

# Econometrics Course: Cost as the Dependent Variable (II)

Paul G. Barnett, PhD

April 10, 2019



## POLL QUESTION #1

# **Which method(s) have you used to evaluate health care costs?**

**(answer all that apply)**

- None yet
- Rank test (non-parametric method)
- Ordinary Least Squares regression with raw cost
- OLS log transformed cost
- GLM model (gamma regression)

# **Health care costs difficult to analyze**

- Skewed by rare but extremely high cost events
- Zero cost incurred by enrollees who don't use care
- No negative values
- Variance can vary with independent variable

# Limitation of Ordinary Least Squares (OLS)

- OLS with raw cost
  - non-normal dependent variable can generate biased parameters
  - can predict *negative* costs
- OLS with log transformation of cost
  - Log cost is normally distributed, can use in OLS
  - Predicted cost is affected by re-transformation bias
  - Can't take log of zero
  - Assumes variance of errors is constant

# Topics for today's course

- What is heteroscedasticity, and what should be done about it?
  - What should be done when there are many zero values?
  - How to test differences in groups with no assumptions about distribution?
  - How to determine which model is best?
-

# Topics for today's course

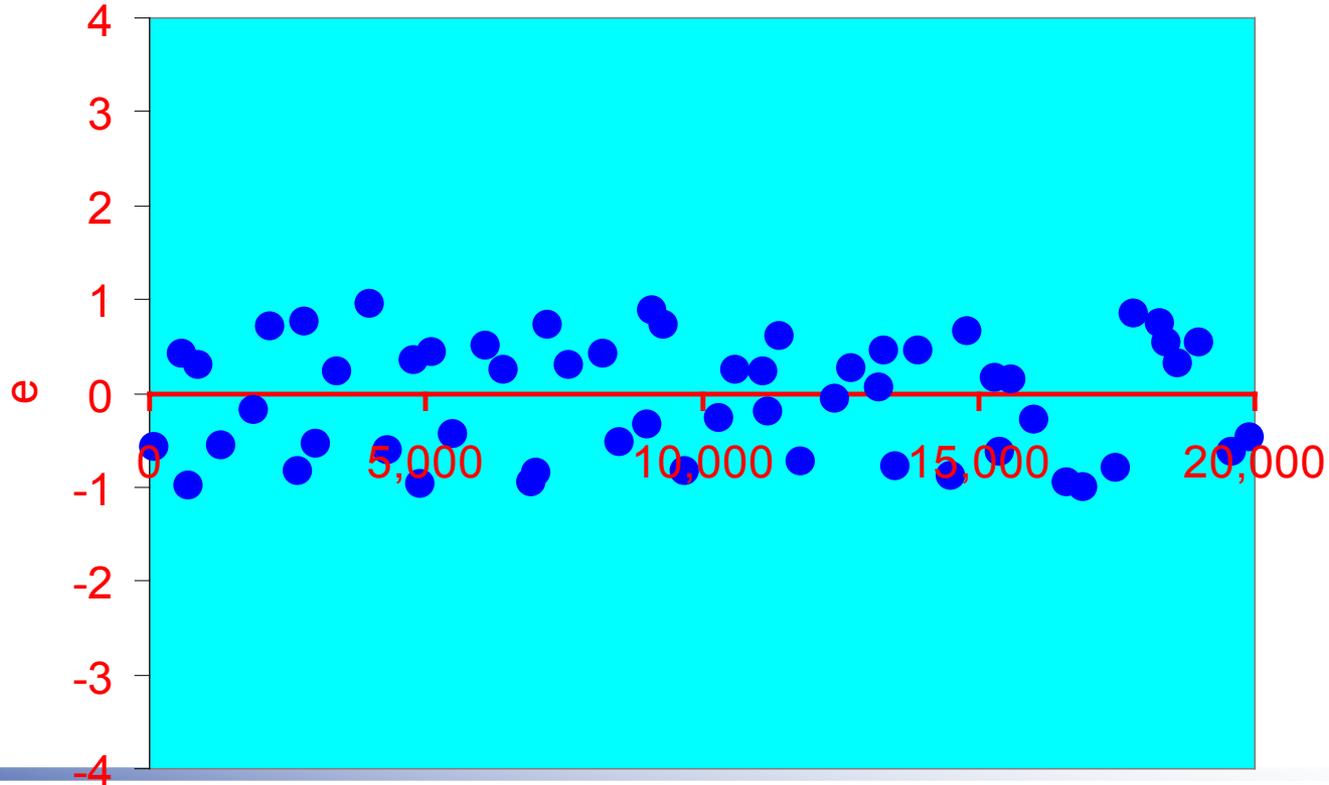
- What is heteroscedasticity and what should be done about it?
  - What should be done when there are many zero values?
  - How to test differences in groups with no assumptions about distribution?
  - How to determine which model is best?
-

# What is heteroscedasticity?

- Heteroscedasticity
  - Variance depends on  $x$  (or on predicted  $y$ )
  - For example, the variation in income increases with age
- OLS assumes Homoscedasticity
  - Identical variance  $E(\varepsilon_i^2) = \sigma^2$

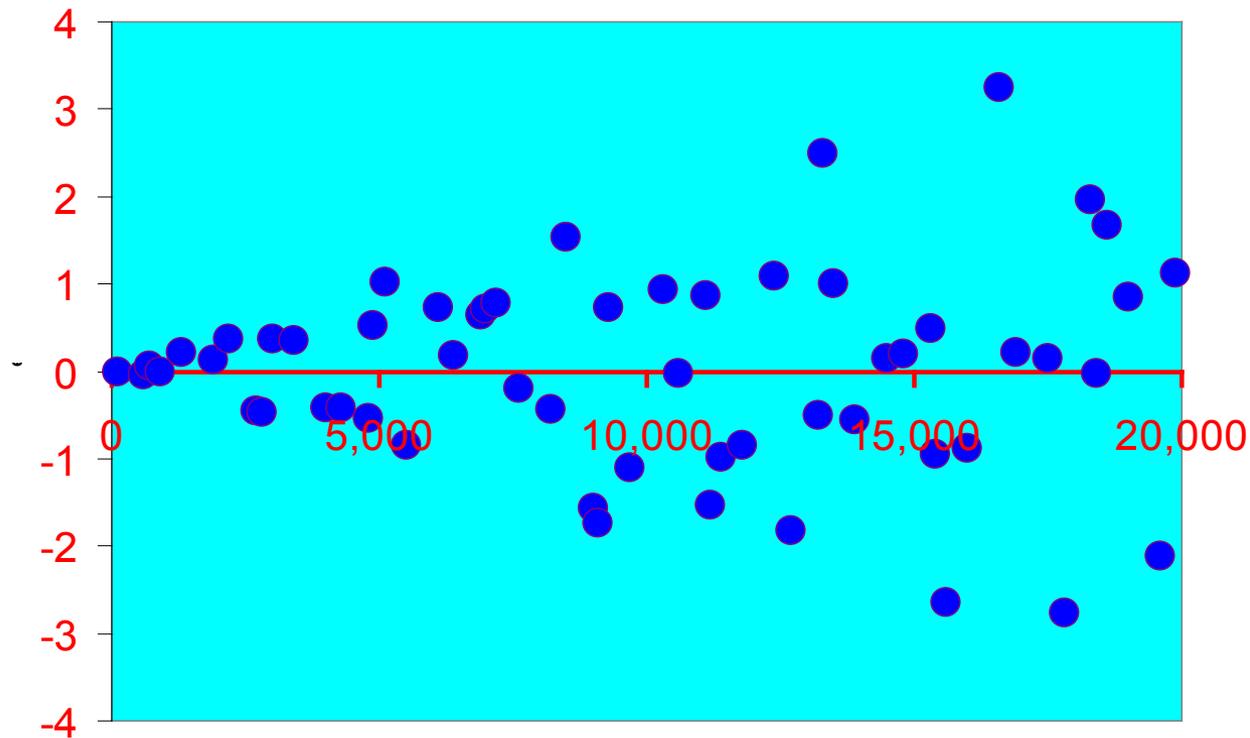
# Homoscedasticity

- Errors have identical variance  $E(\varepsilon_i^2) = \sigma^2$



# Heteroscedasticity

- Errors depend on  $x$  (or on predicted  $y$ )



# Why worry about heteroscedasticity?

- Predictions based on OLS model can be biased
- Re-transformation assumes homoscedastic errors
- Predicted cost when the error is heteroscedastic can be “appreciably biased”

# What should be done about heteroscedasticity?

- Use a Generalized Linear Models (GLM)
- Analyst specifies a link function  $g(\cdot)$
- Analyst specifies a variance function
  - Key reading: “Estimating log models: to transform or not to transform,” Mullahy and Manning J Health Econ 20:461, 2001

# Link function $g(\cdot)$ in GLM

- $g(E(y | x)) = \alpha + \beta x$
- Link function can be natural log, square root, or other function
  - E.g.  $\ln(E(y | x)) = \alpha + \beta x$
  - When link function is natural log, then  $\beta$  represents percent change in  $y$  for a unit change in  $x$

# GLM vs. OLS

- OLS of log estimate:  $E(\ln(y) | x)$
- GLM estimate:  $\ln(E(y | x))$ 
  - Log of expectation of  $y$  is not the same as expectation of  $\log y$ !

# GLM advantages

- Dependent variable can be zero
- No retransformation bias when predicting
  - Smearing estimator is not used
- Does not assume homoscedastic errors

# GLM variance function

- GLM does not assume constant variance
- GLM assumes there is function that explains the relationship between the variance and mean
  - $\text{var}(y | x)$

# Variance assumptions for GLM cost models

- Gamma Distribution (most common)
  - Variance is proportional to the square of the mean
- Poisson Distribution
  - Variance is proportional to the mean

# Estimation methods

- How to specify log link and gamma distribution with dependent variable COST and independent variables  $X_1$ ,  $X_2$ ,  $X_3$

# GLM with log link and gamma distribution in Stata

```
GLM COST X1 X2 X3, FAM(GAM)  
LINK(LOG)
```

# GLM with log link and gamma distribution in SAS

- Basic syntax (*drops* zero cost observations)

```
PROC GENMOD;  
MODEL COST=X1 X2 X3 / DIST=GAMMA LINK=LOG;
```

- Refined syntax (*keeps* zero cost observations)

```
PROC GENMOD;  
A = _MEAN_;  
B = _RESP_;  
D = B/A + LOG(A)  
VARIANCE VAR = A**2  
DEVIANCE DEV = D;  
MODEL COST=X1 X2 X3 / LINK=LOG;
```

# Choice between GLM and OLS of log cost

- GLM advantages:
  - Handles heteroscedasticity
  - Predicted cost is not subject to retransformation error
- OLS of log transform advantages
  - OLS is more efficient (standard errors are smaller than with GLM)

# Dataset for worked examples

- New primary care episodes of non-specific low-back pain in 2016
  - ~10% Sample N=43,909
  - VA and community care costs in the following year (excluding residential, nursing home)
  - Cost difference of episodes that started in CBOC?
-

# Worked example

- Try gamma regression (GLM log link, gamma distribution)
- Evaluate link function with Box-Cox regression
- Evaluate distribution with GLM family test

# GLM Regression

## Log link, gamma distribution

	Parameter Estimate	Standard Error	Wald Chi-sq	Pr >  t
Intercept	7.9401	0.0240	109230	<.0001
Index visit at CBOC	-0.2750	0.0106	675.9	<.0001
Baseline pain score	0.0452	0.0017	726.7	<.0001
Age	0.0026	0.0004	45.5	<.0001
Number of chronic conditions	0.1785	0.0036	2397.4	<.0001
Indicator of diagnosis for substance use or psychiatric diagnosis	0.2246	0.0128	308.2	<.0001
Female	0.1308	0.0208	39.5	<.0001
History of opiate Rx in prior year	0.1566	0.0168	87.3	<.0001
MRI within 6 weeks	0.4245	0.0335	160.7	<.0001
Number of visits for physical therapy within 6 weeks	0.0949	0.0120	62.8	<.0001

# Which GLM link function?

- Maximum likelihood estimation of Box-cox parameter (called  $\theta$  or  $\lambda$ )

$$\frac{COST^\theta - 1}{\theta} = \alpha + \beta x + \varepsilon$$

# Link function depends on Box-Cox parameter

Link function	Box-Cox parameter
Inverse (1/cost)	-1
Log(cost)	0
Square root (cost)	.5
Cost	1
Cost Squared	2

# Box Cox regression

## ■ Stata

**boxcox cost {indep. vars} if cost > 0**

## ■ SAS

```
proc transreg data={dataset} ;  
model boxcox(cost/lambda= -1 to 2 by .5)=  
identity (&ind_vars);
```

# Box-Cox regression with example data

Box-Cox Transformation Information for Cost in year after index stay				
Lambda		R-Square	Log Like	
-1.0		0.05	-409600	
-0.5		0.11	-369122	
0.0	+	0.15	-354795	<
0.5		0.12	-370396	
1.0		0.05	-418368	
1.5		0.01	-496072	
2.0		0.00	-596029	
< - Best Lambda				
* - 95% Confidence Interval				
+ - Convenient Lambda				

# Which variance structure with GLM?

- Is it appropriate to assume the gamma distribution?
  - GLM family test
    - Also called modified Park test
  - Run GLM gamma regression
  - Evaluate residuals with second regression
-

# GLM family test (step 1)

- Run a gamma regression
  - Assume log link, gamma variance
  - Include independent variables
- Find predicted cost (in log scale)
  - $X\beta$  from first regression
- Find residual (in raw scale)
  - $\text{COST} - e^{X\beta}$
- Square these residuals

# GLM family test (step 2)

- Run second gamma regression
  - Dependent variable is squared residuals in raw scale
  - Independent variable is predicted cost in log scale ( $X\beta$  from first regression)

$$(\text{COST} - e^{X\beta})^2 = \gamma_0 + \gamma_1(X\beta) + v$$

# GLM family test (step 3)

- Evaluate the regression coefficient  $\gamma_1$

$\gamma_1$	Variance to be used in GLM
0	Gaussian (Normal)
1	Poisson
2	Gamma
3	Wald (Inverse Normal)

# GLM Family Test with example data

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
$\gamma_0$	1.57	0.1846	72.1	<.0001
$\gamma_1$	1.98	0.0213	8587.6	<.0001

Result:

Since  $\gamma_1 \sim 2$ , the gamma distribution was appropriate

# Other models for skewed data

- Generalized gamma models
  - Estimate link function, distribution, and parameters in single model
  - STATA ado file “pglm”
  - See: Basu & Rathouz (2005)

**Questions?**

# Topics for today's course

- What is heteroscedasticity, and what should be done about it? (*GLM models*)
  - What should be done when there are many zero values?
  - How to test differences in groups with no assumptions about distribution?
  - How to determine which model is best?
-

# What should be done when there are many zero values?

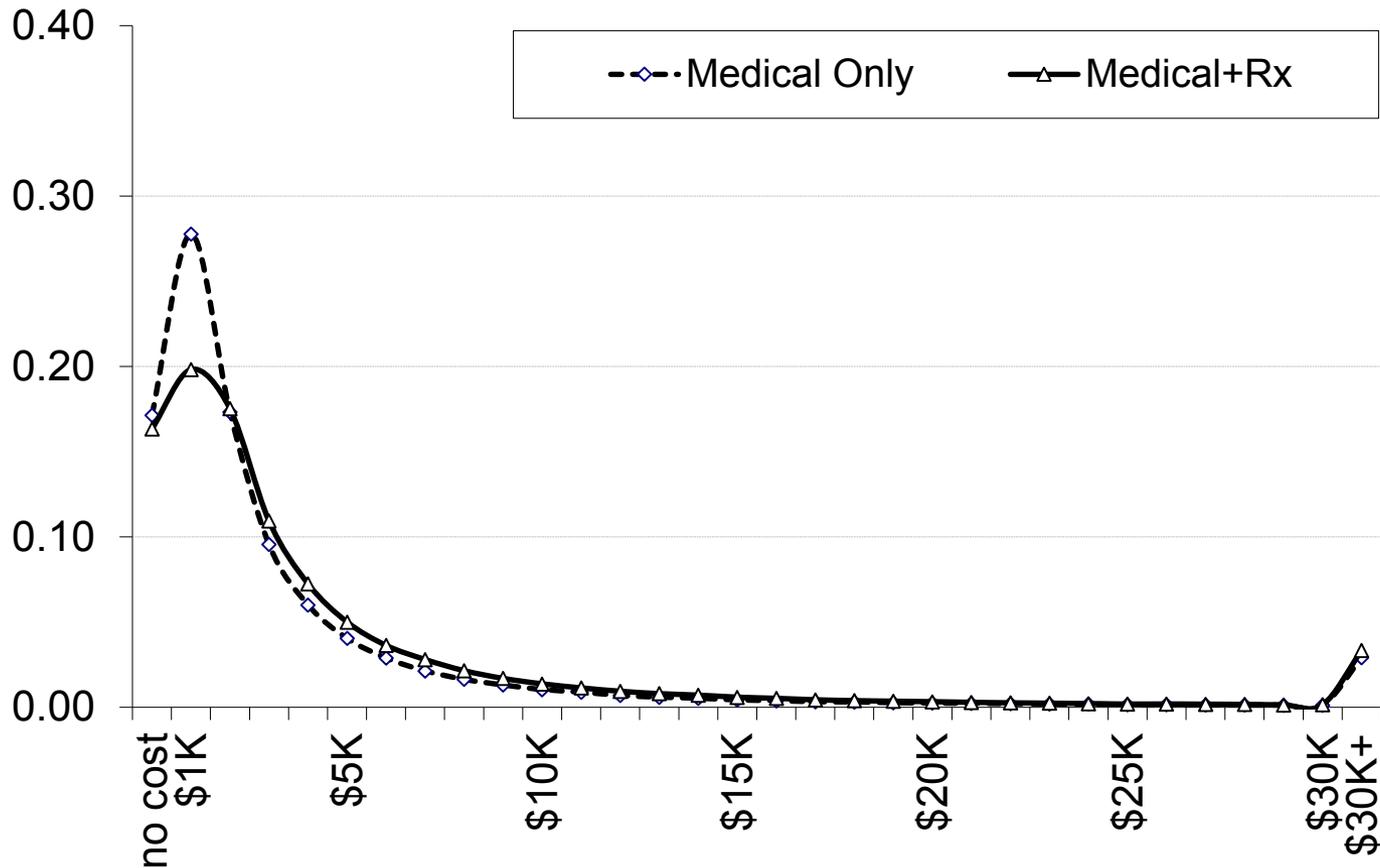
## ■ Examples

- Members of a health plan without utilization
- Clinical trial participants with no cost

# The two-part model

- Part 1: Dependent variable is indicator any cost is incurred
  - 1 if cost is incurred ( $Y > 0$ )
  - 0 if no cost is incurred ( $Y=0$ )
- Part 2: Regression of how much cost, among those who incurred any cost

# Annual per person VHA costs FY10 *among those who used VHA in FY09*



# The two-part model

- Expected value of  $Y$  conditional on  $X$

$$E(Y | X) = P(Y > 0 | X) E(Y | Y > 0, X)$$



Is the product of:

Part 1.

The probability that  
 $Y$  is greater than zero,  
conditional on  $X$

Part 2.

Expected value of  $Y$ ,  
conditional on  $Y$  being  
greater than zero,  
conditional on  $X$

# Predicted cost in two-part model

- Predicted value of Y

$$E(Y | X) = P(Y > 0 | X) E(Y | Y > 0, X)$$



Is the product of:

Part 1.  
Probability of any cost  
being incurred

Part 2.  
Predicted cost  
conditional on  
incurring any cost

# Question for class

$$P(Y > 0) | X)$$

- Part one estimates probability  $Y > 0$ 
  - $Y > 0$  is dichotomous indicator
  - 1 if cost is incurred ( $Y > 0$ )
  - 0 if no cost is incurred ( $Y=0$ )

## POLL QUESTION #2

**Which regression method(s) are used for a dichotomous (zero/one) dependent variable? (check all that apply)**

- Ordinary Least Squares
- Logistic Regression
- Probit Regression
- Survival (Cox) regression

# First part of model

## Regression with dichotomous variable

- Logistic regression or probit
- Logistic regression uses maximum likelihood function to estimate log odds ratio:

$$\log \frac{P_i}{1 - P_i} = \alpha + \beta_1 X$$

# Logistic regression syntax in SAS

**Proc Logistic;**

**Model HASCOST = X1 X2 X3 / Descending;**

**Output out={dataset} prob={variable name};**

- HASCOST an indicator variable
- Output statement saves the predicted probability that the dependent variable equals one (cost was incurred)
- Descending option in model statement is required, otherwise SAS estimates the probability that the dependent variable equals zero

# Logistic regression syntax in Stata

**Logit HASOCOST X1 X2 X3**

**Predict {variable name}, pr**

- Predict statement generates the predicted probability that the dependent variable equals one (cost was incurred)

# Second part of model

## Conditional quantity

- Regression involves only observations with non-zero cost (conditional cost regression)
- Use GLM or OLS with log cost

# Two-part models

- Separate parameters for participation and conditional quantity
  - How independent variables predict
    - participation in care
    - quantity of cost conditional on participation
  - each parameter may have its policy relevance

# Stata TPM command

- Fits two part regressions
    - First part: binary choice (Prob depvar > 0)
    - Second part: distribution of depvar conditional on depvar > 0
  - User developed ADO file
    - must be installed from web
  - Federico Belotti & Partha Deb (2012)
-

# Stata TPM command

- First part options
  - Logit or Probit
- Second part options
  - OLS of raw value, OLS of log, or GLM
- Example syntax
  - TPM COST X1 X2 X3, f(logit) s(glm, fam(gamma) link(log))

# Stata TPM command

- Post-estimation commands
  - Predict values of depvar
  - Allows out of sample predictions
  - Corrects for retransformation bias in OLS models

# Alternatives to two-part model

- OLS with untransformed cost
- OLS with log cost, using small positive values in place of zero
  - not recommended
- GLM models, e.g. gamma regression
  - Cannot have “too many” values with zero

# Topics for today's course

- What is heteroscedasticity, and what should be done about it? (GLM models)
  - What should be done when there are many zero values? (Two-part models)
  - How to test differences in groups with no assumptions about distribution?
  - How to determine which model is best?
-

# Non-parametric statistical tests

- Make no assumptions about distribution, variance
  - Wilcoxon rank-sum test
  - Assigns rank to every observation
  - Compares ranks of groups
  - Calculates the probability that the rank order occurred by chance alone
-

# Extension to more than two groups

- Group variable with more than two mutually exclusive values
- Kruskal-Wallis test
  - is there any difference between any pairs of the mutually exclusive groups?
- If KW is significant, then a series of Wilcoxon tests allows comparison of pairs of groups

# Limits of non-parametric test

- It is too conservative
  - Compares ranks, not means
  - Ignores influence of outliers
  - E.g. all other ranks being equal, Wilcoxon will give same result regardless of whether
    - Top ranked observation is \$1 million more costly than second observation, or
    - Top ranked observation just \$1 more costly
- Doesn't allow for additional explanatory variables

# Topics for today's course

- What is heteroscedasticity, and what should be done about it? (GLM models)
- What should be done when there are many zero values? (Two-part models)
- How to test differences in groups with no assumptions about distribution? (Non-parametric statistical tests)
- How to determine which model is best?

# Which model is best?

- Find predictive accuracy of models
- Estimate regressions with half the data, test their predictive accuracy on the other half of the data
- Find
  - Mean Absolute Error (MAE)
  - Root Mean Square Error (RMSE)

# Mean Absolute Error

- For each observation
  - find difference between observed and predicted cost
  - take absolute value
  - find the mean
- Model with smallest value is best

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

# Root Mean Square Error

- Square the differences between predicted and observed, find their mean, find its square root
- Best model has smallest value

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

# MSE and RMSE in example

(10% sample of low back pain patients)

Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
4 age categorical variables	5,972	14,054
7 additional variables	5,608	13,745

Find in raw scale to be more transparent  
*Lower values are more desirable!*

# Evaluations of residuals

- Mean residual (predicted less observed)  
or
  - Mean predicted ratio (ratio of predicted to observed)
    - calculate separately for each decile of observed Y
    - A good model should have equal residuals (or equal mean ratio) for all deciles
-

# Evaluation of Residuals

(10% sample of low back pain patients)

Decile of Cost	Mean predicted cost in decile	Mean observed cost in decile	Ratio of predicted to observed
1	4,568	333	17.5
2	5,031	792	6.5
3	5,324	1,266	4.2
4	5,645	1,791	3.2
5	5,932	2,443	2.4
6	6,301	3,311	1.9
7	6,697	4,518	1.5
8	7,243	6,382	1.1
9	7,854	9,971	0.8
10	8,894	32,232	0.4

# Formal tests of residuals

- Variant of Hosmer-Lemeshow Test
  - F test of whether residuals in raw scale in each decile are significantly different
- Pregibon's Link Test
  - Tests if linearity assumption was violated
    - See Manning, Basu, & Mullahy, 2005

# Variant of Hosmer-Lemeshow Test

(10% sample of low back pain patients)

Dependent variable: residual in raw scale

F=2,516 p<0.0001

Variable	Parameter Estimate	T-value	p
Intercept	-4,235	-25.1	<.0001
decile2	-5	-0.02	0.984
decile3	176	0.7	0.459
decile4	381	1.6	0.11
decile5	746	3.1	0.002
decile6	1,245	5.2	<.0001
decile7	2,056	8.6	<.0001
decile8	3,374	14.2	<.0001
decile9	6,352	26.7	<.0001
decile10	27,573	115.7	<.0001

**Questions?**

# Review of presentation

- Cost is a difficult dependent variable
  - Skewed to the right by high outliers
  - May have many observations with zero values
  - Cost is not-negative

# When cost is skewed

- OLS of raw cost is prone to bias
  - Especially in small samples with influential outliers
  - “A single case can have tremendous influence”

# When cost is skewed (cont.)

- Log transformed cost
  - Log cost is more normally distributed than raw cost
  - Log cost can be estimated with OLS

# When cost is skewed (cont.)

- To find predicted cost, must correct for retransformation bias
  - Smearing estimator assumes errors are homoscedastic
  - Biased if errors are heteroscedastic

# When cost is skewed and errors are heteroscedastic

- GLM with log link and gamma variance
  - Considers heteroscedastic errors
  - Not subject to retransformation bias
  - May not be very efficient
  - Alternative GLM specification
    - Poisson instead of gamma variance function
    - Square root instead of log link function

# When cost has many zero values

- Two part model
  - Logit or probit is the first part
  - Conditional cost regression is the second part

# Comparison without distributional assumptions

- Non-parametric tests can be useful
- May be too conservative
- Don't allow co-variates

# Evaluating models

- Mean Absolute Error
- Root Mean Square Error
- Other evaluations and tests of residuals

# Key sources on GLM

- MANNING, W. G. (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem, *J Health Econ*, 17, 283-95.
- \* MANNING, W. G. & MULLAHY, J. (2001) Estimating log models: to transform or not to transform?, *J Health Econ*, 20, 461-94.
- \* MANNING, W. G., BASU, A. & MULLAHY, J. (2005) Generalized modeling approaches to risk adjustment of skewed outcomes data, *J Health Econ*, 24, 465-88.
- BASU, A. & Rathouz P.J. (2005) Estimating marginal and incremental effects on health outcomes using flexible link and variance function models, *Biostatistics* 6(1): 93-109, 2005.

# Key sources on two-part models

- \* MULLAHY, J. (1998) Much ado about two: reconsidering retransformation and the two-part model in health econometrics, *J Health Econ*, 17, 247-81
- JONES, A. (2000) Health econometrics, in: Culyer, A. & Newhouse, J. (Eds.) *Handbook of Health Economics*, pp. 265-344 (Amsterdam, Elsevier).

# References to worked examples

- DEB, P & NORTON, EC (2018) Modeling health care expenditures and use, *Ann Rev Public Health* 39:489–505
- FLEISHMAN, J. A., COHEN, J. W., MANNING, W. G. & KOSINSKI, M. (2006) Using the SF-12 health status measure to improve predictions of medical expenditures, *Med Care*, 44, I54-63.
- MONTEZ-RATH, M., CHRISTIANSEN, C. L., ETTNER, S. L., LOVELAND, S. & ROSEN, A. K. (2006) Performance of statistical models to predict mental health and substance abuse cost, *BMC Med Res Methodol*, 6, 53.

# References to worked examples (cont).

- MORAN, J. L., SOLOMON, P. J., PEISACH, A. R. & MARTIN, J. (2007) New models for old questions: generalized linear models for cost prediction, *J Eval Clin Pract*, 13, 381-9.
- DIER, P., YANEZ D., ASH, A., HORNBROOK, M., LIN, D. Y. (1999). Methods for analyzing health care utilization and costs Ann Rev Public Health (1999) 20:125-144 (Also gives accessible overview of methods, but lacks information from more recent developments)

# Link to HERC Cyberseminar HSR&D study of worked example

Performance of Statistical Models to Predict  
Mental Health and Substance Abuse Cost

Maria Montez-Rath, M.S. 11/8/2006

The audio:

- [http://vaww.hsr.d.research.va.gov/for\\_researchers/cyber\\_seminars/HERC110806.asx](http://vaww.hsr.d.research.va.gov/for_researchers/cyber_seminars/HERC110806.asx)

The Power point slides:

- [http://vaww.hsr.d.research.va.gov/for\\_researchers/cyber\\_seminars/HERC110806.pdf](http://vaww.hsr.d.research.va.gov/for_researchers/cyber_seminars/HERC110806.pdf)

# Book chapters

- MANNING, W. G. (2006) Dealing with skewed data on costs and expenditures, in: Jones, A. (Ed.) *The Elgar Companion to Health Economics*, pp. 439-446 (Cheltenham, UK, Edward Elgar).