

Cost as a dependent variable

Mark Bounthavong, PharmD, PhD

26 April 2023



VA



U.S. Department of Veterans Affairs
Veterans Health Administration
Health Services Research & Development Service

Poll # 1

What types of models have you used for cost data as an outcome (dependent) variable?

- A. Ordinary Least Squares (Linear Regression) Model
- B. Log-Transformed (Log-OLS) Model
- C. Generalized Linear Model
- D. Two-part model
- E. I have never modeled cost as an outcome before

Past presentations on cost as a dependent variable

Paul Barnett has done a two-part series on Cost As A Dependent Variable

Part 1 ([link](#))

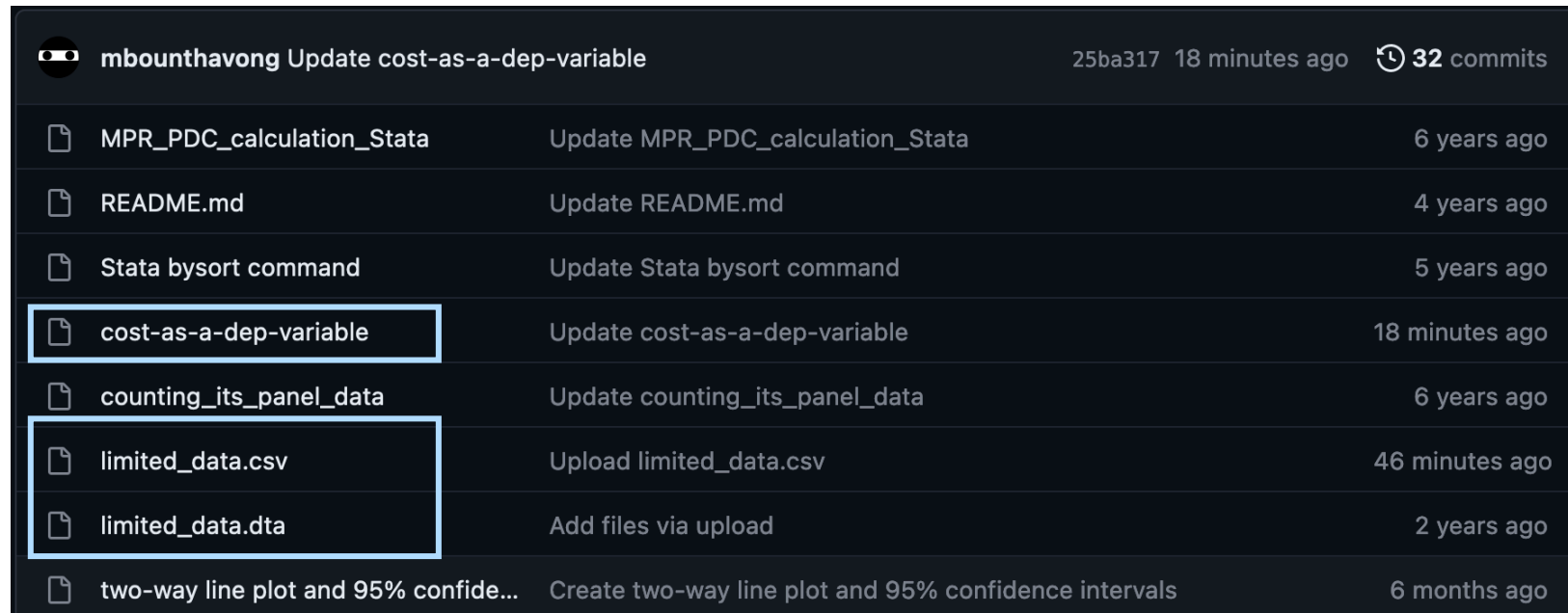
Part 2 ([link](#))

HERC Cyberseminars on Econometric Methods ([Past Sessions](#))

Files for this presentation are located on GitHub

This presentation includes files to perform the analysis with cost as a dependent variable

All files are located on [GitHub](#)



The screenshot shows a GitHub commit history for the repository 'mbounthavong Update cost-as-a-dep-variable'. The commit hash is 25ba317, made 18 minutes ago, with 32 commits in total. The table lists several files that have been updated or added, with the most recent commit (18 minutes ago) updating 'cost-as-a-dep-variable' and 'counting_its_panel_data', and another commit (46 minutes ago) uploading 'limited_data.csv'.

File	Commit Message	Time Ago
MPR_PDC_calculation_Stata	Update MPR_PDC_calculation_Stata	6 years ago
README.md	Update README.md	4 years ago
Stata bysort command	Update Stata bysort command	5 years ago
cost-as-a-dep-variable	Update cost-as-a-dep-variable	18 minutes ago
counting_its_panel_data	Update counting_its_panel_data	6 years ago
limited_data.csv	Upload limited_data.csv	46 minutes ago
limited_data.dta	Add files via upload	2 years ago
two-way line plot and 95% confide...	Create two-way line plot and 95% confidence intervals	6 months ago

Background

Cost distribution is usually skewed with thin right tails

Cost distribution also have a substantial density of zero values

Ordinary Least Squares (OLS) methods are insufficient

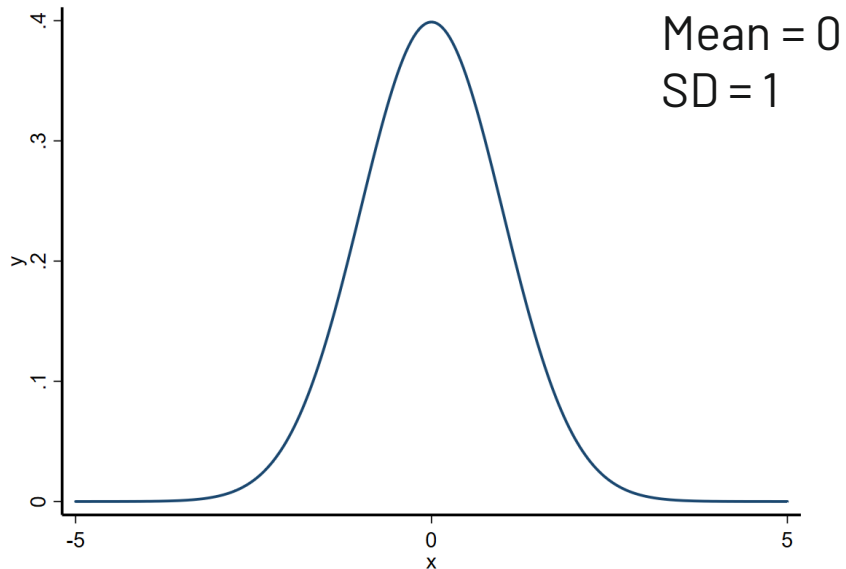
However, other methods take into account the skewness and large point mass at zero

We will explore alternative methods to OLS when modeling costs data as a dependent variable

Characteristics of data

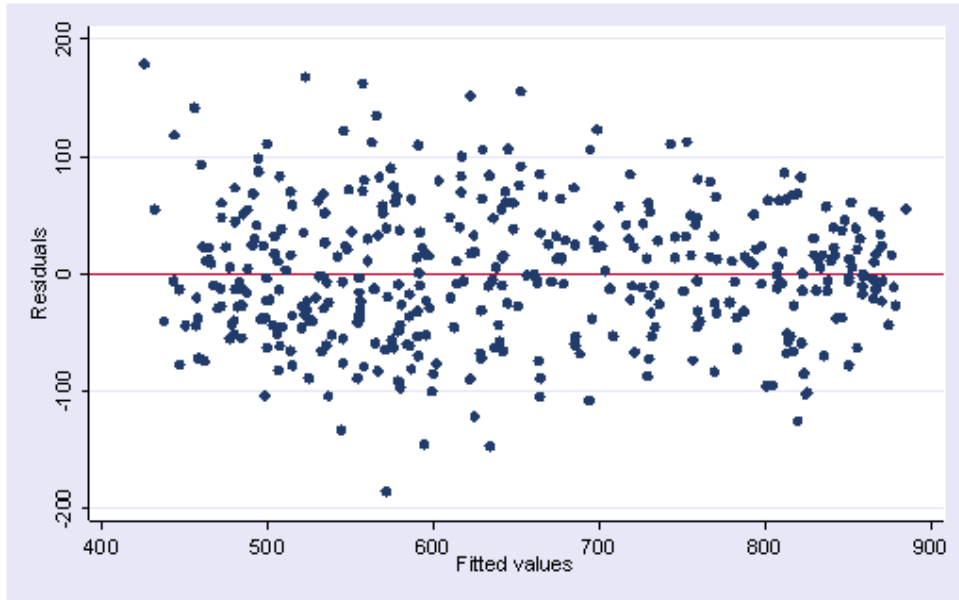
Skewness is a measure of how asymmetric a distribution is around its mean (skewness = 0)

Kurtosis is a measure of how heavy the tail ends of the distributions are (kurtosis = 3)

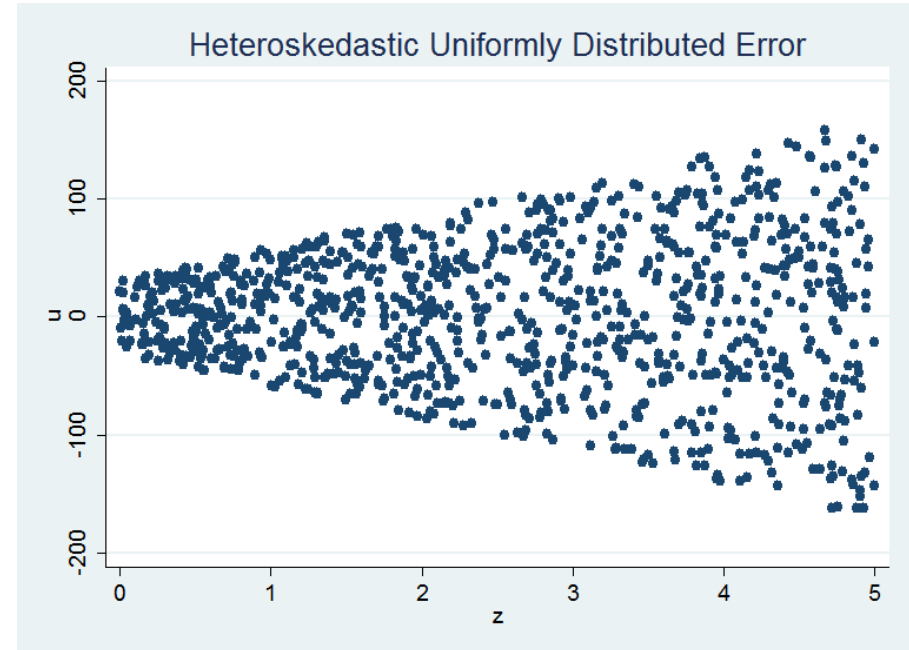


```
**** Plot a normal distribution with x = 0 and sd = 1
graph twoway function y=normalden(x,0,1),
range(-5 5) lw(medthick) legend(off)
xscale(lw(medthick)) yscale(lw(medthick))
graphregion(color(white)) bgcolor(white) ylabel(,
nogrid)
```

Checking Homoscedasticity of Residuals



No pattern to the residuals plotted against the fitted values (\hat{Y})



Variance in the residuals increases with the mean (\hat{Y})

Motivating Example: Total expenditures, MEPS 2017 (1)

Data can be downloaded from [MEPS](#) or [GitHub](#)

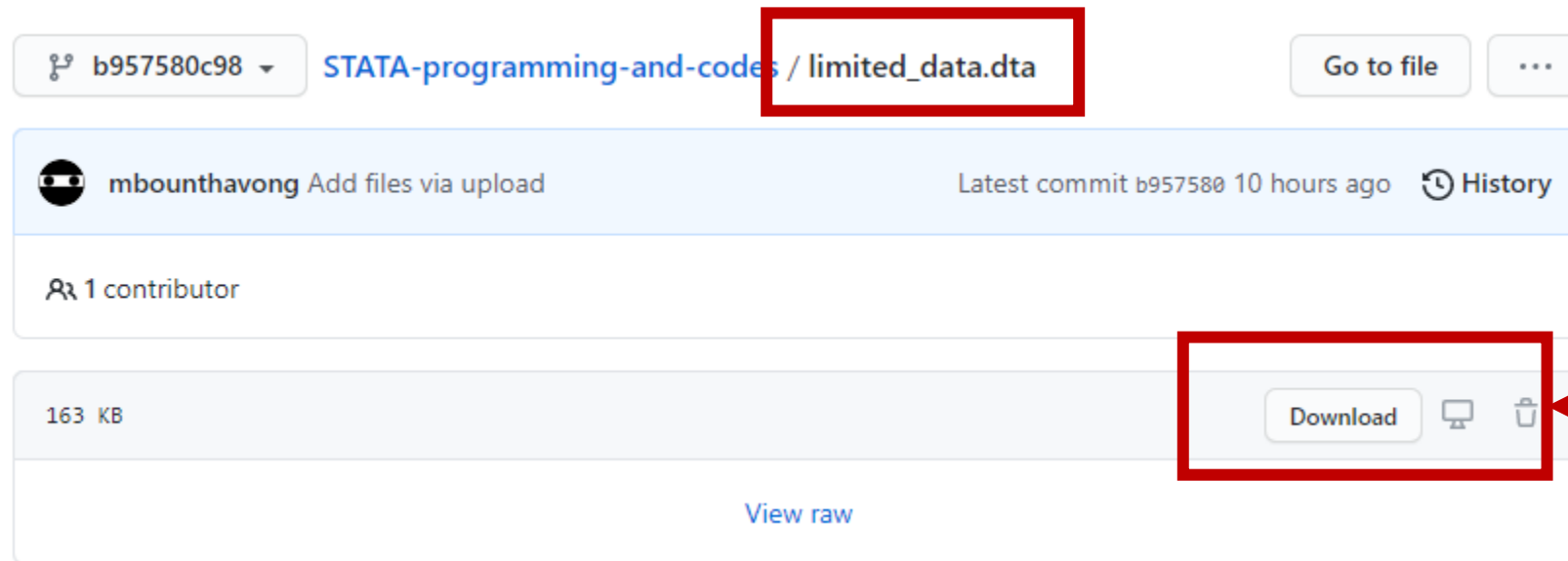
We will use Stata SE version 15 for this exercise

Several ways to download / load example data

Motivating Example: Total expenditures, MEPS 2017 (2)

Method 1: Download data from [GitHub](#) and Load into Stata

For WINDOWS users



The screenshot shows a GitHub file page for the repository 'STATA-programming-and-codes'. The file 'limited_data.dta' is highlighted with a red box. Below the repository header, the file size is listed as '163 KB'. The 'Download' button and its associated icons are also highlighted with a red box. A red arrow points from the text 'Download data onto your computer' to the 'Download' button.

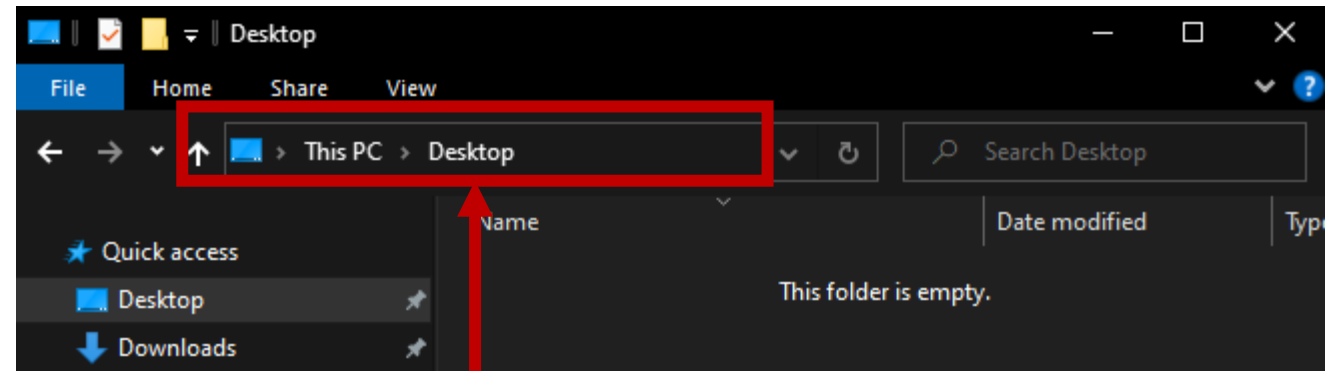
Download data onto your computer

Motivating Example: Total expenditures, MEPS 2017 (3)

Save data and modify the code to the Windows path location

```
***** FOR WINDOWS:  
clear all  
cd "[INSERT FILE LOCATION]"  
use limited_data.dta
```

**Windows file path uses the file explorer
(Make sure to include quotations)**



Motivating Example: Total expenditures, MEPS 2017 (5)

Method 2: Import CSV data from [GitHub](#) directly into Stata

For WINDOWS/MAC users

```
***** FOR WINDOWS or MAC:
```

```
clear all
```

```
import delimited "https://raw.githubusercontent.com/mbounthavong/STATA-programming-and-codes/master/limited_data.csv"
```

Motivating Example: Total expenditures, MEPS 2017 (6)

Goal: To evaluate the average total healthcare expenditures among household respondents diagnosed with high blood pressure

Methods: Use different regression models; Control for baseline demographics (e.g., age, gender, race, ethnicity, poverty status, marital status, and census region)

Motivating Example: Total expenditures, MEPS 2017 (7)

Notations:

Y = Cost

X_i = Independent variables (X_1, X_2, \dots, X_n)

β_i = Coefficients

Analytic Plan:

Models (OLS, Log-OLS, Log-OLS with smearing, GLM, and two-part models)

Goodness of Fit (GOF) tests

Compare mean healthcare expenditures

Goodness of Fit (GOF) tests

Pearson correlation: Correlation between raw scale cost predictions and residuals costs

Pregibon's Link test: Run the same outcome model with XB and XB^2 as covariates. If NS, then the regression equation is properly specified and there are no additional independent variables that are significant except by chance

Hosmer-Lemeshow test:

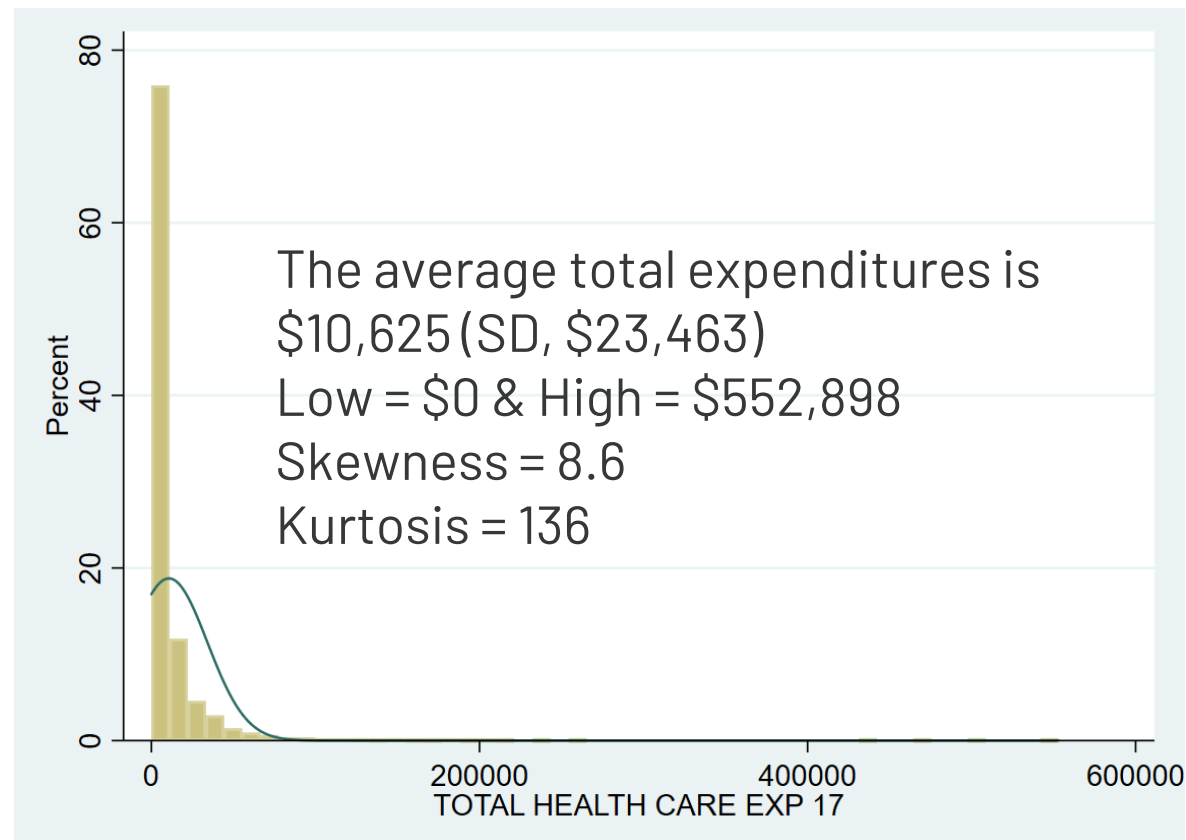
- (1) Plot residuals across deciles of XB
- (2) Joint test to examine whether the mean residuals are zero

Data description: Total expenditures, MEPS 2017

```
. summarize totexp17, detail
```

TOTAL HEALTH CARE EXP 17

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	161	0	Obs	7,872
25%	953.5	0	Sum of Wgt.	7,872
50%	3517		Mean	10625.1
		Largest	Std. Dev.	23462.3
75%	10549.5	474178		
90%	26858	499286	Variance	5.50e+08
95%	43280	506064	Skewness	8.581631
99%	105557	552898	Kurtosis	136.4949



Model 1: OLS (Linear regression)

$$E[Y|X] = \beta_0 + \beta_i(X_i) + \varepsilon$$



X = age, gender, race, ethnicity, poverty status, marital status, and census region

Linear models provide easy interpretation of the coefficients

However, because of the high skewness, any differences in the tails can have a great affect on the mean

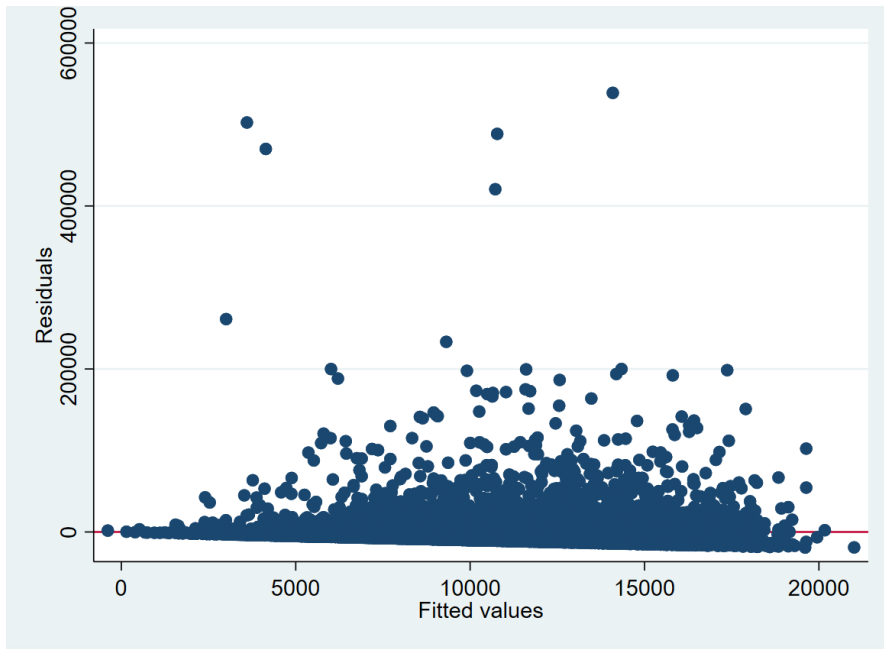
Generates biased estimations due to the non-linearity of Y

Heteroscedasticity (variance increases with mean) generates inefficient standard errors

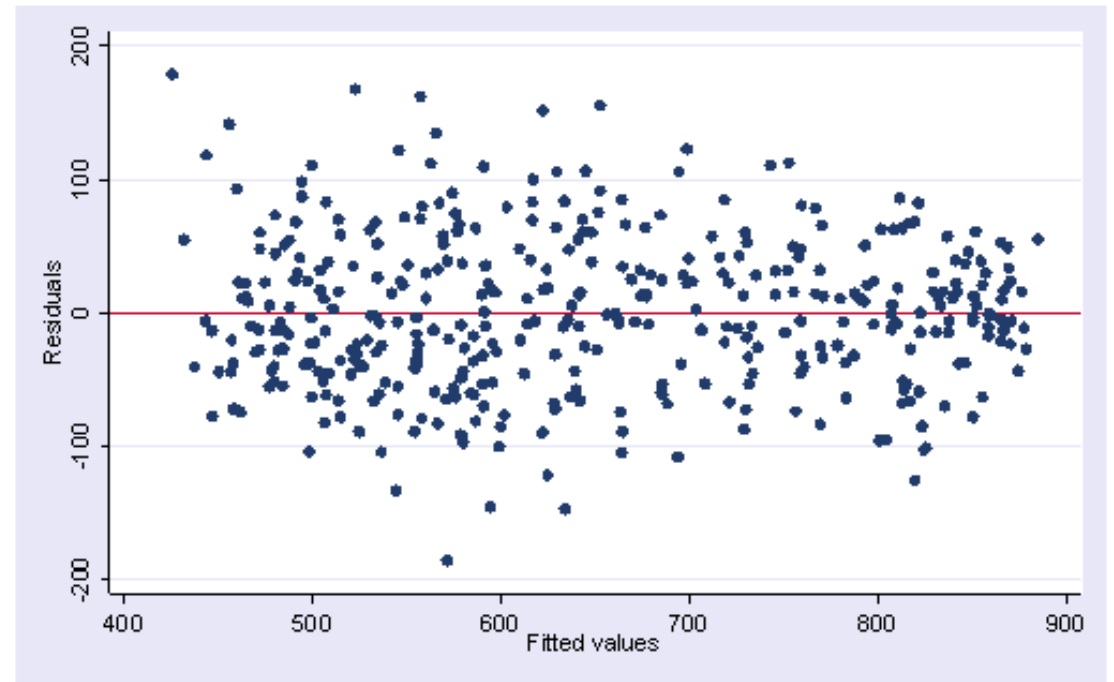
Model 1: OLS (Linear regression)

**** MODEL 1: OLS

```
reg totexp17 age17x sex racev2x hispanx marry17x povcat17 region17
predict yhat /* get the fitted values */
predict error, resid /* get the residuals */
graph twoway scatter error yhat /* plot the residuals to the fitted value */
```



≠



Poll # 2

How different is the OLS regression mean total expenditure compared to the raw mean total expenditure?

- A. OLS regression mean is higher than the raw mean
- B. OLS regression mean is lower than the raw mean
- C. Both means are exactly the same

Model 1: OLS (Linear regression)

```
summarize totexp17 yhat
```

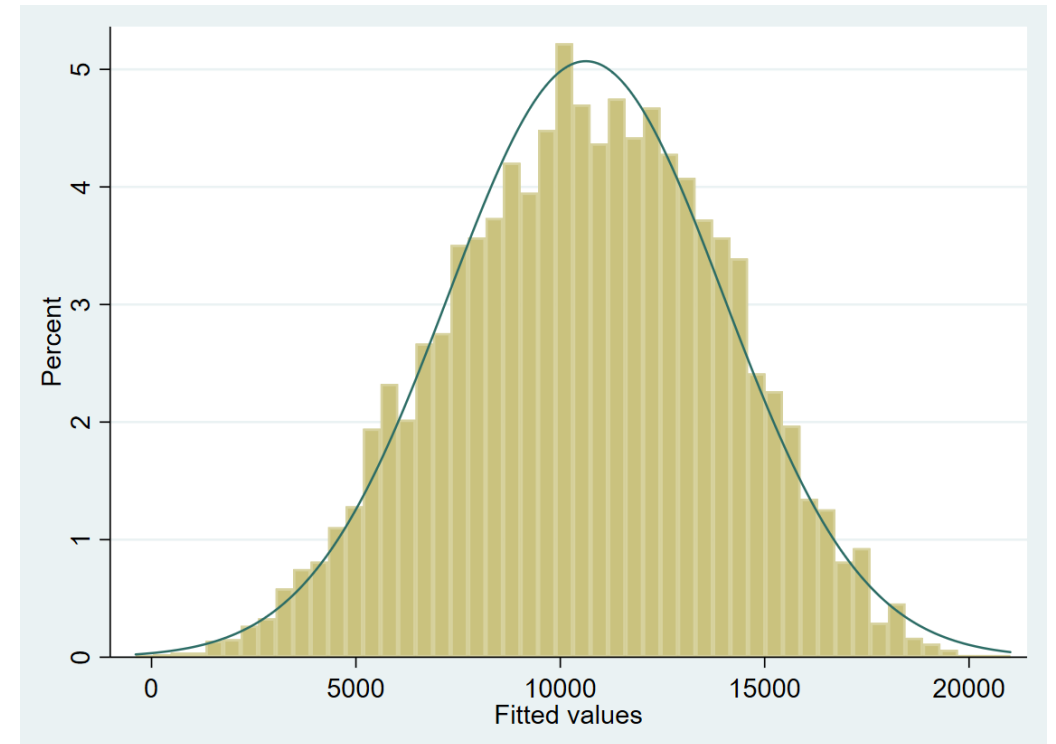
Variable	Obs	Mean	Std. Dev.	Min	Max
totexp17	7,872	10625.1	23462.3	0	552898
yhat	7,872	10625.1	3367.796	-387.9797	21010.64

**Mean costs are the same
But variances are different**

```
. summarize yhat, detail
```

Fitted values

Percentiles	Smallest		
1%	3030.689	-387.9797	
5%	4935.732	147.7056	
10%	6103.418	395.8278	Obs 7,872
25%	8270.418	518.8014	Sum of Wgt. 7,872
50%	10704.85		Mean 10625.1
			Std. Dev. 3367.796
75%	13082.41	19636.35	
90%	14975.01	19948.82	Variance 1.13e+07
95%	15997.13	20169.22	Skewness -.1009199
99%	17779.77	21010.64	Kurtosis 2.585958

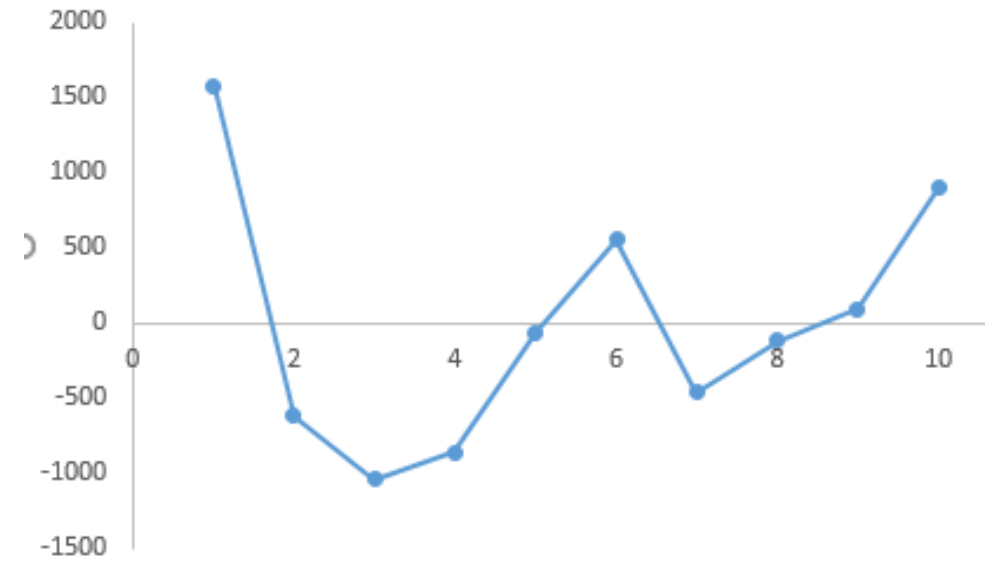


GOF tests: Model 1 (OLS)

Pearson correlation: No correlation between residuals and predicted costs (P = NS)

Pregibon's Link test: Significant association between xb^2 and outcomes (P = 0.003)

Hosmer-Lemeshow test: No significant differences in the mean residuals (P = 0.549)



Comparison: OLS model versus Raw Costs

Features	Raw	OLS
Mean	10,625.10	10,625.10
SD	23,462.30	3,367.80
Min	0.00	-387.98
Max	552,898.00	21,010.64
Median	3,517.00	10,704.85

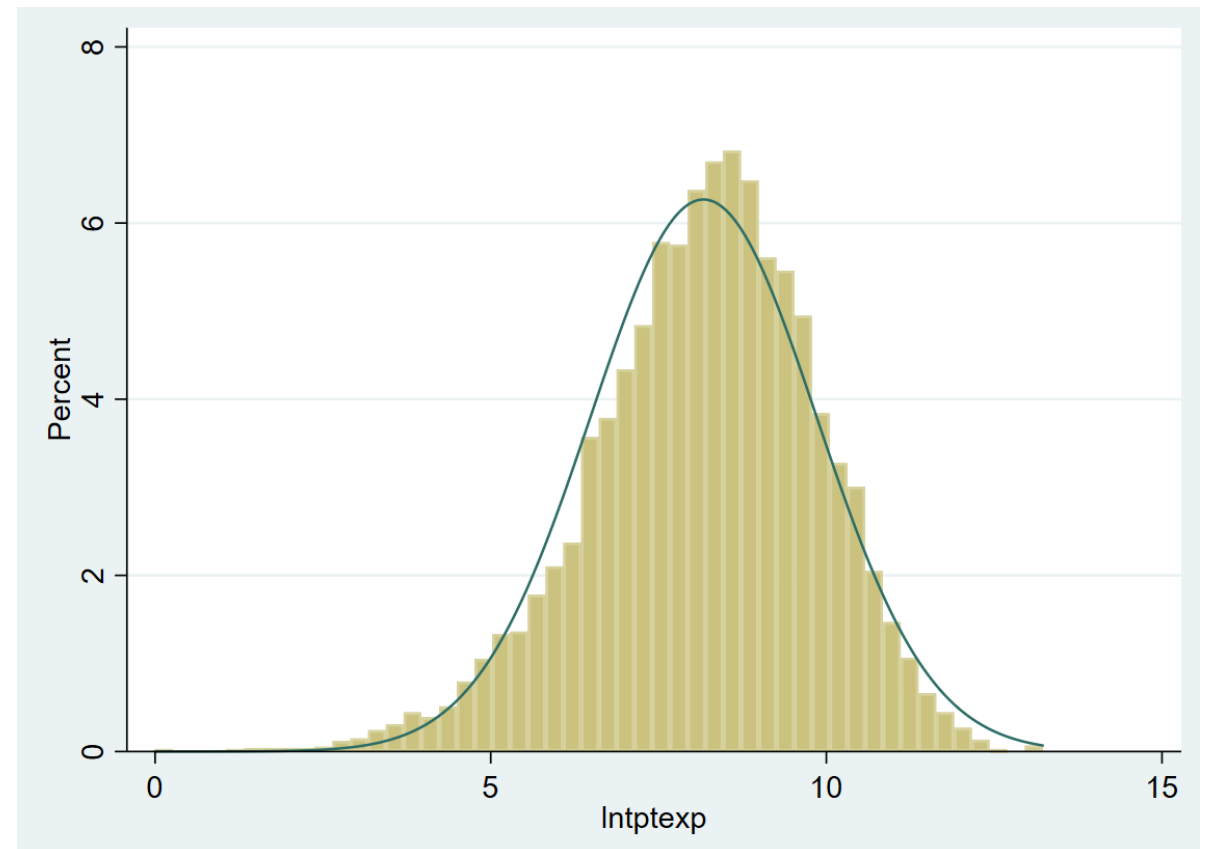
Model 2: Log transformation (Log-OLS)

Log transformation of the cost data can reduce skewness

Log dollars is not easy to interpret

```
. summarize Intptexp, detail
```

Intptexp				
Percentiles		Smallest		
1%	3.688879	0		
5%	5.204007	0		
10%	5.97381	1.098612	Obs	7,419
25%	7.144407	1.098612	Sum of Wgt.	7,419
50%	8.295299		Mean	8.170751
		Largest	Std. Dev.	1.683151
75%	9.339437	13.06934		
90%	10.24658	13.12093	Variance	2.832997
95%	10.71213	13.13442	Skewness	-.4584664
99%	11.58733	13.22293	Kurtosis	3.471834



Model 2: Log transformation (Log-OLS)

$$E[\ln(Y)|X] = \beta_0 + \beta_i(X_i) + \varepsilon$$

$$E[Y|X] = e^{\beta_0 + \beta_i(X_i) + E[\varepsilon]}$$

Expectation of the $\ln(y)$ is not the $\ln[E(y)]$

*** MODEL 2: Log-OLS

```
reg lntptexp age17x sex racev2x  
hispanx marry17x povcat17 region17
```

```
predict lh_yhat, xb  
gen exp_lnyhat = exp(lh_yhat)  
summarize exp_lnyhat, detail
```

```
. summarize exp_lnyhat, detail
```

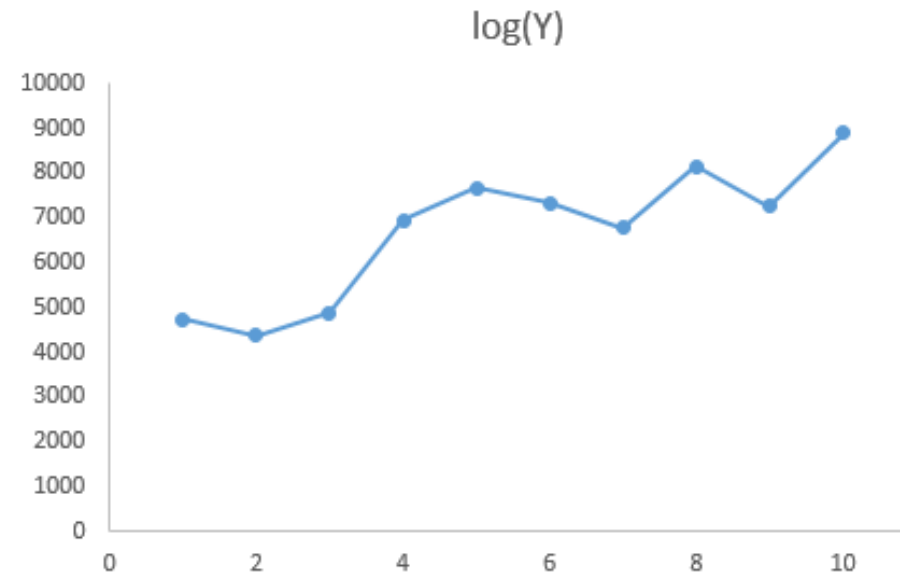
		exp_lnyhat	
		Percentiles	Smallest
1%	984.8097	625.8608	
5%	1401.938	628.6244	
10%	1678.391	650.7963	Obs 7,872
25%	2416.036	665.4033	Sum of Wgt. 7,872
50%	3599.096		Mean 3919.371
		Largest	Std. Dev. 1971.276
75%	4957.952	11840.65	
90%	6731.513	11840.65	Variance 3885930
95%	7754.673	11924.08	Skewness .9096912
99%	9726.771	12892.53	Kurtosis 3.629208

GOF tests: Model 2 (Log-OLS)

Pearson correlation: Significant correlation between residuals and predicted costs ($P < 0.001$)

Pregibon's Link test: Significant association between xb^2 and outcomes ($P = 0.028$)

Hosmer-Lemeshow test: Significant differences in the mean residuals ($P < 0.001$)



Comparison: Log-OLS versus OLS & Raw Costs

Features	Raw	OLS	Log-OLS
Mean	10,625.10	10,625.10	3,919.37
SD	23,462.30	3,367.80	1,971.28
Min	0.00	-387.98	625.86
Max	552,898.00	21,010.64	12,892.53
Median	3,517.00	10,704.85	3,599.10

Model 3: Log transformation (Log-OLS) w/ smearing

$$\left. \begin{aligned} E[Y|X] &= e^{\beta_0 + \beta_i(X_i) + E[\varepsilon]} \\ E[Y|X] &= e^{\beta_0 + \beta_i(X_i) * s} \end{aligned} \right\} \text{S is the smearing factor}$$

Duan's smearing estimator corrects for the retransformation issue with the log-OLS model

Duan's smearing estimator:

$$\ln(Y) = XB + e$$

$$Y = \exp(XB + e)$$

$$Y = \exp(XB) * \exp(e)$$

$$s = \exp(e)$$

$$s = \exp(\ln(Y) - XB)$$

```
**** MODEL 3: Log-OLS w/smearing
```

```
reg lntptexp age17x sex racev2x hispanx marry17x povcat17
```

```
* Smearing estimator
```

```
gen smr = exp(lntptexp - lh_yhat)
```

```
summarize smr
```

```
gen smear = r(mean)
```

```
gen mu_lols = exp(lh_yhat) * smear
```

```
gen res_lols = totexp17 - mu_lols
```

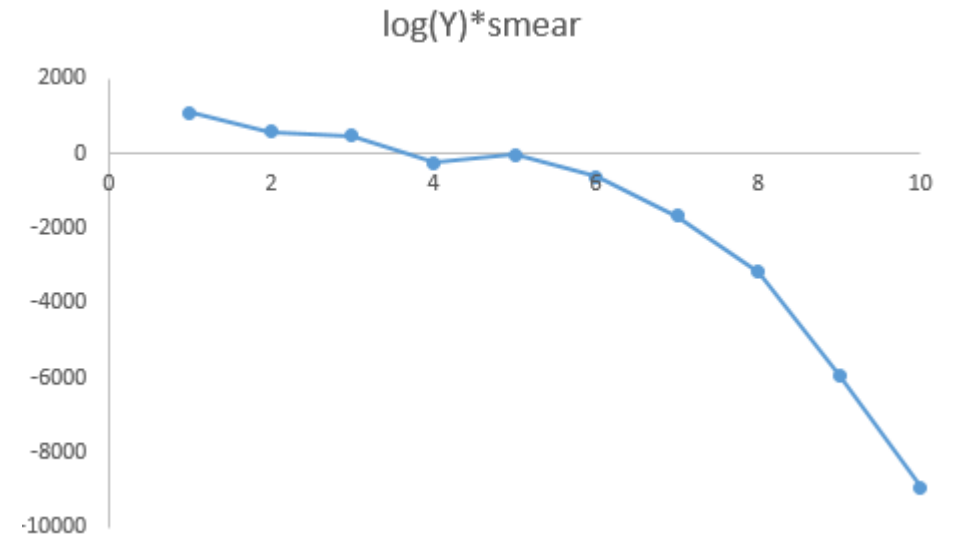
```
summarize mu_lols, detail
```

GOF tests: Model 3 (Log-OLS with smearing)

Pearson correlation: Significant correlation between residuals and predicted costs ($P < 0.001$)

Pregibon's Link test: Significant association between xb^2 and outcomes ($P=0.018$)

Hosmer-Lemeshow test: Significant differences in the mean residuals ($P < 0.001$)



Comparison: Log-OLS w/ smear versus Log-OLS, OLS, & Raw Costs

Features	Raw	OLS	Log-OLS	Log-OLS w/ smearing
Mean	10,625.10	10,625.10	3,919.37	12,462.22
SD	23,462.30	3,367.80	1,971.28	6,267.96
Min	0.00	-387.98	625.86	1,990.02
Max	552,898.00	21,010.64	12,892.53	40,993.70
Median	3,517.00	10,704.85	3,599.10	11,443.86

Model 4: Generalized Linear Model (GLM)

$$\left. \begin{aligned} g(E[Y|X]) &= \beta_0 + \beta_i(X_i) + \varepsilon \\ \ln(E[Y|X]) &= \beta_0 + \beta_i(X_i) + \varepsilon \\ E[Y|X] &= e^{\beta_0 + \beta_i(X_i) + \varepsilon} \end{aligned} \right\}$$

Rather than transform the raw Y, we are transforming the E(Y)
 $\ln(u) = XB$ or $u = \exp(XB)$

GLM uses a link function, $g(\cdot)$

Retransformation is not a problem

Apply a link function to the expectation of Y instead of the raw Y

Family	Link
Gaussian	identity
Binomial	logit, probit, cloglog
Poisson	identity, log, sqrt
Gamma	inverse, identity, log
Inverse Gaussian	inverse squared

Model 4: Generalized Linear Model (GLM)

Family selection is based on the relationship between $\text{Var}[Y|X]$ and $E[Y|X]$

$$\text{Var}[y|x] = \alpha * (E[y|x])^\gamma$$

For $\gamma = 0$ use Gaussian (aka nonlinear least squares; constant variance)

For $\gamma = 1$ use Poisson (variance is proportional to the mean)

For $\gamma = 2$ use Gamma (variance is proportional to the square of the mean)

For $\gamma = 3$ use Wald or inverse Gaussian

Link selection is based on Pregibon's link test

Modified Hosmer-Lemeshow test to assess structural fit

Model 4: Generalized Linear Model (GLM)

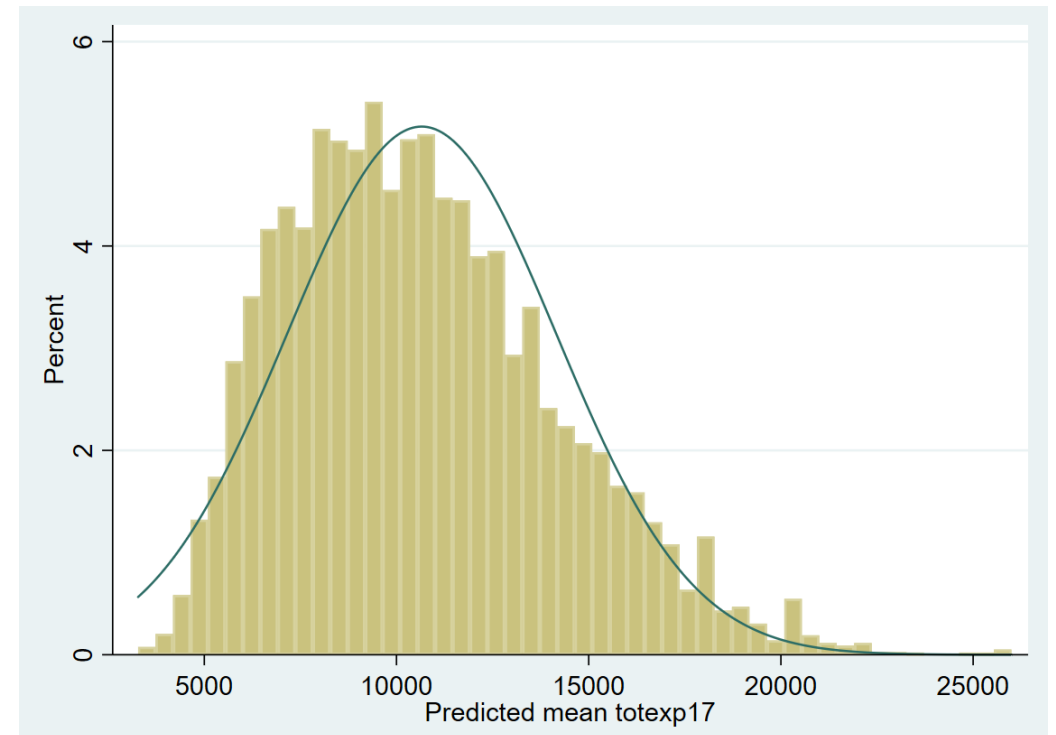
**** MODEL 4: GLM-log (gamma)

```
glm totexp17 age17x sex racev2x hispanx marry17x povcat17, family(gamma) link(log)
predict glm_1
summarize glm_1, detail
```

```
. summarize glm_1, detail
```

Predicted mean totexp17

Percentiles		Smallest		
1%	4710.048	3279.001		
5%	5730.73	3357.663		
10%	6423.879	3477.815	Obs	7,872
25%	8028.091	3578.414	Sum of Wgt.	7,872
50%	10269.17		Mean	10655.88
		Largest	Std. Dev.	3506.746
75%	12816.96	25942.63		
90%	15524.53	25995.67	Variance	1.23e+07
95%	17103.85	25995.67	Skewness	.6116797
99%	20262.99	25995.67	Kurtosis	3.127449

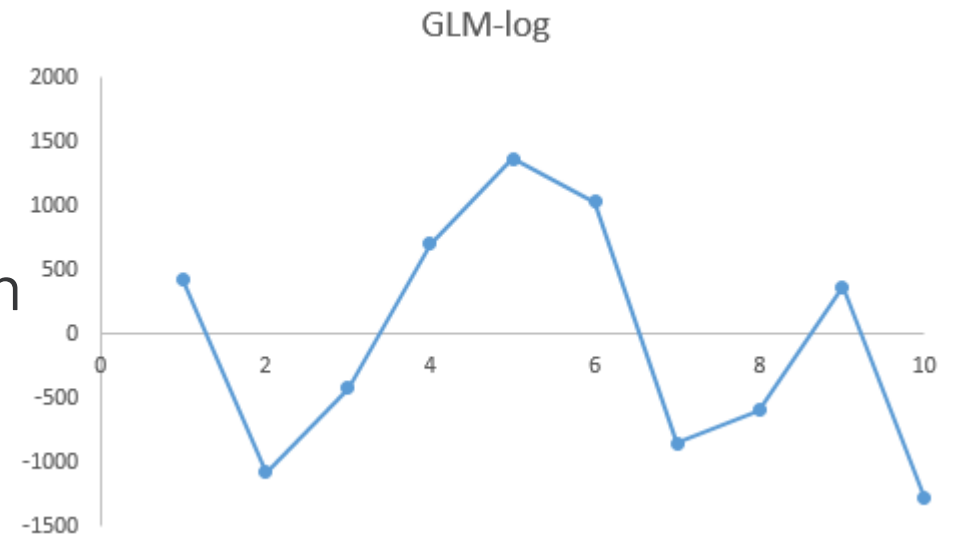


GOF tests: Model 4 (GLM-log)

Pearson correlation: No correlation between residuals and predicted costs ($P = 0.276$)

Pregibon's Link test: No association between xb^2 and outcomes ($P=0.406$)

Hosmer-Lemeshow test: No differences in the mean residuals ($P = 0.182$)



Comparison: GLM-log, Log-OLS w/ smear, Log-OLS, OLS, & Raw Costs

Features	Raw	OLS	Log-OLS	Log-OLS w/ smearing	GLM-log
Mean	10,625.10	10,625.10	3,919.37	12,462.22	10,655.88
SD	23,462.30	3,367.80	1,971.28	6,267.96	3,506.75
Min	0.00	-387.98	625.86	1,990.02	3,279.00
Max	552,898.00	21,010.64	12,892.53	40,993.70	25,995.67
Median	3,517.00	10,704.85	3,599.10	11,443.86	10,269.17

Model 5: Two-Part model

$$E[Y|X] = P(Y > 0|X) * E[Y|Y > 0, X] + P(Y = 0) * E[Y|Y = 0]$$

[First part] * *[Second part]*

First part: logit or probit

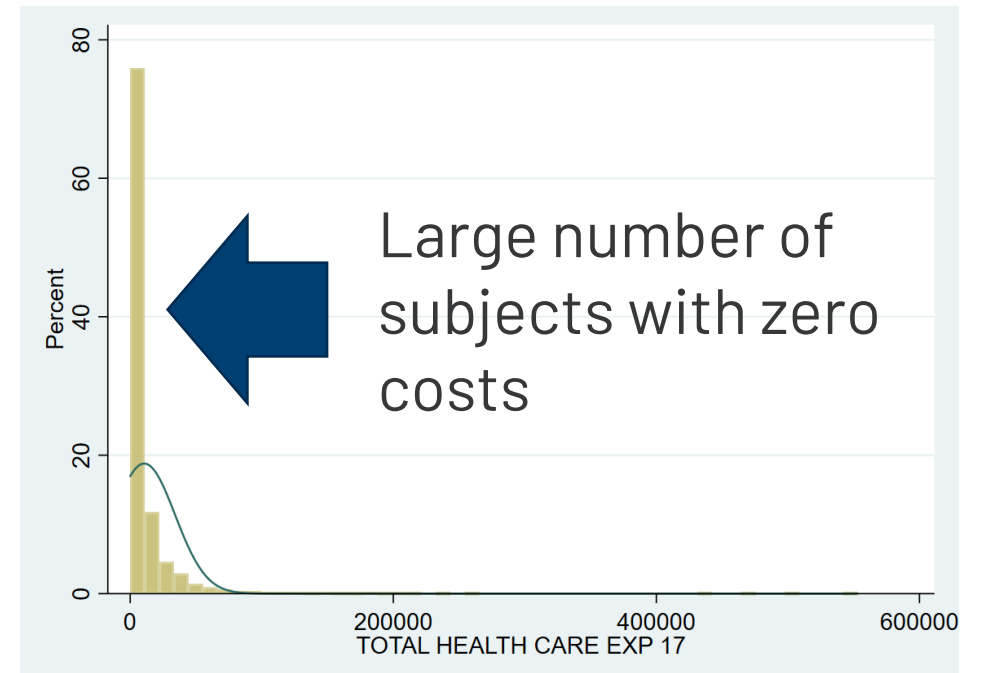
Second part: GLM (gamma dist & log link)

Point mass of subjects with zero costs

Expected value of Y is conditioned on whether the subject has non-zero costs

$P(Y > 0)$ is determined by the logit/probit part

$E[Y|Y > 0]$ is provided by the second part



Model 5: Two-Part model

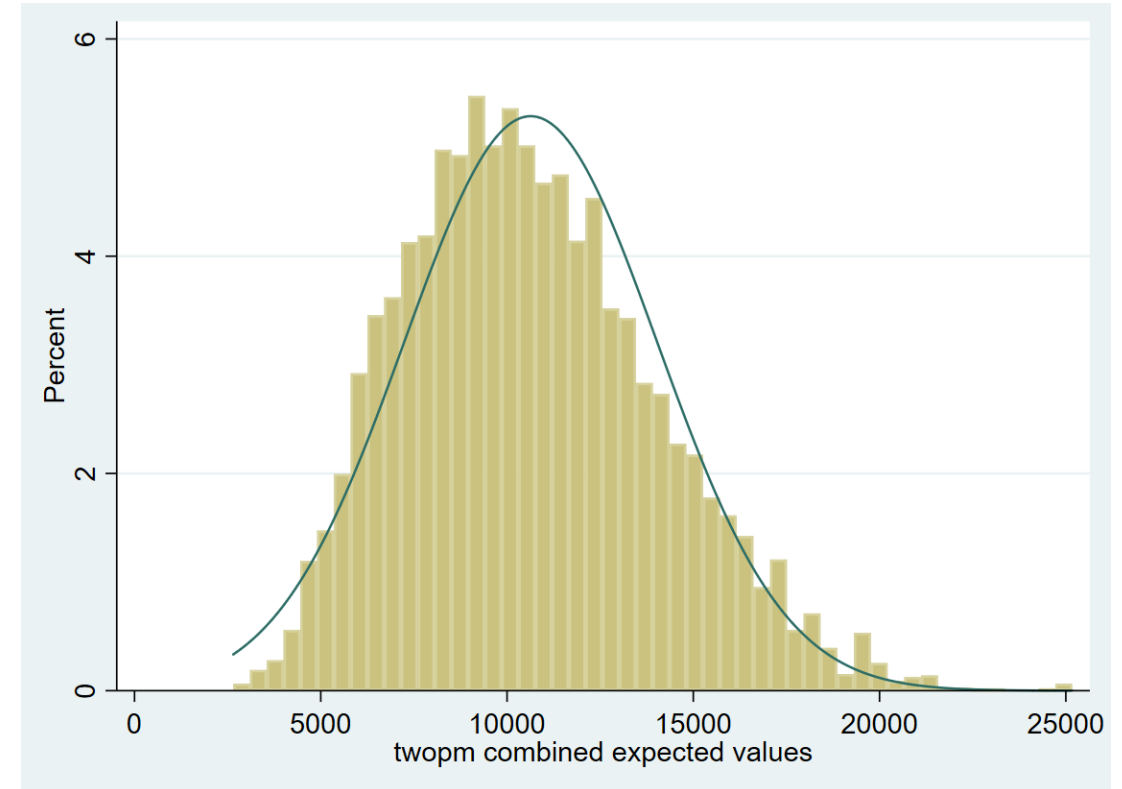
**** MODEL 5: two-part model

```
twopm totexp17 age17x sex racev2x hispanx marry17x povcat17, firstpart(logit)
secondpart(glm, family(gamma) link(log))
predict twopm_xb
summarize twopm_xb, detail
```

```
. summarize twopm_xb, detail
```

twopm combined expected values

Percentiles		Smallest		
1%	4401.528	2651.801		
5%	5663.923	2676.016		
10%	6458.538	2893.538	Obs	7,872
25%	8131.685	2924.518	Sum of Wgt.	7,872
50%		10319.44	Mean	10635.29
		Largest	Std. Dev.	3396.645
75%	12837.85	25102.52		
90%	15261.22	25168.32	Variance	1.15e+07
95%	16656.95	25168.32	Skewness	.4749402
99%	19451.83	25168.32	Kurtosis	3.020862

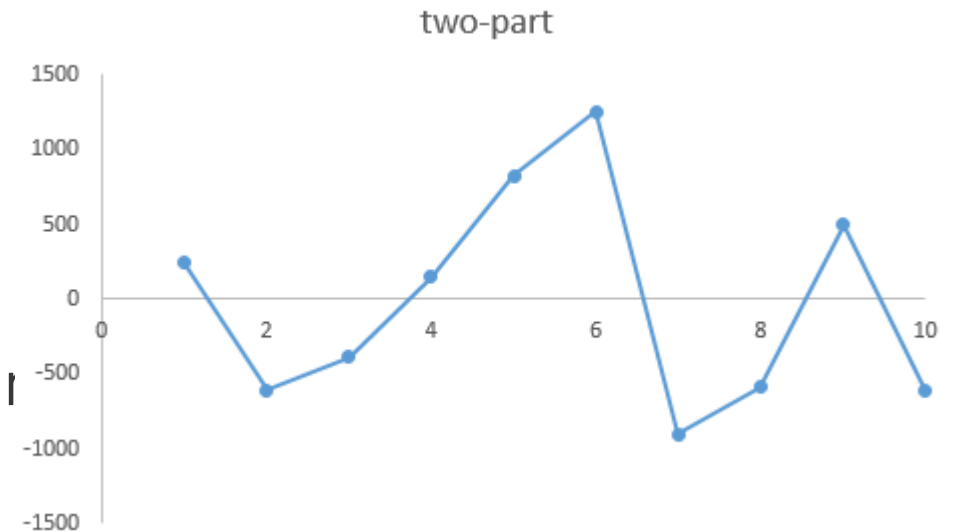


GOF tests: Model 5 (two-part model)

Pearson correlation: No correlation between residuals and predicted costs ($P = 0.591$)

Pregibon's Link test: No association between xb^2 and outcomes ($P = 0.296$)

Hosmer-Lemeshow test: No differences in the mean residuals ($P = 0.658$)



Comparison: two-part, GLM-log, Log-OLS w/ smear, Log-OLS, OLS, & Raw Costs

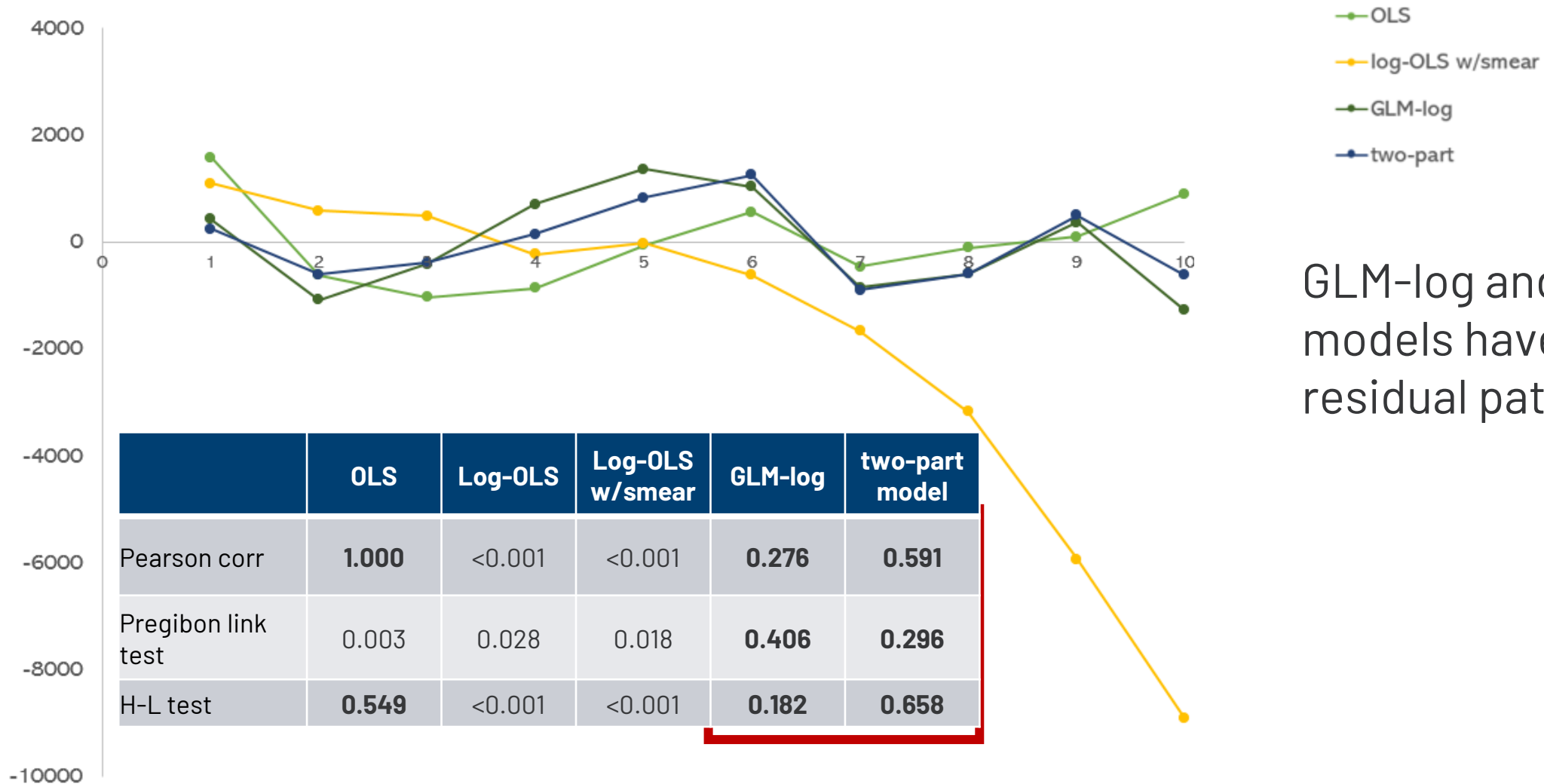
Features	Raw	OLS	Log-OLS	Log-OLS w/ smearing	GLM-log	Two-part
Mean	10,625.10	10,625.10	3,919.37	12,462.22	10,655.88	10,635.29
SD	23,462.30	3,367.80	1,971.28	6,267.96	3,506.75	3,396.65
Min	0.00	-387.98	625.86	1,990.02	3,279.00	2,651.80
Max	552,898.00	21,010.64	12,892.53	40,993.70	25,995.67	25,168.32
Median	3,517.00	10,704.85	3,599.10	11,443.86	10,269.17	10,319.44

Poll # 3

What model would you use for cost as an outcome?

- A. Ordinary Least Squares (Linear Regression) Model
- B. Log-Transformed (Log-OLS) Model
- C. Generalized Linear Model
- D. Two-part model

H-L test: residuals plotted on deciles



GLM-log and two-part models have the best residual patterns

References

GitHub repository of data and Stata codes ([link](#))

Manning WG. [The logged dependent variable, heteroscedasticity, and the retransformation problem](#). J Health Econ. 1998 Jun;17(3):283-95.

Manning WG, Mullahy J. [Estimating log models: to transform or not to transform?](#) J Health Econ. 2001 Jul;20(4):461-94

Basu A, Manning WG. [Issues for the next generation of health care cost analyses](#). Med Care. 2009 Jul;47(7 Suppl 1):S109-14.

Belotti F, Deb P, Manning WG, Norton EC. [Twopm: Two-Part Models](#). The Stata Journal. 2015;15(1):3-20.

References

Paul Barnett has done a two-part series on Cost As A Dependent Variable

Part 1 ([link](#))

Part 2 ([link](#))

Acknowledgements

Many of the codes were from lectures that I attended at the UW Advanced Methods Course Series.

These methods helped me to better understand the nuances associated with skewed data (e.g., costs and counts).

I recreated these codes for Stata as part of this presentation on modeling cost as a dependent variable.

Questions



Questions?

For more information visit
the HERC website at
www.herc.research.va.gov

Email us at HERC@va.gov

Call us at (650) 617-2630

