

# **Econometrics Course: Cost as the Dependent Variable (I)**



**Paul G. Barnett, PhD**

**October 6, 2010**

# HERC SharePoint Discussion Board

The screenshot shows the HERC SharePoint website interface. At the top, there is a navigation bar with 'File Edit View Favorites Tools Help' and a 'Links' section. Below this is the HERC logo and the text 'HEALTH ECONOMICS RESOURCE CENTER'. A sidebar on the left contains a menu with items like 'HERC Home', 'News >', 'Resources >', 'Data >', 'Methods >', 'Research', 'Publications >', 'Training >', and 'About Us >'. The main content area features a large image of a calculator and a stethoscope. To the right of this image is an 'Events' section with two entries: '2010 CYBERCOURSE' (Cost as the Dependent Variable) and '2010 CYBERSEMINARS' (Untangling the Direct and Indirect Effects of Body Mass Dynamics). Below the events is a 'FEATURED' section with the title 'The cost of US health care - a systematic review'. The bottom section is divided into three columns: 'Recent News' (with three news items dated September 14, 2010 and August 10, 2010), 'About Us' (with a paragraph about HERC's location and research focus), and 'Recent Forum Posts' (with three posts dated September 22, 2010, July 28, 2010, and July 28, 2010). Two callout boxes are present: one pointing to the 'Recent Forum Posts' section with the text 'Recent questions posted on discussion board', and another pointing to a link at the bottom of the forum posts section with the text 'Link to HERC SharePoint Discussion Board'. The browser's status bar at the bottom shows 'Local intranet' and '100%' zoom.

Recent questions posted on discussion board

Link to HERC SharePoint Discussion Board

# HERC SharePoint Discussion Board

Ask questions by adding a new forum post

Follow the discussions by setting up email notifications when new posts appear

Contribute to an existing post by clicking on any of the listed items

The screenshot shows the HERC (Health Economics Resource Center) SharePoint Discussion Board. The interface includes a navigation menu on the left with categories like Sites, Documents, Pictures, Lists, Discussions, and Surveys. The main content area displays a list of discussion topics with columns for Subject, Created By, Replies, and Last Updated. A search bar is visible at the top right. Callout boxes with arrows point to specific features: 'Ask questions by adding a new forum post' points to the 'New' button; 'Follow the discussions by setting up email notifications when new posts appear' points to the 'Actions' button; and 'Contribute to an existing post by clicking on any of the listed items' points to a discussion topic in the list.

Subject	Created By	Replies	Last Updated
I read that with large sample sizes (retrospective databases) that p-values do not contribute much information- what should be used to assess the model fit when dealing with large sample sizes?	Fan, Angela	0	9/22/2010 3:24 PM
Can you talk more about the analysis report from HERC 2003. Does it have something for average costs of long term care or for stroke rehab?	Fan, Angela	1	7/28/2010 5:22 PM
Are procedures occurring under sharing agreements (e.g., cardiac surgery at a university hospital paid for by a VA under a sharing agreement) therefore invisible (i.e., not in the VA PTF or operations file, and not in Fee Basis)?	Fan, Angela	1	7/28/2010 5:21 PM
Can Fee Basis care show up during a VA treatment stay (e.g., can a veteran who is an inpatient receive care from a non-VA practitioner while he/she is a patient within a VA hospital)?	Fan, Angela	1	7/28/2010 5:18 PM
How are acute stroke hospitalizations paid for by Fee Basis?	Fan, Angela	1	7/28/2010 5:13 PM
Are you aware of a policy about not referring Veterans who may be experiencing acute stroke symptoms to outside facilities for emergent care?	Fan, Angela	1	7/28/2010 5:11 PM
Any special data considerations for Project HERO (VA & Humana demonstration)?	Fan, Angela	1	7/28/2010 5:11 PM
Can you identify patients receiving palliative or hospice care on Fee Basis?	Fan, Angela	1	7/28/2010 5:10 PM
Can you search the database by a specific procedure? For example can you find all fee records that pertain to fee basis care for cardiac surgery?	Fan, Angela	1	7/28/2010 5:09 PM
How does the zip code field handle the leading 0? If it is a numeric field, it will be dropped.	Fan, Angela	1	7/28/2010 5:07 PM
How does one access the HERC Fee Basis data? What is the HERC intranet address?	Fan, Angela	1	7/28/2010 5:05 PM
Sometimes within VA by providing interventions that result in better quality care, we find unmet needs and end up having them utilize more care. Their long term health might be better, but would CEA of the initial intervention be able to capture that?	Fan, Angela	1	5/26/2010 7:14 PM
Does the VA have a preferred method for measuring cost effectiveness? Are QALYs considered to be the gold standard in the VA?	Fan, Angela	1	5/26/2010 7:12 PM
Would you classify VA CEA studies as a societal perspective or a payer perspective?	Fan, Angela	1	5/26/2010 7:09 PM
Can you give the website portal or SharePoint to find the additional readings?	Fan, Angela	1	5/26/2010 7:07 PM
Can you speak more specifically about how to design CEA studies with a global budget in mind?	Fan, Angela	1	5/26/2010 7:06 PM
Have you ever seen cost effectiveness analysis that compares healthcare systems (payer source, providers, delivery) and not a particular intervention (perhaps the different systems are the intervention)?	Fan, Angela	1	5/26/2010 7:02 PM
Your comments focused on showing the cost effectiveness of pharmaceuticals, are nations as concerned with showing the cost effectiveness findings for devices? Stents, for example.	Fan, Angela	1	5/26/2010 6:53 PM
Can you suggest key articles to read to learn more about cost effectiveness analysis (CEA)?	Fan, Angela	1	5/26/2010 6:42 PM
Can you talk a little about CEA in the oncology area?	Fan, Angela	1	5/26/2010 4:47 PM
One thing that I find difficult to find are average costs for certain items or resource utilization. Is there a way where we can have access to this without filling a form each	Fan, Angela	1	5/26/2010 4:39 PM

# HERC SharePoint Discussion Board: Individual Posts

Initial post

Response posts

The screenshot shows a SharePoint discussion board interface. At the top, there is a navigation bar with 'File', 'Edit', 'View', 'Favorites', 'Tools', and 'Help' menus. Below this is a 'Links' section with a search bar and a 'CyberCourse Discussion' link. The main header features the HERC logo (Health Economics Resource Center) and a search bar. The discussion board is titled 'CyberCourse Discussion' and contains three posts. The first post, by Angela Fan, is the initial post. The second and third posts, by Vilja Joyce and Paul Barnett, are responses. The interface includes a left-hand navigation pane with options like 'View All Site Content', 'Sites', 'Documents', 'Pictures', 'Lists', 'Discussions', and 'Recycle Bin'. A 'View: Flat' dropdown is visible in the top right of the post area. The bottom of the screen shows a 'Local intranet' status and a '100%' zoom level.

HERC  
Health Economics Resource Center

CyberCourse Discussion

Why not use last value carried forward as many clinicians favor?

Posted By: Fan, Angela  
Started: 4/28/2010 4:45 PM

Why not use last value carried forward as many clinicians favor?

Posted: 4/28/2010 7:06 PM  
Joyce, Vilja  
When considering imputation, the last value carried forward method or LVCF may underestimate the variance of the parameter estimates by treating imputed values as if they were observed values. This may bias the results.  
Multiple imputation, which imputes 3+ sets of values for the missing data, helps to solve this problem by incorporating both variability and uncertainty.

Posted: 4/28/2010 7:34 PM  
Barnett, Paul  
The Last Value Carried forward method assumes that the quality of life from the period with missing data is the same it it was in the prior period. If drop out rates (or failure to come to clinic for assessment) is related to treatment, this assumption results in a biased assessment of the mean value for quality of life. The values are not missing at random, there is some information in the state of being missing. Last Value Carrier forward doesn't convey any information about the uncertainty of imputation, so the precision of the imputed value is overstated-- the variance of quality of life is underestimated.

# HERC SharePoint Discussion Board: Shared Documents

The screenshot shows the SharePoint interface for the 'Shared Documents' library. The top navigation bar includes 'File', 'Edit', 'View', 'Favorites', 'Tools', and 'Help'. The breadcrumb trail shows 'Home > Shared Documents'. The left sidebar contains navigation options: 'View All Site Content', 'Sites', 'Documents' (with 'Shared Documents' selected), 'Pictures', 'Lists' (with 'Contacts' and 'Tasks' listed), 'Discussions' (with 'CyberCourse Discussion' listed), and 'Surveys' (with 'Recycle Bin' listed). The main content area features a 'New' button, an 'Upload' button, and an 'Actions' button. Below these is a table of documents:

Type	Name	Modified	Modified By
Folder	Readings on Making CEA more relevant (5-26-2010 seminar)	5/26/2010 4:17 PM	Barnett, Paul
Document	HERC Datasets at Austin Information Technology Center (AITC)	9/22/2010 3:54 PM	Fan, Angela

Four callout boxes provide instructions:

- Create a new document to share in this library**: Points to the 'New' button.
- Upload a single or multiple documents from your computer to this library**: Points to the 'Upload' button.
- Created or uploaded shared documents**: Points to the 'Documents' section in the left sidebar.
- Receive email notifications when items in this library change**: Points to the 'Actions' button.

# HERC SharePoint Discussion Board: Setting Up Notifications

The screenshot displays the HERC (Health Economics Resource Center) SharePoint Discussion Board. The interface includes a top navigation bar with 'File', 'Edit', 'View', 'Favorites', 'Tools', and 'Help'. Below this is a 'Links' section and a search bar. The main content area shows a list of discussion topics, each with a subject line, a snippet of text, and a table of activity (Created By, Replies, Last Updated). A callout box with a black border and white background points to the 'Alert Me' option in the 'Actions' menu. The callout text reads: 'Create email alerts notifying you when there are changes to a specified item, document, list, or library.'

**HERC**  
Health Economics Resource Center

Home

> CyberCourse Discussion  
CyberCourse Discussion

View All Site Content

Sites

Documents

- Shared Documents

Pictures

Lists

- Contacts
- Tasks

Discussions

- CyberCourse Discussion

Surveys

Recycle Bin

New Actions

- Connect to Outlook  
Synchronize items and make them available offline.
- Export to Spreadsheet  
Analyze items with a spreadsheet application.
- Open with Access  
Works with items in a Microsoft Office Access database.
- View RSS Feed  
Syndicate items with an RSS reader.
- Alert Me  
Receive e-mail notifications when items change.

View: Subject

Created By	Replies	Last Updated
Fan, Angela	0	9/22/2010 3:24 PM
Fan, Angela	1	7/28/2010 5:22 PM
Fan, Angela	1	7/28/2010 5:21 PM
Fan, Angela	1	7/28/2010 5:18 PM
Fan, Angela	1	7/28/2010 5:13 PM
Fan, Angela	1	7/28/2010 5:11 PM
Fan, Angela	1	7/28/2010 5:11 PM
Fan, Angela	1	7/28/2010 5:10 PM
Fan, Angela	1	7/28/2010 5:09 PM
Fan, Angela	1	7/28/2010 5:07 PM
Fan, Angela	1	7/28/2010 5:05 PM
Fan, Angela	1	5/26/2010 7:14 PM
Fan, Angela	1	5/26/2010 7:12 PM
Fan, Angela	1	5/26/2010 7:09 PM
Fan, Angela	1	5/26/2010 7:07 PM
Fan, Angela	1	5/26/2010 7:06 PM
Fan, Angela	1	5/26/2010 7:02 PM
Fan, Angela	1	5/26/2010 6:53 PM
Fan, Angela	1	5/26/2010 6:42 PM
Fan, Angela	1	5/26/2010 4:47 PM
Fan, Angela	1	5/26/2010 4:39 PM

# HERC SharePoint Discussion Board: Setting Up Notifications

File Edit View Favorites Tools Help

Links

New Alert

Home Feeds (3) Print Page Tools

Welcome Fan, Angela My Site My Links

**HERC**  
Health Economics Resource Center

Home

> CyberCourse Discussion > New Alert

## New Alert

Use this page to create an e-mail alert notifying you when there are changes to the specified item, document, list, or library.

View my existing alerts on this site.

OK Cancel

**Alert Title**  
Enter the title for this alert. This is included in the subject of the e-mail notification sent for this alert.

CyberCourse Discussion

**Send Alerts To**  
This alert will be sent to the e-mail address indicated.

E-mail address:  
Angela.Fan@va.gov

**Change Type**  
Specify the type of changes that you want to be alerted to.

Only send me alerts when:

All changes  
 New items are added  
 Existing items are modified  
 Items are deleted

**Send Alerts for These Changes**  
Specify whether to filter alerts based on specific criteria. You may also restrict your alerts to only include items that show in a particular view.

Send me an alert when:

Anything changes  
 Someone else changes a post  
 Someone else changes a post created by me  
 Someone else changes a post last modified by me

**When to Send Alerts**  
Specify how frequently you want to be alerted.

Send e-mail immediately  
 Send a daily summary  
 Send a weekly summary

Time:  
Friday 1:00 PM

OK Cancel

Done Local intranet 100%

1. Click on  
“Welcome, user last  
name, user first name”
2. Choose  
“My Settings”
3. Click on  
“My Alerts”
4. Click on  
“Add Alert”
5. Click on  
“Cyber Course  
Discussion”

# Poll

- What is your principal role in the health care system?
    - Principal Investigator
    - Health Economist
    - Research Analyst
    - Clinical
    - Operations
    - Other
-

# Poll

- Are you currently involved in an economic study?
  - Yes
  - No

# Poll

- What is your training?
    - PhD
    - MD & PhD
    - MD & MS
    - MD
    - MS or MA
    - BS or BA
    - Other
-

# Poll

- How many semesters of statistics course work have you taken?
  - 0
  - 1-2
  - 3-4
  - 4+

# What is health care cost?

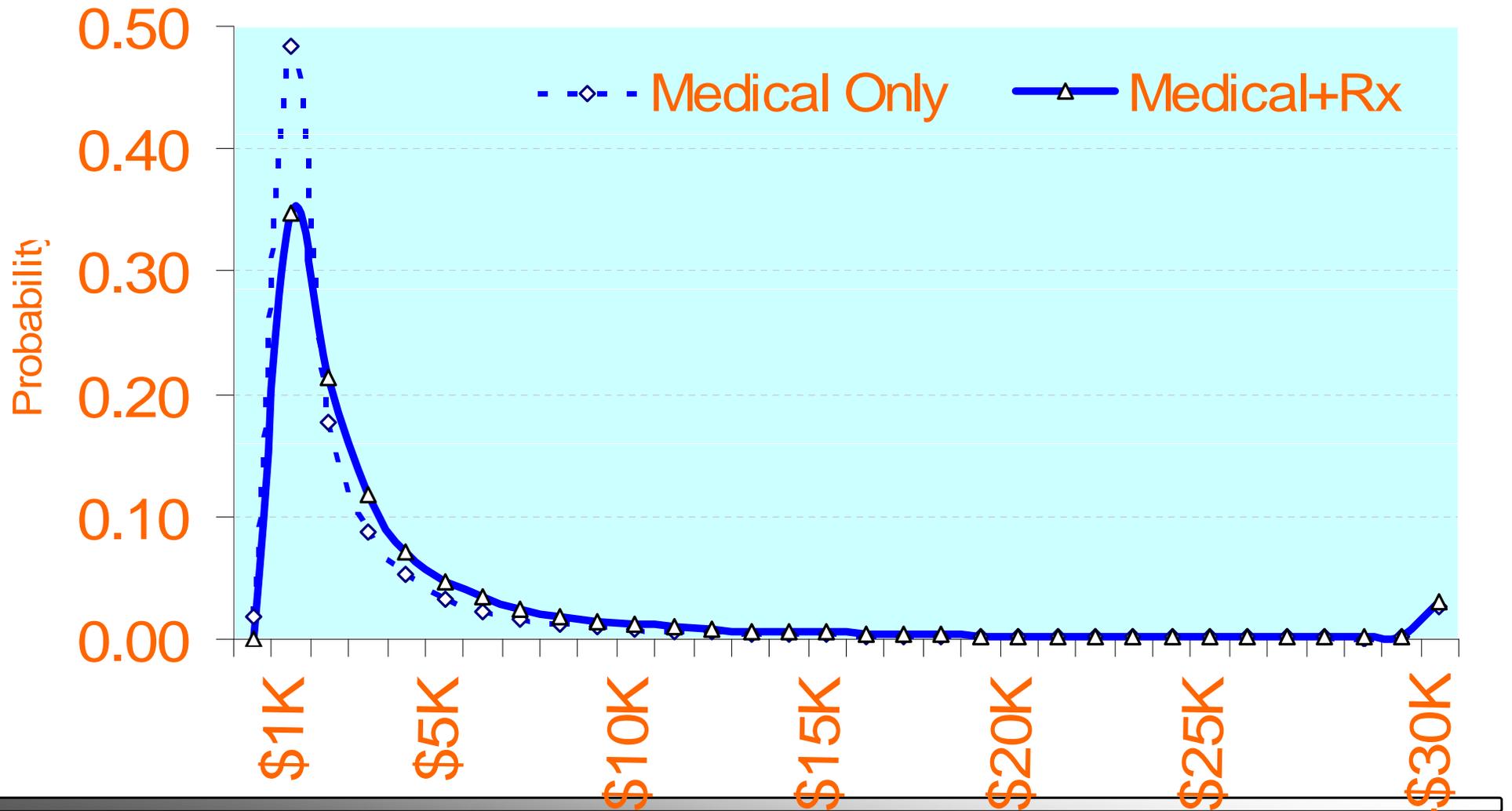
- Cost of an intermediate product, e.g.,
    - chest x-ray
    - a day of stay
    - minute in the operating room
    - a dispensed prescription
  - Cost of a bundle of products
    - Outpatient visit
    - Hospital stay
-

# What is health care cost (cont.)?

- Cost of a treatment episode
  - visits and stays over a time period
- Annual cost
  - All care received in the year

# Annual per person VHA costs FY06

(5% random sample)



# Descriptive statistics: VHA costs FY06

(5% sample, includes outpatient pharmacy)

	Cost
Mean	5,290
Median	1,646
Standard Deviation	16,507
Skewness	11.0
Kurtosis	187.6

# Skewness and kurtosis

- Skewness (3<sup>rd</sup> moment)
  - Degree of symmetry
  - Skewness of normal distribution =0
  - Positive skew: more observations in right tail
- Kurtosis (4<sup>th</sup> moment)
  - Peakness of distribution and thickness of tails
  - Kurtosis of normal distribution=3

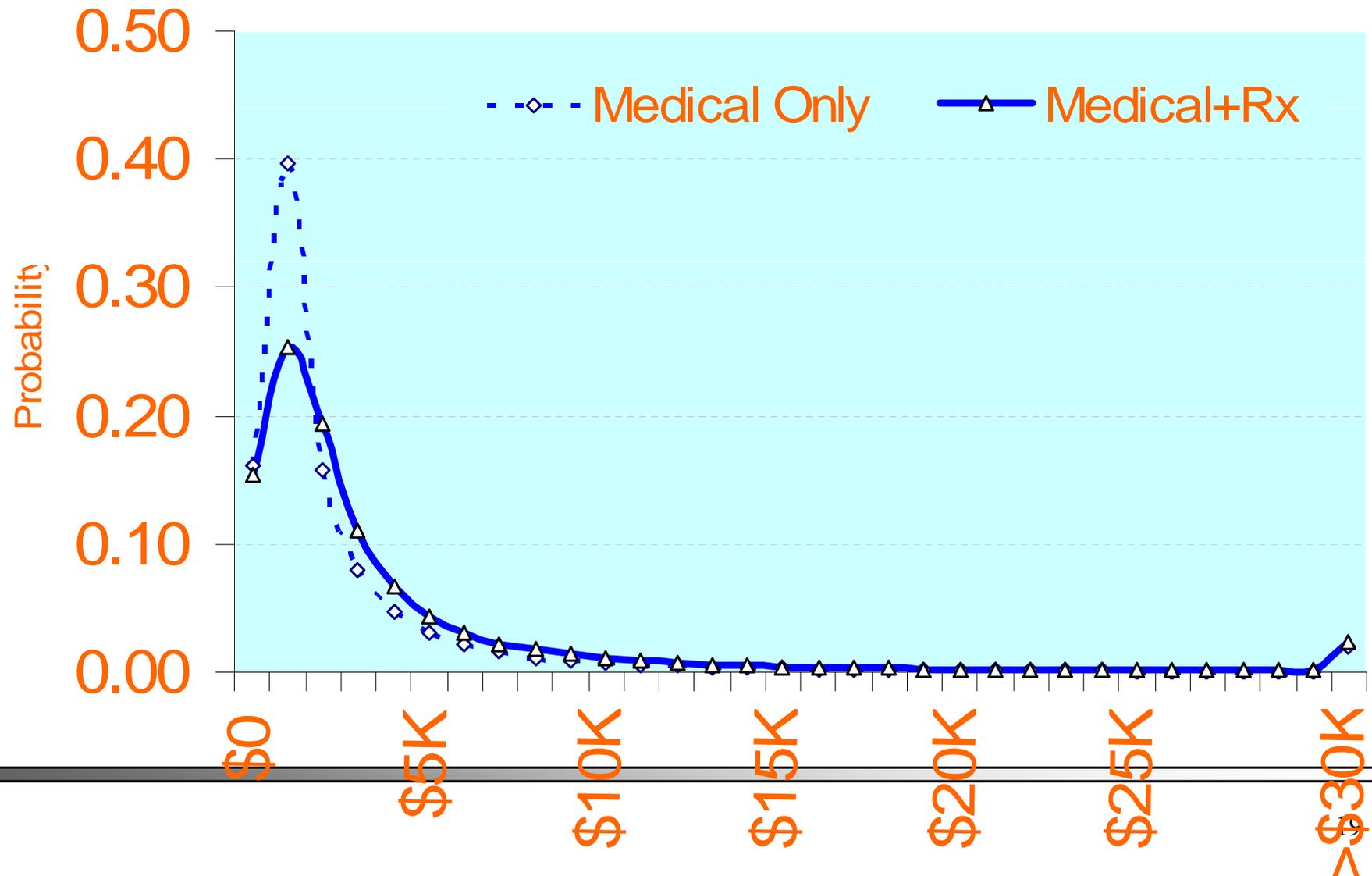
# Distribution of cost: skewness

- Rare but extremely high cost events
  - E.g. only some individuals hospitalized
  - Some individuals with expensive chronic illness
- Positive skewness (skewed to the right)

# Comparing the cost incurred by members of two groups

- Do we care about the mean or the median?

# Annual per person VHA costs FY05 *among those who used VHA in FY06*



# Distribution of cost: zero value records

- Enrollees who don't use care
  - Zero values
  - Truncation of the distribution

**What hypotheses involving cost  
do you want to test?**

# What hypotheses involving cost do you want to test?

- I would like to learn how cost is affected by:
  - Type of treatment
  - Quantity of treatment
  - Characteristics of patient
  - Characteristics of provider
  - Other

# Review of Ordinarily Least Squares (OLS)

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

# Ordinarily Least Squares (OLS)

- Estimates parameters (coefficients)  $\alpha$ ,  $\beta$
- Minimizes the sum of squared errors
  - (the distance between data points and the regression line)

# Linear model

- Regression with cost as a linear dependent variable (Y)
  - $Y_i = \alpha + \beta X_i + \varepsilon_i$
- $\beta$  is interpretable in raw dollars
  - Represents the change of cost (Y) for each unit change in X
  - E.g. if  $\beta=10$ , then cost increases \$10 for each unit increase in X

# Expected value of a random variable

- $E(\text{random variable})$
- $E(W) = \sum W_i p(W_i)$ 
  - For each  $i$ , the value of  $W_i$  times probability that  $W_i$  occurs
  - Probability is between 0 and 1
  - A weighted average, with weights by probability

# 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  $E(X_i\varepsilon_i)=0$

# When cost is the dependent variable

- Which of the assumptions of the classical model are likely to be violated by cost data?
  - Expected error is zero
  - Errors are independent
  - Errors have identical variance
  - Errors are normally distributed
  - Error are not correlated with independent variables

# Compare costs incurred by members of two groups

- Regression with one dichotomous explanatory variable
- $Y = \alpha + \beta X + \varepsilon$
- $Y$  cost
- $X$  group membership
  - 1 if experimental group
  - 0 if control group

# Predicted difference in cost of care for two group

$$Y = \alpha + \beta X + \varepsilon$$

Predicted value of Y conditional on X=0  
(Estimated mean cost of control group)

$$\hat{Y} | (X=0) = \alpha$$

- Predicted Y when X=1  
(Estimated mean cost experimental group)

$$\hat{Y} | (X = 1) = \alpha + \beta_a$$

# Other statistical tests are special cases

- Analysis of Variance (ANOVA) is a regression with one dichotomous independent variable
- Relies on OLS assumptions

# Compare groups controlling for case mix

- Include case-mix variable,  $Z$

$$Y = \alpha + \beta_1 X + \beta_2 Z + \varepsilon$$

# Compare groups controlling for case mix (cont).

- Estimated mean cost of control group controlling for case mix (evaluated at mean value for case-mix variable)

$$\hat{Y} | (X = 0) = \alpha + \beta_2 \bar{Z}$$

*where  $\bar{Z}$  is mean of  $Z$*

# Compare groups controlling for case mix (cont).

- Estimated mean cost of experimental group controlling for case mix (evaluated at mean value for case-mix variable)

$$\hat{Y} | (X = 1) = \alpha + \beta_1 + \beta_2 \bar{Z}$$

*where  $\bar{Z}$  is mean of  $Z$*

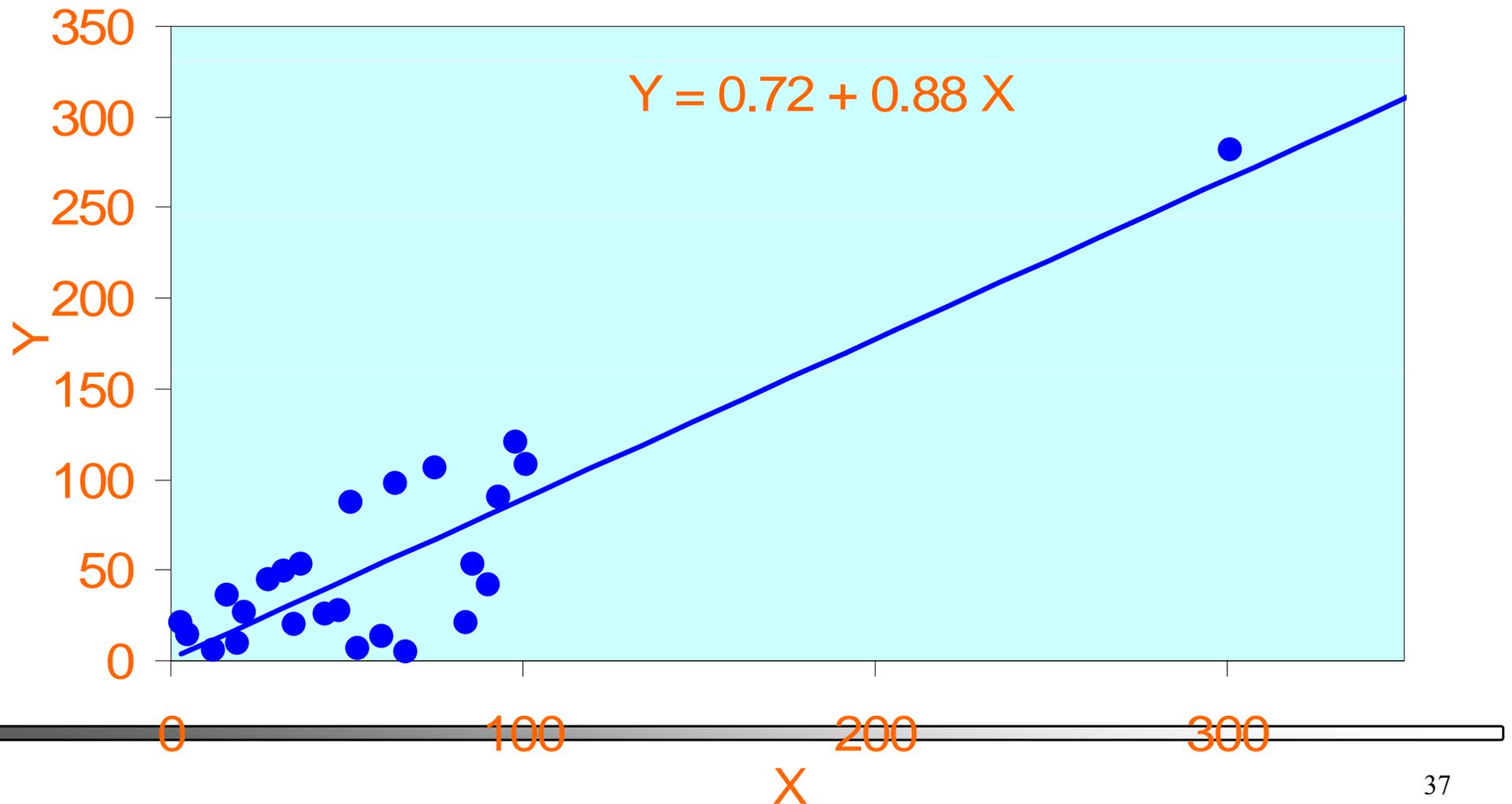
# Assumptions are about error term

- Formally, the OLS assumptions are about the error term
- The residuals (estimated errors) often have a similar distribution to the dependent variable

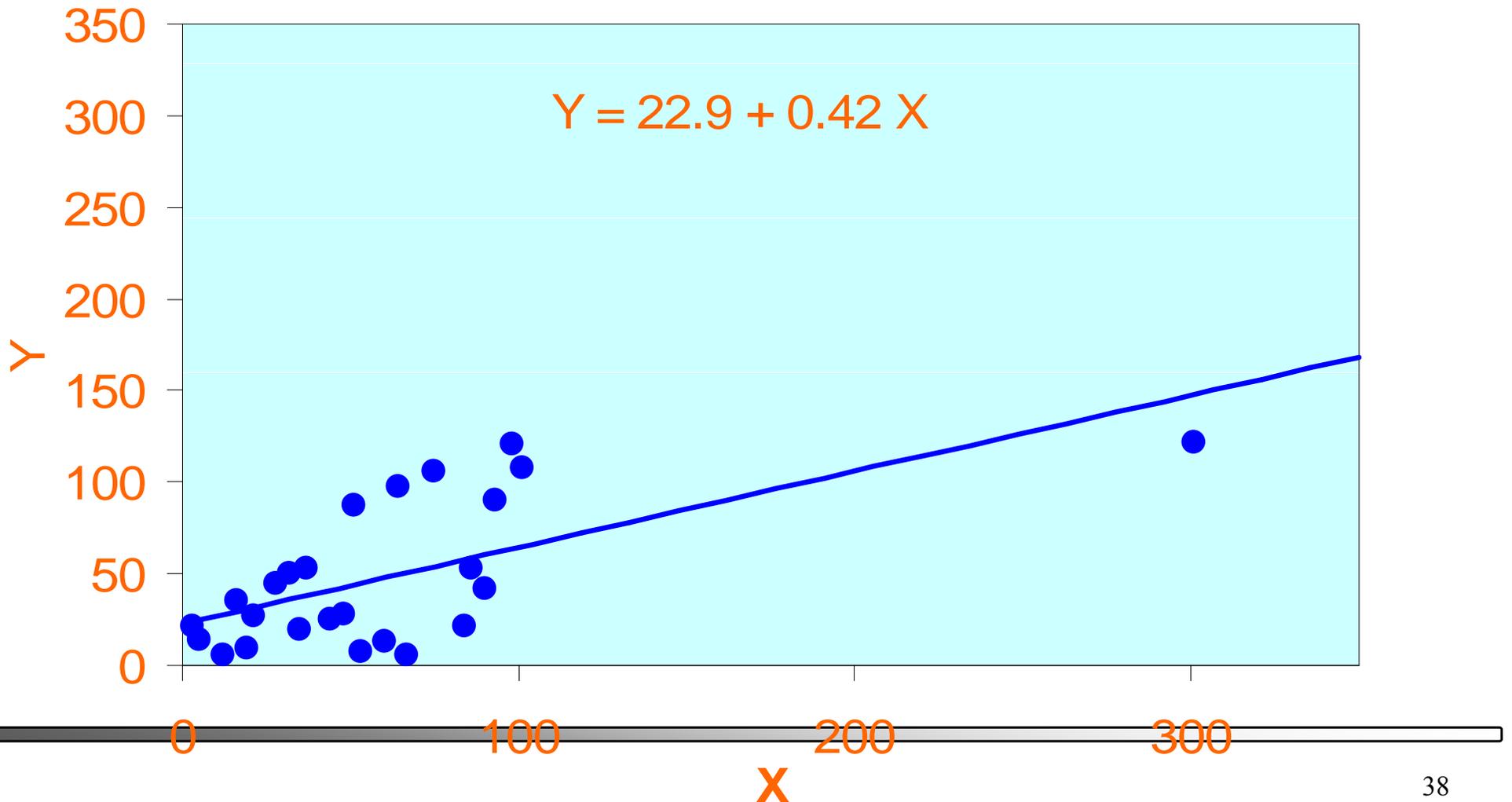
# Why worry about using OLS with skewed (non-normal) data?

- “In small and moderate sized samples, a single case can have tremendous influence on an estimate”
  - Will Manning
  - Elgar Companion to Health Economics AM Jones, Ed. (2006) p. 439
- There are no values skewed to left to balance this influence
- In Rand Health Insurance Experiment, one observation accounted for 17% of the cost of a particular health plan

# The influence of a single outlier observation



# The influence of a single outlier observation

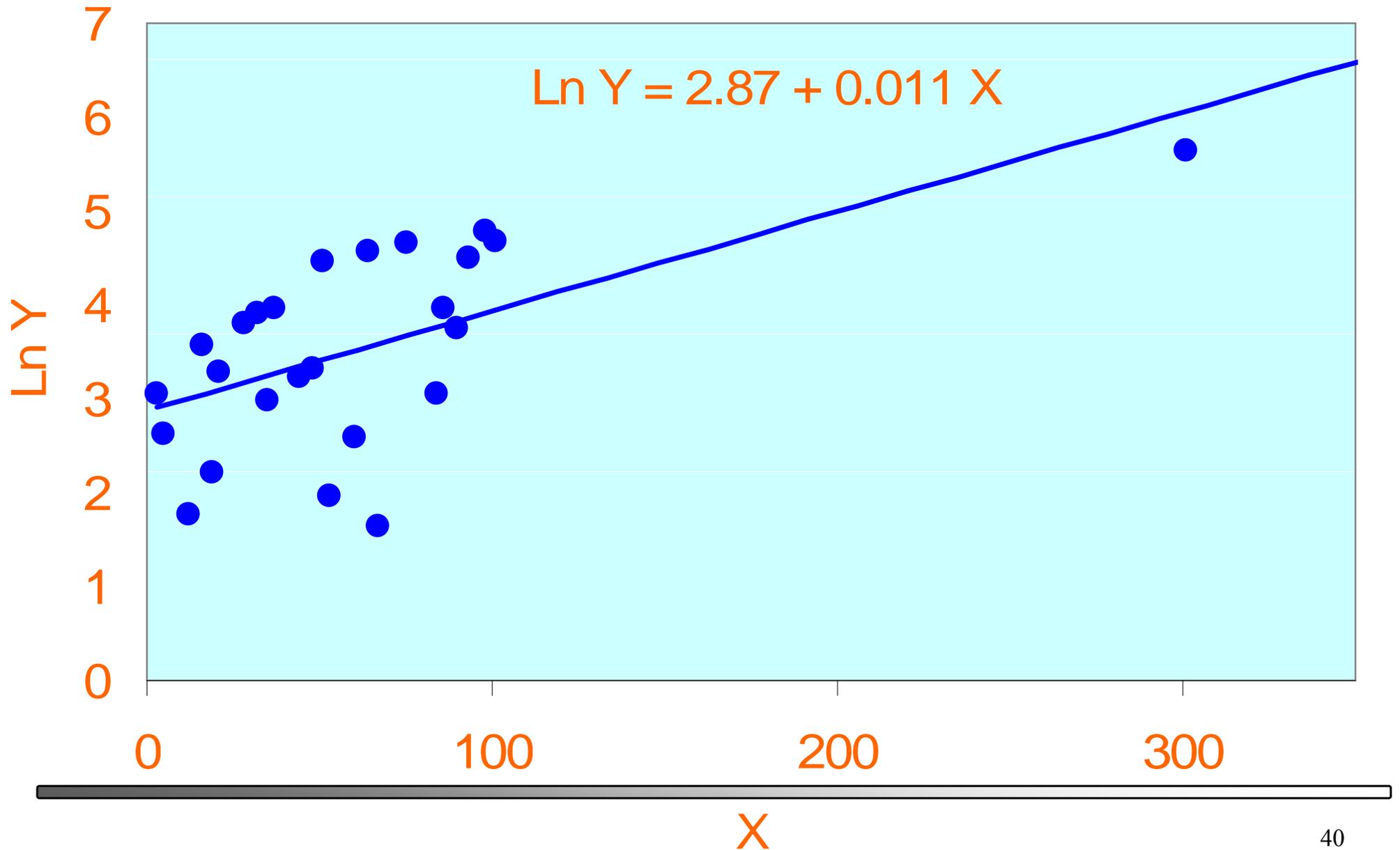


# Log Transformation of Cost

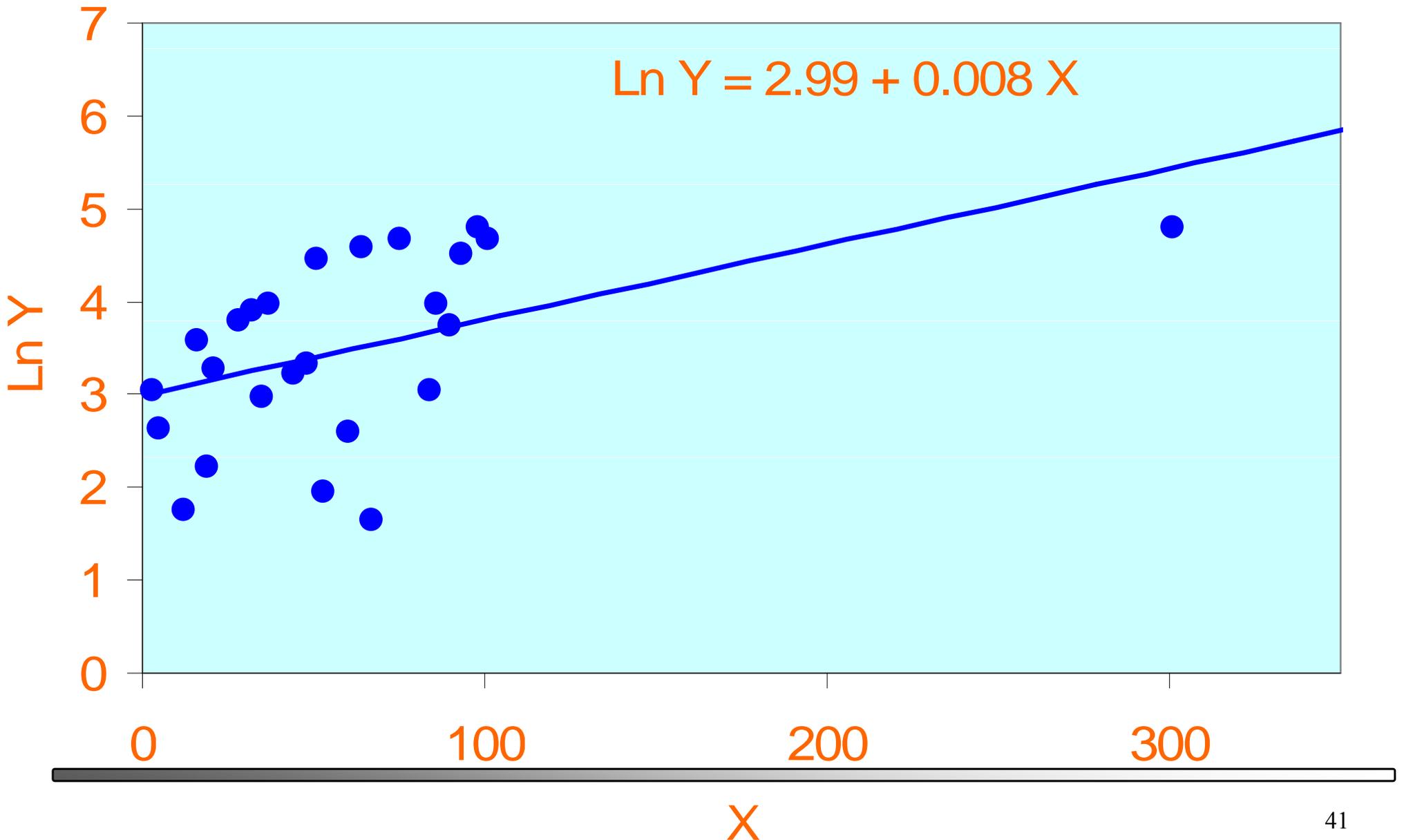
- Take natural log (log with base e) of cost
- Examples of log transformation:

COST	LN(COST)
\$10	2.30
\$1,000	6.91
\$100,000	11.51

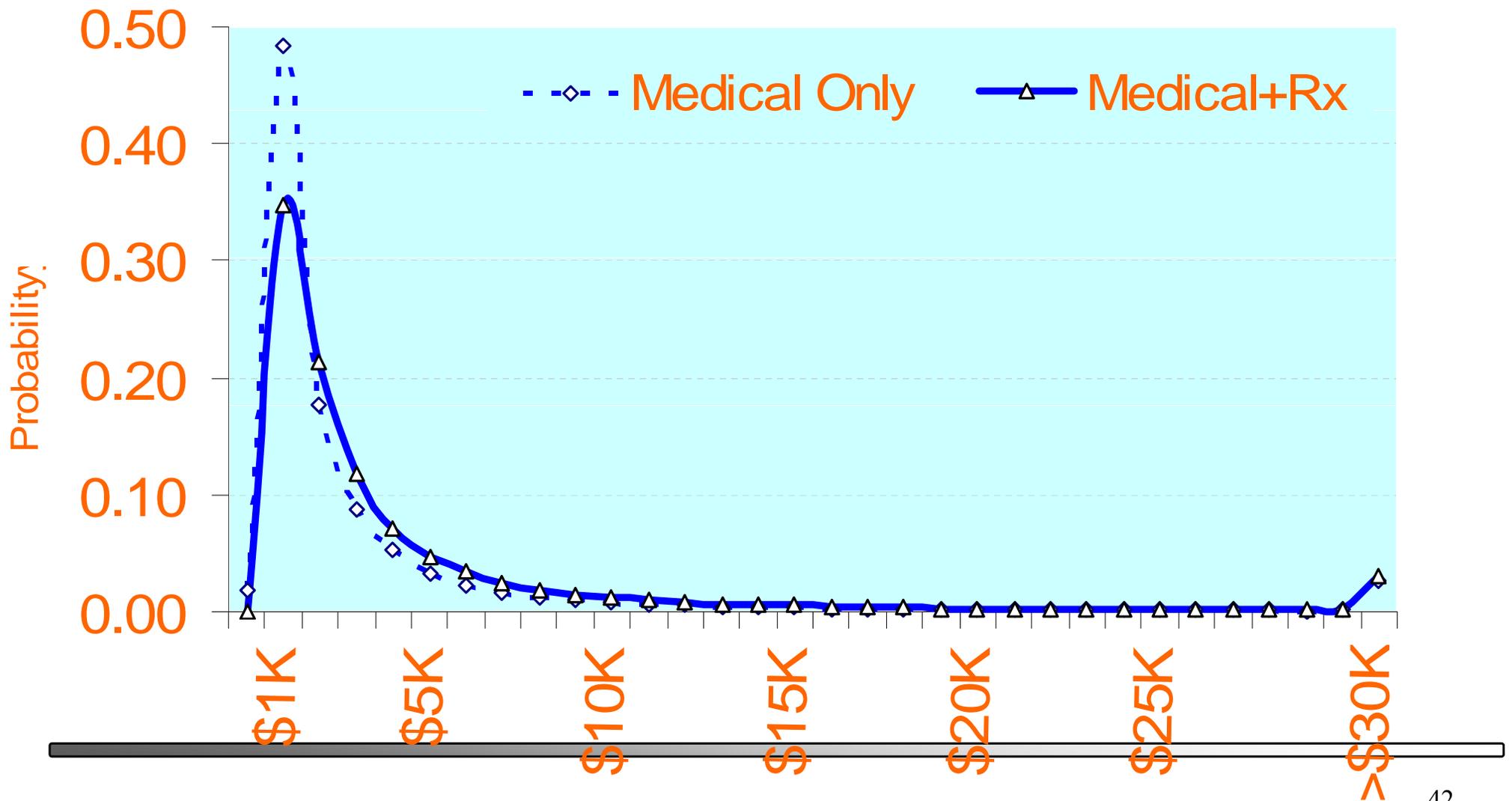
# Same data- outlier is less influential



# Same data- outlier is less influential

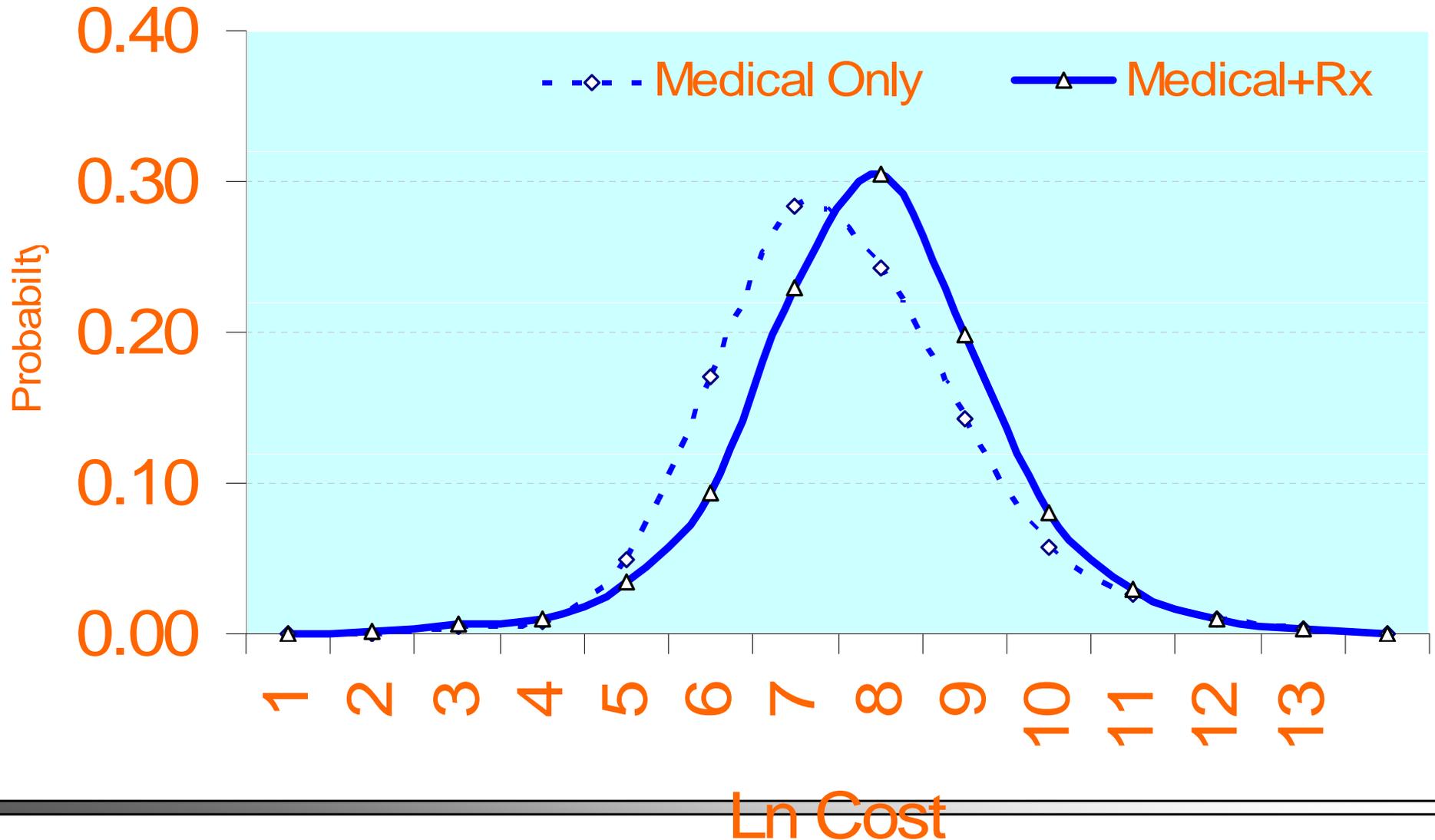


# Annual per person VHA costs FY06



# Effect of log transformation

## Annual per person VHA costs FY06



# Descriptive statistics: VHA costs FY06

(5% sample, includes outpatient pharmacy)

	Cost	Ln Cost
Mean	5,290	7.41
Median	1,646	7.41
Standard Deviation	16,507	1.47
Skewness	11.0	-0.10
Kurtosis	187.6	1.08

# Log linear model

- Regression with log dependent variable

$$\text{Ln } Y = \alpha + \beta X + \mu$$

# Log linear model

- $\text{Ln}(Y) = \alpha + \beta X + \mu$
- Parameters (coefficients) are not interpretable in raw dollars
  - Parameter represents the relative change of cost (Y) for each unit change in X
  - E.g. if  $\beta=0.10$ , then cost increases 10% for each unit increase in X

# What is the mean cost of the experimental group controlling for case-mix?

- We want to find the fitted value of  $Y$
- Conditional on  $X=1$
- With covariates held at the mean

$$\text{Ln}(Y) = \alpha + \beta_1 X + \beta_2 \bar{Z} + \mu$$

*What is  $\hat{Y}$ ?*

# Can we retransform by taking antilog of fitted values?

With the model:

$$\text{Ln}(Y) = \alpha + \beta_1 X + \beta_2 Z + \mu$$

*Does*

$$\hat{Y} = e^{\alpha + \beta_1 X + \beta_2 Z} ?$$

# What is fitted value of Y?

$$\begin{aligned} E(Y) &= E(e^{\alpha + \beta_1 X + \beta_2 Z + \mu_i}) \\ &= e^{\alpha + \beta_1 X + \beta_2 Z} E(e^{\mu_i}) \\ &= e^{\alpha + \beta_1 X + \beta_2 Z} \end{aligned}$$

*only if we can assume :*

$$E(e^{\mu_i}) = 1$$

# Retransformation bias

*Since  $E(\mu_i) = 0$*

*does  $E(e^{\mu_i}) = 1$  ?*

*Does  $e^{E(\mu_i)} = E(e^{\mu_i})$ ?*

# Retransformation bias

*Example of why  $E(e^{\mu_i}) \neq e^{E(\mu_i)}$*

*when  $\mu_1 = 1$  and  $\mu_2 = -1$ :*

$$e^{E(\mu^i)} = e^{+1-1} = e^0 = 1$$

$$E(e^{\mu_i}) = \frac{e^1 + e^{-1}}{2} = \frac{2.72 + 0.37}{2} = 1.5$$

# Retransformation bias

- The expected value of the antilog of the residuals  
does not equal
- The antilog of the expected value of the residuals

$$E(e^{\mu_i}) \neq e^{E(\mu_i)} !$$

# One way to eliminate retransformation bias: the smearing estimator

$$\begin{aligned} E(Y) &= E\left(e^{\alpha + \beta X_1 + \beta Z_2 + \mu_i}\right) \\ &= \left(e^{\alpha + \beta X_1 + \beta Z_2}\right) E\left(e^{\mu_i}\right) \\ &= \left(e^{\alpha + \beta X_1 + \beta Z_2}\right) \frac{1}{n} \sum_{i=1}^n \left(e^{\mu_i}\right) \end{aligned}$$

# Smearing Estimator

$$\frac{1}{n} \sum_{i=1}^n (e^{\mu_i})$$

# Smearing estimator

- This is the mean of the anti-log of the residuals
- Most statistical programs allow you to save the residuals from the regression
  - Find their antilog
  - Find the mean of this antilog
- The estimator is often greater than 1

# Correcting retransformation bias

- See Duan J Am Stat Assn 78:605
- Smearing estimator assumes identical variance of errors (homoscedasticity)
- Other methods when this assumption can't be made

# Retransformation

- Log models can be useful when data are skewed
- Fitted values must correct for retransformation bias

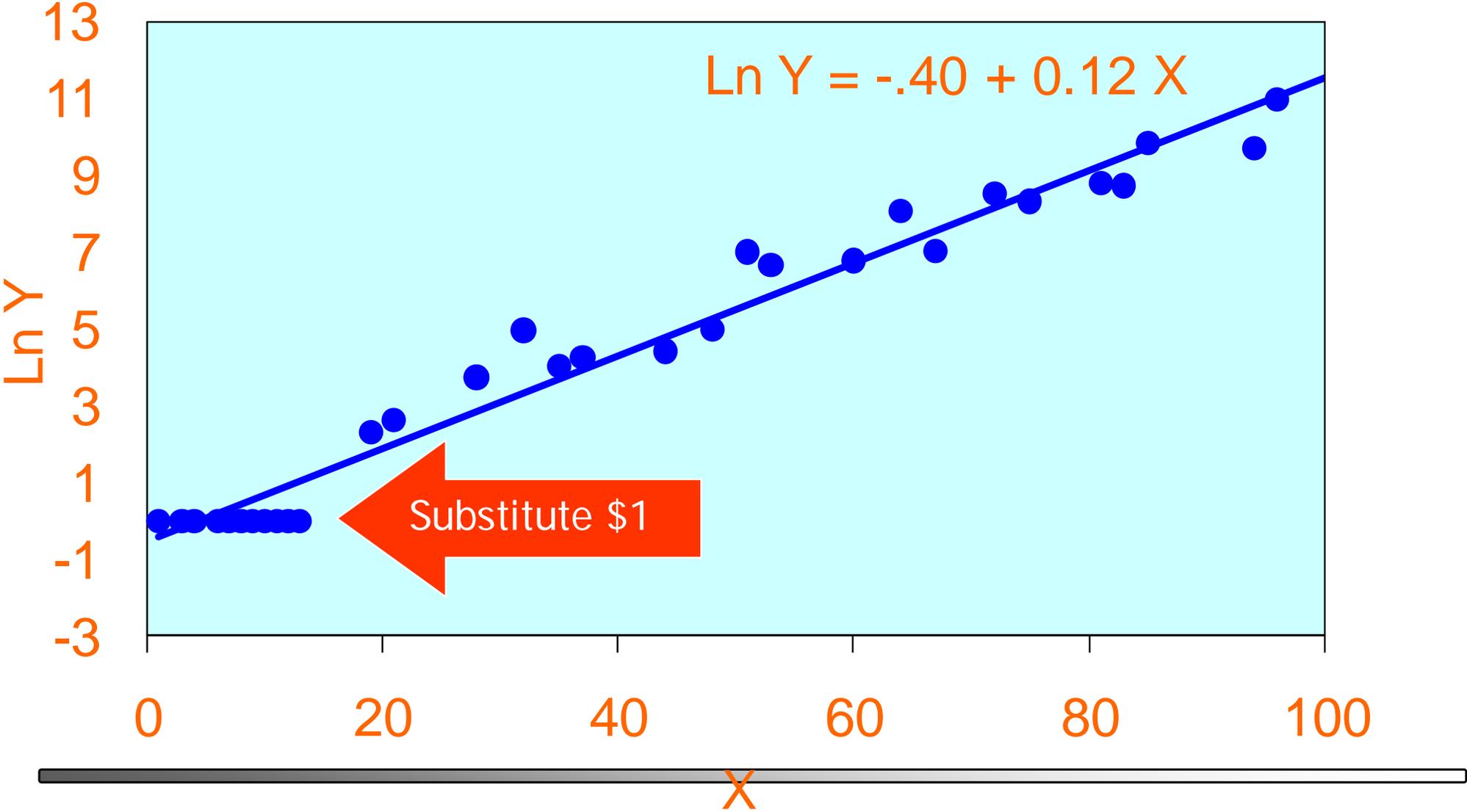
# Zero values in cost data

- The other problem: left edge of distribution is truncated by observations where no cost is incurred
- How can we find  $\text{Ln}(Y)$  when  $Y = 0$ ?
- Recall that  $\text{Ln}(0)$  is undefined

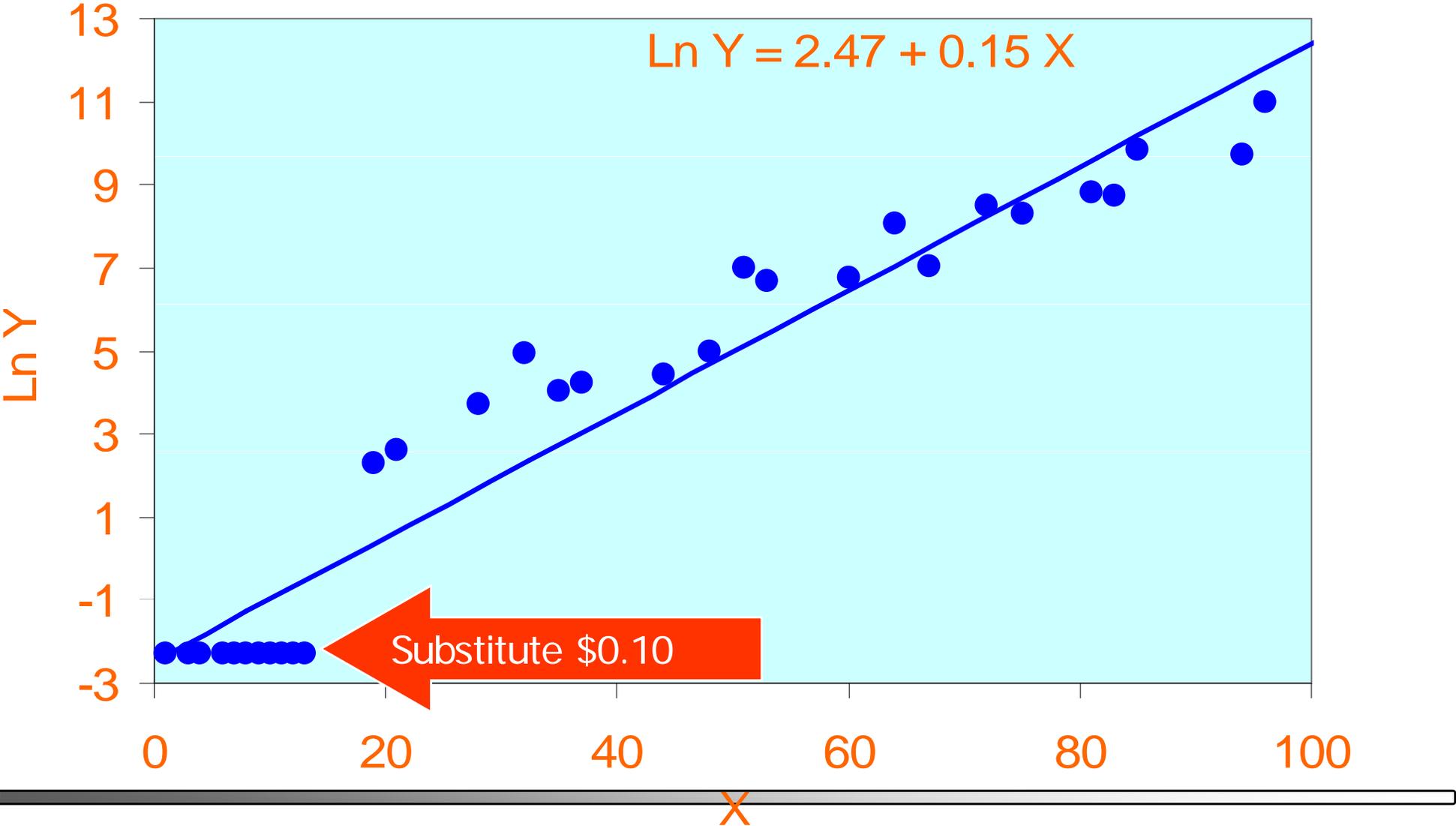
# Log transformation

- Can we substitute a small positive number for zero cost records, and then take the log of cost?
  - \$0.01, or \$0.10, or \$1.00?

# Substitute \$1 for Zero Cost Records



# Substitute \$0.10 for Zero Cost Records



# Substitute small positive for zero cost?

- Log model assumes parameters are linear in logs
- Thus it assumes that change from \$0.01 to \$0.10 is the same as change from \$1,000 to \$10,000
- Possible to use a small positive in place of zeros
  - if just a few zero cost records are involved
  - if results are not sensitive to choice of small positive value
- There are better methods!
  - Transformations that allows zeros (square root)
  - Two-part model
  - Other types of regressions

# Is there any use for OLS with untransformed cost?

- OLS with untransformed cost can be used:
  - When costs are not very skewed
  - When there aren't too many zero observations
  - When there is large number of observations
- Parameters are much easier to explain
- Can estimate in a single regression even though some observations have zero costs
- The reviewers will probably want to be sure that you considered alternatives!

# Review

- Cost data are not normal
  - They can be skewed (high cost outliers)
  - They can be truncated (zero values)
- Ordinary Least Squares (classical linear model) assumes error term (hence dependent variable) is normally distributed

# Review

- Applying OLS to data that aren't normal can result in biased parameters (outliers are too influential) especially in small to moderate sized samples

# Review

- Log transformation can make cost more normally distributed so we can still use OLS
- Log transformation is not always necessary or the only method of dealing with skewed cost

# Review

- Meaning of the parameters depends on the model
  - With linear dependent variable:
    - $\beta$  is the change in *absolute units* of  $Y$  for a unit change in  $X$
  - With logged dependent variable:
    - $\beta$  is the *proportionate change* in  $Y$  for a unit change in  $X$

# Review

- To find fitted value  $\hat{a}$  with linear dependent variable
- Find the linear combination of parameters and variables, e.g.

$$\hat{Y} | (X = 1, Z = \bar{Z}) = \alpha + \beta_1 + \beta_2 \bar{Z}$$

# Review

- To find the fitted value with a logged dependent variable
- Can't simply take anti-log of the linear combination of parameters and variables
- Must correct for retransformation bias

# Review

- Retransformation bias can be corrected by multiplying the anti-log of the fitted value by the smearing estimator
- Smearing estimator is the mean of the antilog of the residuals

$$E(Y | X = 1, Z = \bar{Z}) = \left( e^{\alpha + \beta + \beta_2 \bar{Z}} \right) \frac{1}{n} \sum_{i=1}^n (e^{\mu_i})$$

# Review

- Cost data have observations with zero values, a truncated distribution
- $\ln(0)$  is not defined
- It is sometimes possible to substitute small positive values for zero, but this can result in biased parameters
- There are better methods

# Next session- November 3

- Two-part models
- Regressions with link functions
- Non-parametric statistical tests
- How to determine which method is best?

# Reading assignment on cost models

Basic overview of methods of analyzing costs

- P Dier, D Yanez, A Ash, M Hornbrook, DY Lin. Methods for analyzing health care utilization and costs Ann Rev Public Health (1999) 20:125-144

■ [HERC@va.gov](mailto:HERC@va.gov)

# Supplemental reading on Log Models

- Smearing estimator for retransformation of log models
  - Duan N. Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association (1983) 78:605-610.
- Alternatives to smearing estimator
  - Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics (1998) 17(3):283-295.

# Appendix: Derivation of the meaning of the parameter in log model

$$\text{Ln } Y = \alpha + \beta X + \mu$$

$$\frac{d\text{Ln } Y}{dx} = \beta, \text{ as } d\text{Ln}Y = dY / Y$$

$$\frac{dY / Y}{dx} = \beta$$

$\beta$  is the proportional change in Y for a small change in X