forecast poorly to another subsample from same population

Need more robust methods that

- **Recognize distribution of data**
- **Are less sensitive to right tail**
- **Provide estimates of E(y|x)**

“All models are wrong but some are useful”

No single econometric model is “best” for all cases!

OLS on costs or counts can be biased, very inefficient and overly sensitive to influential outliers

But it is not all bad news!!

We will outline a variety of methods that

- **Work in many disparate situations**
- **Are easy to estimate (generally) in most statistical packages (we use Stata)**
- **Often provide a better fit**
- **Are less sensitive to outliers**
- **Can result in large efficiency gains vis-à-vis linear models**

We will outline approaches to making decisions about model selection, specification, and interpretation

Overview

Studies with skewed outcomes but no zero mass problems

Alternative models

Comparing alternative models

Assessing model fit

Interpretation

Overview (cont'd)

Studies with skewed outcomes but no zero mass problems

Alternative models

- **OLS on untransformed use or expenditures**
- **OLS for log(y)**
- **Box-Cox generalization**
- **Generalized Linear Models (GLM) / GMM**

Studies with No Zero Mass (cont'd)

Concerns

- Robustness to skewness
  - Reduce influence of extreme cases
  - Good forecast performance

- No systematic misfit over range of predictions or range of major covariates (e.g., price, income).

- Efficiency of estimator

OLS of y on x's

Advantages

- **Easy**
- **No retransformation problem**
- **Marginal /incremental effects easy to calculate**

Disadvantages

- **Not robust in small to medium sized data sets**
  or where some subgroups are small
- **Can produce out-of-range predictions:** $\hat{y}_i = x_i'\hat{\beta} < 0$
- **Inefficient (ignores heteroscedasticity)**
- **Poor out-of-sample/forecast performance**

OLS of y on x’s (cont’d)

Why are we concerned with robustness to skewness in OLS?

OLS overemphasizes extreme cases when data are very skewed right or cases have leverage. For OLS,

$$\hat{\beta} = \beta + (X'X)^{-1} X'\varepsilon$$

but some |ε|’s are extremely large, as well as x's extreme or rare

Raises the risk of influential outlier(s) that pull estimate $\hat{\beta}$ away from β

See earlier plots (p. 5-7) for potential problems

Log(y) or Box-Cox Models

Advantages

- **Widely known, especially log(*y*) version**
  - **Reduces robustness problem by focusing on symmetry**
- **Improved precision if *y* is skewed right**
- **May reduce (but not eliminate) heteroscedasticity**

Disadvantages

- **Retransformation problem could lead to bias**
- **Some Box-Cox version’s coefficients are not directly interpretable**
- **May not achieve linearity on estimation scale**

OLS for log(y)

OLS or MLE for $\log(y) = X\beta + \varepsilon$

where $E(\varepsilon) = 0$, $E(X'\varepsilon) = 0$

Estimates for $E(\log(y)|x)$, not $\log(E(y)|x))$

Usually want arithmetic mean, not geometric mean

May be difficult to obtain unbiased estimates of mean response $E(y|x)$ if error ε heteroscedastic in x's or other z's

Dilemma with OLS for log(y)

Logged estimates are often far more precise and robust than direct analysis of unlogged dependent variable y

But, no one interested in log scale results *per se*

Effect of Heteroscedasticity

Untransformed dependent variable (e.g., cost)

Need GLS for efficient estimates and to correct inference statistics

(*<u>Or</u>* use Huber/White/Eicker with OLS to get consistent inference statistics)

Transformed dependent variable (e.g., log(cost))

Need GLS for efficient estimates & correct inference statistics (Or use Huber/White/Eicker with OLS to get consistent inference statistics)

And correction for form of hetero. to yield consistent predictions on the raw (untransformed) scale

OLS on log(y) for comparison of two treatment groups with normal errors

Assume $\log(y)_G \sim N(\mu_G, \sigma^2_G)$ where treatment G = A or B

$$E(y \mid G = A) = e^{(\mu_A + 0.5\sigma_A^2)}$$

Under heteroscedasticity by group

$$\frac{E(y_A)}{E(y_B)} = e^{((\mu_A - \mu_B) + 0.5(\sigma_A^2 - \sigma_B^2))}$$

Under homoscedasticity (σ^2 = a constant)

$$\frac{E(y_A)}{E(y_B)} = e^{(\mu_A - \mu_B)}$$

Note: Same issue applies if error is not normally distributed

Retransformation with Covariate Adjustment

Suppose $y > 0$ and we run OLS regression for $ln(y) = x\beta + \varepsilon$

With $E[\varepsilon|x] = 0$, β and $E[ln(y)|x]$ consistently estimated by linear regression

Policy questions not typically focused on β *per se*, but on how $E[y]$ varies with x

Retransformation (cont'd)

Expectations if E(ε) = 0 and E(X′ε) = 0:

$$E(y_i) = e^{x_i'\beta}\, E(e^{\varepsilon_i} \mid x_i)$$

$E(y_i) \neq e^{x_i'\beta}$ **as is often assumed**

$E(y_i) \neq cons \cdot e^{x_i'\beta}$ **if ε is heteroscedastic in x**

Retransformation (cont'd)

Marginal effects of a covariate x (e.g., income) on expected outcome on the raw scale:

$$\frac{\partial E(y_i)}{\partial x_k} = e^{x_i'\beta}\left(\beta_k E(e^{\varepsilon_i} \mid x_i) + \frac{\partial E(e^{\varepsilon_i} \mid x_i)}{\partial x_k}\right)$$

$$\frac{\partial E(y_i)}{\partial x_k} \neq \{E(y_i \mid x_i)\}\,\beta_k$$ **if heteroscedastic in x**

$$\frac{\partial E(y_i)}{\partial x_k} \neq \{e^{x_i'\beta}\}\beta_k$$ **as is often assumed**

Examples

Health Insurance Experiment (HIE) error variance on log scale for users increasing in cost sharing for outpatient and total medical expenses. Use of homoscedastic model overstates effect of cost sharing. (Manning, JHE, 1998)

Visits from National Health Interview Survey. Response heteroscedastic in gender and education. (Mullahy, JHE, 1998)

MEPS 2004 response heteroscedastic in income and education (see below)

Box-Cox Models

Log transform is not only solution to skewness

Assume transform of y such that:

$$[(y_i^{\lambda} - 1)/\lambda] = x_i' \beta + \varepsilon_i \quad \text{if } \lambda \neq 0$$

$$\log(y_i) = x_i' \beta + \varepsilon_i \qquad \text{if } \lambda = 0$$

where ε_i is distributed iid as N(0, σ^2).

Estimate by MLE

Tends to minimize skewness in residuals

Log is not always "best" transform; depends on degree, sign of skewness.

GEP Box and DR Cox, *JRSS, Series B* (1964)

Example: Square Root Model by OLS

Assume that $\sqrt{y}$ is linear in β and additive in ε

$$\sqrt{y_i} = x_i{}'\beta + \varepsilon_i$$

with E(ε) = 0 and E(x'ε) = 0. Then,

$$E(\hat{\beta}_{OLS}) = \beta$$

Thus, OLS or least squares unbiased on square-root scale.

Heteroscedasticity only raises efficiency and inference problems on square-root scale

Square Root Model by OLS (cont'd)

Back to the raw scale:

$$y_i = (x_i'\beta)^2 + 2\,(x_i'\beta)\,\varepsilon_i + \varepsilon_i^2$$

Thus

$$E(y_i \mid x_i) = (x_i'\beta)^2 + \sigma_\varepsilon^2(x)$$

What is the marginal effect of x?

$$\frac{\partial E(y_i)}{\partial x_k} = 2(x_i'\beta)\beta_k + \frac{\partial \sigma_\varepsilon^2(x_i)}{\partial x_k} \neq 2\,(x_i'\beta)\,\beta_k$$

Heteroscedasticity on square-root scale raises bias issues on raw-scale if not properly retransformed

FYI: General Box-Cox Case

Box-Cox Model

$$[(y_i^{\lambda} - 1)/\lambda] = x_i'\beta + \varepsilon_i \quad \text{if } \lambda \neq 0$$

$$\log(y_i) = x_i'\beta + \varepsilon_i \quad \text{if } \lambda = 0$$

Raw scale value of y

$$y = [\lambda(x'\beta + \varepsilon) + 1]^{\left(\lambda/(1-\lambda)\right)}$$

Marginal effect of covariate x_j

$$\frac{\partial}{\partial x_j} E(y \mid x) = \frac{\partial}{\partial x_j} \int [\lambda(x'\beta + \varepsilon) + 1]^{\left(\lambda/(1-\lambda)\right)} \, dF(\varepsilon \mid x)$$

Abrevaya, *Econometric Reviews*, 2002

More on Retransformation Issues

Normal assumption is not innocuous!

Although estimates of β's may be insensitive, the expectation of untransformed value can be quite sensitive to departures from normality, esp. in right tail

Solutions

Use Duan's (JASA, 1983) smearing estimator by subgroup, which is non-parametric. Difficult if heteroscedastic in a continuous covariates or in multiple covariates

Use an appropriate Generalized Linear Model (GLM)

Retransformation Issues (cont'd)

Retransforming model results for log(y) by least squares

$$\ln(y) = x\beta + \varepsilon$$

Homoscedastic case

$$E(y \mid x) = e^{(x\beta + 0.5\sigma^2)}$$ **if ε is normally distributed**

$$E(y \mid x) = \left(e^{x\beta}\right) s,$$ **if not normally distributed**

$$\hat{s} = \frac{1}{N}\sum e^{(\log(y) - x\hat{\beta})}$$ **smearing (Duan, *JASA* 1983)**

Heteroscedastic by group

Different variances by group if ε normally distributed

Different smearing by group if ε not normal

Duan's Smearing Estimator

Sample Stata code for homoscedastic case

```
regress lny $x
     predict double resid, residual
egen Dsmear = mean(exp(resid))
display Dsmear
```

Consistent estimate of E(exp(ε)); Duan (*JASA*, 1983)

The smearing factor is typically between 1 and 4

Separate smearing by group if heteroscedastic by group

Retransformation Issues (cont'd)

Generally error ε is not normally distributed, heteroscedasticity may be complex, or may be heteroscedastic in several variables

Normal theory retransformation methods can be biased

Heteroscedastic smearing by group is:

- Inefficient
- Difficult if covariate continuous

Alternative: model E(y|x) directly using GLM

Generalized Linear Models (GLM)

Goal

estimate mean of y, conditional on covariates x's

Specify

a distribution that reflects mean - variance relationship

a link function between linear index xβ and mean μ = E(y|x)

Example

Gamma regression with log link

V(y|x) proportional to $[E(y \mid x)]^2$

$$Log(E(y \mid x_i)) = x_i\beta \quad \Rightarrow \quad E(y_i \mid x_i) = e^{x_i'\beta}$$

′

GLM (cont'd)

Use data to find distributional family and link

Family "down weights" noisy high mean cases

Link can handle linearity in sense of no systematic misfit

Note difference in roles from Box-Cox model

- **Box-Cox power transforms to gain symmetry in error (residual)**
- **GLM with power link function addresses linearity of response on scale-of-interest (raw-scale)**

GLM (cont'd)

GLM/GEE/GMM modeling approach's estimating equations

$$\sum_{i=1}^{N} \frac{\partial \mu(\mathbf{x}_i'\beta)}{\partial \beta} \times V(\mathbf{x}_i)^{-1} \times (y_i - \mu(x_i'\beta)) = 0$$

Given correct specification of E[y|x] = μ(x'β), the key issues relate to second-order or efficiency effects

This requires consideration of the structure of V(y|x)

GLM Variance Structure

Accommodates skewness and related issues via variance-weighting rather than transform/retransform methods

Assumes $\mathrm{Var}[y|x] = \alpha \times [E(y|x)]^{\delta}$

$$= \alpha \times [\exp(x\beta)]^{\delta}$$

This implies moment restriction:

$$E[\{y - \exp(x'\beta)\}^2 - \{\alpha \times [\exp(x'\beta)]^{\delta}\} \mid x] = 0$$

GLM Variance Structure (cont'd)

For GLM, can

1. Adopt alternative "standard" parametric distributional assumptions,

$\delta = 0$ (e.g. Gaussian NLLS)
$\delta = 1$ (e.g. Poisson)
$\delta = 2$ (e.g. Gamma)
$\delta = 3$ (e.g. Wald or inverse Gaussian)

Estimation and inference available in Stata's glm or xtgee procedures

If $\hat{\delta}$ not near integer, consider extended GLM (see below) or use closest parametric case and take an efficiency loss

GLM Variance Structure (cont'd)

2. Estimate δ via:

- **modified "Park test" estimated by GLM)** **preferred**

  gamma regression of $(y-\hat{y})^2$ on $[1, x'\hat{\beta}]$

- **modified Park Test by least squares**

  linear regression of $\log((y-\hat{y})^2)$ on $[1, x'\hat{\beta}]$

- **nonlinear regression of**

  $(y-\hat{y})^2$ on $\alpha(\exp(x'\hat{\beta}))^{\delta}$

Use the estimates to construct working V(x) and conduct (more efficient) *second-round* estimation and inference

Overview

Studies with skewed outcomes but no zero mass problems

Alternative models

Comparing alternative models

Assessing model fit

Interpretation

Performance of Alternative Estimators

Examine alternative estimators of log(E(y|x)) for consistency and precision

Determine sensitivity to common data problems in health economics applications

- Skewness
- Heavy tailed, even with log transform
- Heteroscedasticity
- Different shapes to pdf

Results: no dominant estimator

See Manning and Mullahy (JHE, 2001) for details of Monte Carlo simulation

FYI: Monte Carlo Simulation

Data generation

Skewness in dependent measure

Log normal with variance 0.5, 1.0, 1.5, 2.0

Heavier tailed than normal on the log scale

Mixture of log normals

Heteroscedastic responses

Std. dev. proportional to x

Variance proportional to x

Alternative pdf shapes

monotonically declining or bell-shaped

Gamma with shapes 0.5, 1.0, 4.0

FYI: Estimators Considered

Log-OLS with

- **homoscedastic retransformation**
- **heteroscedastic retransformation**

Generalized Linear Models (GLM), log link

- **Nonlinear Least Squares (NLS)**
- **Poisson**
- **Gamma**

Figure 1
Effect of Skewness on the Raw Scale

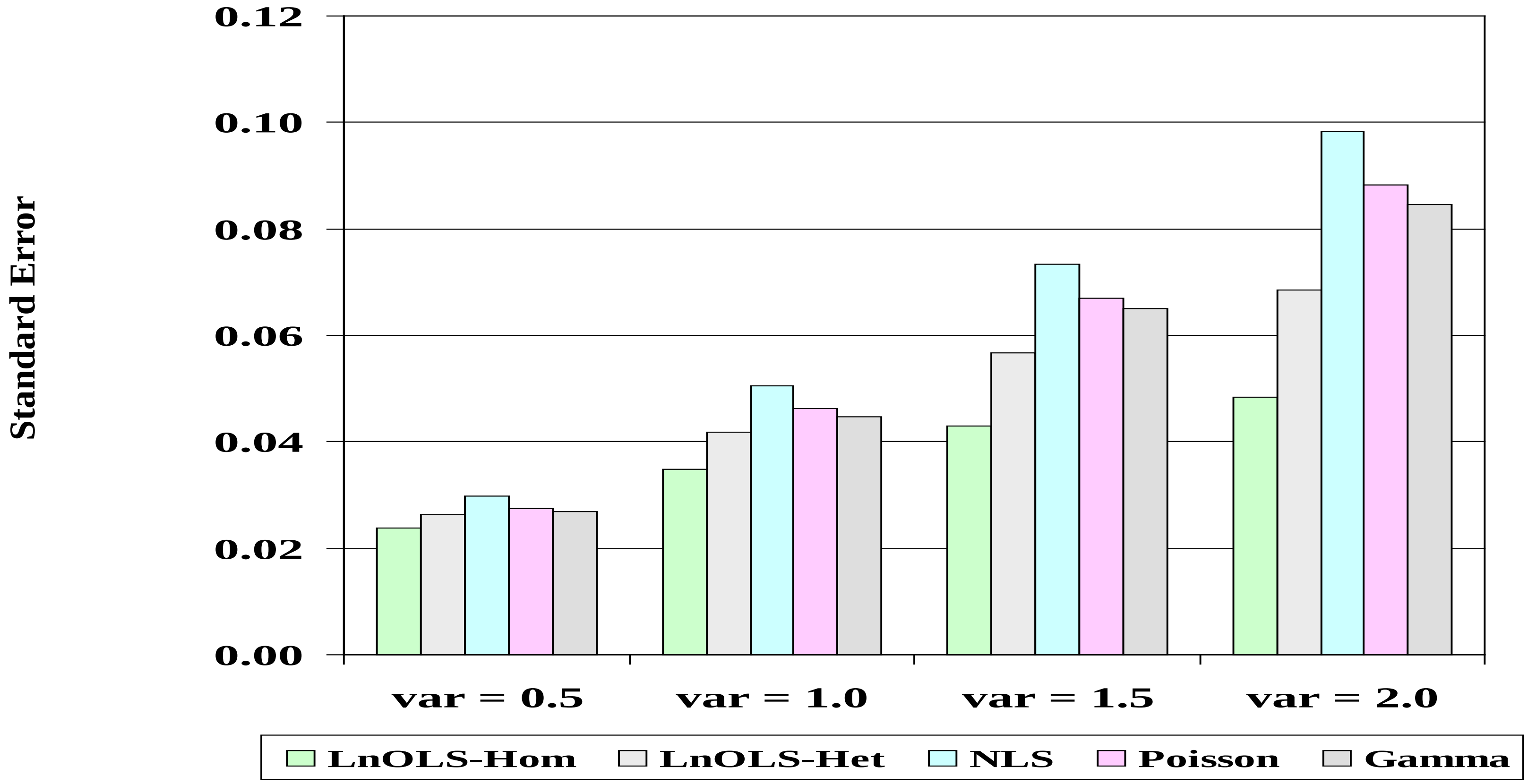


Figure 2
Effect of Heavy Tails on Log Scale

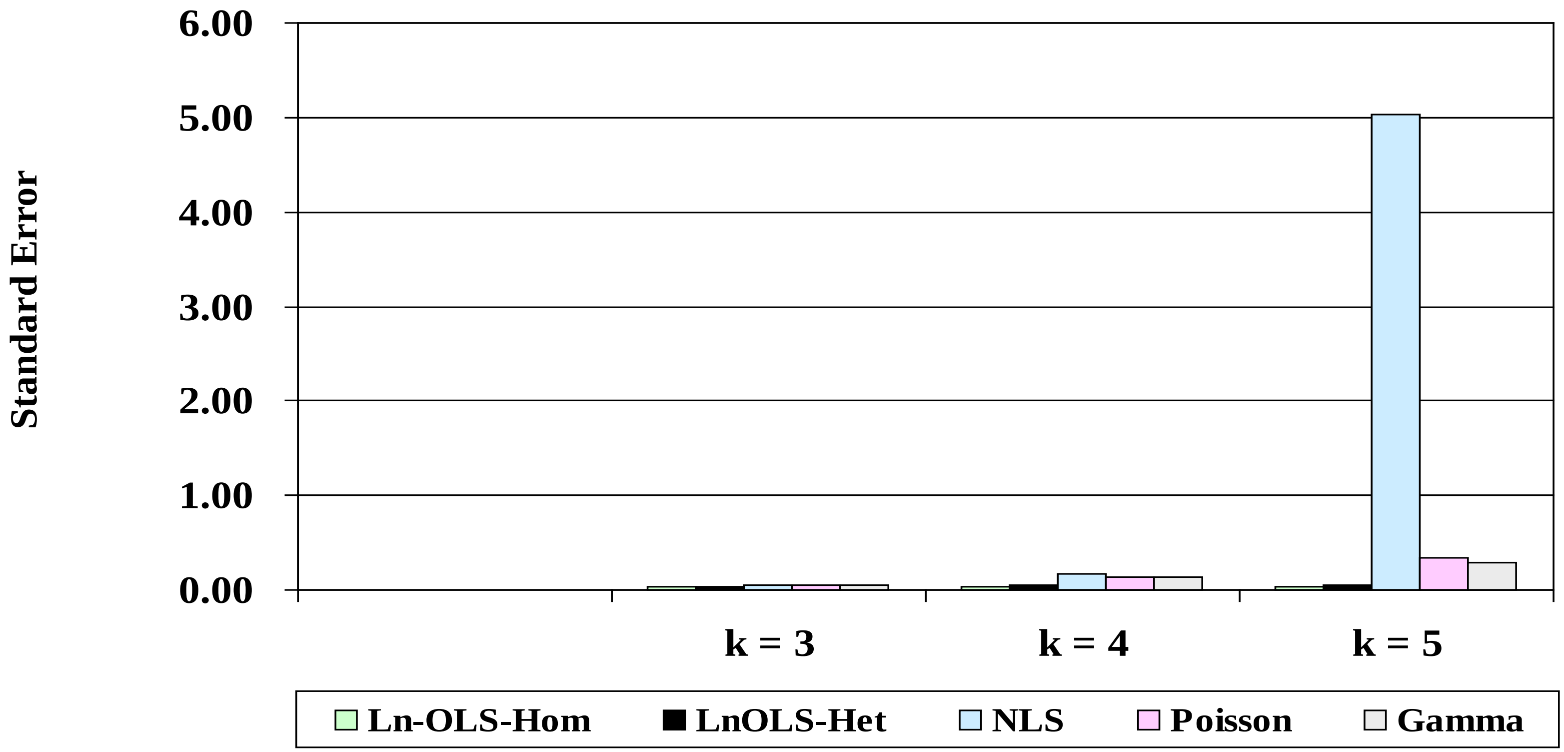


Figure 3
Effect of Shape

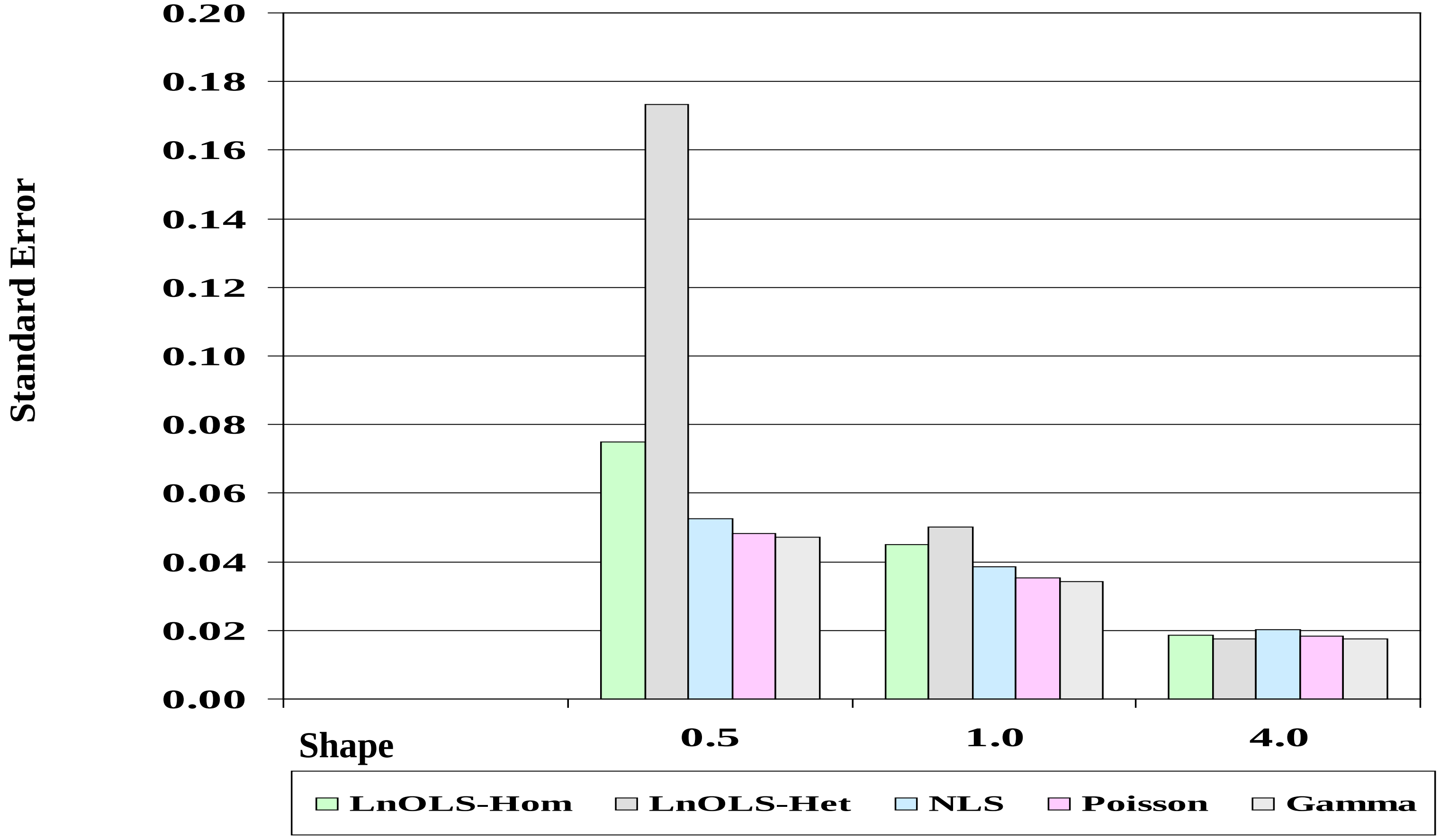
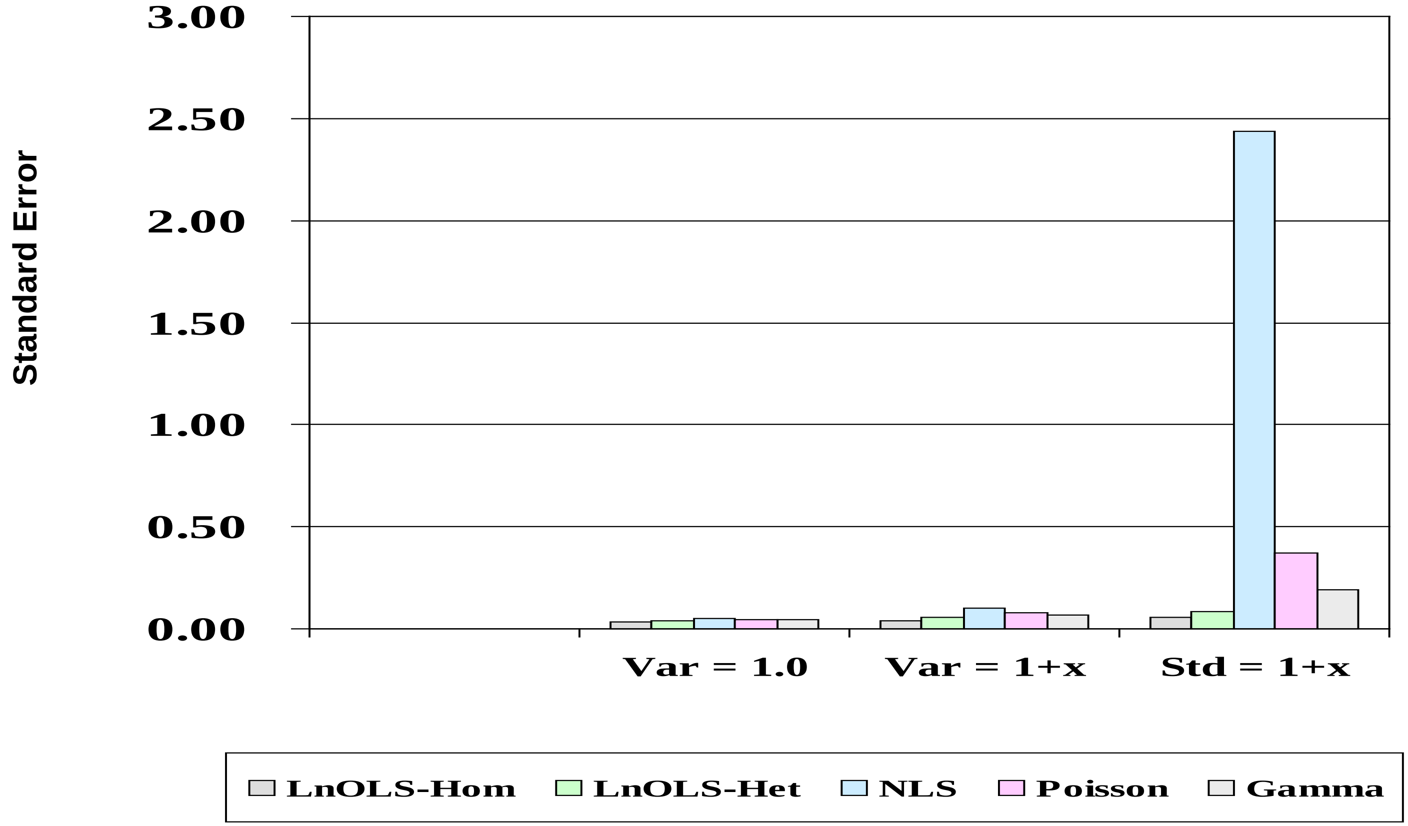


Figure 4
Effect of Heteroscedasticity
on the Log Scale



Summary of Simulation Results

All are consistent, except Log-OLS with homoscedastic retransformation if the log-scale error is actually heteroscedastic

GLM models suffer substantial precision losses in face of heavy-tailed (log) error term. If kurtosis > 3, substantial gains from least squares or robust regression.

Substantial gains in precision from estimator that matches data generating mechanism

Overview

Studies with skewed outcomes but no zero mass problems

Alternative models

Comparing alternative models

Assessing model fit

Interpretation

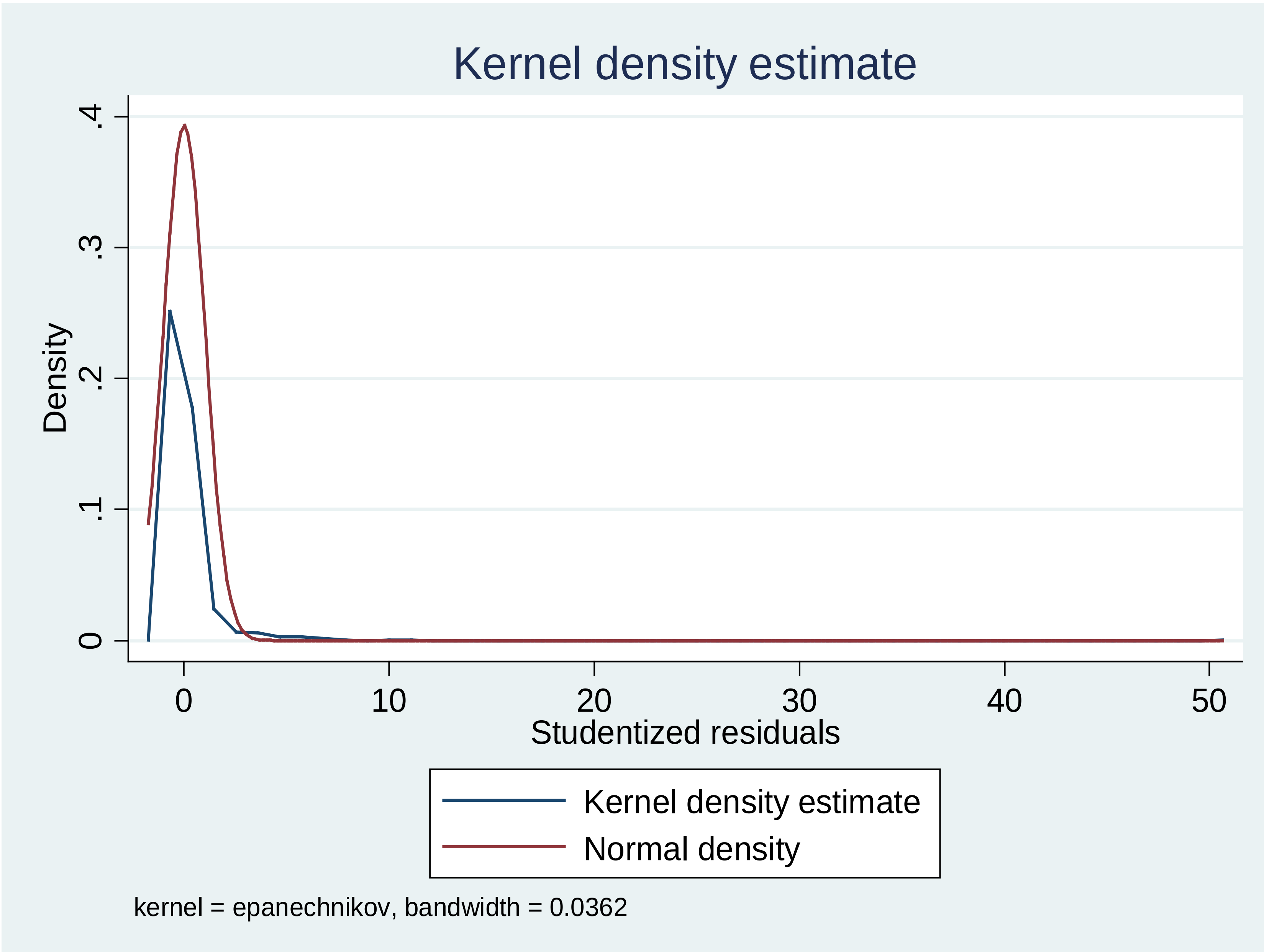
MEPS Data for Examples

Medical Expenditure Panel Survey (MEPS) data
- **Representative of non-institutionalized US population**
- **Subsample of NHIS**
- **Available to public**

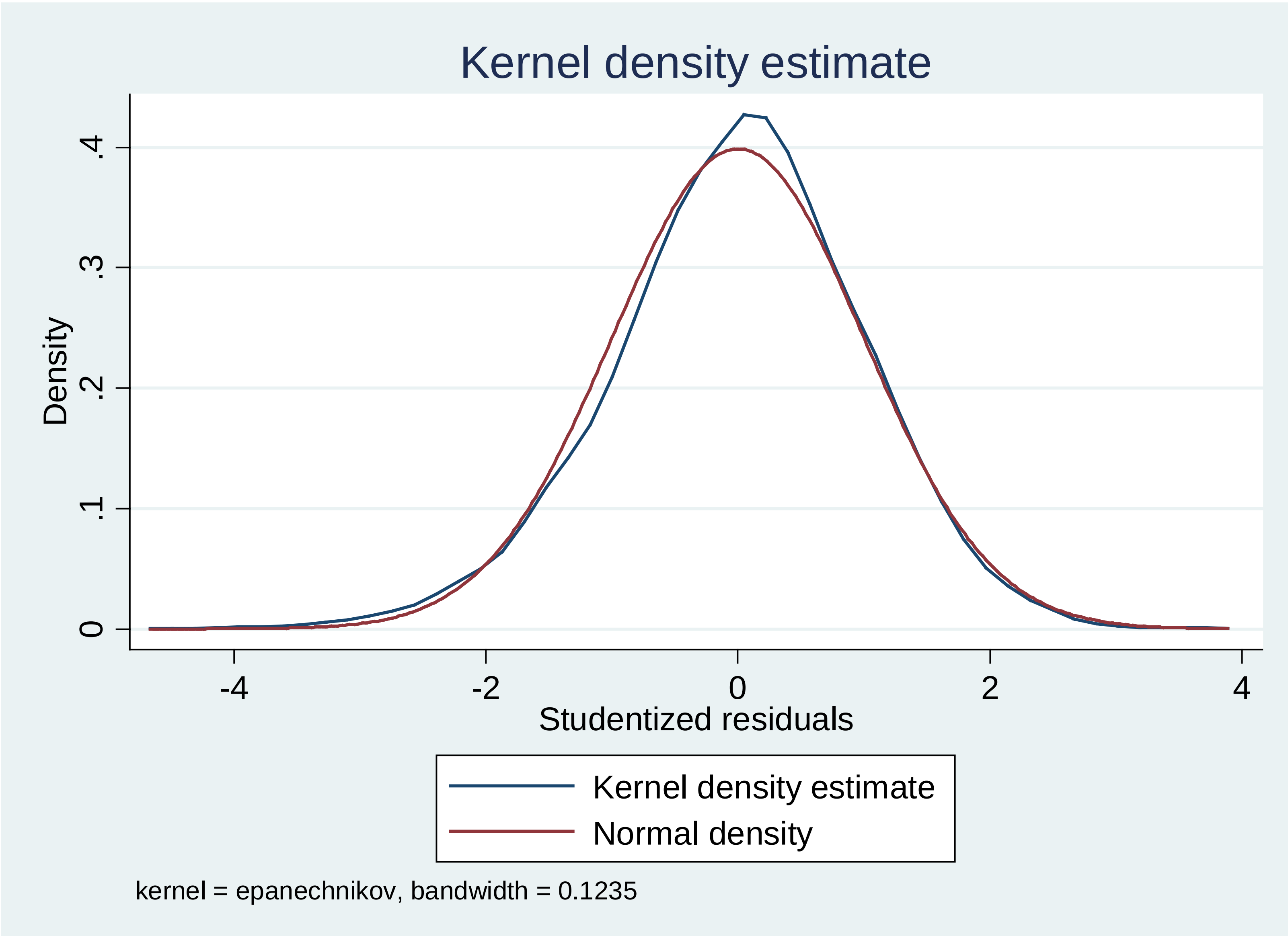
Information on
- **Health expenditures and utilization**
- **Health status**
- **Insurance**
- **Demographics, income, education, family**

MEPS sample for these examples
- **Observations at person-year level, N = 19,386**
- **Adults (ages 18+) without missing data**
- **Year = 2004**

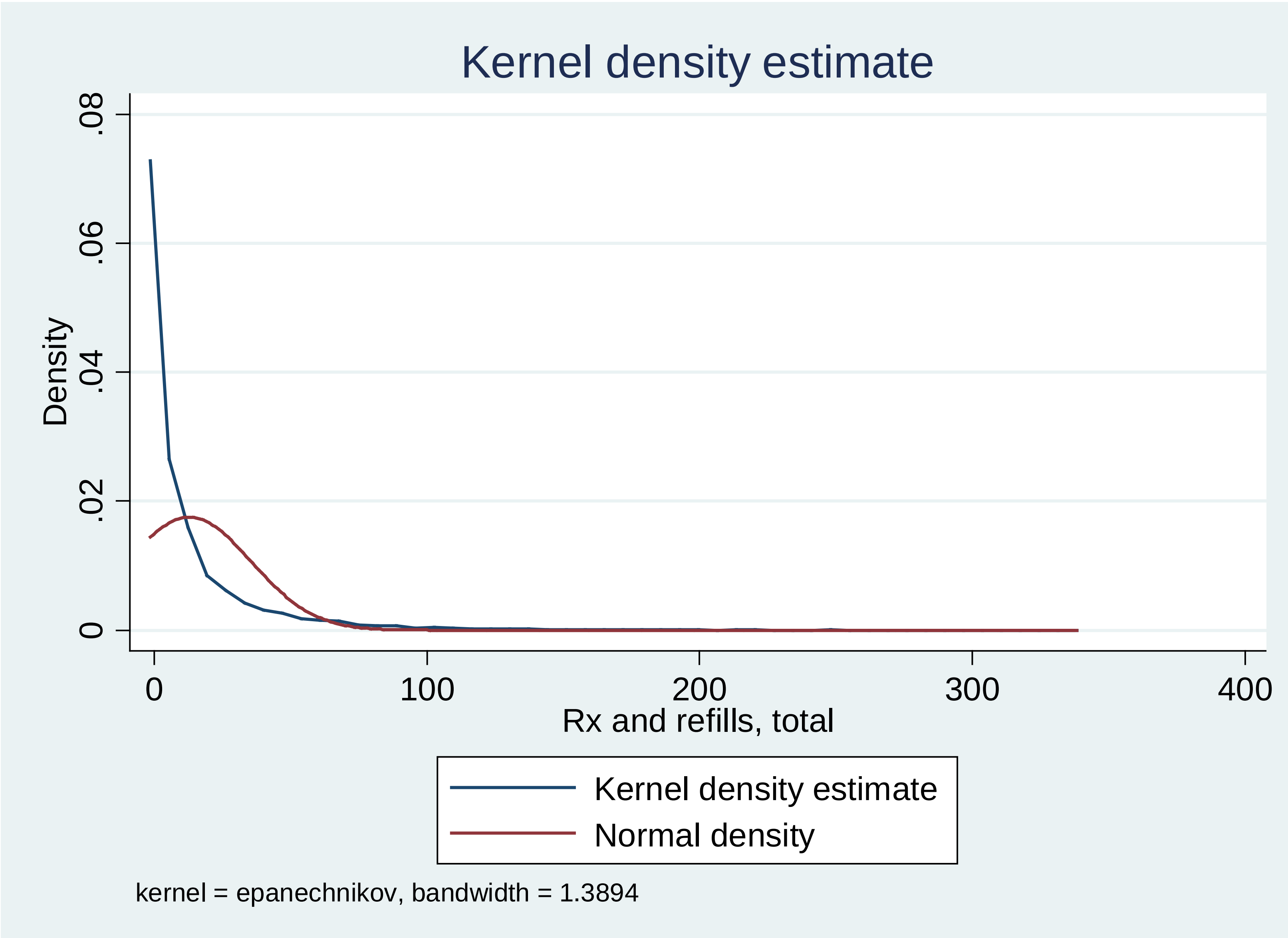
Density for OLS Studentized Residuals Total Medical



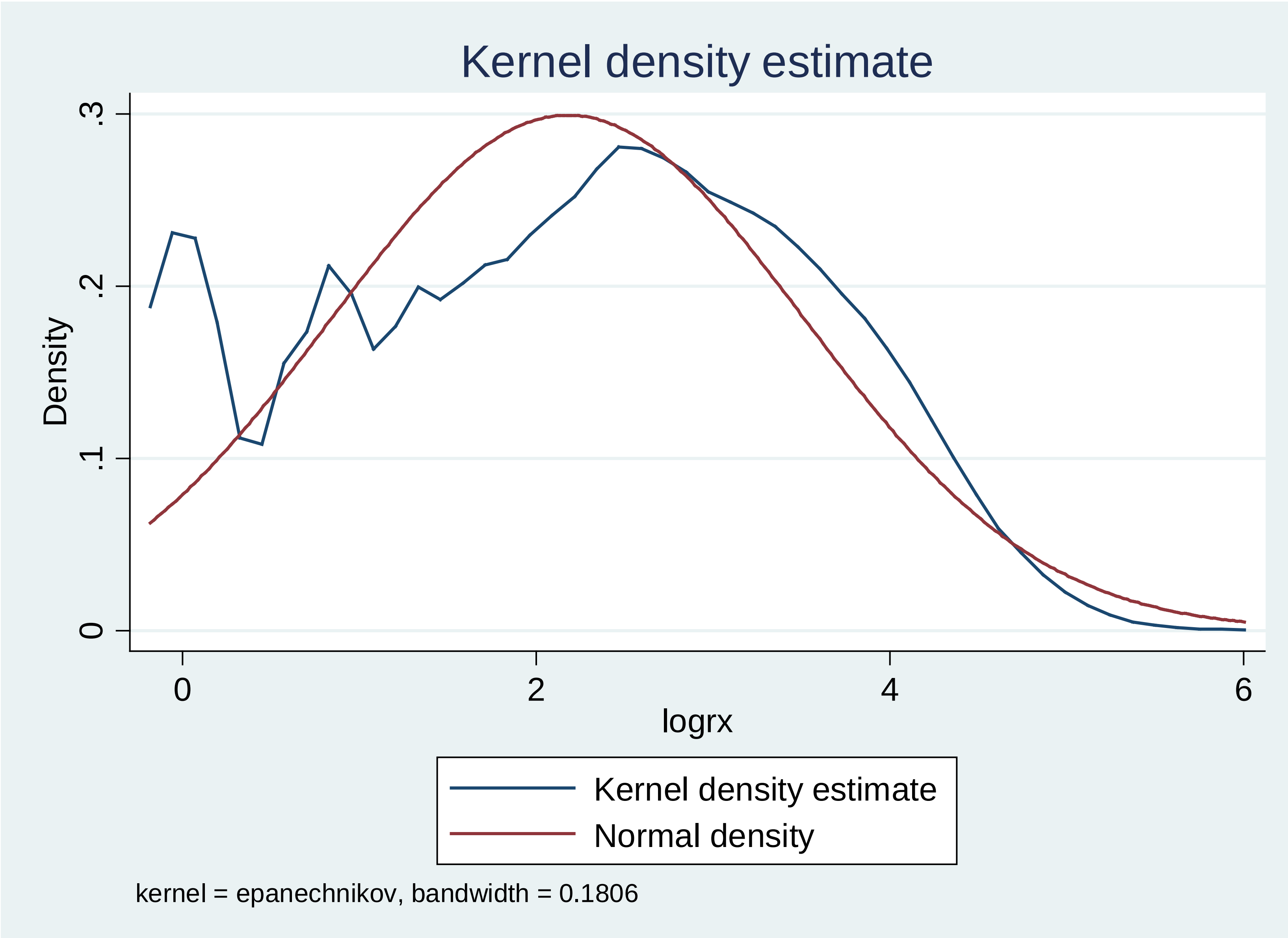
Density for OLS Studentized Residuals, Log Scale Total with $ > 0



FYI: Density for Number of Rx fills and refills



FYI: Density for log(# Rx fills | Rx > 0)



Overview

Statistical issues and potential problems

Skewness

Studies with skewed outcomes but no zero mass problems

Model checks

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Overview

Studies with skewed outcomes but no zero mass

- **Assessing model fit**
  - **Picking a model**
    - **Box-Cox test**
    - **GLM family test**
  - **Checking for heteroscedasticity**
  - **Checking model fit**
    - **Pregibon's Link Test and Ramsey's RESET test**
    - **Modified Hosmer-Lemeshow test**
  - **Checking for overfitting**
    - **Copas style tests**

Model Checks

Primary Concern

Systematic bias as a function of covariates *x*

Secondary Concern

Efficiency

Tertiary Concern

Ease of use

Tests can be modified for most models considered here

Most are easily implemented in Stata 12

2004 MEPS Data Examples

Dependent variables		if y > 0		Pos %
Variable	Mean	Mean	Std Dev	%
Total medical $	3386	4480	10604	82.2
Total dental $	211	566	978	37.3
# Prescriptions	13	19	25	66.6

Box-Cox Test

Purpose

To determine relationship between $x\beta$ and E(y|x)

Box-Cox test

Find MLE value of λ $$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda}$$

Stata command `boxcox y $x if y > 0`

Conclude

If $\hat{\lambda} = -1$ inverse $\Rightarrow$ $(1/y) = X\beta + \varepsilon$

If $\hat{\lambda} = 0$ ln(y) $\Rightarrow$ $\ln(y) = X\beta + \varepsilon$

If $\hat{\lambda} = .5$ square root $\Rightarrow$ $\sqrt{y} = X\beta + \varepsilon$

If $\hat{\lambda} = 1$ linear $\Rightarrow$ $y = X\beta + \varepsilon$

If $\hat{\lambda} = 2$ square $\Rightarrow$ $y^2 = X\beta + \varepsilon$

(if skewed left)

Box-Cox Test
Examples

Variable	$\hat{\lambda}$ (std)	Conclusion
Total medical	.0519 (.0041)	Close to log**
Total dental	-.1071 (.0080)	Close to log*
# Prescriptions	.0340 (0064)	Close to log * *

*** but significantly different from zero (log) at p < 0 .10**

*** but significantly different from zero (log) at p < 0 .05**

Note: λ is called *theta* in Stata for some versions of the test (LHS only)

Checking for Heteroscedasticity

Concern is retransformation bias under log(y) or Box-Cox transformation

Use one of standard tests for heteroscedasticity on log-scale

Breusch-Pagan-Godfrey-White test

Park test – GLM version

Consider $\ln(y_i) = x_i{}'\beta + \varepsilon_i$

Use least square residuals on log- scale to create

$$\text{logvar}_i = \left(\ln(y_i) - x_i{}'\hat{\beta}\right)^2$$

Estimate response logvar to x's by GLM (gamma, log link)

```
glm logvar $x,family(gamma)link(log)robust
    test $x
```

Or use alternative test for heteroscedasticity

Total Medical Expenditures, if Positive
MEPs 2004, Adults

Significantly heteroscedastic in

- o **Decreasing variance in age ($p < 0.001$) but being female (NS)**
- o **Higher variance for blacks ($p = 0.010$)**
- o **Complex variance in income and education ($p < 0.001$)**

- **Increasing variance for uninsured ($p = 0.006$)**
- **Not significant in health status / functioning**

Complex heteroscedasticity probably rules out OLS on log(total medical expenditures) in favor of GLM

- **Studentized residuals too skewed and heavy-tailed for normal theory model→ bias in retransformation**
- **Group-wise smearing will have major precision losses**

GLM Family Test

Purpose

Determine relationship between raw-scale mean and variance functions, E(y|x), and Var(y|x)

Use a GLM family test that is modified Park test with GLM

```
glm  y $x, family(gamma)link(log)
     predict xbetahat, xb
gen  rawresid = y – exp(xbetahat)
gen  rawvar = rawresid^2
glm  rawvar xbetahat, f(gamma)link(log)
```

coefficient on `xbetahat` indicates distribution

Stata: see iHEA2013_ *sample_programs.zip*

FYI: GLM Family Test (alternative)

OLS alternative

1. **Regress y (raw scale) on x, predict $\hat{y}$**
2. **Save raw-scale residuals $\hat{r}$ = y-$\hat{y}$**
3. **Regress $\ln(\hat{r}^2)$ on $\ln(\hat{y})$ and a constant**

Because the use of log transform of residual squared raises a retransformation bias issue,

the GLM version is preferred over the OLS version of the Family Test

GLM Family Test (cont'd)

Coefficient on xbetahat = $\ln(\hat{y})$ gives the family

- **If $\hat{\gamma} = 0$ Gaussian NLLS (variance unrelated to mean)**
- **If $\hat{\gamma} = 1$ Poisson (variance equals mean)**
- **If $\hat{\gamma} = 2$ Gamma (variance exceeds mean)**
- **If $\hat{\gamma} = 3$ Wald or inverse Gaussian**

Variable	$\hat{\gamma}$	Std. Error	Conclusion
Total medical	0.9065	0.4593	Poisson**
Total dental	1.1611	0.5879	Gamma
# Prescriptions	1.2605	0.1321	Either Gamma or Poisson*

***For total medical, # Prescriptions, Gamma is consistent, but not efficient. Issues with inference in two-step process.**

**** Results sensitive to right hand side specification. Fuller specification suggests gamma, but also rejects it.**

FYI Sample Family test code

```
use  meps_ashe_subset5.dta
drop if exp_tot== 0 | exp_tot ==.
quietly {
    glm exp_tot age female, link(log) family(gamma)
      predict double rawyhat, mu
      predict double xbeta1,   xb
    generate  double rawvar = (exp_tot - rawyhat)^2
    generate  double xbeta2 = xbeta1^2
}
** family test
glm rawvar xbeta1, link(log) family(gamma) nolog robust
     test xbeta1 - 0 = 0   /* NLLS or Gaussian family */
     test xbeta1 - 1 = 0   /* Poisson family */
     test xbeta1 - 2 = 0   /* Gamma family */
     test xbeta1 - 3 = 0   /* Inverse Gaussian family */
** check fit for family test using Pregibon's Link Test
glm rawvar xbeta1 xbeta2, link(log) family(gamma) nolog robust
```

FYI Sample Family test results

```
** family test
. glm rawvar xbeta1, link(log) family(gamma) nolog robust
------------------------------------------------------------------------------
             |               Robust
      rawvar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      xbeta1 |   .9065369   .4593718     1.97   0.048     .0061848    1.806889
       _cons |   10.87347   3.992582     2.72   0.006     3.048155    18.69879
------------------------------------------------------------------------------
.       test xbeta1 - 0 = 0   /* NLLS or Gaussian family */
 ( 1)  [rawvar]xbeta1 = 0
           chi2(  1) =    3.89
         Prob > chi2 =    0.0484

.       test xbeta1 - 1 = 0   /* Poisson family */
 ( 1)  [rawvar]xbeta1 = 1
            chi2(  1) =    0.04
         Prob > chi2 =    0.8388

.       test xbeta1 - 2 = 0   /* Gamma family */

 ( 1)  [rawvar]xbeta1 = 2
            chi2(  1) =    5.67
         Prob > chi2 =    0.0173

.       test xbeta1 - 3 = 0   /* Inverse Gaussian family */
( 1)  [rawvar]xbeta1 = 3
            chi2(  1) =   20.77
         Prob > chi2 =    0.0000
```

FYI Sample Family test results (cont'd)

```
. ** check fit for variance function via family test using Pregibon's Link Test
. glm rawvar xbeta1 xbeta2, link(log) family(gamma) nolog robust
------------------------------------------------------------------------------
             |               Robust
      rawvar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      xbeta1 |  -2.137975   10.63596    -0.20   0.841    -22.98407    18.70812
      xbeta2 |   .1821556   .6136742     0.30   0.767    -1.020624    1.384935
       _cons |   23.54713   45.98447     0.51   0.609    -66.58078     113.675
------------------------------------------------------------------------------
```

Can reject NLLS/Gaussian, Gamma and Inverse Gaussian

Cannot reject Poisson family

Model does not fail Link Test

Results have not been corrected for two-step approach, `xbeta1` treated as fixed.

- **Bootstrap whole?**
  - **Edward will discuss this later in two-part models**
- **I will also discuss EEE extension to GLM *next***

Results sensitive to right hand side specification

EEE extension to GLM

May need more flexible GLM setup with richer set of link and family to avoid bias if incorrect link or loss of efficiency if incorrect family

Estimate link and variance power functions to estimate λ, θ's, and β's jointly

$$E(y_i \mid x_i) = \mu_i = g^{-1}(x_i'\beta)$$

$$g(\mu_i) = (\mu_i^{\lambda} - 1) / \lambda$$

$$V(y_i) = \theta_1 (\mu_i)^{\theta_2}$$

EEE extension to GLM (cont'd)

As $\lambda \to 0$, we have log link and ECM (exponential conditional mean) model

Allows for $\lambda \neq 0$ (non-log) models and $\theta_2 \neq$ integer

More efficient than choosing link and family separately

Avoids need to correct Family Test for two-stage process

Avoids bias from wrong link

Basu and Rathouz's extended estimating equation or GLM approach (*Biostatistics*, 2005).

See code and discussion in Basu paper in *The STATA Journal* 5(4). This helps to avoid major numerical issues

Install pglm from Basu website at: http://faculty.washington.edu/basua/index.html

FYI: Sample code for EEE

```
use meps_ashe_subset5
drop if tot_exp == 0 | totexp == .
** renormalize to reduce numerical problems
summarize tot_exp
generate double newraw = exp_tot/(r(mean))
** simpler specification
pglm newraw age female, vf(q)
     * tests for link
     test [lambda]_cons = 0
     test [lambda]_cons - 1 = 0
     * tests for variance functions
     test [theta2]_cons - 0 = 0
     test [theta2]_cons - 1 = 0
     test [theta2]_cons - 2 = 0
     test [theta2]_cons - 3 = 0
     * test for log link and gamma family
     test [lambda]_cons = 0
     test [theta2]_cons - 2 = 0, accum
```

FYI: simple EEE for positive total medical expenditure,MEPS 2004 adult sample

exp_tot	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
exp_tot						
age	.0274713	.0010639	25.82	0.000	.025386	.0295566
female	.1555225	.0404298	3.85	0.000	.0762815	.2347634
_cons	-1.475407	.0669523	-22.04	0.000	-1.606631	-1.344183
lambda						
_cons	.3325981	.131289	2.53	0.011	.0752764	.5899198
theta1						
_cons	5.562189	3.452875	1.61	0.107	-1.205322	12.3297
theta2						
_cons	-.1643946	2.315348	-0.07	0.943	-4.702394	4.373605

Example: EEE estimates for total medical expenditures if positive

Parameter	*coef.*	*Std. err.*	*z*	*p-value*
λ	0.3326	0.1333	2.53	0.011
θ_2	-1.1643	2.3153	-0.07	0.943

Reject log link ($\lambda = 0$) at $p = 0.0113$
Reject identity link (($\lambda = 1$) at $p < 0.001$

Cannot reject Gaussian (NLLS) ($\theta_2 = 0$)
Cannot reject Poisson family ($\theta_2 = 1$)
Cannot reject Gamma family ($\theta_2 = 2$)
Cannot reject inverse gaussian ($\theta 2 = 3$)

But can reject log link and gamma family at $p = 0.0089$

EEE Summary

Reminder to be careful about correcting inferences in Family Test for two-step approach. In this case, the inferences for conclusions could be dramatically different

Conclusions about the link function do not change qualitatively with richer age-gender specifications or with added other covariates.

- **Log link still rejected**
- **Variance function estimate of $\hat{\delta}$ still imprecise**

Assessing the Model Fit for Linearity

Pregibon's Link Test (scale of estimation)

Ramsey's RESET Test (scale of estimation)

Modified Hosmer-Lemeshow (on scale of estimation or scale of interest)

Link and RESET Tests

Purpose

To determine linearity of response on scale of estimation

These tests work for any model (e.g., OLS, logit, probit)

Pregibon's Link test for OLS

$$y = \delta_0 + \delta_1 (x'\hat{\beta}) + \delta_2 (x'\hat{\beta})^2 + \nu$$

Test $\hat{\delta}_2 = 0$

Stata: `linktest`

Ramsey's RESET test (one version, as implemented in Stata)

$$y = \delta_0 + \delta_1 (x'\hat{\beta}) + \delta_2 (x'\hat{\beta})^2 + \delta_3 (x'\hat{\beta})^3 + \delta_4 (x'\hat{\beta})^4 + \nu$$

Test $\hat{\delta}_2 = \hat{\delta}_3 = \hat{\delta}_4 = 0$

Stata: `estat ovtest`

Link and RESET Tests (cont'd)

For alternative estimators with linear index: $x'\beta$

Use original estimator with functions of $x'\hat{\beta}$**, and** $(x'\hat{\beta})^2$ **as covariates**

STATA example:

```
logit $depv $indv
      predict xbeta1, xb
gen   xbeta2 = xbeta1^2
logit $depv xbeta1 xbeta2, robust
      test xbeta2
```

Similarly for RESET

Link and RESET Tests (cont'd)

Conclude

These tests are diagnostic, not constructive

If do not reject null, keep model the same

If reject null, there could be problem with functional form or influential outliers for either OLS or GLM

Example for log(y) by OLS version

Variable	*p*-values Link	*p*-values RESET	Conclusion
Total medical	< 0.001	<0.001	Fails Both Tests
Total dental	<0.001	<0.001	Fails Both Tests
# Prescriptions	<0.001	<0.001	Fails Both Tests

Similar conclusions for gamma GLM with log link.

Sensitive to specification of covariates!

Stata: see sample programs (`chklinols.ado` and `chklinglm.ado`)

Link and RESET Tests (cont'd)

Advantages

- **Easy**
- **Omnibus tests**

Disadvantages

- **Incomplete for multipart models**
- **Sensitive to influential outliers, especially RESET**

Modified Hosmer-Lemeshow Test

Purpose

To check fit on scale of interest or raw scale for systematic bias

Modified Hosmer-Lemeshow test

Estimate model (e.g., GLM y or OLS $\ln(y) = x\beta + \varepsilon$)

Retransform to get $\hat{y}$ on raw scale

Compute raw-scale residual $\hat{r} = y - \hat{y}$

Create 10 groups, sorted by specific *x* (or by $x\hat{\beta}$)

***F*-test of whether all 10 mean residuals different from zero**

Look for systematic patterns (e.g., U-shaped pattern)

Stata: see iHEA2013_*sample_programs.zip*

Modified Hosmer-Lemeshow Test (cont'd)

Conclude

This test is also non-constructive

No problem if there is no systematic pattern

If reject null, there could be problem with

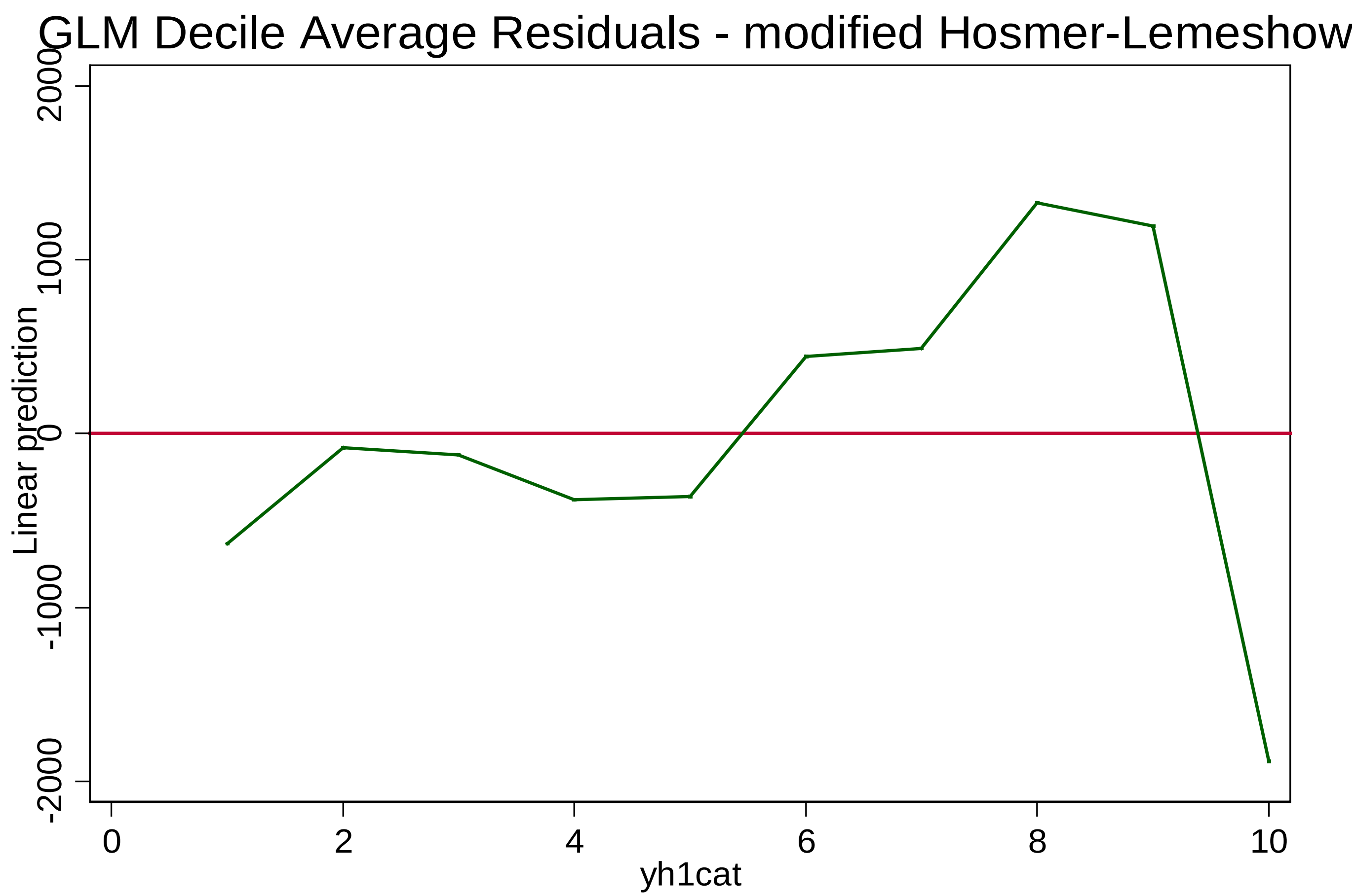
- **left side (wrong power or link function)**
- **right side (wrong functional form of *x's*)**
- **or both**

Example for gamma with log link

Variable	*p*-value for *F*-test	Conclusion
Total medical	< 0.001	Problem*
Total dental	< 0.001	Problem*
# Prescriptions	< 0.001	Problem*

**** Problem partially in age-gender specification***

Modified Hosmer-Lemeshow Estimates by Deciles of Prediction for Total Medical Expenditures



Modified Hosmer-Lemeshow Test (cont'd)

Test linearity fit for power function matters for total medical expenditures `(exp_tot)`, positive cases only.

Use modified Hosmer-Lemeshow, Pregibon's Link, and Ramsey's RESET tests on estimation scale. OLS example.

Variable	Power Function	Hosmer-Lem. *F*-test	Pregibon Link Test F	Ramsey RESET F
Totexp	1.0	4.17	19.48	11.83
Totexp^0.5	0.5	4.28	9.34	13.35
Log(Totexp)	0.0	3.82	6.03	9.69
(1/Totexp)^0.5	-0.5	4.86	38.363	14.83

All significant at $p < 0.001$, except ln(y) Link test at $p = 0.01$

Modified Hosmer-Lemeshow Test (cont'd)

All models considered fail specification tests, except log transform using Link Test

For all tests best fit for transforms considered is log model

Remaining specification failure is largely due to age-gender specification, not fine-tuning transformation

Modified Hosmer-Lemeshow Test (cont'd)

Advantages

Works on scale of ultimate interest, as well as on scale of estimation; can choose scale concerned about

Works for any model (including logit, probit, 2-part, NB)

Can detect problems missed by omnibus Link and RESET tests, because can look at fit for key covariates

Disadvantage

Lacks power

Individual coefficients sensitive to influential observations if done on scale of interest (raw scale)

FYI: Sample Stata code

Stata 12 code can be found in

iHEA2013_sample_programs.zip

Code for linearity tests for OLS or transformed y with outlier diagnostics: `chklinols.ado`

Code for linearity tests for GLM: `chklinglm.ado`

Code for *.ado, *.sthlp, with test programs in
`Chklinpgm  in iHEA2013_sample_programs.zip`

Overfitting Tests

Overfitting can be a problem

Tailoring the model to the specific data set

But at the expense of explaining other similar data sets

Overemphasis on explaining a few outliers when data are very skewed or cases have leverage. For OLS,

but some ε's are extremely large as well as x's extreme

Combined risk of influential outlier

Overfitting is often a major problem for expenditure data esp. for small to moderate sample sizes or rare covariates

Maximizing R-squared leads to overfitting

FYI: Copas Style Tests (cont'd)

Purpose

- **To test for over-fitting and misspecification using split sample cross validation**

Copas test (original version of it)

- **Randomly split sample into two equal groups A and B**
- **Estimate model on sample A, retain coefficients**
- **Forecast to sample B**

$$y_i^A = \left(x_i^A\right)' \beta + \varepsilon_i^A$$

$$\hat{\beta}^A = \left(X'_A X_A\right)^{-1}\left(X'_A Y_A\right)$$

$$y_i^B = \delta_0 + \delta_1\left(\left(x_i^B\right)' \hat{\beta}^A\right) + \varepsilon_i^B$$

$$= \delta_0 + \delta_1\left(\hat{y}_i^{AB}\right) + \varepsilon_i^B$$

FYI: Copas Style Tests (cont'd)

If there is no overfitting, we expect $E(\hat{\delta}_1) = 1$

Test $\hat{\delta}_1 = 1$

- **But Expect $\hat{\delta}_1 < 1$ due to sampling variance in $\hat{\delta}_1$ or overfitting**
- **Distance $1 - \hat{\delta}_1$ is measure of overfitting in large samples**

If using GLM, both sides of cross-validation done by GLM with same link and distribution

Generally same scale of estimation or estimation approach is used for both splits for the original Copas style tests

FYI: Copas Style Tests (cont'd)

Copas test (common health econometric use)

- **Not interested in scale of estimation per se. It solves a statistical issue**
- **Interest is in scale-of-interest or raw-scale behavior($ or €)**
- **Difference from standard Copas is that B sample estimation conducted on scale of interest or raw scale ($ or €)**
- **See Veazie et al (2003) or Basu et al (2006)**
- **Interpretation and expectations for $\hat{\delta} = 1$ and $1 - \hat{\delta}$ are still the same**

Split Sample Tests

See sample code in ihea2013_sample_programs under either Manning_programs for Copas style or Deb_programs for 10-fold (K-fold) splits

Conclusions:

1. **If $\hat{\delta}$ significantly different from one, consider outliers or pruning model in terms of covariates or more parsimouious specification**
2. **If $\hat{\delta}$ quite imprecise, consider more efficient or robust methods**
3. **Also consider other methods for split sampling which may be stronger tests (more precise) than 50-50 splits. Rich literature in Statistics**
4. **Results depend on sample size and complexity of the specification**

Summary of MEPS modeling
Positive Total Medical Expenditures

Standard OLS log($) subject to complex heteroscedasticity

- **Error is not normally distributed**
- **Normal theory models will be biased on retransformation by failure of normality**
- **Potential bias for estimates of impact of x on E($ | x, $ > 0) due to heteroscedasticity**

Log transform overcorrects in Box-Cox family and log link is not optimal for GLM

- **Potential for bias in either case**

Summary of MEPS modeling (cont'd)

Evidence on GLM Family is mixed and depends on test and specification

- **Distribution for GLM is neither Identity link nor Inverse Gamma family**
- **Evidence mixed on Poisson vs. Gamma**
- **Efficiency gains from using EEE or iteratively reweighted least squares**

Simple age and gender specification is inadequate

- **Over-predicts most expensive group – the elderly**
- **Needs more complex age function interacted with gender**

Summary of MEPS modeling (cont'd)

GLM (log link, gamma) more precise than OLS on raw dollars

Log link too severe to achieve linearity

- **Specific solution depends on specification of covariates**
- **All have $\hat{\lambda} > 0$ in MEPS 2004**

OLS more susceptible to influential outliers

- **Here issue is expensive cases with any health limitation**
- **Important but uncommon subgroup**

Overview

Statistical issues - skewness and the zero mass

Studies with skewed outcomes but no zeroes

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Top Ten Urban Myths of Health Econometrics

Overview

Studies with skewed outcomes but no zero mass problem

- **Alternative models**
- **Comparing alternative models**
- **Assessing model fit**
- **Interpretation**

Overview

Studies with skewed outcomes but no zero mass problem

Interpretation

Marginal and incremental effects
OLS
GLM with Log Link
Four Models for ln(*y*)
Square Root

Single-Equation Models for $y > 0$

Interpretation

- $\hat{y}$
- **Marginal and incremental effects**

Models

- **OLS**
- **GLM with log link**
- **ln(*y*): four versions depending on error assumptions**
  - **normal or non-normal; homo- or heteroskedastic**
- **Square root of *y*, as example of Box-Cox**

Marginal and Incremental Effects (1)

Compare seven different single-equation models

Use the same MEPS 2004 data

Compute

- $\hat{y}$
- **Marginal and incremental effects**
- **Include interaction between age and female**

Show formulas for general models
Show basic Stata code
Compare results across models

Marginal and Incremental Effects (2)

Marginal effects

For continuous variables

Take partial derivative

Incremental effects

For dummy variables

Also for discrete change in continuous variables

Take discrete difference

Marginal and Incremental Effects (3)

Marginal effects in linear models (OLS) are easy

Marginal effects in nonlinear models are more complicated
Several ways to compute them

- **For full sample**
  - **Recycled or standardized predictions**
  - **Average-of-the-probabilities approach**
- **For a single, typical observation**
- **Can change value for subsample or whole sample**
- **Compute treatment effect**
  - **For the treated, the untreated, or standardized pop.**

The appropriate method depends on the research question

Marginal and Incremental Effects (4)

Stata's `margins` command makes predictions and marginal effects easier

Main points about `margins`

- **Don't let the name fool you, not just marginal effects**
- **Computes predicted values and probabilities**
- **Computes marginal and incremental effects (`i.var`)**
- **Computations for single obs., or averaged, or subsample**
- **Track nonlinearities and interactions if use # notation**
- **Can plot relationships quickly with `marginsplot`**
- **Computes standard errors (delta method)**

Warnings!!!

Just because margins calculates standard errors easily does not mean that they are correct

- **Normal theory may not apply**
- **Delta method standard errors are biased in certain cases**
- **Especially a problem for retransformed models**
- **Margins does not account for all sources of uncertainty**
- **We will show examples and explain why**

Marginal and Incremental Effects (6)

Use proper syntax so Stata knows variable type

Continuous variable: `c.varname`

Incremental variable: `i.varname`

Example: `regress y c.age i.female`

Stata takes complicated 1st derivatives, not simple 2nd derivatives if show relationship between variables using #

Example with interaction: `regress y c.age##i.female`

The Stata manual has extensive examples

Marginal and Incremental Effects (7)

Basic model

$$y = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 age \times female + \varepsilon$$

Focus on 82 percent with positive expenditures (*N* = 15,946)

Mean expenditures (if expenditures > 0) is $4,480
Mean age is 47.4 [range is 18 to 85]
Women are 59 percent of the sample

OLS (1)

Model

$$\boldsymbol{y = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 age \times female + \varepsilon}$$

Interpretation

$$\boldsymbol{\hat{y} = x\hat{\beta}}$$

$$\boldsymbol{\frac{\partial \hat{y}}{\partial age} = \hat{\beta}_1 + \hat{\beta}_3 female}$$

$$\boldsymbol{\frac{\Delta \hat{y}}{\Delta female} = \hat{\beta}_2 + \hat{\beta}_3 age}$$

OLS (2)

regress $y c.age##i.female, vce(robust)

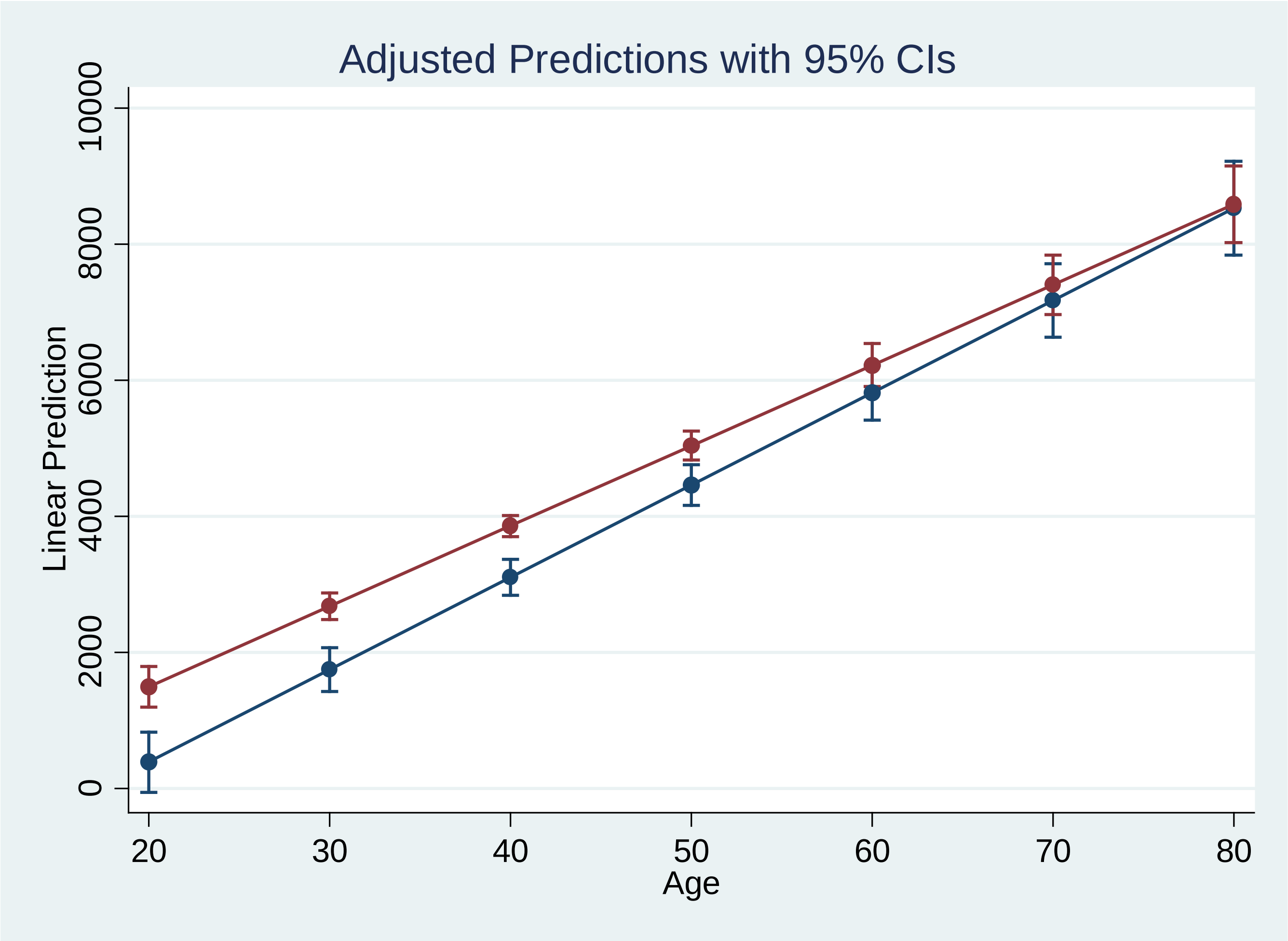
```
. regress $y $x, vce(robust)

Linear regression                                      Number of obs =   15946
                                                       F(  3, 15942) =  191.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0439
                                                       Root MSE      =   10370

------------------------------------------------------------------------------
             |               Robust
     exp_tot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   135.7795   8.457908    16.05   0.000     119.2011     152.358
    1.female |     1456.8   466.2623     3.12   0.002     542.8731    2370.727
             |
female#c.age |
          1  |   -17.5167   10.80951    -1.62   0.105    -38.70457    3.671164
             |
       _cons |  -2330.034   377.0775    -6.18   0.000    -3069.148   -1590.919
------------------------------------------------------------------------------
```

FYI: OLS (3)

```
/* predicted values */
margins                          /* overall mean  */
margins, at((mean) _all)         /* at mean of x  */
margins, at(age=(20(20)80))      /* change age    */
margins female                   /* change sex    */
margins, over(female)            /* subset by sex */

/* marginal effects*/
margins, dydx(age female)
margins, dydx(age) at(female=(0 1))
margins, dydx(age) at(age=(20(10)80) female=(0 1))
```

OLS (4)

```
margins, at(age=(20(10)80) female=(0 1))
marginsplot, legend(off)
```



OLS (5)

Margins: `at() vs. over()`

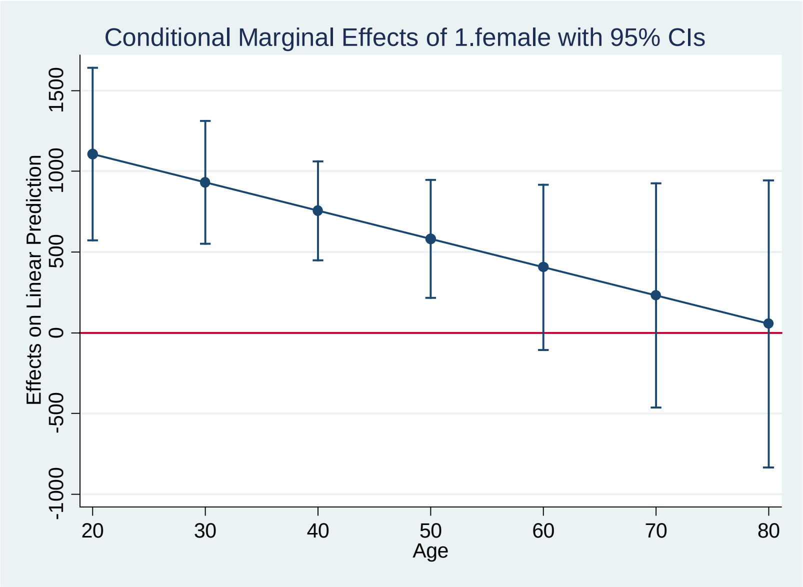
```
. margins female      /* change sex    */

-----------------------------------------------------
             |             Delta-method
             |     Margin   Std. Err.      z    P>|z|
-------------+---------------------------------------
         _at |
          1  |   4108.596   141.8317    28.97   0.000
          2  |   4734.758    98.4763    48.08   0.000
-----------------------------------------------------

. margins, over(female)       /* subset by sex */
-----------------------------------------------------
             |             Delta-method
             |     Margin   Std. Err.      z    P>|z|
-------------+---------------------------------------
      female |
          0  |   4144.966     142.62    29.06   0.000
          1  |   4712.789   97.75823    48.21   0.000
-----------------------------------------------------
```

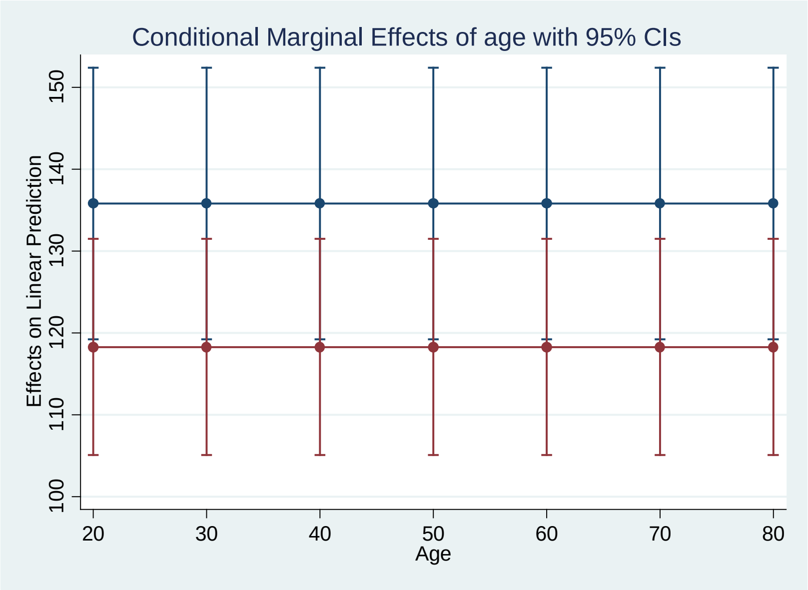
OLS (6)

```
. margins, dydx(age female)   /* marginal effects */
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   125.4361   5.272087    23.79   0.000      115.103    135.7692
    1.female |   626.1626   172.6667     3.63   0.000      287.742    964.5832
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

. margins, dydx(age) at(female=(0 1)) /* me of age by sex */

Average marginal effects                          Number of obs   =      15946
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age

1._at        : female          =           0

2._at        : female          =           1

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   135.7795   8.457908    16.05   0.000     119.2023    152.3567
          2  |   118.2628   6.731224    17.57   0.000     105.0699    131.4558
------------------------------------------------------------------------------
```

OLS (7)

```
margins, dydx(female) at(age=(20(10)80))
marginsplot, legend(off) yline(0)

margins, dydx(age) at(age=(20(10)80) female=(0 1))
marginsplot, legend(off)
```





GLM with Log Link (1)

Model

$$ln[\hat{y}|x] = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 age \times female$$

Interpretation (for GLM models with log link)

$$\hat{y} = exp(x\hat{\beta})$$

$$\frac{\partial \hat{y}}{\partial age} = (\hat{\beta}_1 + \hat{\beta}_3 female) \times \hat{y}$$

$$\frac{\Delta \hat{y}}{\Delta female} = exp(\hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 + \hat{\beta}_3 age) - exp(\hat{\beta}_0 + \hat{\beta}_1 age)$$

GLM with Log Link (2)

```
glm $y $x, link(log) family(gamma) nolog

margins
margins, at((asobserved))
margins, at(age=(65) female=(1))

margins, dydx(age female)
margins, dydx(female) at(age=(20(15)80)
margins, dydx(age) at(female=(0 1))
margins, dydx(age) at(age=(20(10)80) ///
         female=(0 1))

margins, at(age=(20(10)80) female=(0 1))
marginsplot, legend(off)
```

GLM with Log Link (3)

```
. glm $y $x, link(log) family(gamma) nolog

-----------------------------------------------------
             |                 OIM
     exp_tot |      Coef.   Std. Err.      z    P>|z|
-------------+---------------------------------------
         age |   .0345881   .0020504    16.87   0.000
    1.female |   .7164142   .1305455     5.49   0.000
             |
female#c.age |
          1  |  -.0106117   .0025837    -4.11   0.000
             |
       _cons |   6.513084   .1035501    62.90   0.000
-----------------------------------------------------
```

GLM with Log Link (4)

```
. margins                          /* overall mean  */

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   4497.587   110.4282    40.73   0.000     4281.152    4714.022
------------------------------------------------------------------------------

. margins, at((asobserved))   /* at observed x  */

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   4497.587   110.4282    40.73   0.000     4281.152    4714.022
------------------------------------------------------------------------------

. margins, at(age=(65) female=(1)) /* 65-yo woman */

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   6554.849   261.0338    25.11   0.000     6043.232    7066.466
------------------------------------------------------------------------------
```

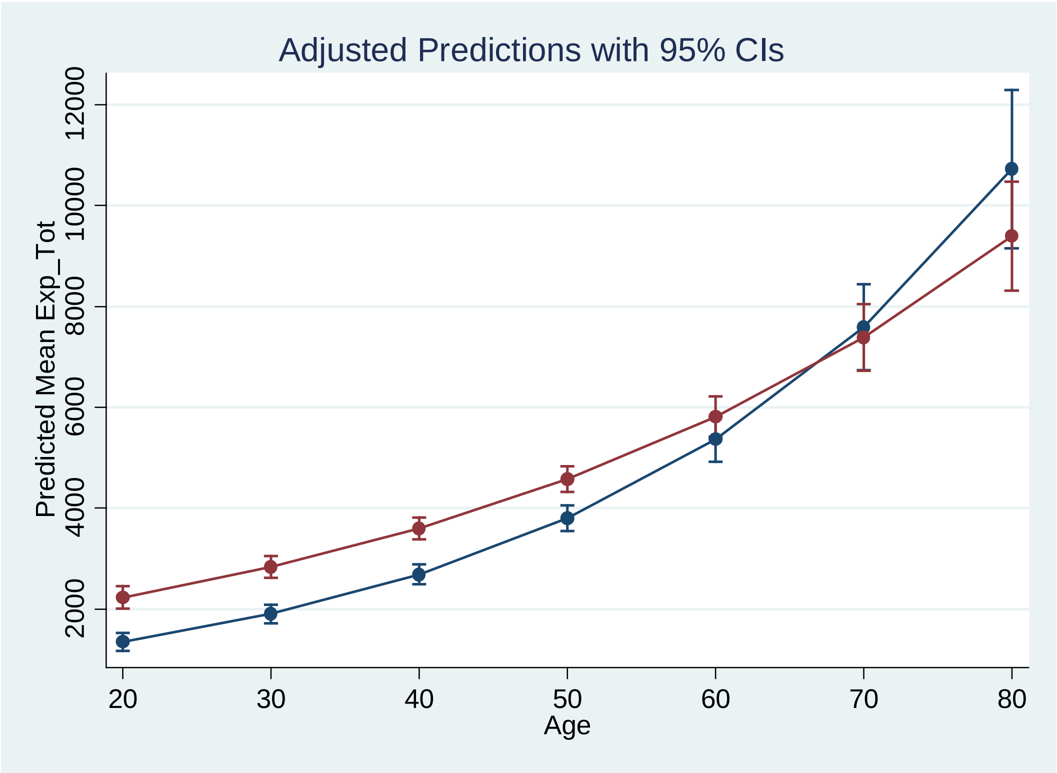
GLM with Log Link (5)

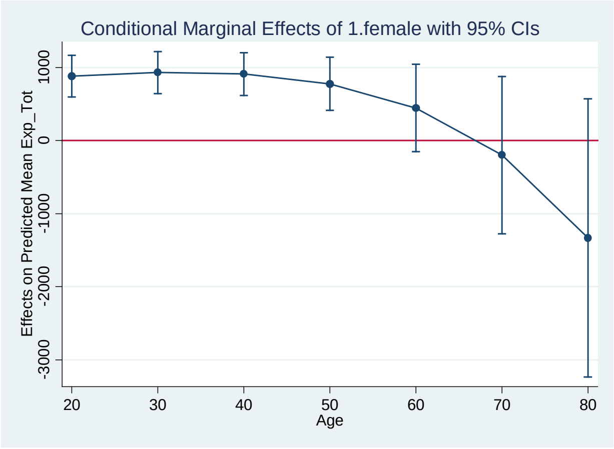
```
. margins, dydx(age female)   /* marginal effects */

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   126.0603   7.591198    16.61   0.000     111.1818    140.9388
    1.female |   507.6144   224.4612     2.26   0.024     67.67859    947.5502
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

. margins, dydx(female) at(age=(20(10)80)) /* me of sex by age*/

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.female     |
         _at |
          1  |   882.4317   145.0748     6.08   0.000     598.0903    1166.773
          2  |   930.0174   145.9933     6.37   0.000     643.8757    1216.159
          3  |   911.3586   149.3602     6.10   0.000      618.618    1204.099
          4  |   775.8044   185.3521     4.19   0.000      412.521    1139.088
          5  |    445.466   304.5039     1.46   0.143    -151.3506    1042.283
          6  |  -197.7472   548.9424    -0.36   0.719    -1273.655    878.1601
          7  |  -1330.912   971.3012    -1.37   0.171    -3234.628    572.8028
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

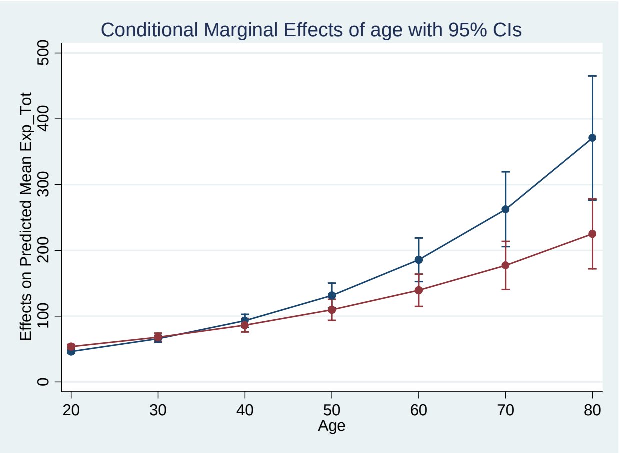
GLM with Log Link (6)

```
. margins, dydx(age) at(female=(0 1))     /* me of age by sex */
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |    145.406    12.8201    11.34   0.000     120.2791     170.533
          2  |   112.9657   9.366857    12.06   0.000     94.60698    131.3244
------------------------------------------------------------------------------

. margins, dydx(age) at(age=(20(10)80) female=(0 1))
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |    46.5534   1.591205    29.26   0.000     43.43469     49.6721
          2  |   53.42813   1.943758    27.49   0.000     49.61843    57.23783
          3  |   65.79085   2.705936    24.31   0.000     60.48731    71.09439
          4  |   67.90439   3.246718    20.91   0.000     61.54094    74.26784
          5  |   92.97788   5.139742    18.09   0.000     82.90417    103.0516
          6  |   86.30298    5.27994    16.35   0.000     75.95449    96.65147
          7  |   131.3995   9.530568    13.79   0.000       112.72    150.0791
          8  |   109.6866    8.27676    13.25   0.000     93.46448    125.9088
          9  |   185.6983   16.92464    10.97   0.000     152.5266      218.87
         10  |    139.406   12.57692    11.08   0.000     114.7557    164.0564
         11  |   262.4352   28.97887     9.06   0.000     205.6376    319.2327
         12  |   177.1779   18.64907     9.50   0.000     140.6264    213.7294
         13  |   370.8824   48.24422     7.69   0.000     276.3254    465.4393
         14  |   225.1839   27.12764     8.30   0.000     172.0147    278.3531
------------------------------------------------------------------------------
```

GLM with Log Link (6)







Four Models for ln(*y*) (1)

Model

$$\boldsymbol{ln(y) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 age \times female + \varepsilon}$$

Four assumptions about the error term

1. **Homoskedastic and Normal**
2. **Homoskedastic and Non-normal**
3. **Heteroskedastic and Normal**
4. **Heteroskedastic and Non-normal**

These assumptions matter when making calculations about dollars on the raw scale, instead of log dollars

FYI

Error Assumptions	Retransform-ation Factor $\mathrm{E}\left[e^{\varepsilon_i}\right]$	Expected Value of y_i $\mathrm{E}\left[y_i \vert x_i\right]$	Population Average $\mathrm{E}\left[y \vert x\right]$
General theory	$\int\limits_{-\infty}^{\infty} e^{\varepsilon_i} d\varepsilon_i$	$e^{x_i'\beta}\mathrm{E}\left[e^{\varepsilon_i}\right]$	$\frac{1}{N}\sum\limits_{i=1}^{N} e^{x_i'\beta}\mathrm{E}\left[e^{\varepsilon_i}\right]$
Homoskedastic & Normal	$e^{.5\sigma^2}$	$e^{x_i'\beta}e^{.5\sigma^2}$	$\left(\frac{1}{N}\sum\limits_{i=1}^{N} e^{x_i'\beta}\right)e^{.5\sigma^2}$
Homoskedastic & Non-normal	$\frac{1}{N}\sum\limits_{i=1}^{N} e^{\hat{\varepsilon}_i}$	$e^{x_i'\beta}D_{smear}$	$\left(\frac{1}{N}\sum\limits_{i=1}^{N} e^{x_i'\beta}\right)D_{smear}$
Heteroskedastic & Normal	$e^{.5\sigma^2(g_i)}$	$e^{x_i'\beta}e^{.5\sigma^2(g_i)}$	$\frac{1}{N}\sum\limits_{g=1}^{G}\sum\limits_{i=1}^{N_g} e^{x_i'\beta}e^{.5\sigma^2(g_i)}$
Heteroskedastic & Non-normal	$\frac{1}{N_A}\sum\limits_{i\in g^A} e^{\hat{\varepsilon}_i}$	$e^{x_i'\beta}D^A_{smear}$	$\frac{1}{N}\sum\limits_{g=1}^{G}\sum\limits_{i=1}^{N_g} e^{x_i'\beta}D^g_{smear}$

FYI: Duan's Smearing Estimator

Corrects for non-normality in log models
Does *not* directly correct for heteroscedasticity
Can be adapted for heteroskedasticity by subgroup

Stata code

```
regress lny $x
predict double resid, residual
egen Dsmear = mean(exp(resid))
```

The smearing factor is always greater than 1.0, and is typically less than 4.0.

Four Models for ln(*y*) (2)

All based off same regression (only differ in retransformation)

```
. generate ln_y = ln(exp_tot)

. regress ln_y $x

------------------------------------------------------------------------------
        ln_y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0436052   .0010853    40.18   0.000     .0414779    .0457324
    1.female |   .9564238   .0702239    13.62   0.000     .8187771     1.09407
             |
female#c.age |
          1  |  -.0127392   .0013876    -9.18   0.000     -.015459   -.0100194
             |
       _cons |   4.957389   .0549901    90.15   0.000     4.849602    5.065176
------------------------------------------------------------------------------
```

Four Models for ln(*y*) (3)

```
* Generate error retransformation terms
predict double ehat, residual
predict double h_ii, hat
generate double etilde = ehat/sqrt(1 - h_ii)
```

Homoskedastic retransformation factors

```
generate Normal_hom = exp(.5*e(rmse)^2)
egen     Dsmear_hom = mean(exp(ehat))
```

Heteroskedastic retransformation factors (if het. by gender)

```
sort female
by female: egen s2_hetN = mean(etilde^2)
generate        Normal_het = exp(.5*s2_hetN)
by female: egen Dsmear_het = mean(exp(etilde))
```

Four Models for ln(*y*) (4)

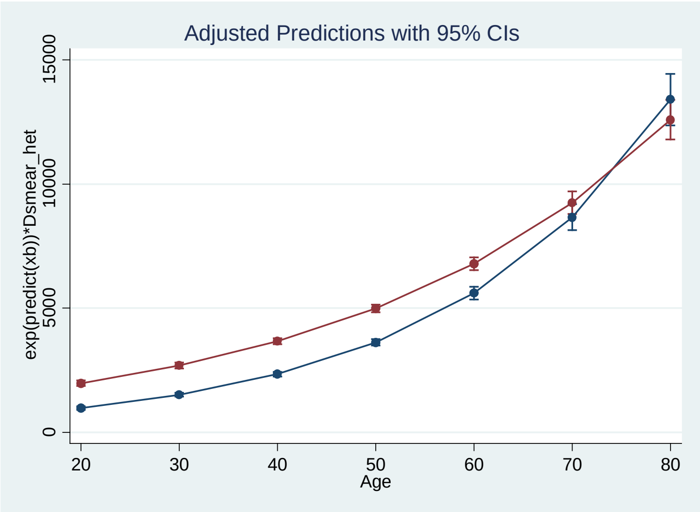
Std. Dev. = 0 means same for all obs. (not zero variance!)

```
. summarize Normal_hom Dsmear_hom

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  Normal_hom |     15946    3.089722           0   3.089722   3.089722
  Dsmear_hom |     15946     2.87812           0    2.87812    2.87812

. by female: summarize Normal_het Dsmear_het

----------------------------------------------------------------------
-> female = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  Normal_het |      6530    3.129854           0   3.129854   3.129854
  Dsmear_het |      6530     3.11787           0    3.11787    3.11787

----------------------------------------------------------------------
-> female = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  Normal_het |      9416    3.062133           0   3.062133   3.062133
  Dsmear_het |      9416    2.713122           0   2.713122   2.713122
```

Four Models for ln(*y*) (5)

Can use `margins` to predict $\hat{y}$ for $y>0$

```
margins, expression(exp(predict(xb))*Normal_hom)
margins, expression(exp(predict(xb))*Dsmear_hom)
margins, expression(exp(predict(xb))*Normal_het)
margins, expression(exp(predict(xb))*Dsmear_het)

margins, expression(exp(predict(xb))*Dsmear_het)
   at(age=(20(10)80) female=(0 1))
marginsplot, legend(off)
```



Four Models for ln(*y*) (6)

BUT … `margins` only takes uncertainty of estimated coefficients in `predict(xb)` into account when computing standard errors

It ignores uncertainty in the retransformation factor

Standard errors can be computed by

- **Delta method if heteroskedasticity is by group (see HIE), or**
- **Bootstrapping**

Problem is that `margins` easily computes the wrong standard errors

Four Models for ln(*y*) (7)

Can use `margins` to compute predicted mean for sample, one for each type of retransformation

	Margin
yhat_Nhom	5309.599
yhat_Dhom	4945.968
yhat_Nhet	5304.076
yhat_Dhet	4912.771

But to get correct standard errors, must bootstrap

FYI: Bootstrapped Standard Errors

Bootstrap the standard errors

Draw repeated samples, with replacement
Re-estimate the model
Re-compute all statistics of interest (means, marginal effects)
Do this 1000 times

In Stata use the `bootstrap` command with `percentile`

Benefits of bootstrapping include

- **Generates asymmetric CIs (e.g., for small samples)**
- **Accounts for uncertainty in retransformation for ln(*y*) models**

FYI: Four Models for ln(*y*) (8)

Bootstrap example for homoscedastic normal error

```
* Bootstrap yhat for homoskedastic normal error model
capture program drop yhat_Normal_hom
program define yhat_Normal_hom, rclass
    tempvar xbeta yhat_Nhom

    regress ln_y $x
    predict double `xbeta', xb
    generate `yhat_Nhom' = exp(`xbeta')*exp(.5*e(rmse)^2)
    summarize `yhat_Nhom', meanonly
    return scalar yhat_Nhom = r(mean)
end
bootstrap yhat_Nhom=r(yhat_Nhom), ///
    reps(1000) seed(14): yhat_Normal_hom
estat bootstrap, all
```

FYI: Four Models for ln(*y*) (9)

Bootstrap example for homoscedastic normal error

```
* Bootstrap yhat for homoskedastic Duan smearing model
capture program drop yhat_Dsmear_hom
program define yhat_Dsmear_hom, rclass
    tempvar xbeta ehat yhat_Dhom Dsmear_hom

    regress ln_y $x
    predict double `xbeta', xb
    predict double `ehat', residual
    egen `Dsmear_hom' = mean(exp(`ehat'))
    generate `yhat_Dhom' = exp(`xbeta')*`Dsmear_hom'
    summarize `yhat_Dhom', meanonly
    return scalar yhat_Dhom = r(mean)
end
bootstrap yhat_Dhom=r(yhat_Dhom), ///
    reps(1000) seed(14): yhat_Dsmear_hom
estat bootstrap, all
```

Four Models for ln(*y*) (10)

Bootstrapped standard errors are much larger!
By 29% to 81% in this example

	Observed Coef.	Bootstrap Std. Err.
yhat_Nhom	5309.599	97.105
yhat_Dhom	4945.968	127.019
yhat_Nhet	5304.076	96.925
yhat_Dhet	4912.771	120.555

Taking account of all sources of uncertainty matters
Biggest source of variation is retransformation

(Except for GLM models with low powers for variance functions, or ln(*y*) models with low log-scale variances)

Four Models for ln(*y*) (11)

Error Assumptions	Marginal Effect $\partial \mathrm{E}[y_i \vert x_i] / \partial x_k$	Incremental Effect $\Delta \mathrm{E}[y_i \vert x_i] / \Delta x_k$
General theory	$\beta_k e^{x_i' \beta} \mathrm{E}[e^{\varepsilon}] + e^{x_i' \beta} \frac{\partial \mathrm{E}[e^{\varepsilon}]}{\partial x_k}$	$\mathrm{E}[y_i \vert x_k = 1] - \mathrm{E}[y_i \vert x_k = 0]$
Homoskedastic & Normal	$\beta_k \mathrm{E}[y_i \vert x_i]$	$\left(e^{x_i^1 \beta} - e^{x_i^0 \beta} \right) e^{.5\sigma^2}$
Homoskedastic & Non-normal	$\beta_k \mathrm{E}[y_i \vert x_i]$	$\left(e^{x_i^1 \beta} - e^{x_i^0 \beta} \right) D_{smear}$
Heteroskedastic & Normal		$e^{x_i^1 \beta} e^{.5\sigma^2(g_i^1)} - e^{x_i^0 \beta} e^{.5\sigma^2(g_i^0)}$
Heteroskedastic & Non-normal		$e^{x_i^1 \beta} D_{smear}^1 - e^{x_i^0 \beta} D_{smear}^0$

Four Models for ln(*y*) (12)

Can use margins for marginal and incremental effects

```
estimates restore lnymodel
margins, expression((_b[age] +
   _b[1.female#c.age]*female)*exp(predict(xb))*Normal_hom)

margins, expression(exp(predict(xb))*Normal_hom) at(female=(0 1))
   post
lincom  _b[2._at] - _b[1bn._at]
```

Heteroskedastic models even more complicated
The standard errors still wrong, must bootstrap
Now betas in expression also held constant

Four Models for ln(y) (13)

Marginal effect of age with delta-method and bootstrapped standard errors

	Margin	Bootstrap Std. Err.
Homosk. & Normal	188.232	5.690
Homosk. & Non-N	175.341	6.192
Hetero. & Normal	188.377	5.686
Hetero. & Non-N	176.205	6.331

Differences between standard errors from 3% to 23%.

Square Root (1)

Model

$$\sqrt{y} = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 age \times female + \varepsilon$$

Interpretation (assuming homoscedasticity)

$$\hat{y} = \left(x\hat{\beta}\right)^2 + \hat{V} \qquad \hat{V} = \frac{1}{N-k-1}\sum \hat{\varepsilon}^2$$

$$\frac{\partial \hat{y}}{\partial age} = 2\left(\hat{\beta}_1 + \hat{\beta}_3 female\right)\left(x\hat{\beta}\right)$$

$$\frac{\Delta \hat{y}}{\Delta female} = \left(\hat{\beta}_0 + \hat{\beta}_2 + \left(\hat{\beta}_1 + \hat{\beta}_3\right)age\right)^2 - \left(\hat{\beta}_0 + \hat{\beta}_1 age\right)^2$$

Square Root (2)

```
* Square Root model:  Homoskedastic non-normal error
generate sqrt_y = sqrt($y)
regress sqrt_y $x
scalar s2_sqrt = e(rmse)^2

margins, expression(predict(xb)^2 + s2_sqrt)
margins, expression(2*(_b[age] +
   _b[1.female#c.age]*female)*predict(xb))
margins, expression(predict(xb)^2 + s2_sqrt)
   at(female=(0 1)) post
lincom _b[2._at] - _b[1._at]
```

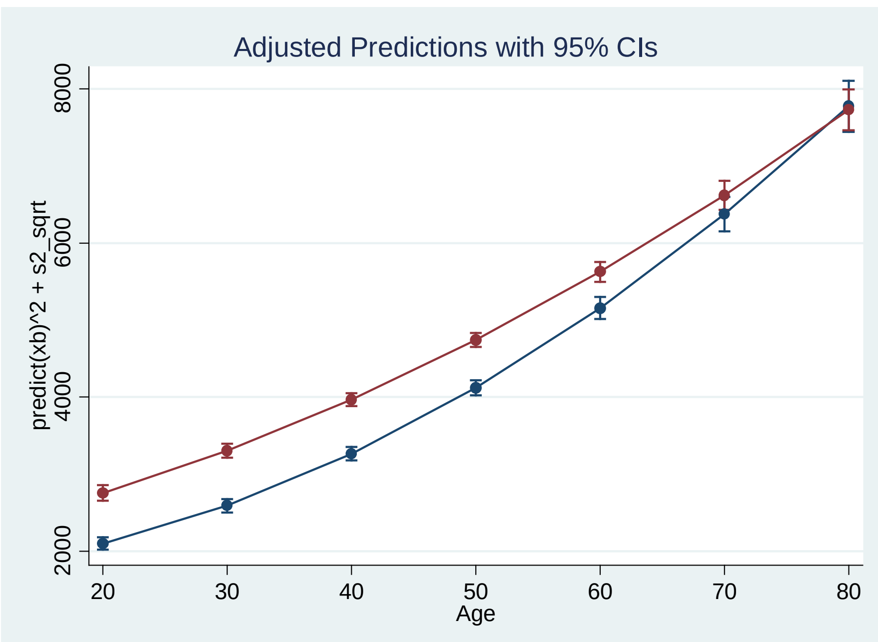
Square Root (3)

```
. regress sqrt_y $x

------------------------------------------------------------------------------
      sqrt_y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9536138   .0294372    32.39   0.000     .8959135    1.011314
    1.female |   16.25427   1.904798     8.53   0.000     12.52065    19.98789
             |
female#c.age |
          1  |  -.2067755   .0376374    -5.49   0.000    -.2805491   -.1330019
             |
       _cons |   1.910251   1.491589     1.28   0.200    -1.013432    4.833933
------------------------------------------------------------------------------
```

Square Root (4)

```
. scalar s2_sqrt = e(rmse)^2
. margins, expression(predict(xb)^2 + s2_sqrt)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   4480.678    34.27149   130.74   0.000     4413.507    4547.849
------------------------------------------------------------------------------
```

```
margins, expression(predict(xb)^2 + s2_sqrt) at(age=(20(10)80) female=(0 1))
marginsplot, legend(off)
```



Square Root (5)

```
. margins, expression(2*(_b[age] + _b[1.female#c.age]*1.female)*predict(xb))
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   84.14436   1.946344    43.23   0.000      80.3296    87.95913
------------------------------------------------------------------------------

. margins, expression(predict(xb)^2 + s2_sqrt) at(female=(0 1)) post
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   4162.637   50.35074    82.67   0.000     4063.952    4261.323
          2  |   4703.314   46.39187   101.38   0.000     4612.387     4794.24
------------------------------------------------------------------------------

. lincom _b[2._at] - _b[1._at]

 ( 1)  - 1bn._at + 2._at = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   540.6764   68.46461     7.90   0.000     406.4882    674.8646
------------------------------------------------------------------------------
```

Heteroskedastic models even more complicated
The standard errors still wrong, must bootstrap
Betas in expression held constant

Overview

Statistical issues - skewness and the zero mass

Studies with skewed outcomes but no zeroes

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Top Ten Urban Myths of Health Econometrics

Overview

Studies with zero mass and skewed outcomes

- **Brief overview of the Zero problem**

- **Two-part models**
  - **Overview**
  - **Stata code**
  - **Predictions**
  - **Marginal and incremental effects**
  - **Complications**

The Zero Problem (1)

Statement of the problem

- Often a large mass at zero
- These are true zeros, not censored values
- Zero mass may respond differently to covariates

Examples

- Expenditures or use
  - Inpatient, outpatient, nursing home, Rx
- Cigarette smoking
- Alcohol consumption

The Zero Problem (2)

Alternative estimators

OLS (ignore the problem)

ln(*y* + *c*), or Box-Cox with two parameters

GLM

Tobit (assume censored normal distribution)

Heckman selection (adjusted or generalized Tobit)

Two-part model

Conditional Density Estimation

Overview

Studies with zero mass and skewed outcomes

Brief overview of the Zero problem

Two-part models

- Overview
- Stata code
- Predictions
- Marginal and incremental effects
- Complications

Two-Part Model

Takes advantage of basic rule of probability

$$\mathrm{E}(y|x) = \Pr\left(y>0\right) \times \mathrm{E}\left(y|y>0\right)$$

Splits consumption into two parts

1. **Pr(any use or expenditures)**
   Full sample
   Estimate with logit or probit model

2. **Level of expenditures or use**
   Conditional on $y > 0$
   - **Estimates are on subsample with** $y > 0$
   - **Predictions are for entire sample**

   Use appropriate continuous or count model
   (e.g., OLS, ln(*y*), GLM, zero truncated count)

Stata Code for Two-Part Model (1)

Example

`y` is dependent variable on raw scale
`ydum` is dummy variable indicating if `y>0`
`lny` is logarithm of `y` if `y>0`

Part 1

```
   logit  ydum $x
or probit ydum $x
```

Part 2

```
   regress y   $x if y>0
or regress lny $x if y>0
or glm     y   $x if y>0, l(log) f(gamma)
```

or for count models: `tnbreg y $x if y>0`

Stata Code for Two-Part Model (2)

Use new Stata command `tpm`, written by Belotti, Deb, Manning, and Norton (WP 2013)

Install using the following Stata command:

```
ssc install tpm
```

Choose first part logit or probit
Choose second part OLS, ln(y), or GLM

Integrated with `predict` (including Duan) and `margins`

Examples

```
tpm $y $x, f(logit, nolog) s(regress) vce(robust)
tpm $y $x, f(logit, nolog) s(regress, log)
tpm $y $x, f(probit, nolog) s(glm, family(gamma)
    link(log) nolog) vce(robust)
tpm $y $x, f(probit, nolog) s(glm, family(igaussian)
    link(identity) nolog) vce(robust)
```

Stata Code for Two-Part Model (3)

The `tpm` command is limited

Specifically, it does NOT do

- **Margins for either part separately (but easy to do)**
- **Count models as second part**
- **General Box-Cox transformations (other than ln(*y*))**
- **EEE**
- **Heteroskedasticity in ln(*y*) retransformation**
- **Uncertainty in Duan smearing retransformation (std. err.)**

Warning!!!

`Margins` incorrectly calculates standard errors for all ln(*y*) and Box-Cox models because it is systematically incomplete on the retransformation

Predictions in Two-Part Model (1)

Predictions depend on both parts of the model

$$E(y|x) = \Pr(y>0) \times E(y|y>0)$$

First part

Probit $\Pr(y) = \Phi(x\alpha)$

or Logit $\Pr(y) = \dfrac{1}{1+\exp(-X\alpha)}$

Predictions in Two-Part Model (2)

Second part

If the log-scale error term is not Normal

$$\hat{y} = \Phi\left(X\hat{\alpha}\right) \times \exp\left(X\hat{\beta}\right) \times \hat{D}$$

where $\hat{D}$ is Duan's (1983) smearing estimator

If GLM with log link, so ln(E(*y*))=*X*δ

$$\hat{y} = \Phi\left(X\hat{\alpha}\right) \times \exp\left(X\hat{\delta}\right)$$

Predictions in Two-Part Model (3)

Example of Stata code for GLM with log link

```
. tpm $y $x, f(probit, nolog) s(glm, family(gamma) link(log) nolog) vce(robust)
------------------------------------------------------------------------------
             |               Robust
     exp_tot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
probit       |
         age |   .0303793   .0010029    30.29   0.000     .0284137     .032345
      female |
          1  |   .8773617    .061366    14.30   0.000     .7570864    .9976369
female#c.age |
          1  |  -.0086631   .0014161    -6.12   0.000    -.0114386   -.0058877
       _cons |  -.5964252   .0433537   -13.76   0.000    -.6813969   -.5114536
-------------+----------------------------------------------------------------
glm          |
         age |   .0345881   .0024311    14.23   0.000     .0298233    .0393529
      female |
          1  |   .7164142   .1614244     4.44   0.000     .4000283      1.0328
female#c.age |
          1  |  -.0106117   .0027302    -3.89   0.000    -.0159628   -.0052607
       _cons |   6.513084   .1468077    44.36   0.000     6.225347    6.800822
------------------------------------------------------------------------------
```

Predictions in Two-Part Model (4)

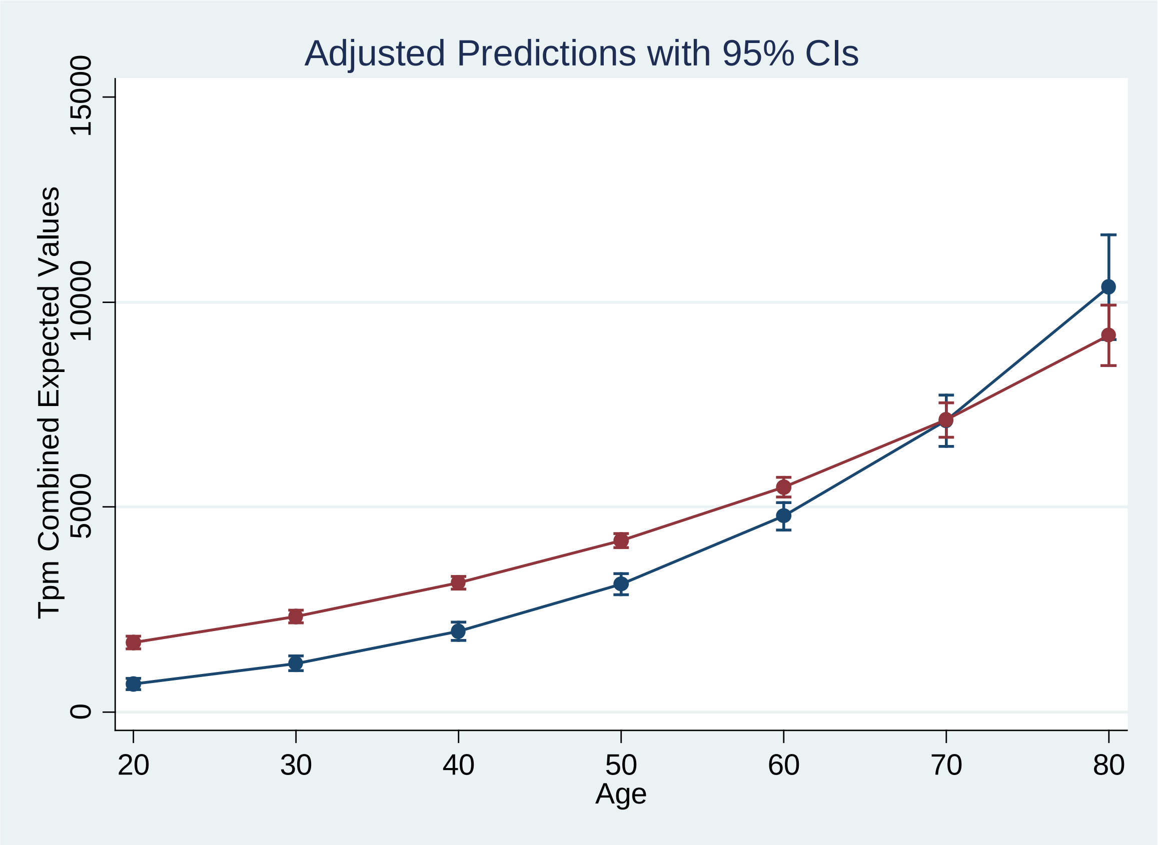
Example of Stata code for GLM with log link

```
. margins                      /* overall mean  */
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   3695.711   68.74228    53.76   0.000     3560.979    3830.443
------------------------------------------------------------------------------

. margins, at(age=(65) female=(1)) /* 65-yo woman */
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   6258.069   164.4266    38.06   0.000     5935.799    6580.339
------------------------------------------------------------------------------
```

The standard errors do not have the same problem as before because no retransformation factor

Predictions in Two-Part Model (5)

```
margins, at(age=(20(10)80) female=(0 1))
marginsplot, legend(off)
```



Marginal Effects in Two-Part Model (1)

For continuous variable x_c

$$\frac{\partial \hat{y}_i}{\partial x_c} = \frac{\partial\left(\hat{p}_i \times (\hat{y}_i | y_i > 0)\right)}{\partial x_c}$$

$$\frac{\partial \hat{y}_i}{\partial x_c} = \hat{p}_i \frac{\partial(\hat{y}_i | y_i > 0)}{\partial x_c} + (\hat{y}_i | y_i > 0) \frac{\partial \hat{p}_i}{\partial x_c}$$

If there is heteroskedasticity, then $\frac{\partial(\hat{y}_i | y_i > 0)}{\partial x_c}$ may need to account for heteroskedasticity (Duan smearing by group), or use GLM

FYI: Marginal Effects in Two-Part Model (2)

Example: Logit first part, ln(*y*) second part

$$\frac{\partial \hat{y}_i}{\partial x_c} = \left(\hat{p}_i \hat{\beta}_c (\hat{y}_i | y_i > 0)\right) + \left((\hat{y}_i | y_i > 0) \hat{\alpha}_c \hat{p}_i (1 - \hat{p}_i)\right)$$

$$\frac{\partial \hat{y}_i}{\partial x_c} = \left(\hat{\beta}_c + \hat{\alpha}_c (1 - \hat{p}_i)\right) \hat{p}_i \hat{y}_i$$

Where $\hat{p}_i$ is the $Pr(y_i > 0)$, α are first-part parameters, β are second-part parameters, subscript *c* indicates the continuous variable, subscript *i* indicates individual, $(\hat{y}_i | y_i > 0)$ is the conditional mean, $\hat{y}_i$ is the unconditional mean, and there are no interaction or higher-order terms

FYI: Marginal Effects in Two-Part Model (3)

Example: Probit first part, GLM with log link second part

$$\frac{\partial \hat{y}_i}{\partial x_c} = \left(\hat{p}_i \hat{\beta}_c (\hat{y}_i | y_i > 0)\right) + \left((\hat{y}_i | y_i > 0) \hat{\alpha}_c \varphi(x_i \hat{\alpha})\right)$$

$$\frac{\partial \hat{y}_i}{\partial x_c} = \left(\hat{\beta}_c + \hat{\alpha}_c \frac{\varphi(x_i \hat{\alpha})}{\Phi(x_i \hat{\alpha})}\right) \hat{y}_i$$

Where all notation is as is the previous slide, φ is the Normal pdf, and there are no interaction or higher-order terms

FYI: Incremental Effects in Two-Part Model

Example: Logit first part, ln(*y*) second part

$$\frac{\Delta\hat{y}_i}{\Delta x_d} = \hat{p}_i\big((\hat{y}_i|y_i > 0, x_d = 1) - (\hat{y}_i|y_i > 0, x_d = 0)\big)$$
$$+ (\hat{y}_i|y_i > 0)\big((\hat{p}_i|, x_d = 1) - (\hat{p}_i|, x_d = 0)\big)$$

$$\frac{\Delta\hat{y}_i}{\Delta x_d} = \hat{p}_i \times \Delta(\hat{y}_i|y_i > 0) + (\hat{y}_i|y_i > 0) \times \Delta\hat{p}_i$$

Where all notation is as in the previous slides, the subscript *d* indicates the dichotomous variable, and there are no interaction or higher-order terms

Marginal Effects in Two-Part Model (3)

```
. margins, dydx(age female)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   123.2596   4.932306    24.99   0.000     113.5924    132.9267
    1.female |   797.6211   142.5512     5.60   0.000     518.2258    1077.016
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

. margins, dydx(age) at(female=(0 1)) /* me of age by sex */

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   138.8681   8.888547    15.62   0.000     121.4469    156.2893
          2  |   112.5524   5.740319    19.61   0.000     101.3015    123.8032
------------------------------------------------------------------------------
```

The standard errors have the same problem as before

Complications for Marginal Effects

Interaction terms in first part, nonlinear model

- **Interaction terms in nonlinear models are complicated**
- **Must take full derivative or double difference**
- **See Ai and Norton (*Economics Letters* 2003)**

Heteroskedasticity in the second part

- **Heteroskedasticity affects retransformation**
- **Alternatives to scalar smearing factor**
  - **Multiple smearing factors, by group (Manning, 1998)**
  - **Model the heteroskedasticity (Ai & Norton, 2000, 2008)**

Complications for Marginal Effects

Bootstrap standard errors of marginal and incremental effects

Important for two reasons

- **To get around limitations of `margins`**
- **To be careful about finite sample issues**

`tpm` is new command that helps estimate two-part models

- **It makes programming more tractable**
- **It makes predictions more tractable**
- **Still need to be careful of standard errors**

Overview

Statistical issues - skewness and the zero mass

Studies with skewed outcomes but no zeroes

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Top Ten Urban Myths of Health Econometrics

Examples and Characteristics

Examples

- Number of visits to the doctor
- Number of ER visits
- Number of cigarettes smoked per day

Like expenditures / costs

- Many zeros
- Very skewed in non-zero range
- Intrinsically heteroskedastic (variance increases with mean)

Differences

- Integer valued
- Concentrated on a few low values (0, 1, 2)
- Prediction of event probabilities often of interest

Overview

Studies with count data

Poisson (canonical model)

Estimation

Prediction – Mean, Events

Interpretation – Marginal effects, Incremental Effects

Goodness of fit

Negative Binomial

Hurdle Models (Two Part Models for Counts)

Zero Inflated Models

Model Selection - Discriminating among nonnested models

Poisson

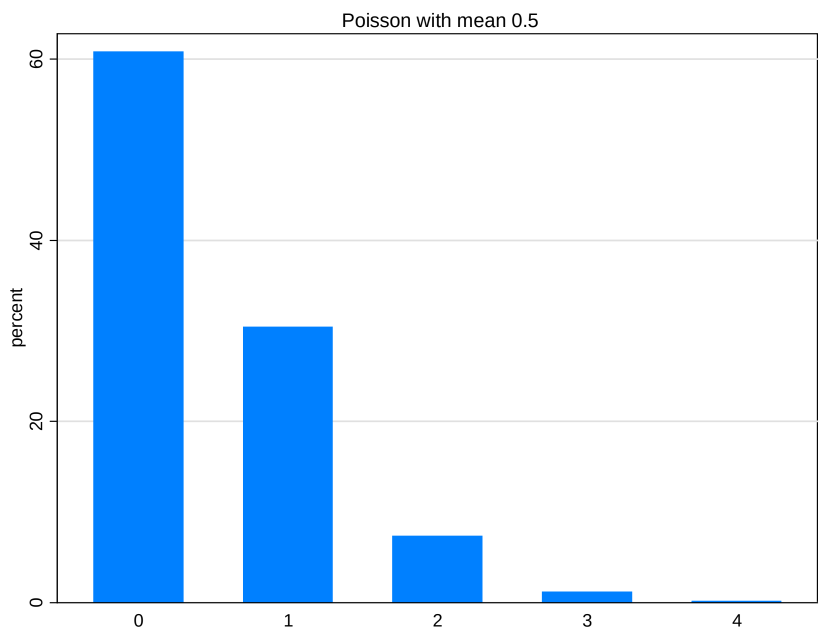
Mean $\mu = \exp(X\beta)$

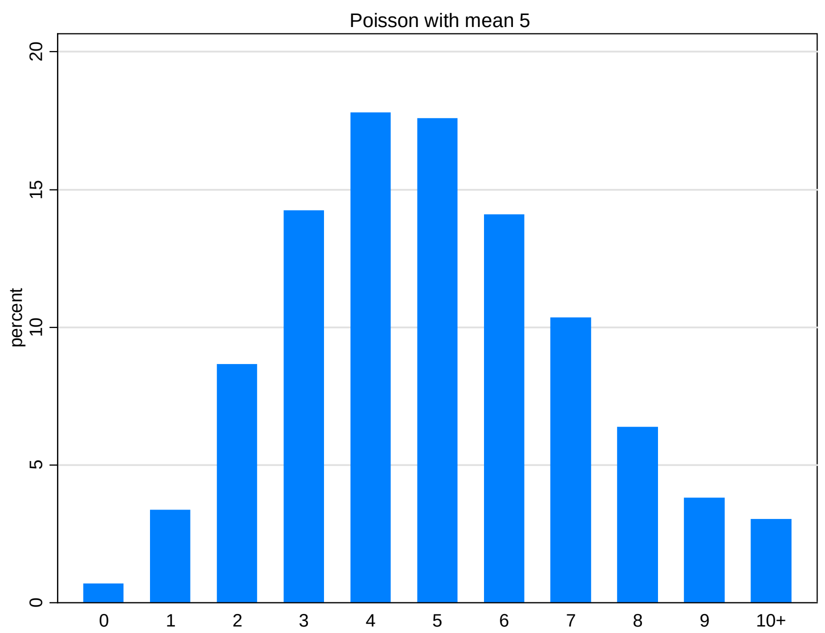
Variance $\sigma^2 = \exp(X\beta)$

Density

$$\Pr(Y = y|X) = \frac{\exp(-\mu)\,\mu^{y}}{\Gamma(y+1)}$$

Note that $\Gamma(y+1) = y!$





Estimation

Estimation is usually conducted using Maximum Likelihood

First Order Condition for MLE

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} \left(y_i - \mu_i \right) X_i = 0$$

But it is very unlikely that mean = variance property of the Poisson distribution is satisfied for most health count outcomes

The Quasi MLE for a Poisson regression relaxes the mean = variance assumption
But has the same first order condition as the MLE

So

$$\hat{\beta}_{MLE} = \hat{\beta}_{QMLE}$$

Estimation

Poisson MLE is robust to misspecification of variance of *y*, i.e.
$\hat{\beta}_{MLE} = \hat{\beta}_{QMLE}$

In other words, it is okay to estimate a Poisson regression in terms of point estimates even if the dgp is not Poisson but the weaker QMLE assumptions are satisfied

But standard errors for $\hat{\beta}_{MLE}$ are not correct unless the true dgp is Poisson (mean = variance)

The sandwich form for $Cov(\hat{\beta})$ ("robust") is appropriate because it uses only the QMLE assumptions (mean need not be equal to variance)

Stata command: `poisson use_off age i.female, robust`

Prediction

The typical prediction of interest is the conditional mean.

But, in nonlinear models, predictions of quantities other than the conditional mean are often of interest.

In the context of count data, we might be interested in predictions of the distribution of the count variable

$\Pr(Y = 0 \mid X)$

$\Pr(Y = 12 \mid X)$

We might also be interested in predictions of certain events of interest

$\Pr(Y > 5 \mid X) = 1 - \Pr(Y \leq 5 \mid X)$

Substantively

Probability of exceeding a benefit cap (mental health)

Probability of a "drive-through" delivery

Prediction in Poisson

Conditional Mean: $\hat{\mu} = \exp(X\hat{\beta})$

Stata command: `predict muhat` **(default)**

Distribution and events:

$$\Pr(Y = y | X) = \frac{\exp(-\hat{\mu})\,\hat{\mu}^{y}}{\Gamma(y+1)} \qquad \forall\, y = 0, 1, 2, 3, \ldots$$

Stata commands:

```
predict prhat0, pr(0)
predict prhat12, pr(12)

predict prhat0to5, pr(0,5)
generate prhatgt5 = 1 – prhat0to5
```

Interpretation

Marginal Effects - for continuous variables

$$\frac{\partial E(y_i | X)}{\partial X^k} = \beta^k \ \mu_i$$

Examples: Income, Price, Health status

Incremental Effects - for binary variables

$$E(y_i | X, X^k = 1) - E(y_i | X, X^k = 0)$$
$$= \left[\mu_i | X^k = 0 \right] \left[\exp(\beta^k) - 1 \right]$$

Examples: Treatment/ Control, Insurance, Gender, Race

FYI: Predictions (at specific X), Marginal & Incremental Effects

Approach depends on research question. How one does it can make a big difference

1. Evaluate for hypothetical individuals

   a. Mean (or Median, other quantiles) of X in sample

   b. Mean (or Median, other quantiles) of X in sub-sample of interest

   c. Hypothetical individual of interest

2. Evaluate for each individual

   a. Average over sample

   b. Average over sub-samples of interest

3. For Incremental Effects – (Treatment vs. Control)

   a. Switch all individuals from control to treatment

   b. Switch controls to treatment

FYI: Predictions (at specific X), Marginal & Incremental Effects

Stata command for predictions at specific values of X:

```
margins female
margins, at(age=27)
margins female, at(age=32)
```

Stata command for marginal / incremental effects:

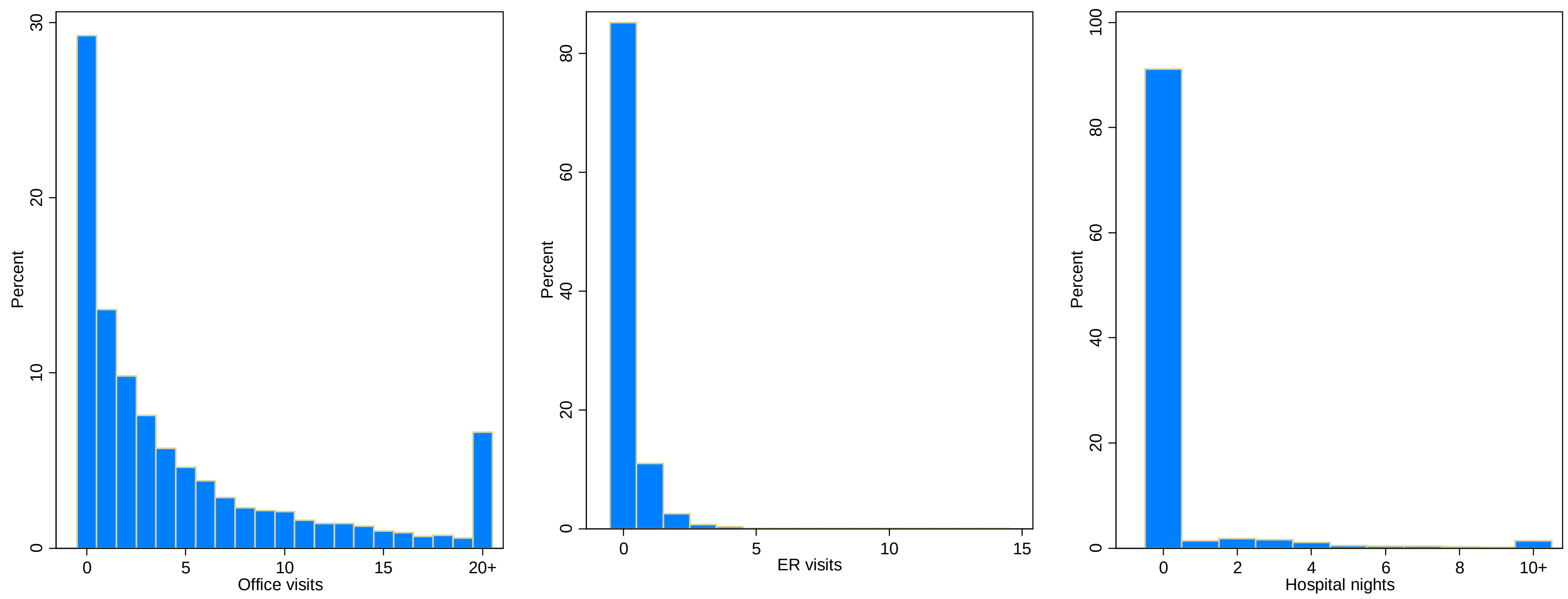
Be sure to code indicator variables using factor notation (`i.female`)

```
margins, dydx(age)
margins, dydx(*)
margins, dydx(*) at(age=27)
margins female, dydx(*) at(age=27)
```

Examples

Data from MEPS

1. Number of office-based visits
2. Number of emergency room visits
3. Number of hospital nights



Poisson Estimates

Poisson Coefficients

	Office visits	ER visits	Hospital nights
Age	0.005**	-0.018**	0.001
	(0.001)	(0.002)	(0.004)
1.female	0.328**	0.171**	-0.044
	(0.027)	(0.044)	(0.138)

* *p*<0.05; ** *p*<0.01

Predictive margins from Poisson

```
. margins female

Predictive margins                              Number of obs   =      19386
Model VCE    : Robust

Expression   : Predicted number of events, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
          0  |   4.737476   .1074384    44.09   0.000       4.5269    4.948051
          1  |   6.577042   .0963255    68.28   0.000     6.388248    6.765837
------------------------------------------------------------------------------
```

Predictive margins from Poisson

```
. margins, at(age=(30 50 70))

Predictive margins                                Number of obs   =      19386
Model VCE    : Robust

Expression   : Predicted number of events, predict()

1._at        : age             =          30
2._at        : age             =          50
3._at        : age             =          70
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   5.170575   .1441185    35.88   0.000     4.888108    5.453042
          2  |   5.729337   .0719968    79.58   0.000     5.588226    5.870448
          3  |   6.348482   .1421247    44.67   0.000     6.069923    6.627041
------------------------------------------------------------------------------
```

Marginal Effects from Poisson

```
. margins, dydx(age female)

Average marginal effects                          Number of obs   =      19386
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age 1.female

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0297709   .0063503     4.69   0.000     .0173246    .0422171
    1.female |   1.839567   .1465079    12.56   0.000     1.552416    2.126717
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Marginal Effects from Poisson

```
. margins female, dydx(age)

Average marginal effects                          Number of obs   =      19386
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  age        |
      female |
          0  |    .024307   .0052005     4.67   0.000     .0141142    .0344998
          1  |   .0337455    .007205     4.68   0.000      .019624     .047867
------------------------------------------------------------------------------
```

Marginal Effects from Poisson

Poisson Marginal Effects: Office visits

	Average	Mean of X	Median of X
age	0.030**	0.022**	0.021**
	(0.002)	(0.001)	(0.001)
1.female	1.840**	1.402**	1.294**
	(0.035)	(0.027)	(0.025)

Poisson Marginal Effects: ER visits

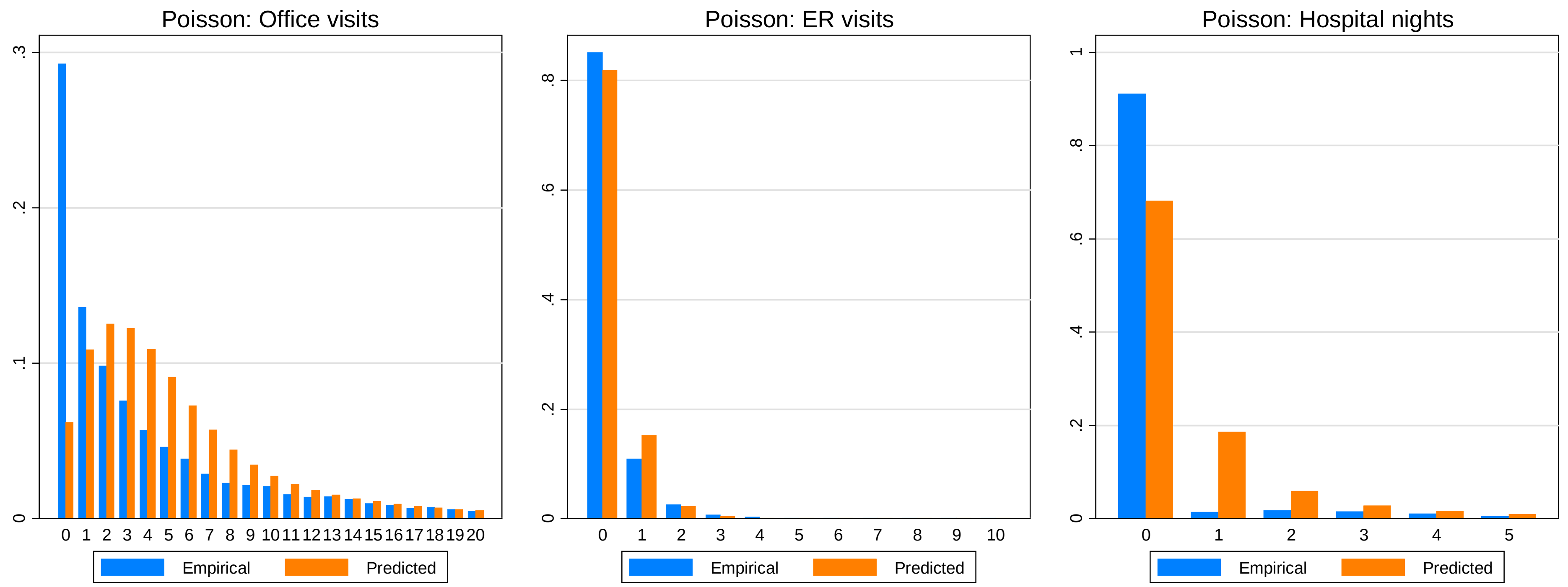
	Average	Mean of X	Median of X
age	-0.004**	-0.003**	-0.003**
	(0.000)	(0.000)	(0.000)
1.female	0.036**	0.029**	0.028**
	(0.007)	(0.005)	(0.005)

Poisson Marginal Effects: Hospital nights

	Average	Mean of X	Median of X
Age	0.001	0.000	0.000
	(0.000)	(0.000)	(0.000)
1.female	-0.028*	-0.013*	-0.013*
	(0.012)	(0.006)	(0.006)

In-sample Goodness of fit

Informal / Graphical - compare empirical distribution of *y* to predicted distribution



In-sample Goodness of Fit

Mean Prediction (of distribution) Error

$$MPE = \frac{1}{J}\sum_{j=0}^{J}(f_j - \hat{f}_j)$$

Mean Square Prediction (of distribution) Error

$$MSPE = \frac{1}{J}\sum_{j=0}^{J}(f_j - \hat{f}_j)^2$$

J should be chosen to cover most of the support (but not all the values of the count variable)

	Office visits (0-20)	ER visits (0-10)	Hospital nights (0-5)
MPE	-0.155	-0.002	-0.129
MSPE	30.615	2.705	139.367

FYI: Stata code for Poisson goodness of fit measures

```
preserve
forvalues j=0/20 {
   gen byte y_`j' = `e(depvar)' == `j'
   predict pr_`j', pr(`j')
}
collapse (mean) y_* pr_*
gen i=_n
reshape long y_ pr_, i(i) j(y)

graph bar (asis) y_ pr_

generate pr_diff = (y_ - pr_)*100
generate pr_diff2 = pr_diff^2

mean pr_diff pr_diff2
restore
```

Poisson - Summary

Advantages

Robust (asymptotic) to misspecification of variance

Easy to compute marginal effects and predictions

Disadvantages

Not robust in finite samples

Possibly sensitive to influential observations and outliers

Not efficient if variance is misspecified

Overview

Studies with count data

- **Poisson (canonical model)**

- **Negative Binomial**
  - **Estimation**
  - **Prediction – Mean, Events**
  - **Interpretation – Marginal effects, Incremental Effects**
  - **Goodness of fit**

- **Hurdle Models (Two Part Models for Counts)**

- **Zero Inflated Models**

Model Selection - Discriminating among nonnested models

Negative Binomial

Canonical model for overdispersed data

Mean $\mu = \exp(X\beta)$

Overdispersion – variance exceeds the mean

$$Var(y|X) = \mu + \alpha\, g(\mu) > \mu$$

Negative Binomial-1 $Var(y|X) = \mu + \alpha\,\mu$

Negative Binomial-2 $Var(y|X) = \mu + \alpha\,\mu^2$

Estimation

Maximum Likelihood

Stata command for NB-2:

```
nbreg use_off age i.female, dispersion(mean)
nbreg use_off age i.female
```

Note: dispersion(mean) is not required – it is the default

Stata command for NB-1:

```
nbreg use_off age i.female, dispersion(constant)
```

Choosing between NB-1 and NB-2

These are non-nested models

Use model selection criteria

FYI: Negative Binomial-2: Estimates

```
. nbreg use_off $X, robust

<snip>

Negative binomial regression                    Number of obs   =      19386
Dispersion           = mean                     Wald chi2(21)   =    4900.92
Log pseudolikelihood = -49111.723               Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |               Robust
     use_off |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0108457   .0010707    10.13   0.000     .0087473    .0129442
    1.female |   .4911042   .0264055    18.60   0.000     .4393504    .5428581
<snip>
-------------+----------------------------------------------------------------
    /lnalpha |   .3581475   .0176417                      .3235703    .3927246
-------------+----------------------------------------------------------------
       alpha |   1.430677   .0252396                      1.382053    1.481011
------------------------------------------------------------------------------

.
. estimates store nb2
```

FYI: Negative Binomial-1:Estimates

```
. nbreg use_off $X, disp(constant) robust

<snip>

Negative binomial regression                         Number of obs   =      19386
Dispersion           = constant                      Wald chi2(21)   =    7663.19
Log pseudolikelihood = -48824.428                    Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |               Robust
     use_off |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0077678   .0006543    11.87   0.000     .0064855    .0090501
    1.female |   .3710594   .0153733    24.14   0.000     .3409282    .4011906
<snip>
-------------+----------------------------------------------------------------
    /lndelta |   2.144767   .0246959                      2.096363     2.19317
-------------+----------------------------------------------------------------
       delta |   8.540047   .2109045                      8.136526     8.96358
------------------------------------------------------------------------------

. estimates store nb1
```

Negative Binomial: Choosing Between NB2 and NB1

```
. estimates stats nb2 nb1
```

Office visits

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
nb2	19386	-52504.62	-49111.72	23	98269.45	98450.51
nb1	19386	-52504.62	-48824.43	23	97694.86	97875.92

Note: N=Obs used in calculating BIC; see [R] BIC note

ER visits

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
nb2	19386	-10671.19	-9995.218	23	20036.44	20217.5
nb1	19386	-10671.19	-10020.41	23	20086.83	20267.89

Hospital nights

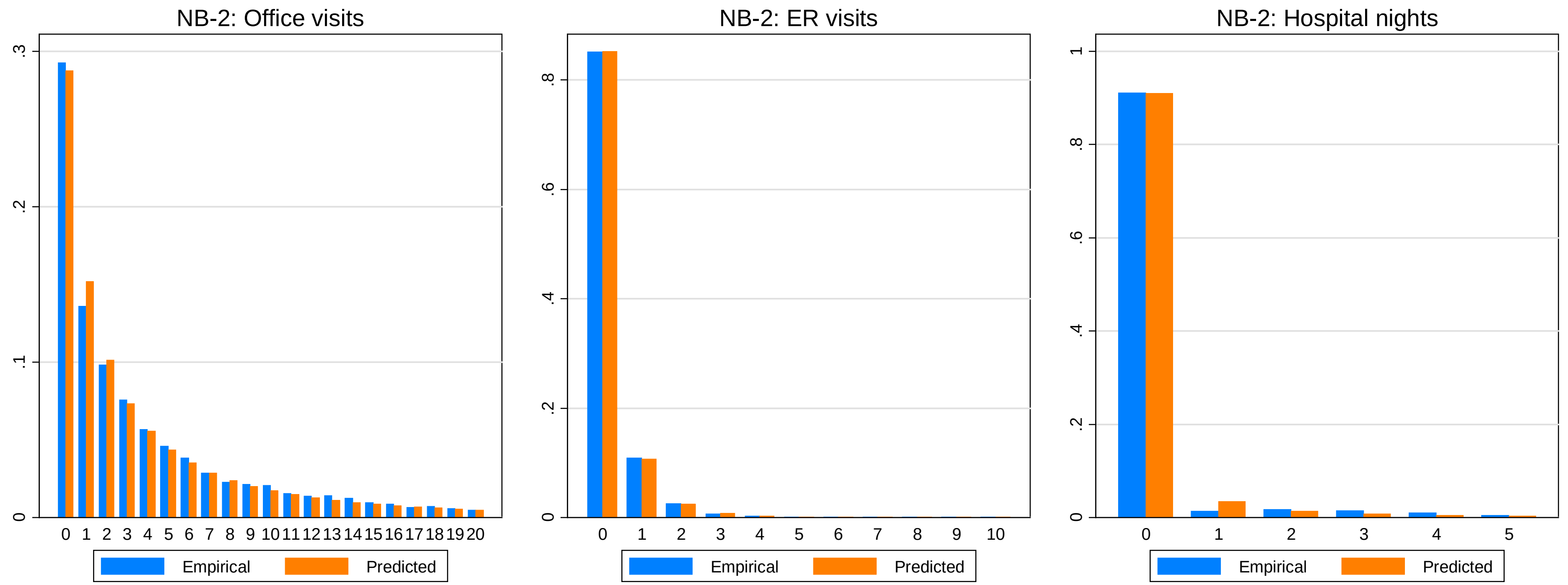
Model	Obs	ll(null)	ll(model)	df	AIC	BIC
nb2	19386	-10635.45	-10033.9	23	20113.8	20294.86
nb1	19386	-10635.45	-9884.171	23	19814.34	19995.41

NB Marginal Effects

	Office visits		ER visits		Hospital nights	
	NB-2	NB-1	NB-2	NB-1	NB-2	NB-1
age	0.068**	0.045**	-0.004**	-0.003**	-0.002	-0.003*
	(0.007)	(0.004)	(0.000)	(0.000)	(0.004)	(0.001)
1.female	2.909**	2.073**	0.031**	0.033**	0.214**	0.252**
	(0.153)	(0.085)	(0.009)	(0.008)	(0.082)	(0.033)

* $p<0.05$; ** $p<0.01$

In-sample Goodness of Fit



	Office visits	ER visits	Hospital nights
MPE	0.046	-0.001	-0.043
MSPE	0.167	0.005	0.883

Negative Binomial - Summary

Advantages

Much less sensitive to influential observations and outliers

Mean is robust in finite samples

Disadvantages

Distribution is not robust to misspecification of variance

Not efficient if variance is misspecified

Overview

Studies with count data

- **Poisson (canonical model)**

- **Negative Binomial**

- **Hurdle Models (Two Part Models for Counts)**
  - **Estimation**
  - **Prediction – Mean, Events**
  - **Interpretation – Marginal effects, Incremental Effects**
  - **Goodness of fit**

- **Zero Inflated Models**
  - **Estimation**
  - **Prediction – Mean, Events**
  - **Interpretation – Marginal effects, Incremental Effects**
  - **Goodness of fit**

Model Selection - Discriminating among nonnested models

Hurdle Model

Two Part Model for counts - Zeros are from different process
No demand / demand in sample period

$$\Pr(Y = 0 \mid X) = f_1(0 \mid \theta_1)$$

$$\Pr(Y = y > 0 \mid X) = \frac{(1 - f_1(0 \mid \theta_1))}{(1 - f_2(0 \mid \theta_2))} f_2(y \mid \theta_2)$$

where

$f_1(. \mid \theta_1)$ **is a Logit / Probit Model**

$f_2(. \mid \theta_2)$ **is a Poisson / NB Model**

$\frac{1}{(1 - f_2(0 \mid \theta_2))} f_2(y \mid \theta_2)$ **is a Truncated Count Density**

Stata command:

```
logit(probit) / tpoisson (tnbreg)
```

FYI: Estimation and Prediction from Hurdle Models

Because model is constructed "manually", `predict` does not work directly

```
. quietly probit use_off $X, nolog
. predict prgt0
(option pr assumed; Pr(use_off))

. quietly tnbreg use_off $X if use_off>0, ll(0) nolog
. predict yhat_cm, cm

. gen yhat = prgt0 * yhat_cm
. sum prgt0 yhat_cm yhat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       prgt0 |      19386    .7067938    .2280265   .0373487   .9995626
     yhat_cm |      19386    7.309756    5.249475   1.753717   53.40609
        yhat |      19386    5.971018    5.699017   .0722289   53.38273
```

Predictive Margins and Marginal Effects

1. **Conditional Mean**

2. **Distribution**

3. Marginal / Incremental Effects

See do file: `deb-countdata.do`

Standard errors of marginal / incremental effects calculated via nonparametric bootstrap

Marginal Effects from Hurdle Models

	Office visits			
	NB-2	Probit	Trunc. NB-2	Hurdle NB-2
Age	0.068**	0.003**	0.040**	0.023**
	(0.005)	(0.000)	(0.007)	(0.004)
1.female	2.909**	0.146**	2.063**	2.259**
	(0.121)	(0.006)	(0.160)	(0.173)
Log likelihood	-49,112	-9,110	-39,292	-48,402

* p<0.05; ** p<0.01

Marginal Effects from Hurdle Models

	ER visits			
	NB-2	Probit	Trunc. NB-2	Hurdle NB-2
Age	-0.004**	-0.002**	-0.007**	-0.002**
	(0.000)	(0.000)	(0.002)	(0.000)
1.female	0.031**	0.017**	0.036	0.029**
	(0.009)	(0.005)	(0.037)	(0.009)
Log likelihood	-9,995	-7,534	-2,407	-9,941

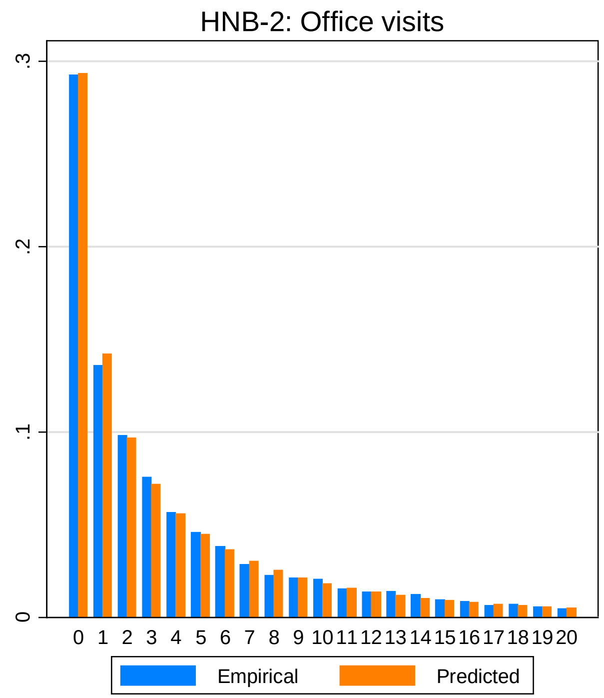
* *p*<0.05; ** *p*<0.01

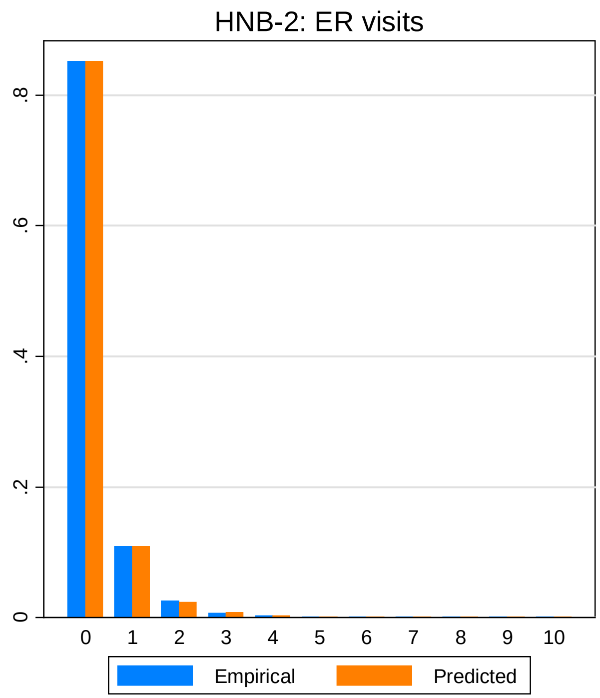
Marginal Effects from Hurdle Models

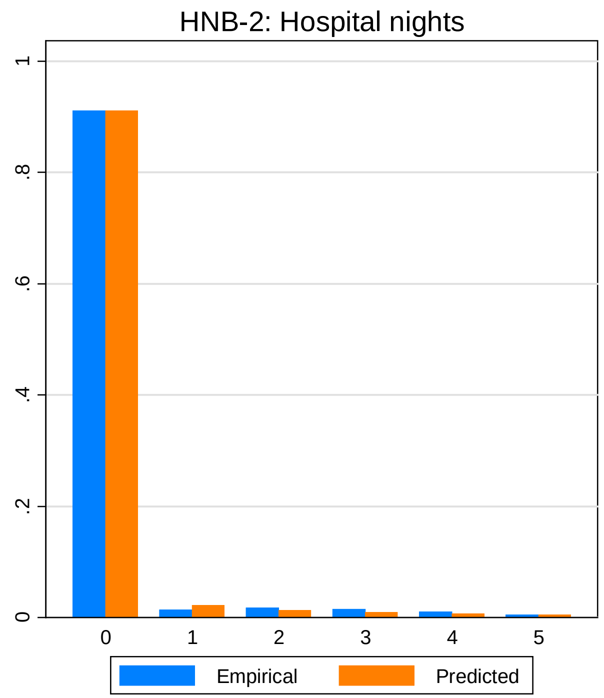
	Hospital nights			
	NB-2	Probit	Trunc. NB-2	Hurdle NB-2
Age	-0.002	-0.001**	0.023	-0.001
	(0.003)	(0.000)	(0.020)	(0.001)
1.female	0.214**	0.036**	-2.499**	0.027
	(0.061)	(0.004)	(0.537)	(0.074)
Log likelihood	-10,034	-5,077	-4,614	-9,691

* p<0.05; ** p<0.01

In-sample Goodness of Fit







	Office visits	ER visits	Hospital nights
MPE	0.024	-0.001	0.098
MSPE	0.040*	0.002	0.251

Hurdle Models

Advantages

- **Estimation in 2 parts**
- **Same variables in both parts not a problem**
- **Numerically well behaved**

Disadvantages

- **Many have a strong prior belief that zeros from different process than positives**
- **Even when marginal / incremental effects from each process are “sensible”, overall effects may be “odd” (composition effects)**

FYI: Zero Inflated Models

Zeros are from two processes

No demand / No demand in sample period

$$\Pr(Y = 0 \mid X) = f_1(0 \mid \theta_1) + (1 - f_1(0 \mid \theta_1))\ f_2(0 \mid \theta_2)$$

$$\Pr(Y = y > 0 \mid X) = (1 - f_1(0 \mid \theta_1))\ f_2(y \mid \theta_2)$$

where

$f_1(. \mid \theta_1)$ **is a Logit / Probit Model**

$f_2(. \mid \theta_2)$ **is a Poisson / NB Model**

Stata command: `zip / zinb`

Usually, same covariates used in $f_1(. \mid \theta_1)$ **and** $f_2(. \mid \theta_2)$

Sometimes $f_1(. \mid \theta_1)$ **specified as a constant**

FYI: Example

Office visits

	NB-2	Hurdle NB-2	ZINB-2
Age	0.068**	0.023**	0.044**
	(0.005)	(0.004)	(0.005)
1.female	2.909**	2.259**	2.180**
	(0.121)	(0.173)	(0.111)
Log likelihood	-49,112	-48,402	-48,605

FYI: Zero-Inflated Models

Advantages

Natural way to introduce extra zeros

Disadvantages (Especially if both parts have same covariates)

Computationally complex – likelihood function can have plateaus and multiple maxima

Weak identification of Binary and Count Model parameters in finite samples

Even when marginal / incremental effects from each processes are "sensible", overall effects may be "odd" (composition effects)

Overview

Studies with count data

Poisson (canonical model)

Negative Binomial

Hurdle Models (Two Part Models for Counts)

Zero Inflated Models

Model Selection - Discriminating among non-nested models

Model Selection

In Sample

Akaike Information Criterion

$$AIC = -2\log(L) + 2k$$

Bayesian Information Criterion

$$BIC = -2\log(L) + \log(N)k$$

Graphical check of distribution fit

Model Selection

Out-of Sample

50% split cross-validation

1. Randomly split sample into 2 parts
2. Estimation – using 1 (training) sample
3. Prediction – using remaining (prediction) sample
4. Evaluate model performance in prediction sample
5. Repeat steps 1 - 4

K-fold cross-validation

1. Randomly split sample into K (10) parts
2. Estimation – using (K-1) parts
3. Prediction – Remaining K_{th} part
4. Evaluate model performance in prediction sample

Examples: In-sample Fit

Office visits

	Poisson	NB-2	NB-1	HNB-2	HNB-1	ZINB-2
K	22	23	23	45	45	45
AIC	197,522	98,269	97,695	96,893	97,091	97,300
BIC	197,695	98,451	97,876	97,247	97,445	97,654

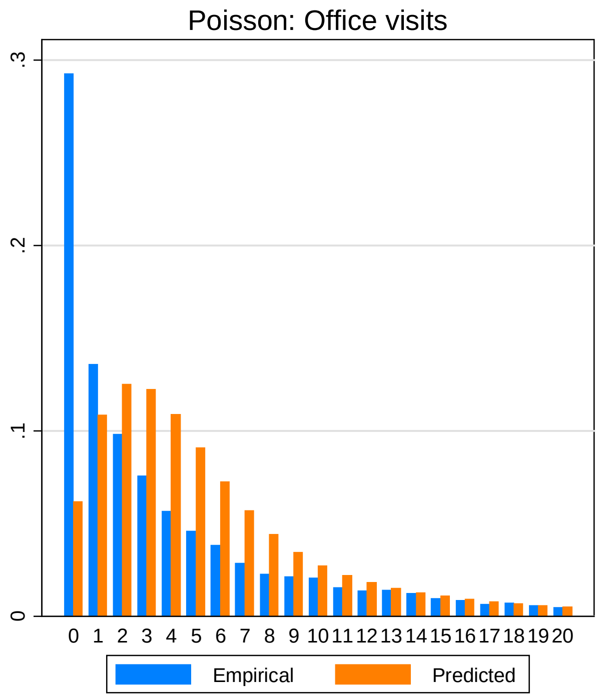
ER visits

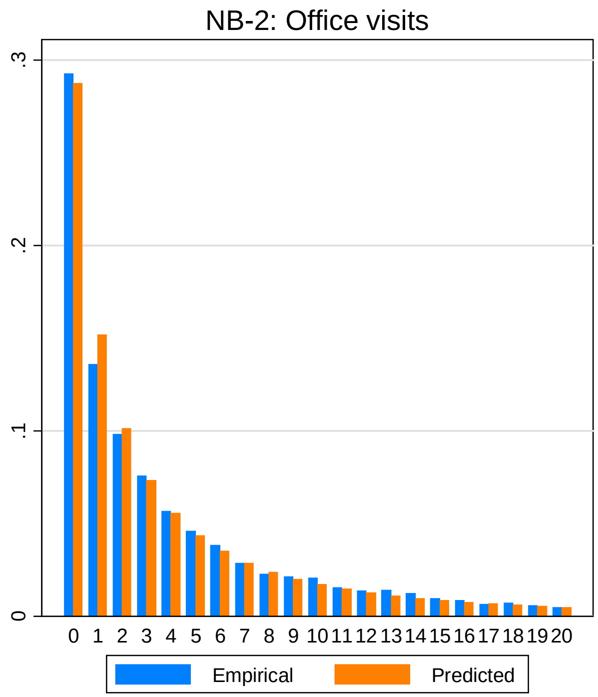
	Poisson	NB-2	NB-1	HNB-2	HNB-1	ZINB-2
K	22	23	23	45	45	45
AIC	21,408	20,036	20,087	19,973	20,192	19,966
BIC	21,581	20,217	20,268	20,327	20,547	20,320

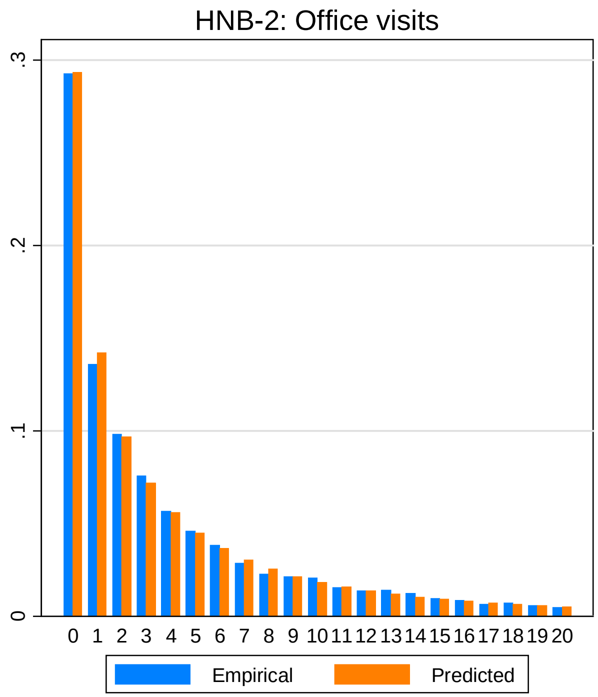
Hospital nights

	Poisson	NB-2	NB-1	HNB-2	HNB-1	ZINB-2
K	22	23	23	45	45	45
AIC	66,723	20,114	19,814	19,472	19,658	19,487
BIC	66,896	20,295	19,995	19,826	20,012	19,841

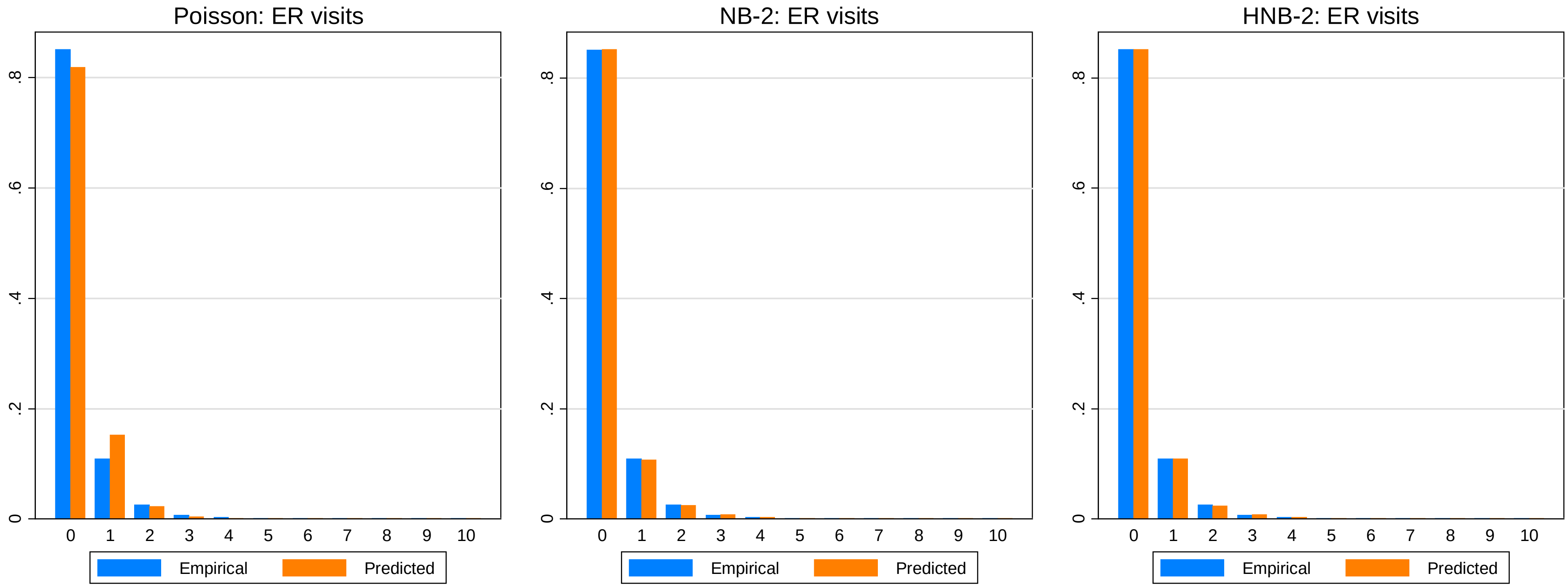
Examples: In-sample Fit



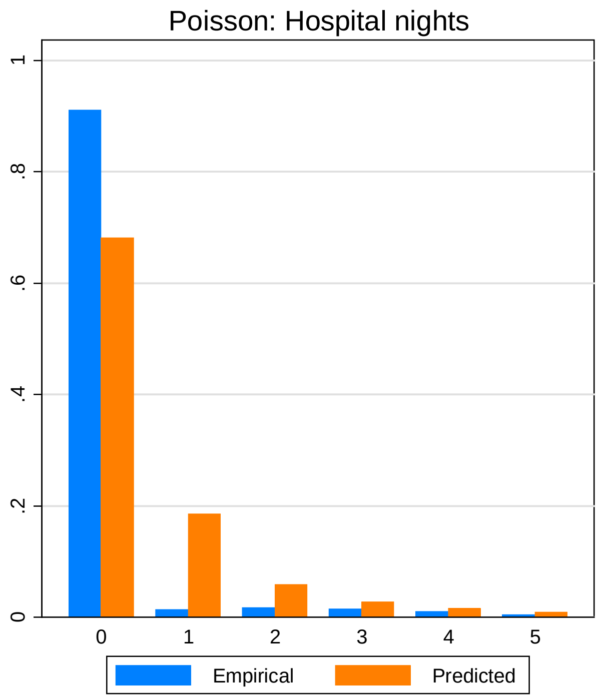


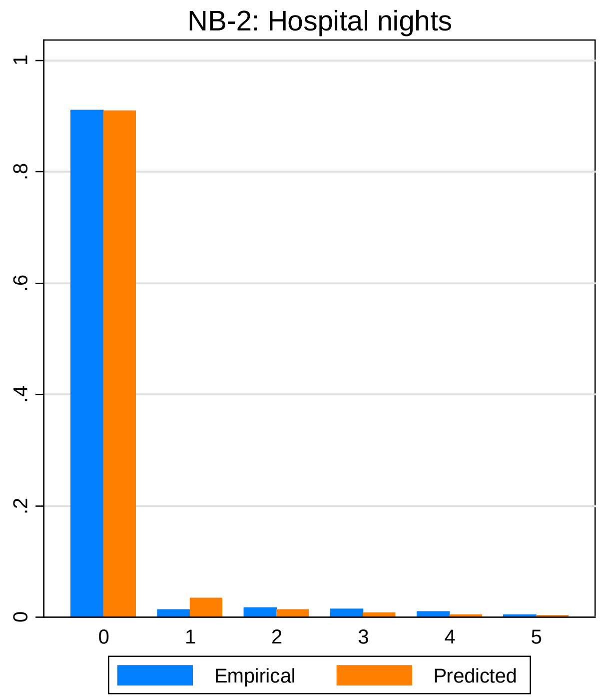


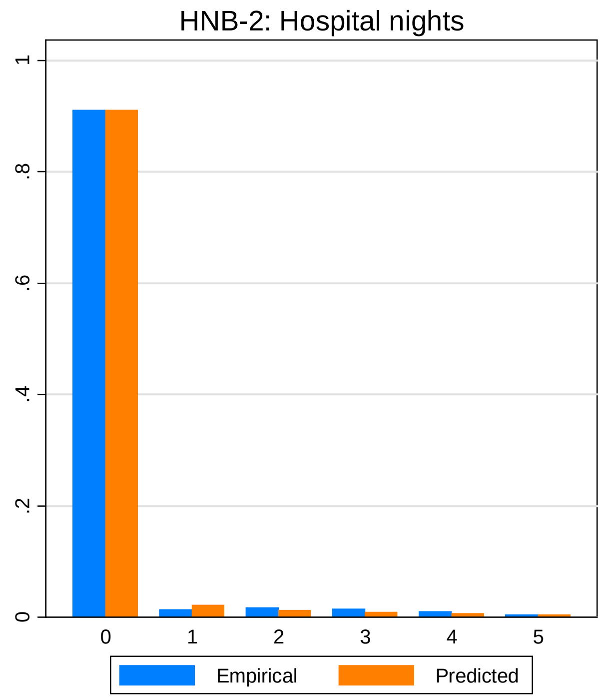
Examples: In-sample Fit



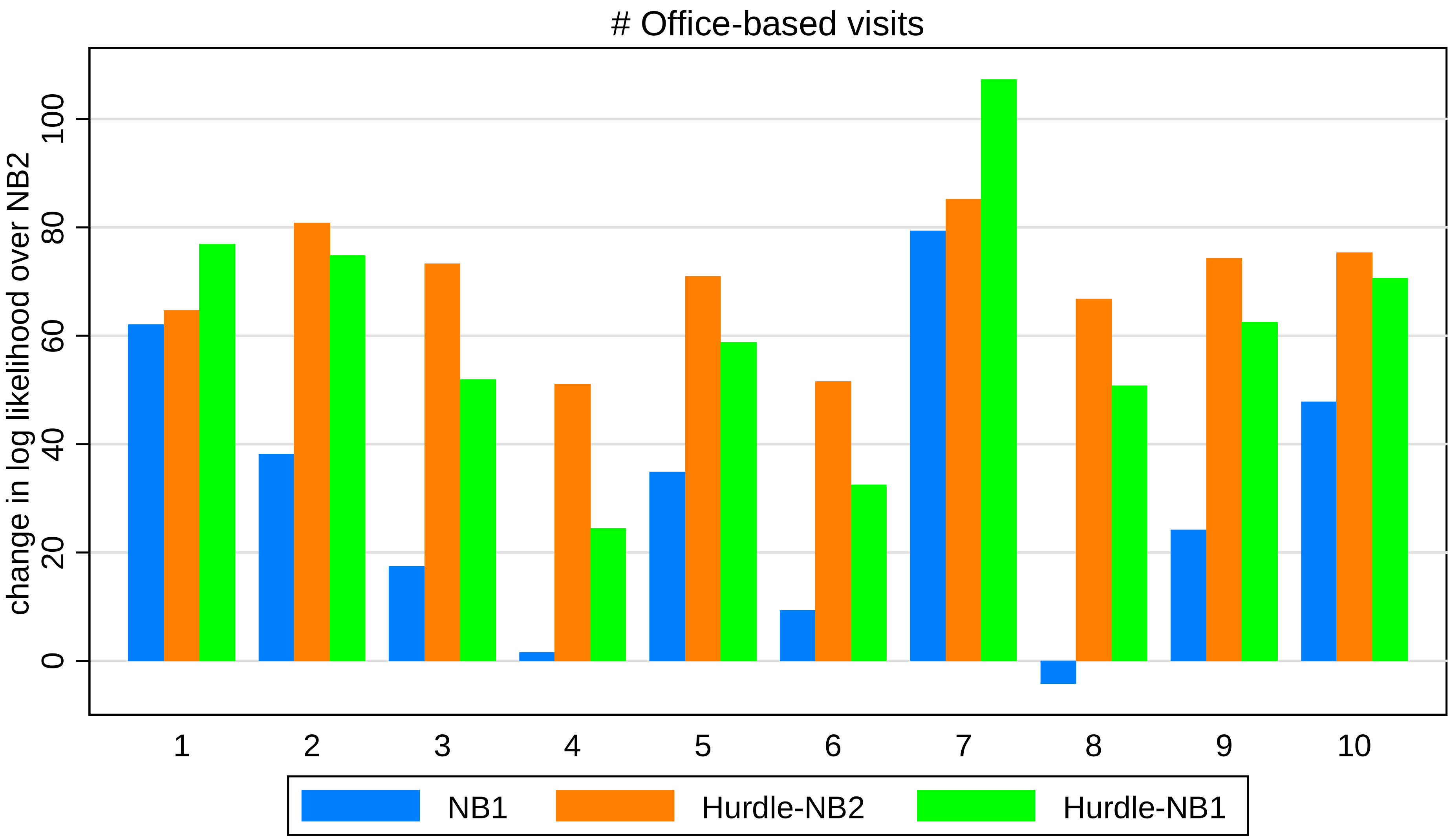
Examples: In-sample Fit



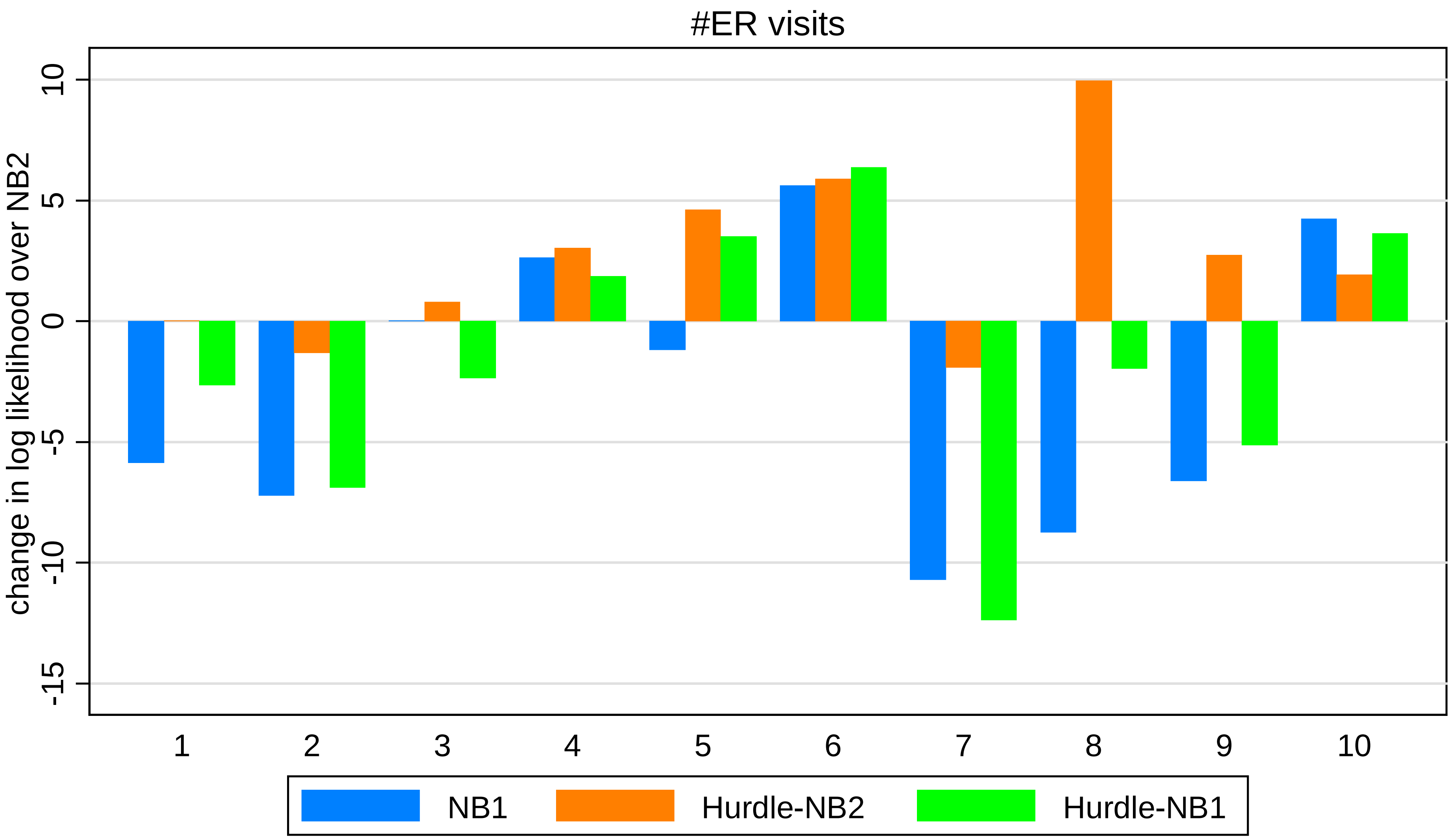




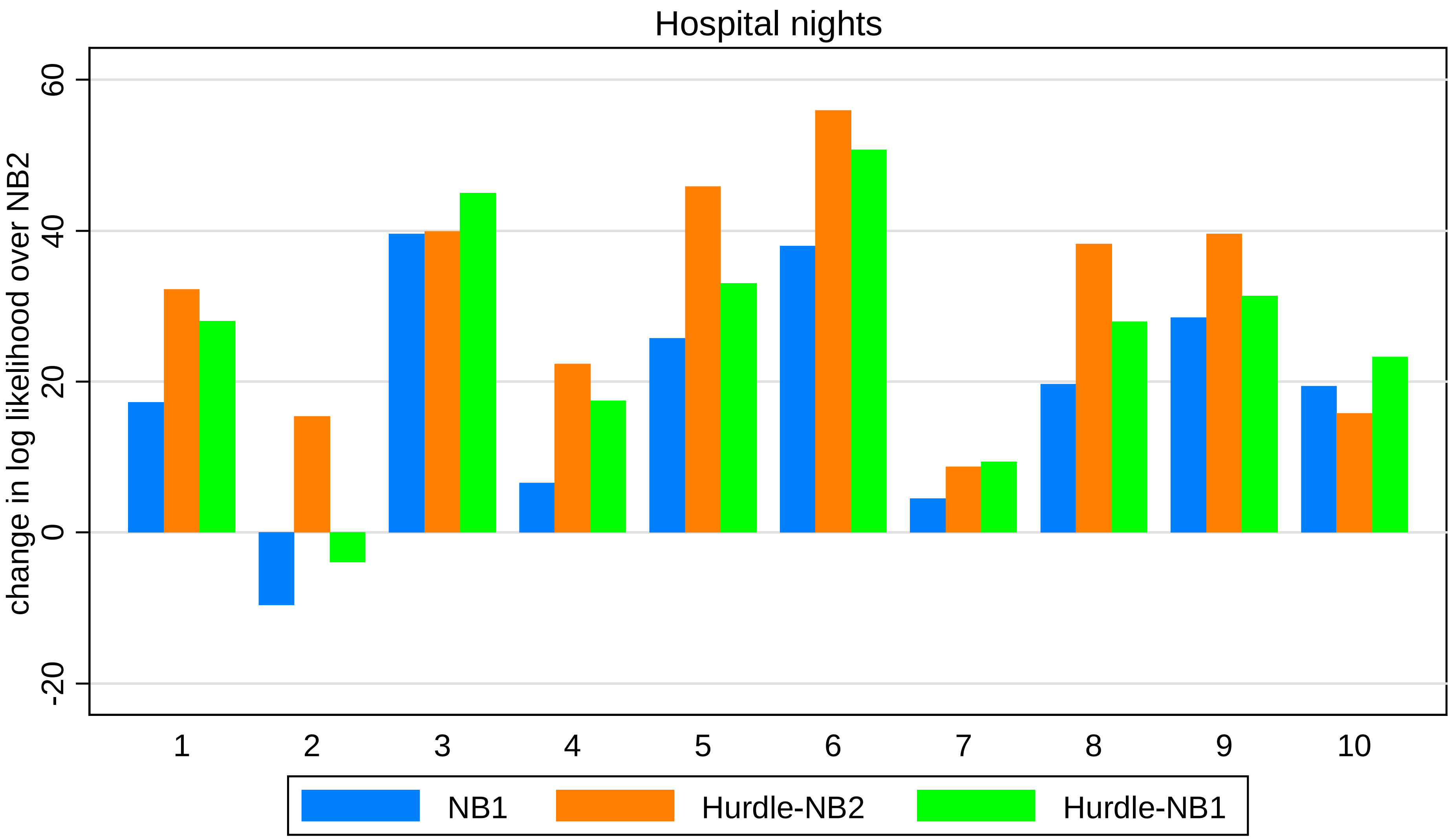
Examples: Out-of-sample Fit (10-fold Cross-validation)



Examples: Out-of-sample Fit (10-fold Cross-validation)



Examples: Out-of-sample Fit (10-fold Cross-validation)



Overview

Statistical issues - skewness and the zero mass

Studies with skewed outcomes but no zeroes

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Top Ten Urban Myths of Health Econometrics

Conclusions

Health Care Outcomes

- **Are pervasively skewed to the right with long right tails**
- **Have substantial fraction of observations with zeroes**
- **Display heteroscedasticity even after transformation**
- **Display different responsiveness to covariates at different parts of the distribution**

No single model is “best” for all cases or for all populations

Conclusions (cont'd)

Log transform is not the only, nor the best solution to skewness

Retransformation is more complicated than meets the eye

- **We do not care about *ln*($) or *ln*(€) *per se* but $ or €**
- **Model E(y|x) or some $g^{-1}(x'\beta)$ instead**

Comprehensive model checking is recommended

In-sample checks are not always reliable

- **Overfitting is a very real danger**
- **Cross-validation checks are strongly recommended**

Conclusions (cont'd)

But it is not all bad news – Cox, Draper: "All models are wrong but some are useful"

We have outlined a variety of methods that

Work in many disparate situations

Are easy to estimate (generally)

Often provide a better fit

- **Are less sensitive to outliers**
- **Can result in large efficiency gains vis-à-vis linear models**

Also outlined approaches to making decisions about models

Conclusions (cont'd)

We have provided a working two-part model *tpm* software that allows for:

1. **Marginal and incremental effects**
2. **Stratification and survey weights, clustering**
3. **Compatible with bootstrap to capture the full uncertainty and deal with finite sample issues**

Covers many of the different *but not all* types of two-part models. Hurdle models are on our to-do list.

Provided ado's for checking linearity and testing for influential outliers.

Promising methodological work continues in the literature

Websites for handouts, recommended reading and programming code

Willard Manning
mailto:w-manning@uchicago.edu
http://harrisschool.uchicago.edu/faculty/web-pages/willard-manning.asp
follow link to iHEA materials

Edward Norton
mailto:ecnorton@umich.edu
http://www.sph.umich.edu/iscr/faculty/profile.cfm?uniqname=ecnorton
follow link to Health Econometrics

Partha Deb
mailto:partha.deb@hunter.cuny.edu
http://econ.hunter.cuny.edu/parthadeb/home/health-econometrics-minicourse/

Overview

Statistical issues - skewness and the zero mass

Studies with skewed outcomes but no zeroes

Studies with zero mass and skewed outcomes

Studies with count data

Conclusions

Top Ten Urban Myths of Health Econometrics

1. **OLS is fine**

2. **Solution to being skewed to the right is to analyze**
   a. ***ln(y)* if *y > 0* or *ln(y+1)* if *y = 0***
   b. **Trimming is OK because things will be better behaved ; symmetric trimming is better**

3. **Heteroscedasticity is innocuous**
   a. **Does not affect predictions or marginal effects**
   b. **Can always be fixed by invoking sandwich estimator (Stata-ese)**

4. **The standard GLM should be gamma with log link**

5. **Counts are Poisson**

Top Ten Urban Myths of Health Econometrics (cont'd)

6. **Overdispersion implies that the correct model is NB (in Econ., NB2 specifically)**

7. **Zeroes always require 2-part models or hurdles**

8. **In-sample measures (R-squared) are perfectly fine for decision-making .**

9. **Model selection by citation is cost-effective and safe (as in the FDA sense)**

10. **"Robust" is Stata-ese often used as a shield (protection against zombie reviewers, editors, and thesis committees) - against a host of econometric illnesses. Not limited to just intracluster correction or sandwich estimators.**

The term "robust" means different things to different audiences (economists vs. statisticians) and to being out-of-sampling / resampling approaches. Use has been evolving and sometimes deals with forms of contamination or being resilient to….