

Econometrics Course: Cost as the Dependent Variable (I)



Paul G. Barnett, PhD

May 27, 2015

What is health care cost?

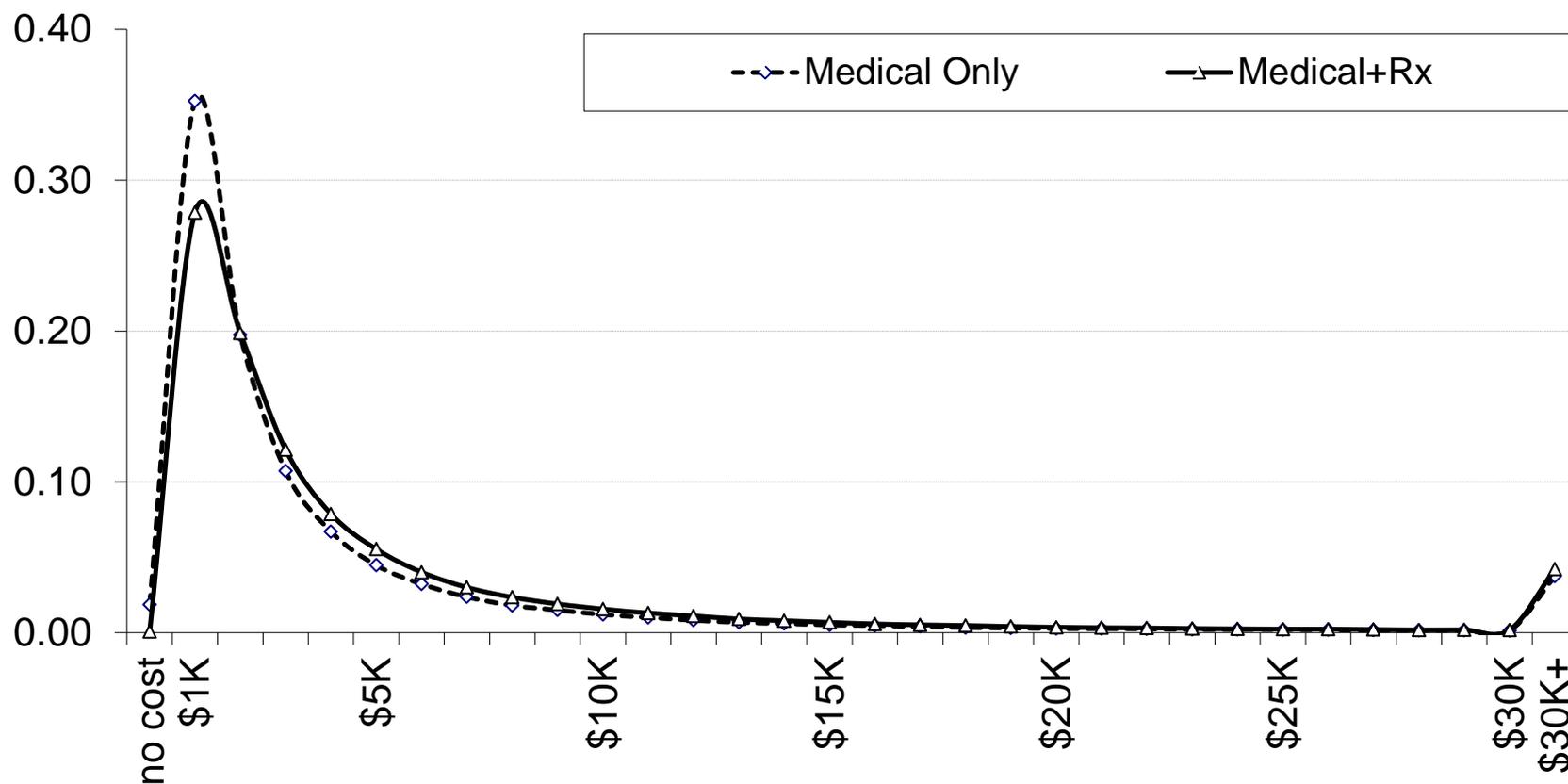
- Cost of an intermediate product, e.g.,
 - chest x-ray
 - a day of stay
 - minute in the operating room
 - a dispensed prescription
 - Cost of a bundle of products
 - Outpatient visit
 - Hospital stay
-

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 FY10

(5% random sample)



Descriptive statistics: VHA costs FY10

(5% sample, includes outpatient pharmacy)

Cost

Mean 5,768

Median 1,750

Standard Deviation 18,874

Skewness 13.98

Kurtosis 336.3

Skewness and kurtosis

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

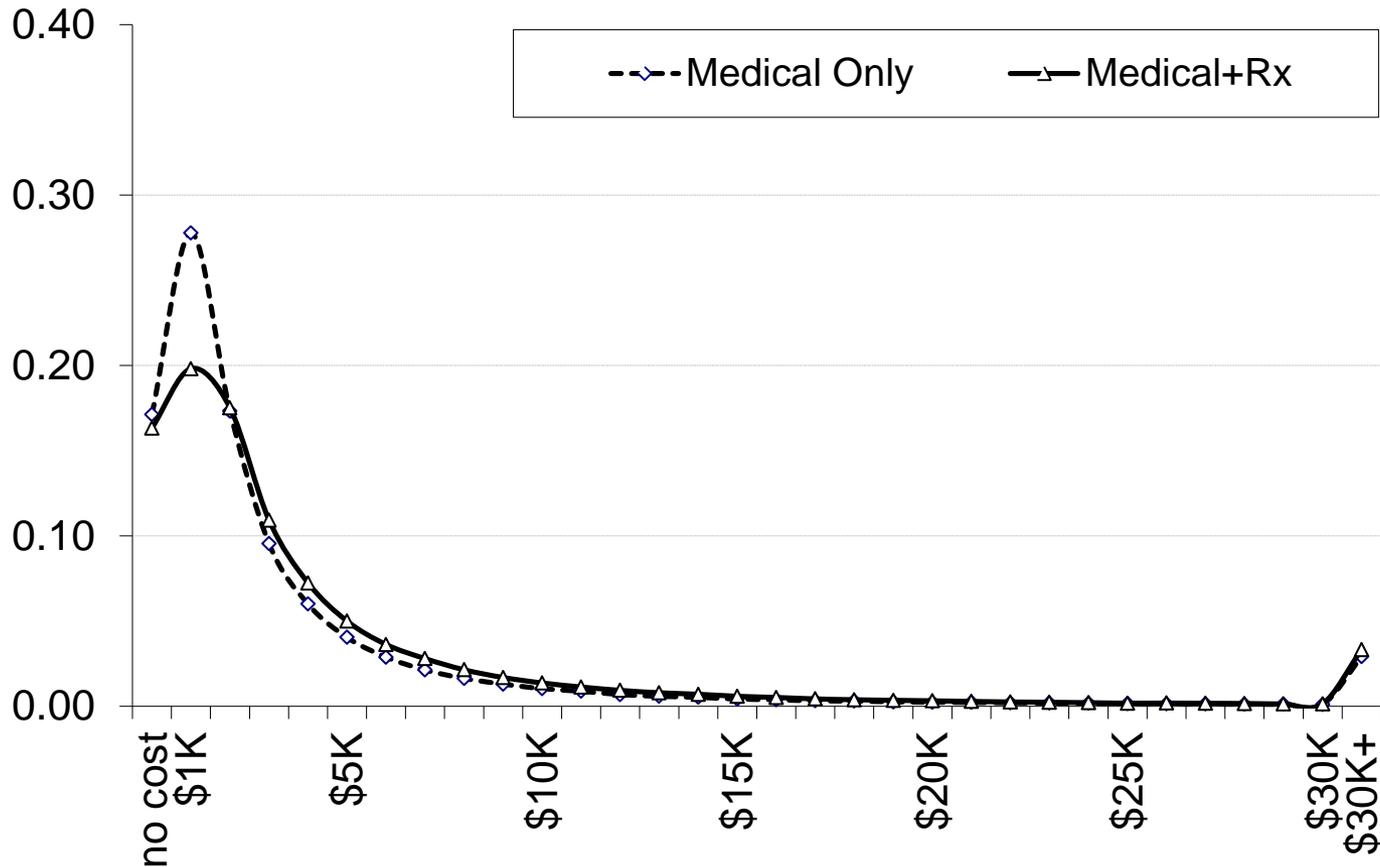
Distribution of cost: skewness

- 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 FY10 *among those who used VHA in FY09*



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) α , β
- 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)
 - $Y_i = \alpha + \beta X_i + \varepsilon_i$
- β 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

Expected value of a random variable (cont)

- Expected value from a roll of the die
- $E(W) = \sum W_i p(W_i)$

$$1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) \\ = 3.5$$

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)

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

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

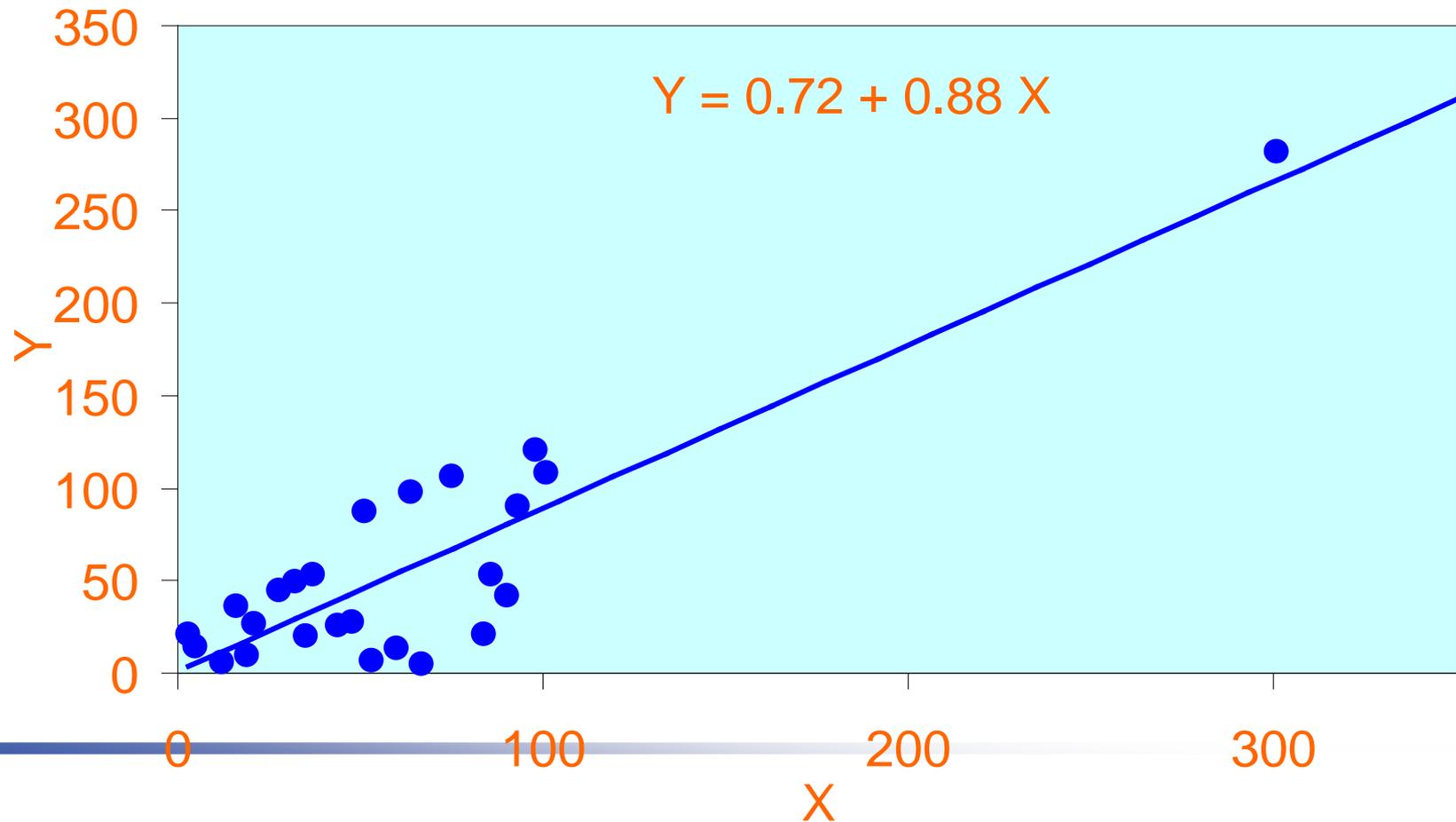
Assumptions are about error term

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

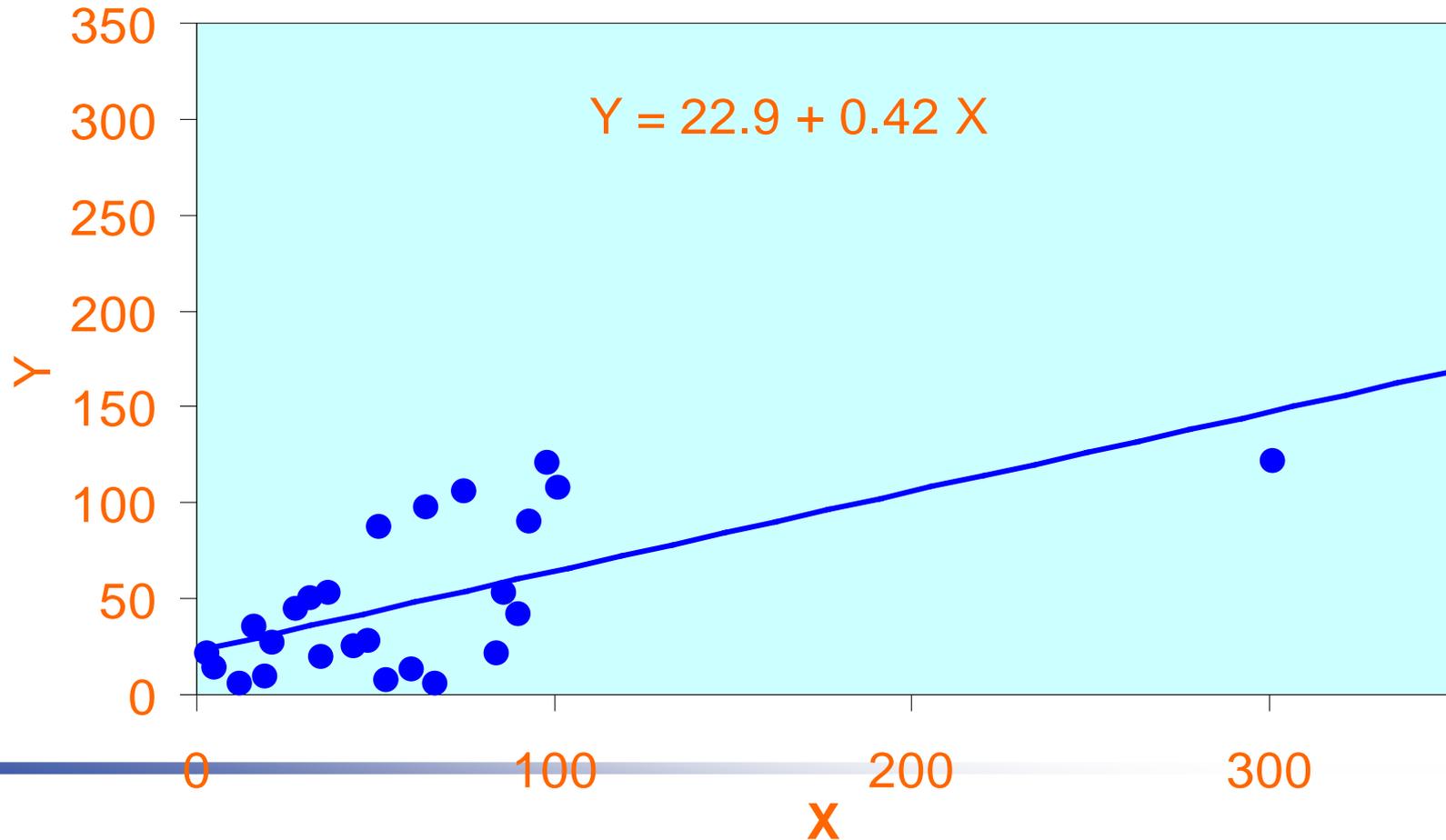
Why worry about using OLS with skewed (non-normal) data?

- “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

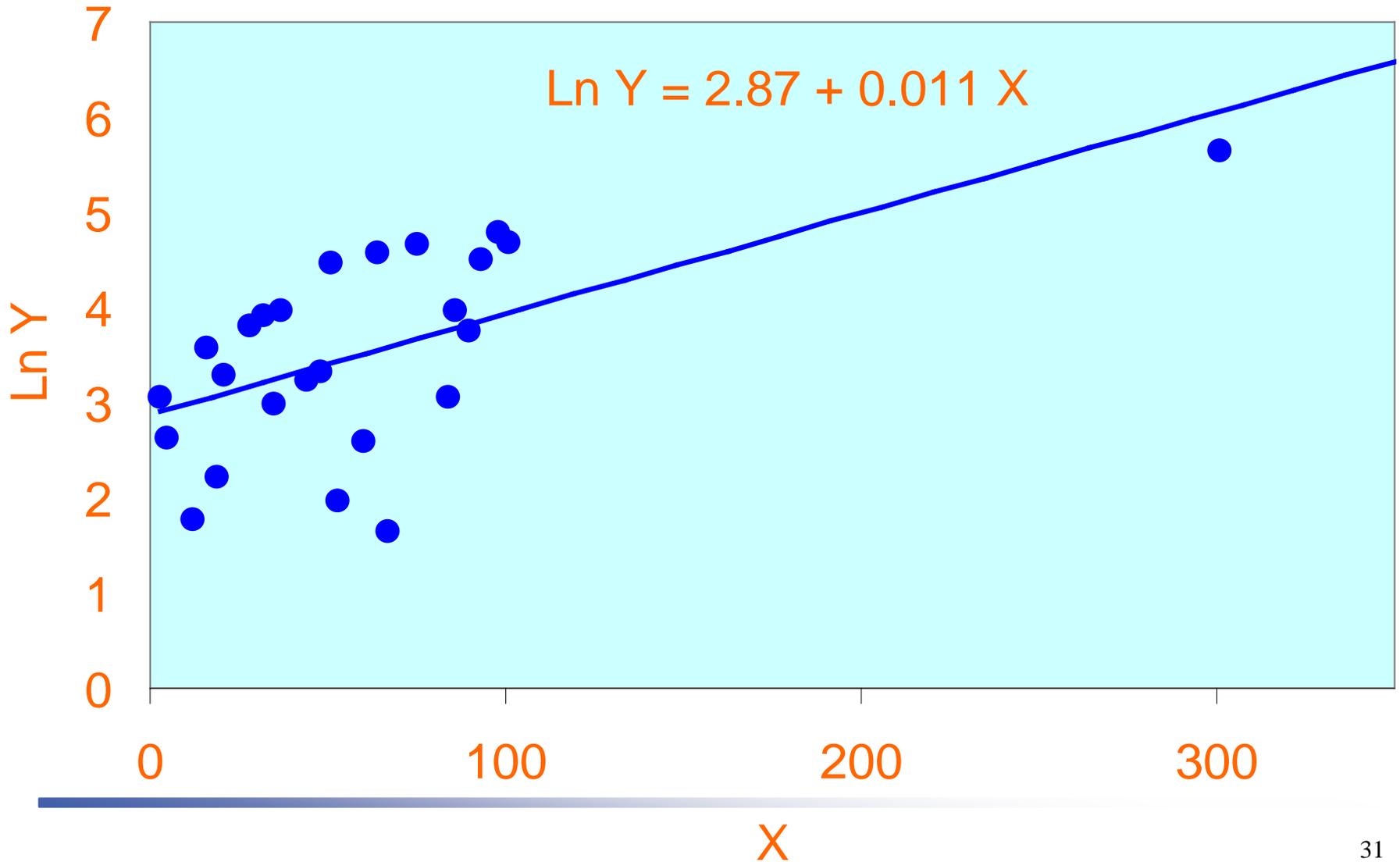


Log Transformation of Cost

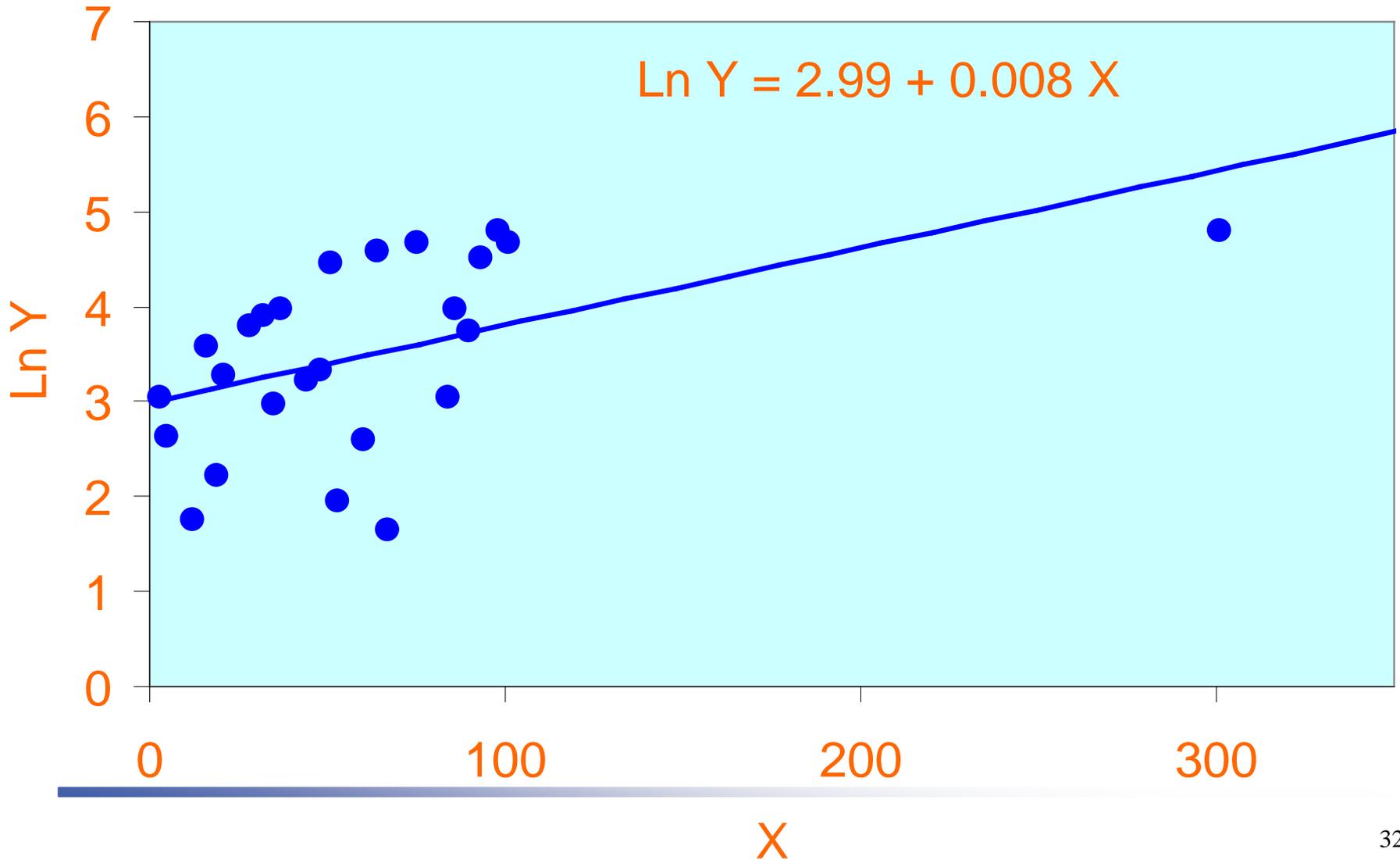
- Take natural log (log with base e) of cost
- Examples of log transformation:

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

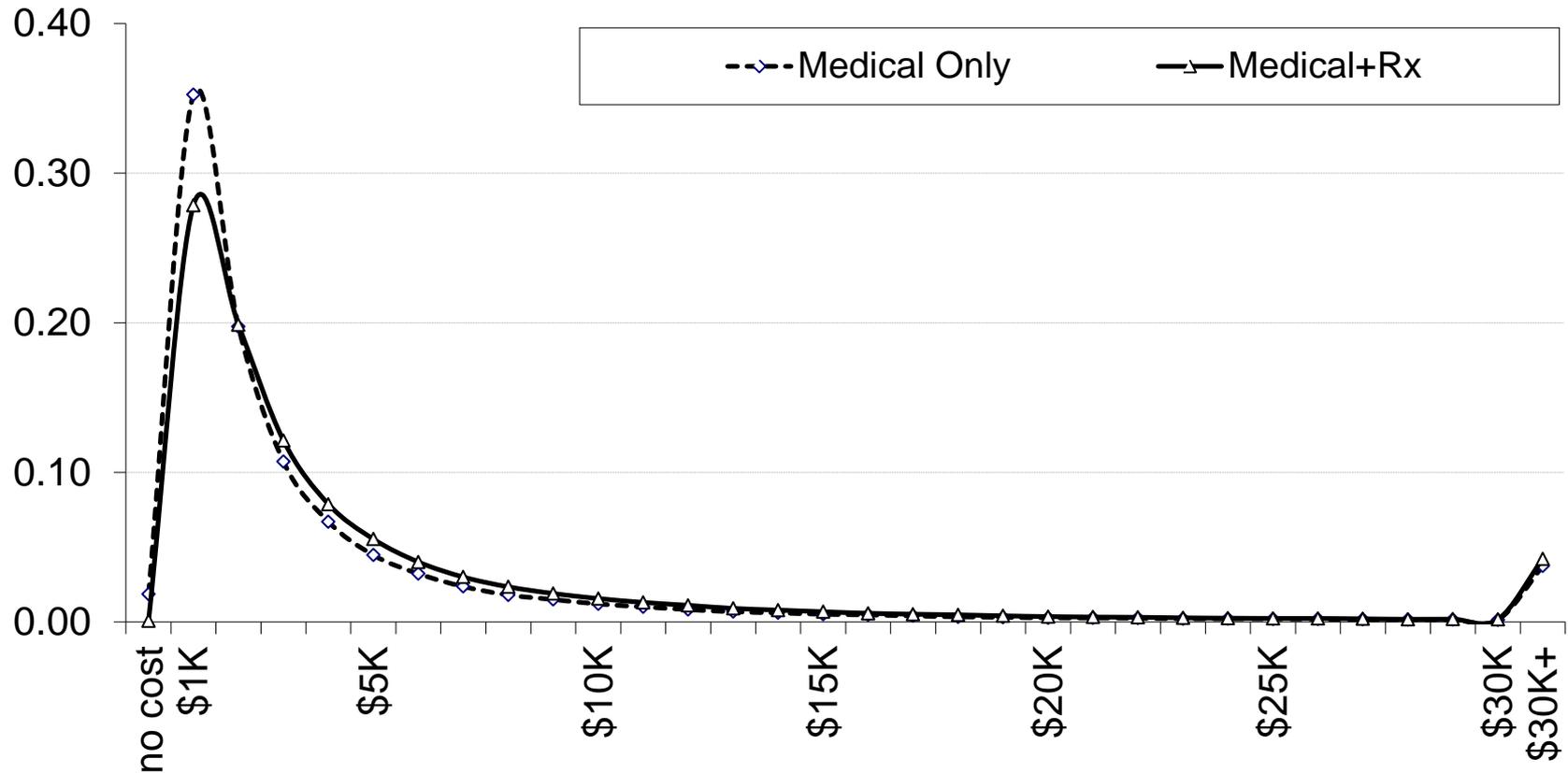
Same data- outlier is less influential



Same data- outlier is less influential

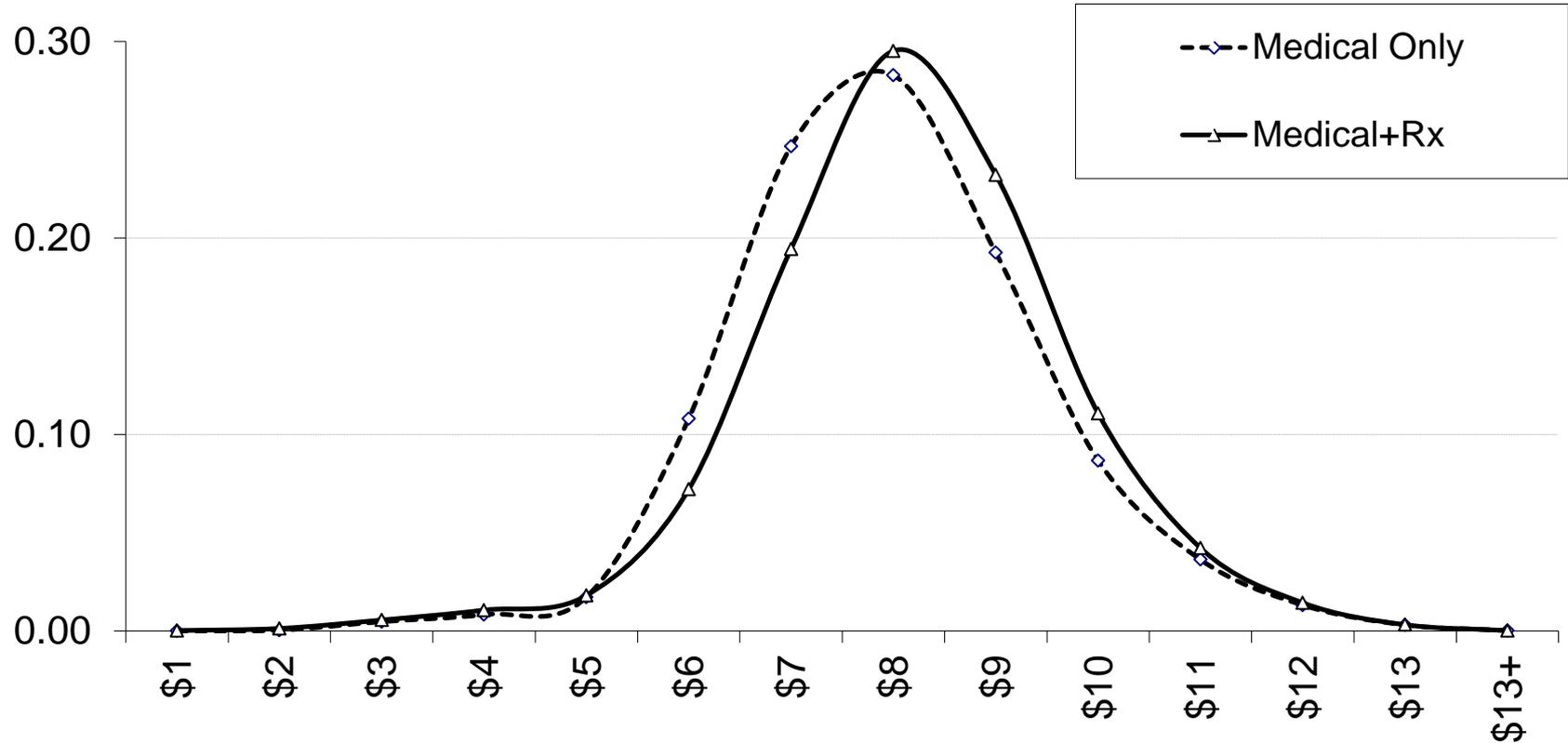


Annual per person VHA costs FY10



Effect of log transformation

Annual per person VHA costs FY10



Descriptive statistics: VHA costs FY10

(5% sample, includes outpatient pharmacy)

| | Cost | Ln Cost |
|--------------------|--------|---------|
| Mean | 5,768 | 7.68 |
| Median | 1,750 | 7.67 |
| Standard Deviation | 18,874 | 1.50 |
| Skewness | 13.98 | -0.18 |
| Kurtosis | 336.3 | 1.12 |

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

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

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
-

Smearing estimator in SAS

- Save the residuals from the regression
 - PROC REG;
 - MODEL LNCOST=GROUP;
 - OUTPUT OUT=my_file RESIDUALS=my_residuals
- Find their antilog
 - DATA my_file; SET my_file;
 - my_smear=EXP(my_residuals)
- Find the mean
 - PROC MEANS DATA=my_file MEAN;
 - VAR my_smear;

Smearing estimator in SAS

- Predict the log cost (PREDICTED)
- Transform back into natural units (dollars of cost)
- $\text{Predicted} = \text{EXP}(\text{Predicted_lncost}) * \text{my_smear};$

Predicting the log cost

- Simple approach: use coefficients
 - INTERCEPT (predicted log cost if not in the group)
 - INTERCEPT + BETA (predicted log cost if in the group)
 - This yields two predict log cost
 - Retransform to dollars cost with smearing correction and compare them
- What if there are other covariates besides GROUP?
 - Evaluate parameters with covariates set to their mean?
 - This can lead to bias
 - The antilog of the mean is not the mean of the antilog!

Predicting the log cost (continued)

- When there are other variables in the regression
- Predict log cost for each scenario given the values of other variables for that observation
 - “as if” observation was in the group (set GROUP to ONE)
 - “as if” observation wasn’t in the group (set GROUP to ZERO)
- Retransform each predicted cost to “natural units of cost” with smearing estimator
- Find mean of the N observations for each scenario
- Compare the means of the corrected predictions

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

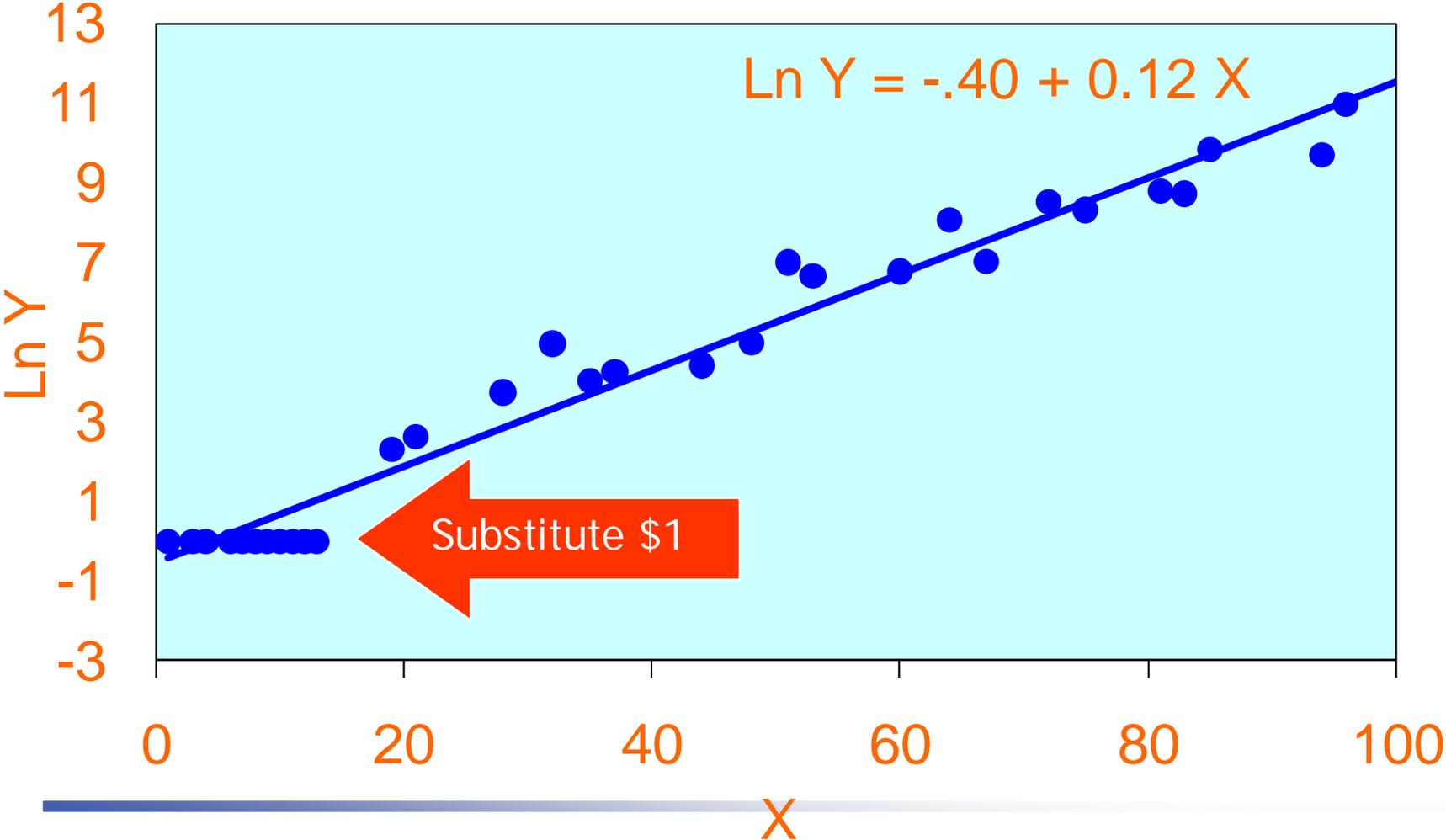
Zero values in cost data

- 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

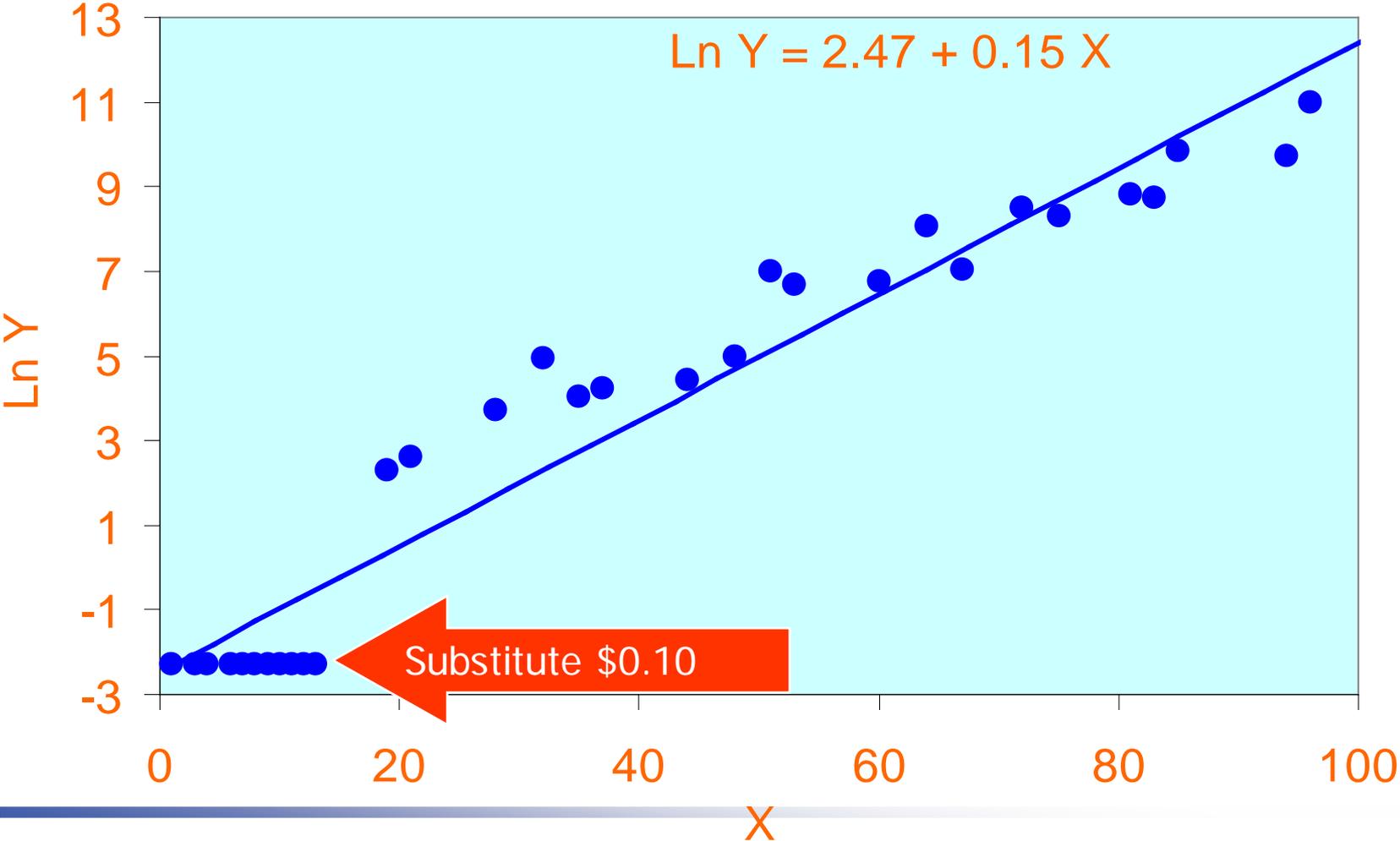
Log transformation

- 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:
 - β is the change in *absolute units* of Y for a unit change in X
 - With logged dependent variable:
 - β 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

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- December 4

- 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

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

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

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

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

β is the proportional change in Y for a small change in X