

Econometrics Course: Cost at the Dependent Variable (II)



Paul G. Barnett, PhD

June 3, 2015

Review of Ordinarily Least Squares (OLS)

- Classic linear model
- Assume dependent variable can be expressed as a linear function of the chosen independent variables, e.g.:
- $Y_i = \alpha + \beta X_i + \varepsilon_i$

Review of OLS assumptions

- Expected value of error is zero $E(\varepsilon_i)=0$
- Errors are independent $E(\varepsilon_i\varepsilon_j)=0$
- Errors have identical variance $E(\varepsilon_i^2)=\sigma^2$
- Errors are normally distributed
- Errors are not correlated with independent variables $E(X_i\varepsilon_i)=0$

Cost is a difficult variable

- Skewed by rare but extremely high cost events
- Zero cost incurred by enrollees who don't use care
- No negative values

Review from last session

- Applying Ordinary Least Squares OLS to data that aren't normal can result in biased parameters
 - OLS can predict *negative* costs
- Log of cost
 - is normally distributed, can use in OLS
- Limits of OLS of log cost
 - Predicted cost is affected by re-transformation bias
 - Assumes constant error (homoscedasticity)
 - Can't take log of zero

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 method is best?
-

Topics for today's course

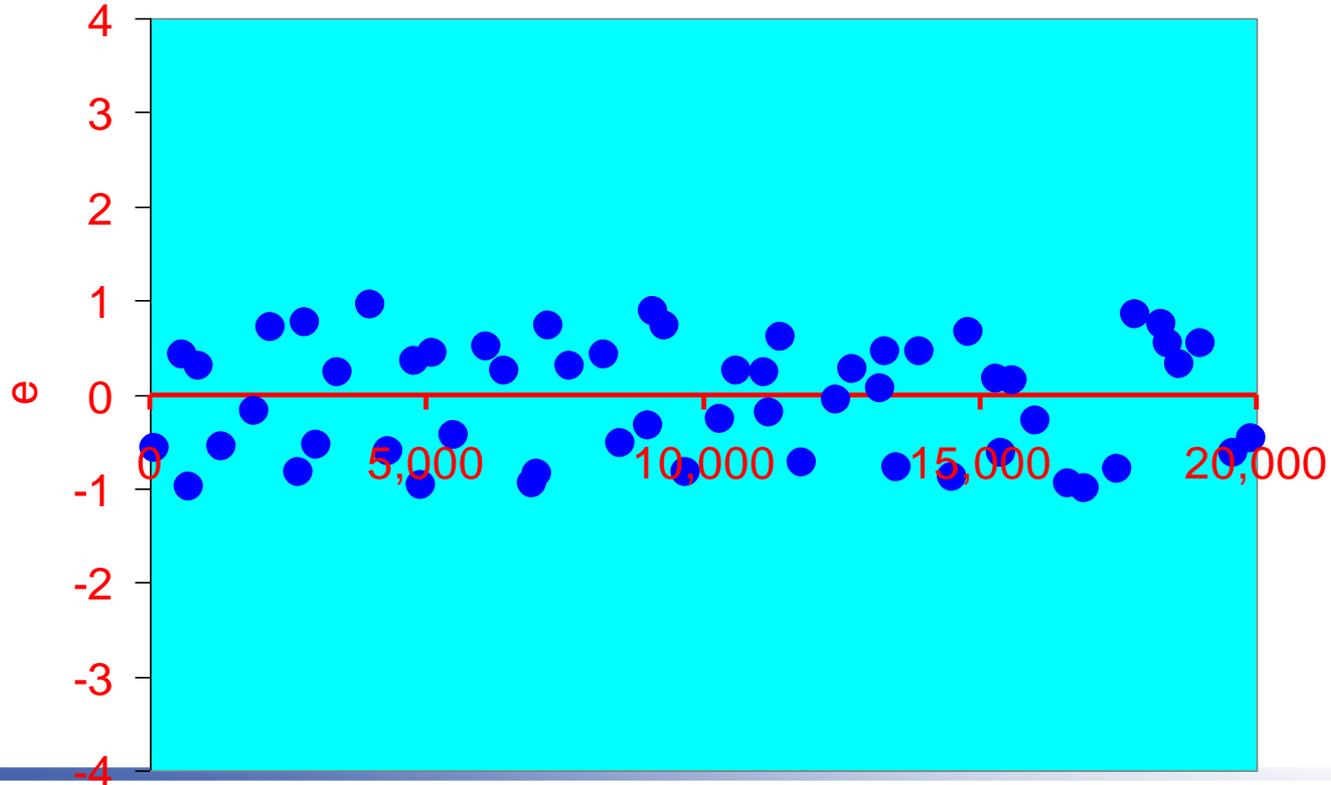
- What is heteroscedasticity and what should be done about it?
 - What should be done when there are many zero values?
 - How to test differences in groups with no assumptions about distribution?
 - How to determine which method is best?
-

What is heteroscedasticity?

- A violation of one of the assumptions needed to use OLS regarding the error terms.
- OLS assumes Homoscedasticity
 - Identical variance $E(\varepsilon_i^2)=\sigma^2$
- Heteroscedasticity
 - Variance depends on x (or on predicted y)

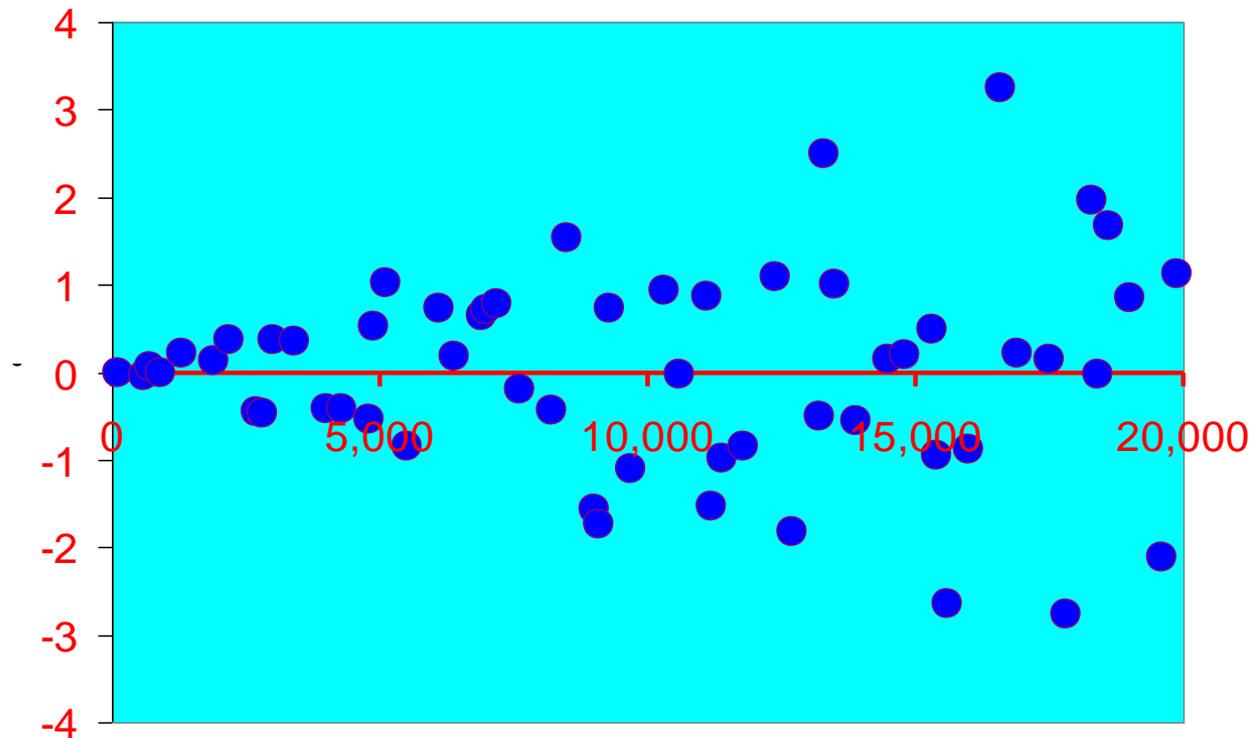
Homoscedasticity

- 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 JHE 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 β 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)

Which GLM link function?

– Box-Cox regression

– Stata command:

boxcox cost {indep. vars} if y > 0

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

Which link function?

■ Box-Cox parameter

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

Which variance structure with GLM?

Modified Park test

- GLM regression gamma family log link & find residual
- Square the residuals
- Second regression by OLS
 - Dependent variable squared residuals
 - Independent variable predicted y

$$(Y_i - \hat{Y}_i)^2 = \gamma_0 + \gamma_1 \hat{Y}_i + \nu_i$$

Which variance structure with GLM?

- Parameter from GLM family test (modified Park test)

$$(Y_i - \hat{Y}_i)^2 = \gamma_0 + \gamma_1 \hat{Y}_i + \nu_i$$

γ_1	Variance
0	Gaussian (Normal)
1	Poisson
2	Gamma
3	Wald (Inverse Normal)

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?

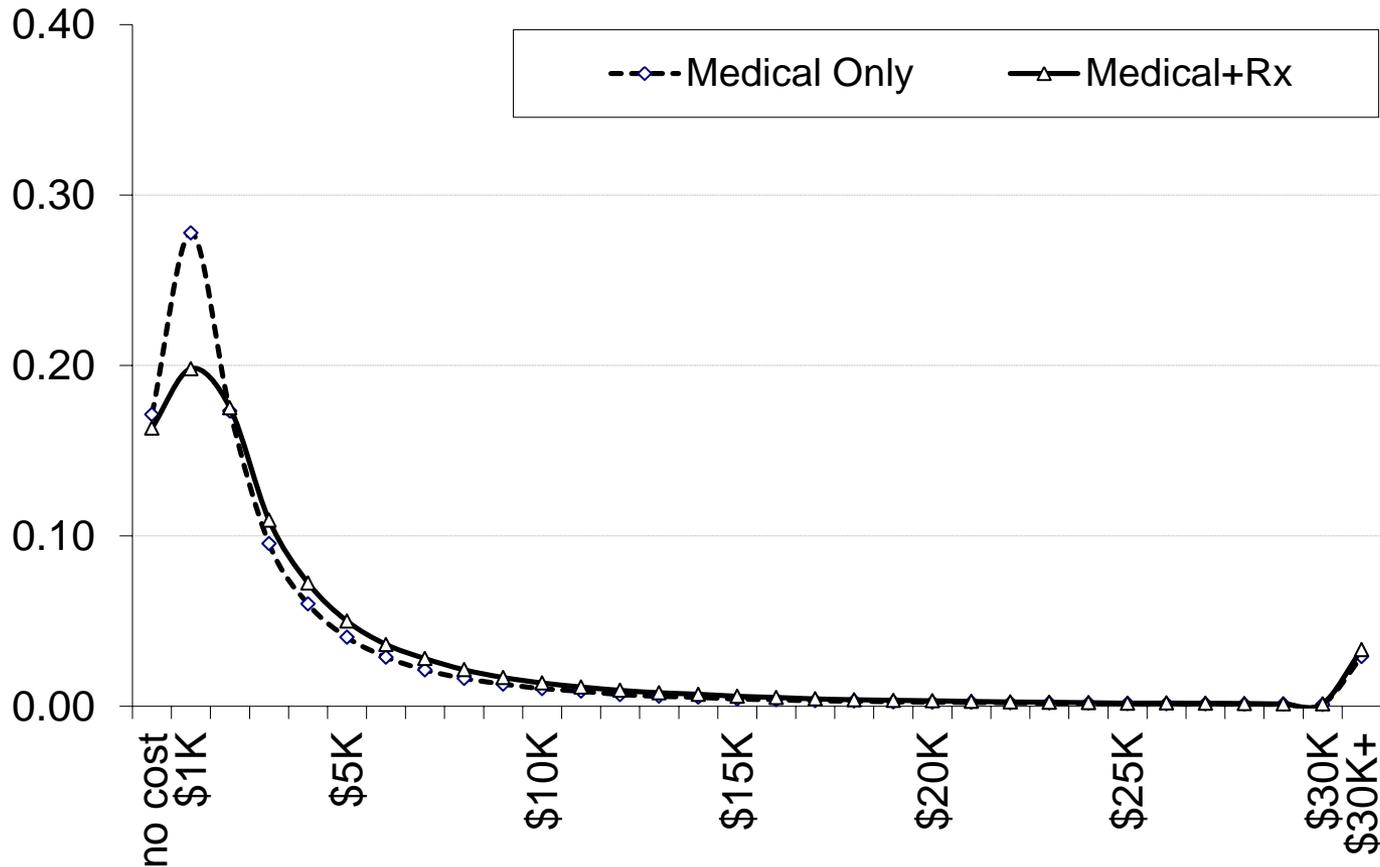
Topics for today's course

- 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 method is best?
-

What should be done when there are many zero values?

- Example of participants enrolled in a health plan who have no utilization

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



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

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 ($Y=0$)

Question

Which regression method(s) are used for a dichotomous (zero/one) dependent variable?

- Ordinary Least Squares
- Generalized Linear Model
- Logistic Regression
- Probit
- 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
- Certain GLM models

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 method 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 method is best?
-

Which method 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

Root Mean Square Error

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

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
-

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

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

- 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.hsrdr.research.va.gov/for_researchers/cyber_seminars/HERC110806.asx

The Power point slides:

- http://vaww.hsrdr.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).