Econometrics Course: Cost as the Dependent Variable (I)

Paul G. Barnett, PhD

March 6, 2019

HERC
Health Economics Resource Center

HSR&D
HEALTH SERVICES RESEARCH & DEVELOPMENT

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
- Economic analyses are not limited to cost; they almost always considers non-cost outcomes

POLL QUESTION #1

What hypotheses involving cost do you want to test? (select all that apply)

- Compare VA cost to non-VA cost of a service
- Compare cost of intervention group to control group
- Compare cost of between providers, clinics, or sites
- Find extra cost associated with a condition
- Other/None

Typical goals of cost analysis

- Find difference in cost between two or more groups of patients
  - E.g. groups defined by intervention, provider, setting
- Simulate the cost of a particular group of patients
  - Predict cost given characteristics of patient

Find Difference in Costs

- Cost (Y) modeled as a function of independent variable(s) (X) with errors in prediction $Y_i = \alpha + \beta X_i + \varepsilon_i$
- We estimate parameter(s) (β)
- When X is an indicator variable (e.g. 1 if treated, 0 if control), (β) is the difference in costs

Simulate costs

- Regression model is used to predict cost ($\hat{Y}$) for a given value of the independent variable(s)

$$\hat{Y}_i = \alpha + \beta X_i + \varepsilon_i$$

- The residual is the error in prediction

$$\varepsilon_i = Y_i - \hat{Y}_i$$

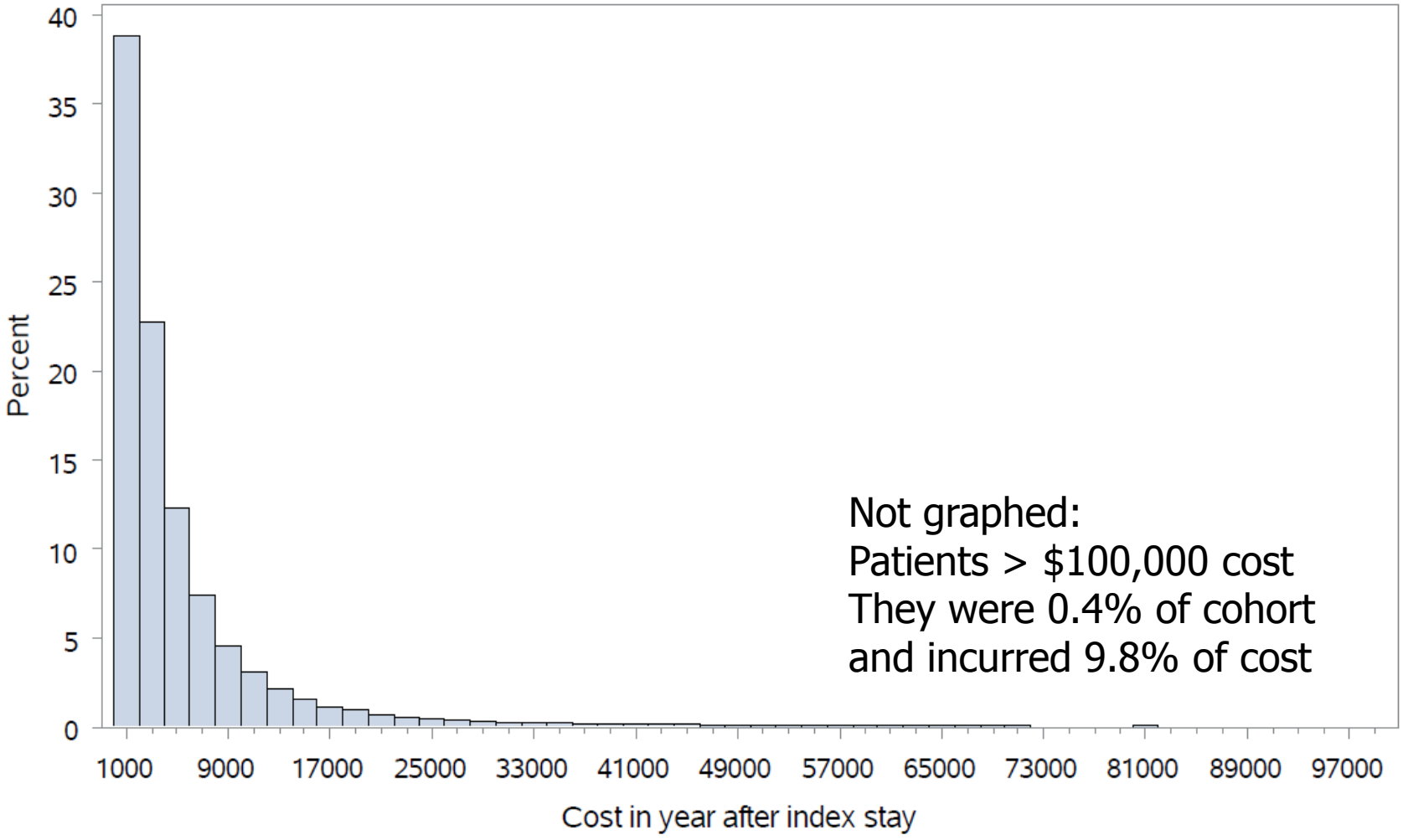
Dataset for worked examples

- VA Primary care patients new episode of non-specific low-back pain in 2016
- ~10% Sample N=43,909
- Costs in the following year
  - VA & Purchased Care
  - National Average Cost (HERC)
  - Excludes residential, nursing home

Dataset for worked examples

- Explanatory Variables
  - Baseline patient diagnoses, demographics
  - Limited attributes of provider & site
  - Indicators of early MRI vs early PT
- We'll compare CBOC vs VAMC
- Acknowledge: Jeanie Lo, Jo Jacobs, HSR&D IIR 15-450

One year cost of primary care patients with new episode of non-specific low back pain in 2016



Descriptive statistics

One year cost of patients with new episode of non-specific low back pain

	Cost
Mean	6,304
Median	2,840
Standard Deviation	14,084
Skewness	13.8
Kurtosis	467.0

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)

Cost by type clinic where episode began

	CBOC	Hospital Based	Difference
Mean	5,493	7,438	-1,945
Median	2,493	3,406	-913

POLL QUESTION #2

What is the better measure for comparing cost incurred by members of two groups?

- Mean
- Median

Compare groups while controlling for case mix

- Multivariate regression
- Cost is the dependent variable
- Independent variables to represent differences in patients and setting

Review of Ordinarily Least Squares (OLS)

- Dependent variable assumed to be a linear function of the chosen independent variables $Y_i = \alpha + \beta X_i + \varepsilon_i$
- OLS Estimates parameters (coefficients) α, β by minimizing the sum of squared errors (the distance between data points and the regression line)

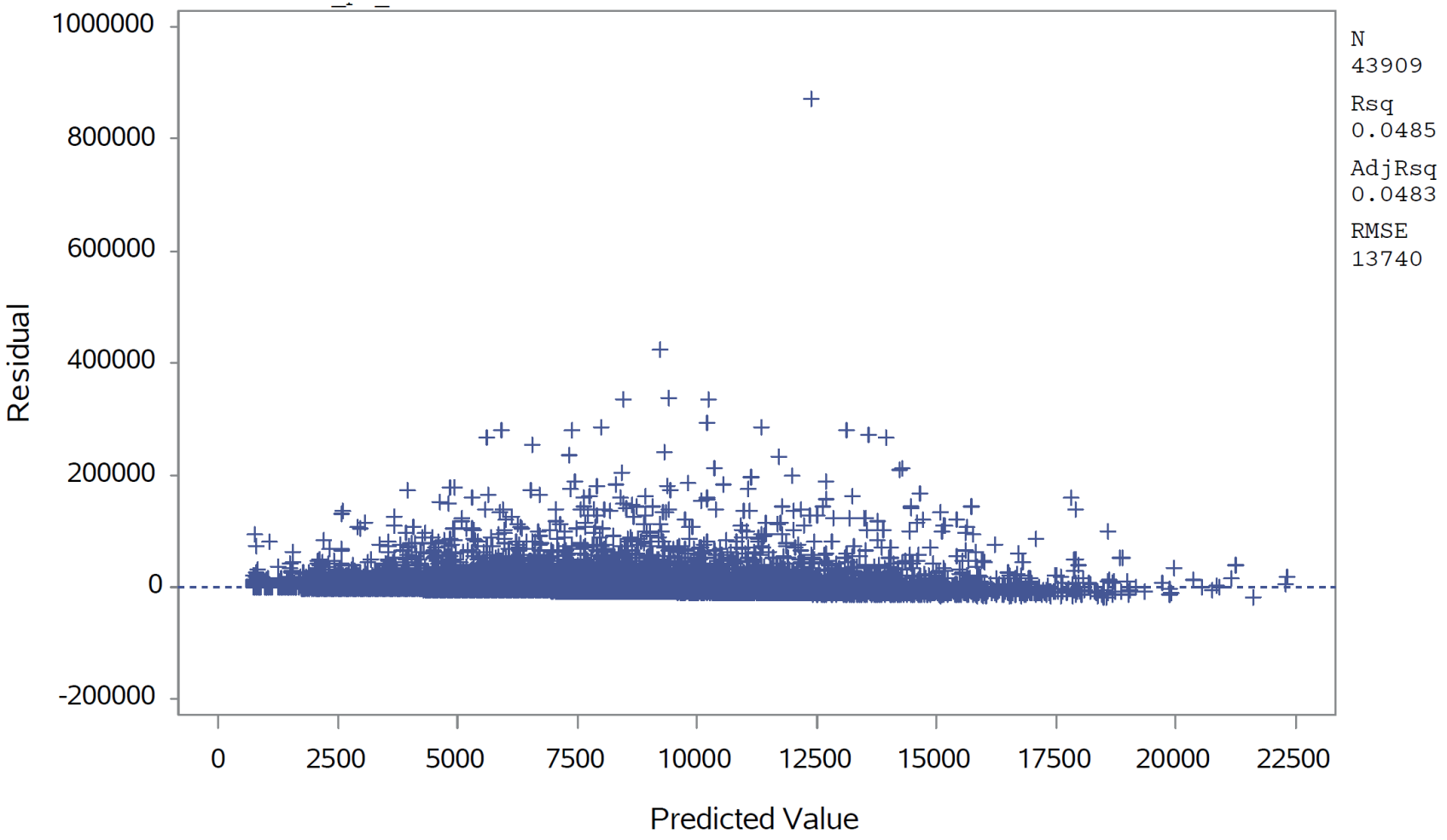
OLS model

- Regression with raw cost as 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 β=1000, then cost increases $1,000 for each unit increase in X
  - If X is an indicator variable, then group with X=1 has $1,000 greater cost

Regression Results

	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	2434	289	8.41	<.0001
Index visit at CBOC	-1757	133	-13.17	<.0001
Baseline pain score	256	21	12.01	<.0001
Age	2.0	4.7	0.44	0.6629
Number of chronic conditions	1370	47	29.31	<.0001
Indicator of diagnosis for substance use or psychiatric diagnosis	949	163	5.81	<.0001
Female	691	262	2.64	0.0082
History of opiate Rx in prior year	1099	211	5.2	<.0001
MRI within 6 weeks	2810	422	6.65	<.0001
Number of visits for physical therapy within 6 weeks	572	141	4.06	<.0001

Plot of residuals



POLL QUESTION #3

Which OLS assumptions aren't appropriate when raw cost is the dependent variable? (select all that apply)

- Expected error is zero
- Errors are independent
- Errors have identical variance
- Errors are normally distributed
- Errors not correlated with independent variables

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

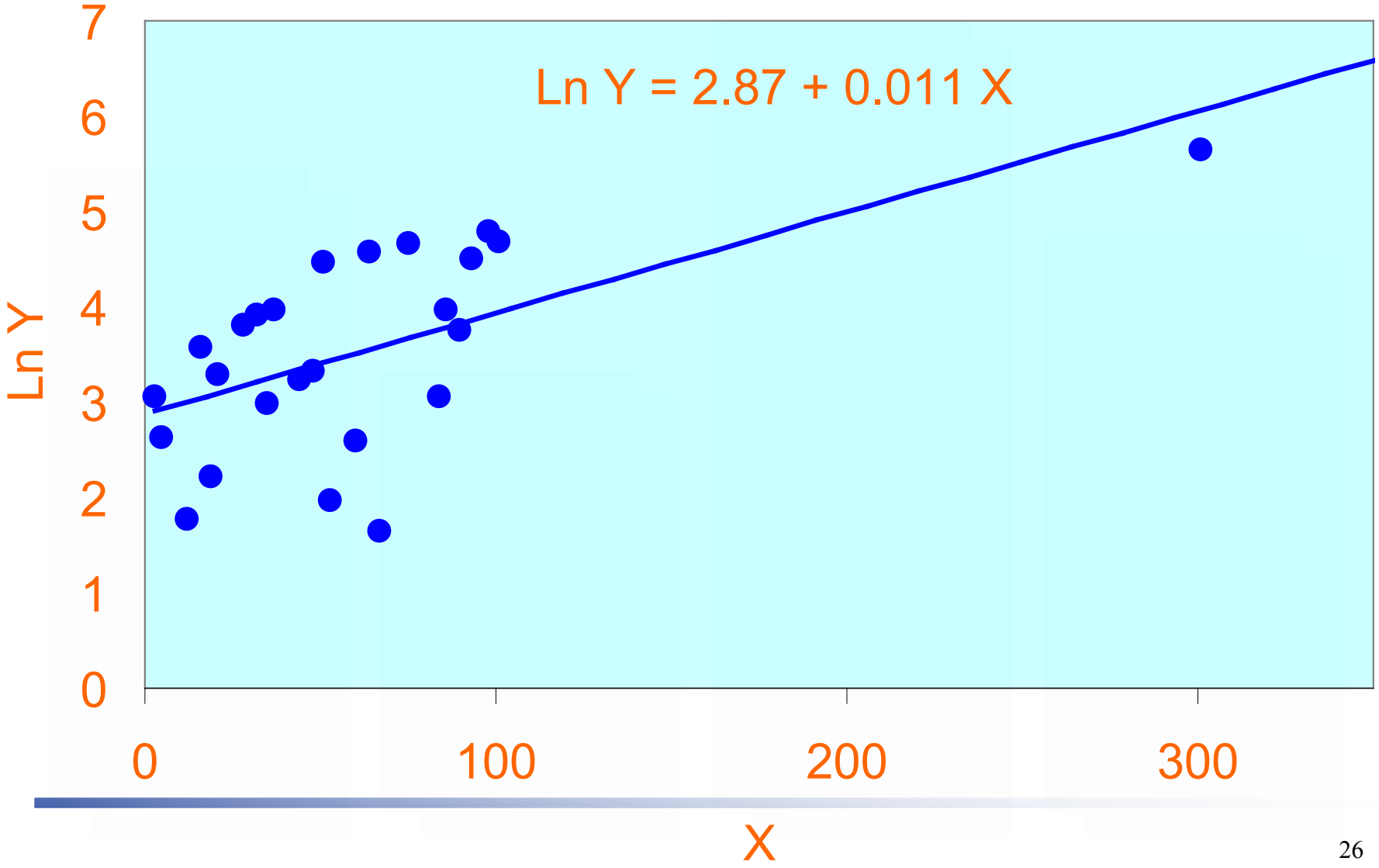
- "In small and moderate sized samples, a single case can have tremendous influence on an estimate"
  - Will Manning
  - Elgar Companion to Health Economics AM Jones, Ed. (2006) p. 439
- There are no values skewed to left to balance the influence of values skewed to the right
- In Rand Health Insurance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

Log Transformation of Cost

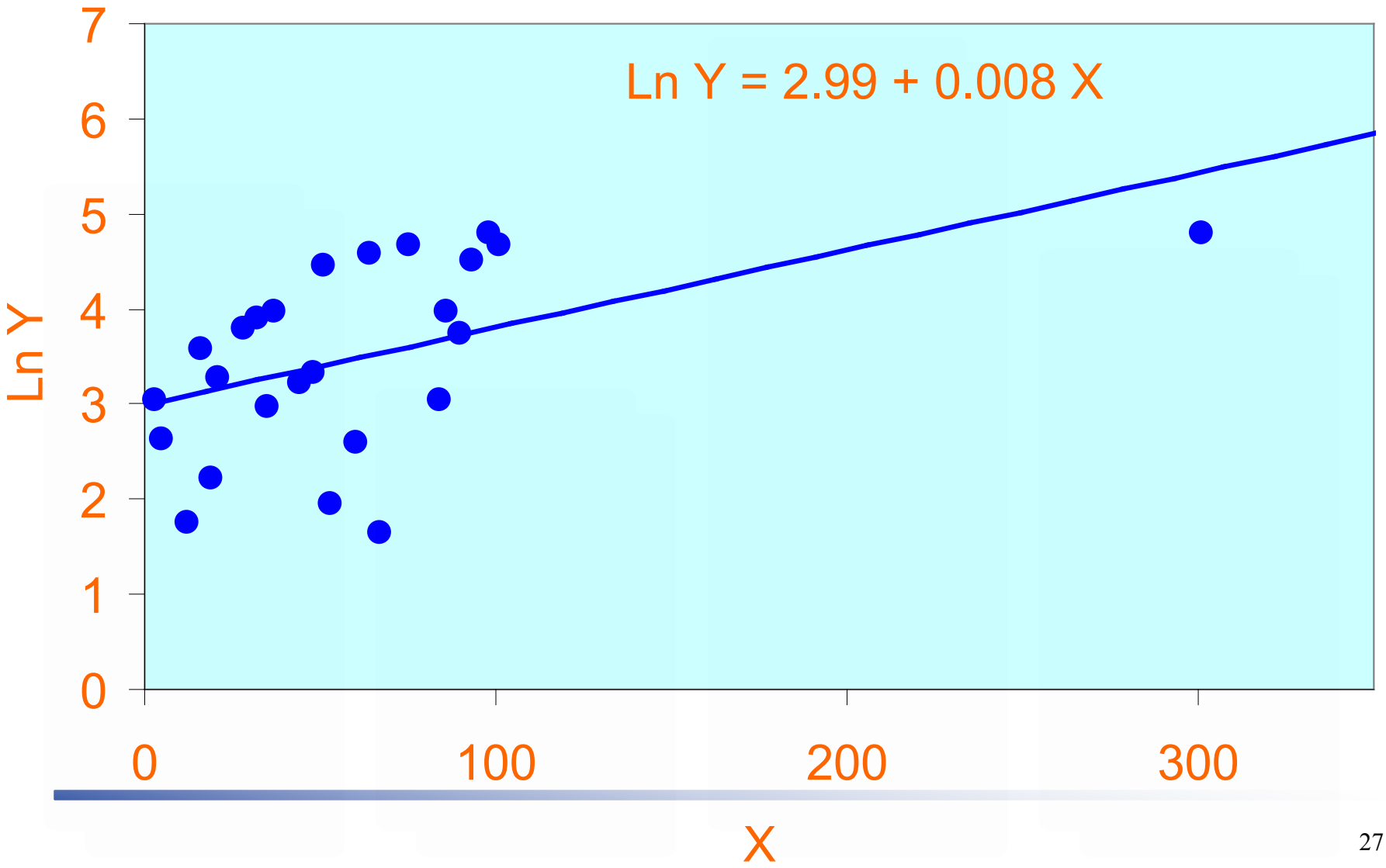
- Take natural log (log with base e) of cost
- Examples of log transformation:

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

Same data- outlier is less influential

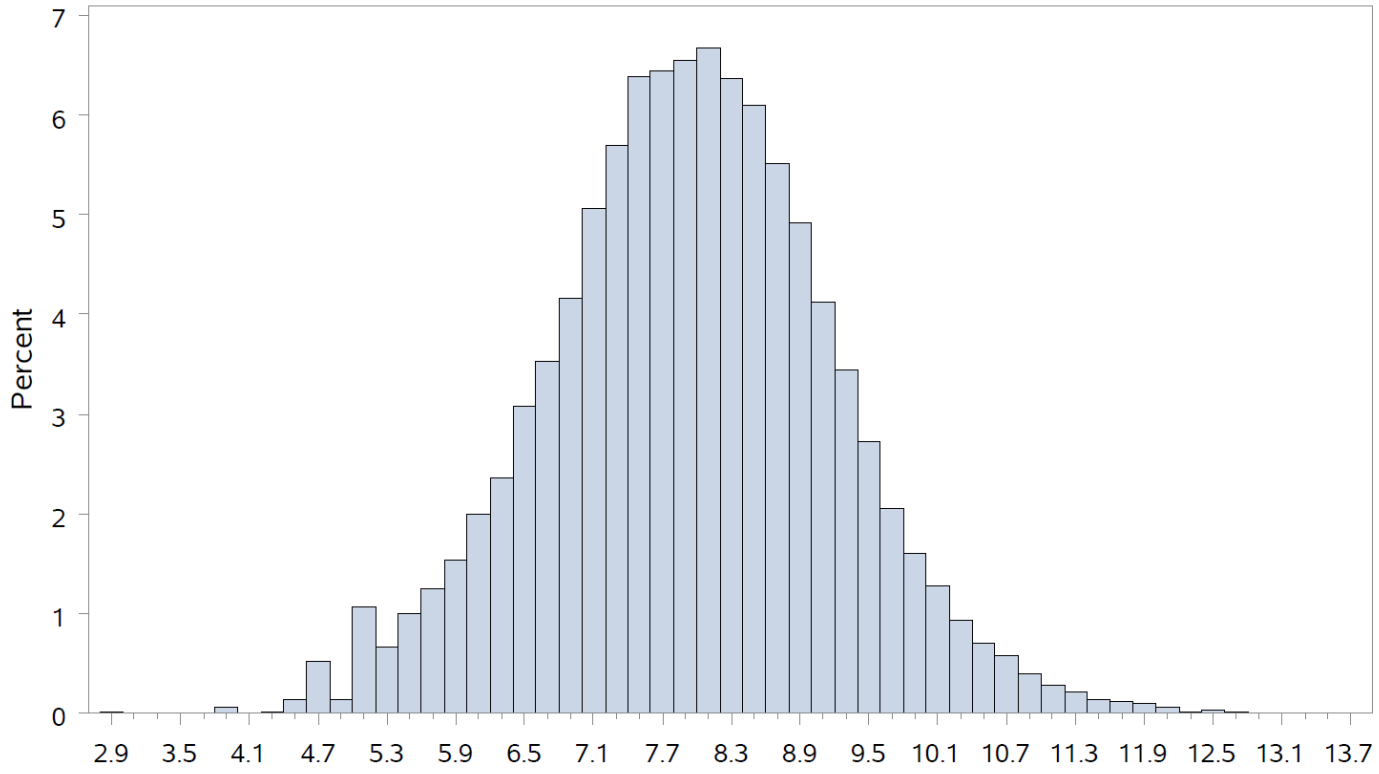


Same data- outlier is less influential



Log of one year cost

Primary care patients with new episode of non-specific low back pain in 2016



Descriptive statistics

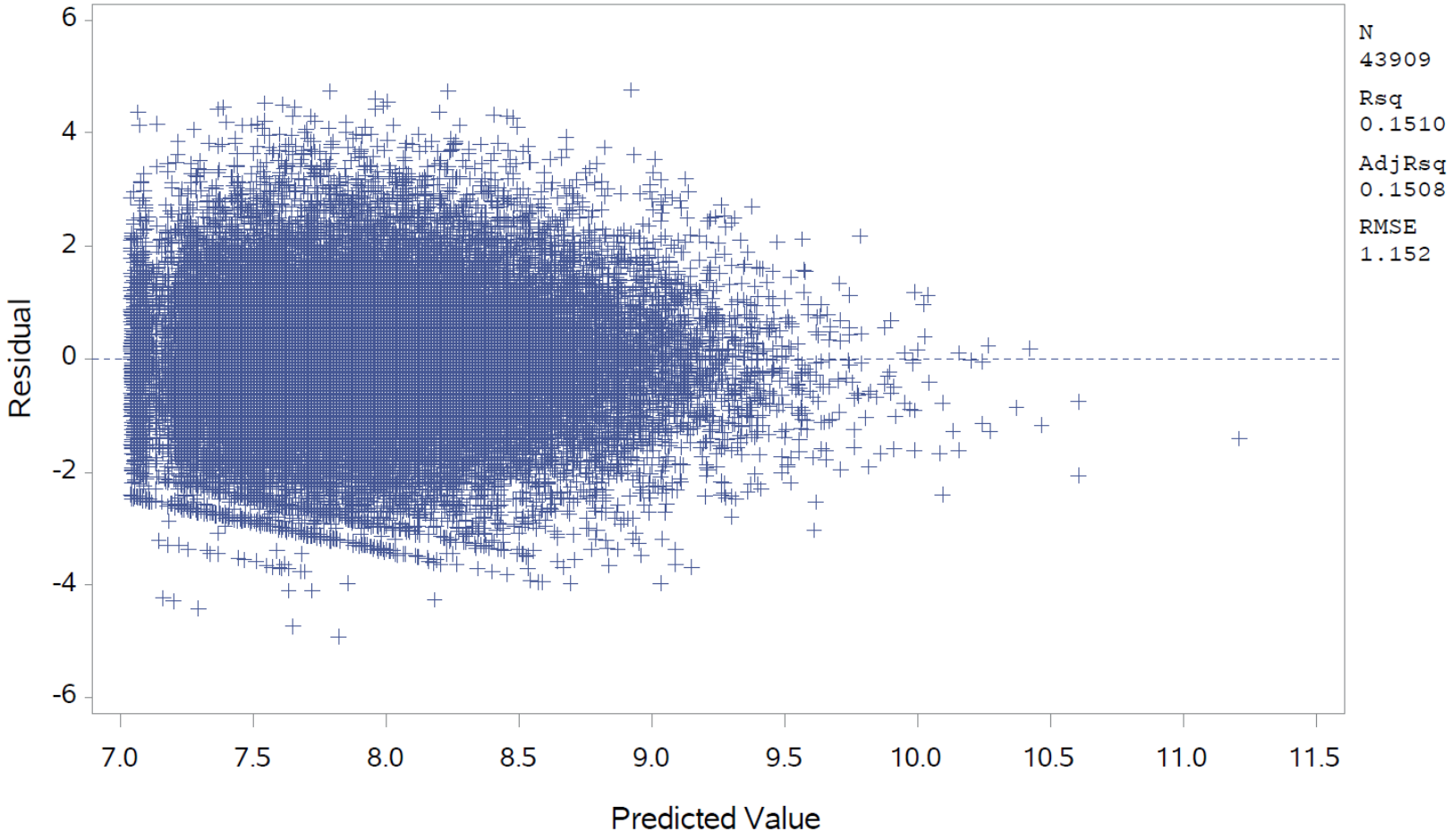
One year cost of patients with new episode of non-specific low back pain

	Cost	LnCost
Mean	6,304	7.94
Median	2,840	7.95
Standard Deviation	14,084	1.25
Skewness	13.8	0.005
Kurtosis	467.0	0.223

Ln cost OLS regression results

	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.297	0.024	300.7	<.0001
Index visit at CBOC	-0.284	0.011	-25.4	<.0001
Baseline pain score	0.044	0.002	24.6	<.0001
Age	0.001	0.000	3.0	0.0027
Number of chronic conditions	0.173	0.004	44.2	<.0001
Indicator of diagnosis for substance use or psychiatric diagnosis	0.286	0.014	20.9	<.0001
Female	0.232	0.022	10.6	<.0001
History of opiate Rx in prior year	0.173	0.018	9.8	<.0001
MRI within 6 weeks	0.629	0.035	17.8	<.0001
Number of visits for physical therapy within 6 weeks	0.174	0.012	14.7	<.0001

Plot of residuals



Variance of residual depends on predicted y

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

Interpretation of parameters

- $\text{Ln}(Y) = \alpha + \beta X + \mu$
- β is not interpretable in raw dollars
  - It represents the relative change of cost (Y) for each unit change in X (see appendix)
  - E.g. if $\beta=0.04$, then cost increases 4% for each unit increase in X

Simulation in original scale with log linear model

- We usually want to express findings in dollars, not as proportionate change
- How to express $\hat{Y}$ on original scale ($)?
  - We estimated $\text{Ln}\ \hat{Y} = \alpha + \beta X$
  - Since $Y = e^{\text{LnY}}$ should we find $e^{\alpha + \beta X}$?

Simulation with log linear model

Must include expected value of errors

E is the expectation operator

$$E(Y) = E(e^{\alpha + \beta X + \mu})$$

$$E(Y) = e^{\alpha + \beta X} E(e^{\mu})$$

Although $E(\mu)=0$; $E(e^{\mu}) \neq 1$

How to estimate $E(e^{\mu})$?

The Smearing Estimator

- An estimate of $E(e^{\mu})$: mean of antilog of residuals:

$$E(e^{\mu}) = \frac{1}{n}\sum e^{\mu}$$

- Smearing estimator
  - assumes identical variance of errors (homoscedasticity)
  - Alternatives adjustments, but:
- Models without retransformation problem

Smearing Estimator Applied to Low Back Pain Cohort

	Mean
Raw cost (original data)	6,304
$e^{\hat{Y}}$ (Not smearing adjusted)	3,193
$e^{Y-\hat{Y}}$ (Smearing estimator)	2.02
$e^{\hat{Y}}$*2.02 (Smearing adjusted)	6,449

Estimating absolute differences in cost with log model

- Consider model $\text{Ln}(Y) = \alpha + \beta_1 X + \beta_2 Z$
  - X is indicator of group membership
  - Z is case-mix variable(s)
- β_1 is the proportional difference from group membership
- How can we state the case-mix adjusted effect of group membership in dollars?

Estimating absolute differences in cost with log model

- Method one (not recommended for log model)
- Evaluate parameter at mean value
  - Find cost with X=1 all other variables set to their mean
  - Find cost with X=0 all other variables set to their mean
  - Find this difference
- This can lead to bias!
  - The effect of β is different for every observation
  - $e^{\text{Mean}(\hat{Y})} \neq \text{Mean}(e^{\hat{Y}})$

Estimating absolute differences in cost with log model

- Method 2 (recommended)
  - Compare the means of the smearing adjusted predictions
- For each observation predict log cost
  - "as if" observation was in the group (set X to ONE)
  - "as if" observation wasn't in the group (set X to ZERO)
- For each observation, retransform these predicted costs to raw cost ($) with smearing estimator
- Find mean cost for each scenario (over N observations)

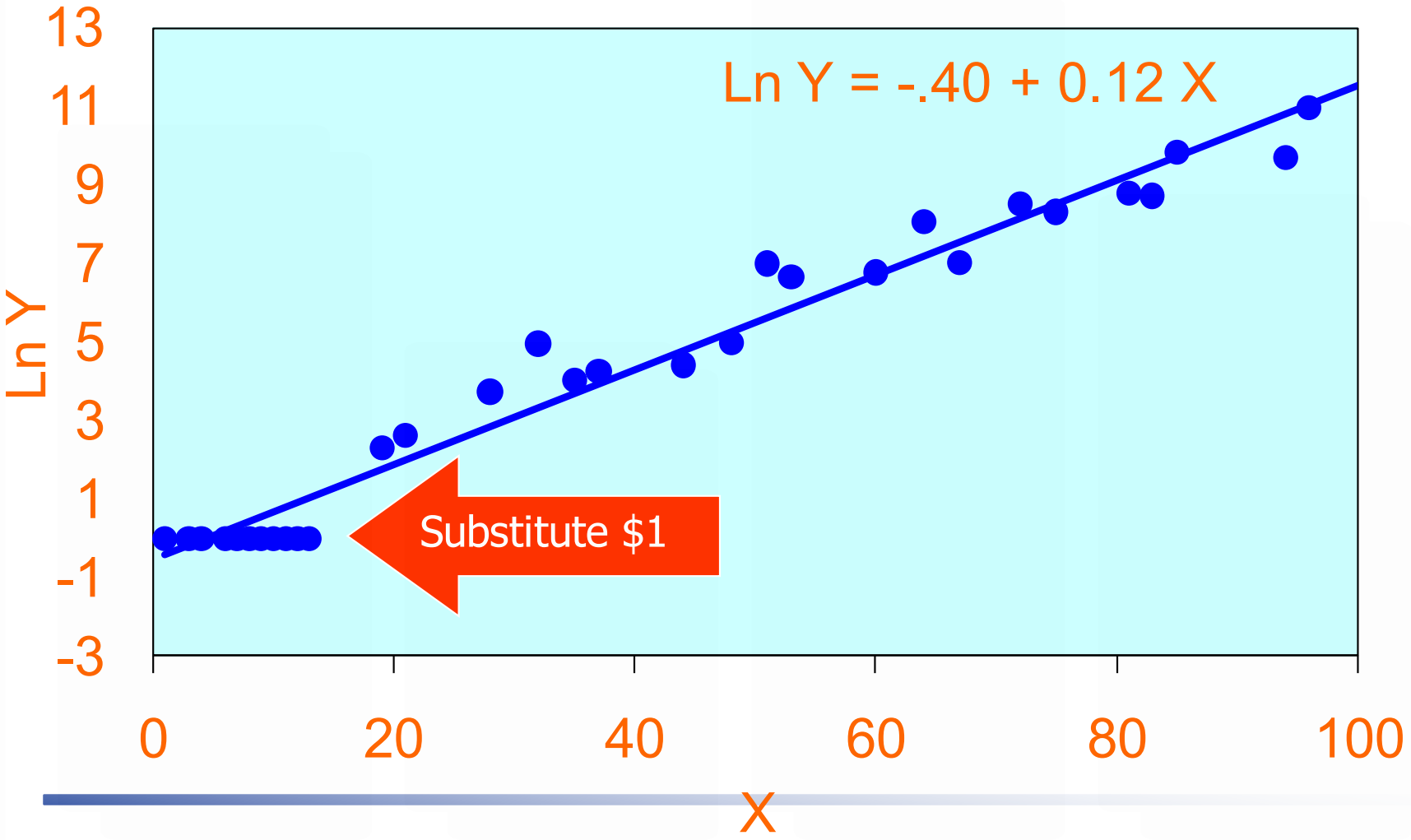
Limitation of log models: observations with zero cost

- Many cohorts have members with no cost
- How can we find Ln(Y) when Y = 0?
- Recall that Ln (0) is undefined

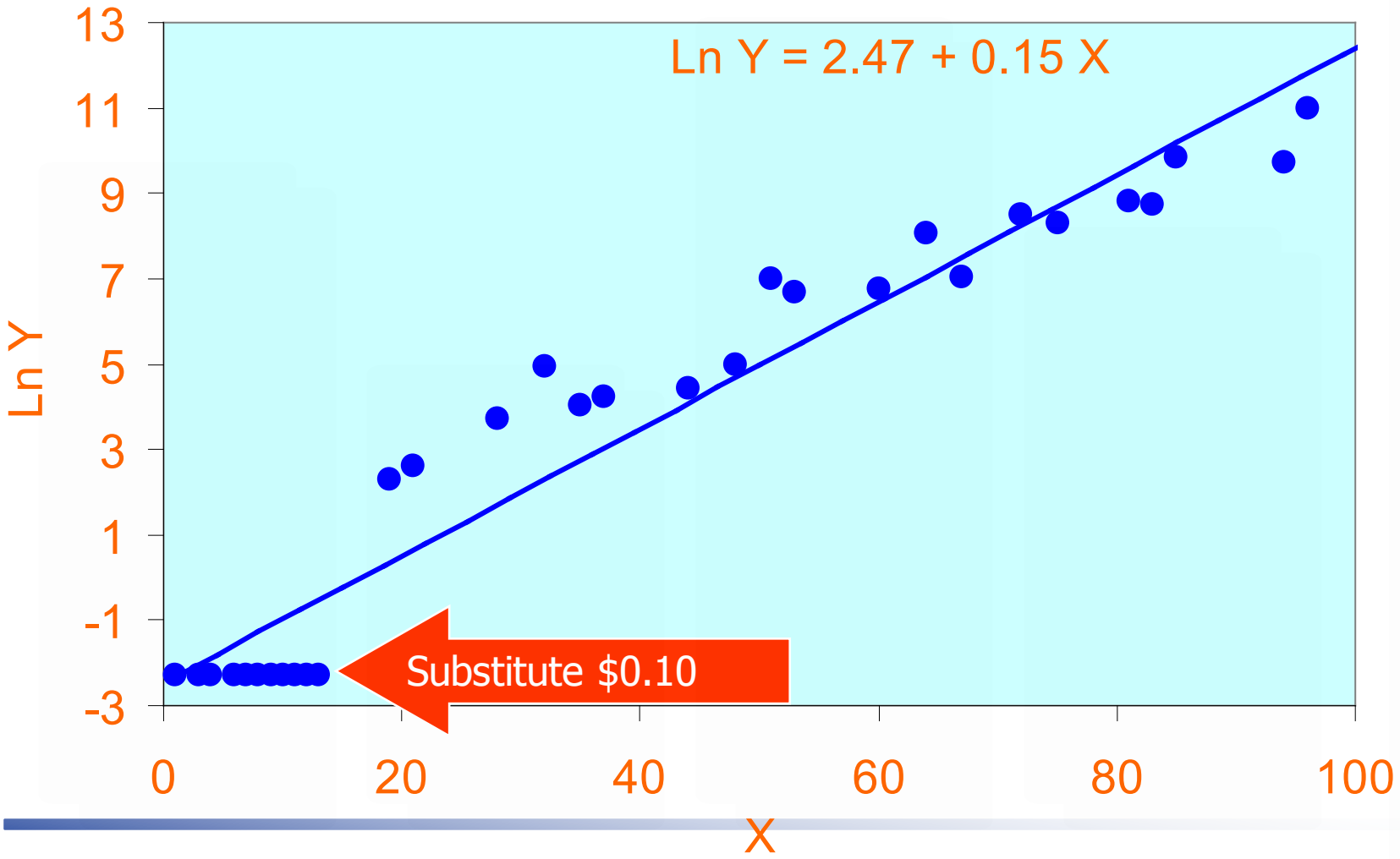
Log transformation

- 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 number 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
  - General Linear Model regressions allow zero observations, even with log link (transformation)

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: Use for OLS with raw cost?

- OLS with raw cost can be used only if:
  - 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 want to know that alternatives were considered!

Review

- Log transformation can make cost more normally distributed so we can still use OLS

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

- Predicting costs in dollars with a log model
  - Correct for retransformation bias
  - Smearing estimator is one way to do this
- Simulating absolute effect of a parameter
  - Evaluate its effect in all observations

Review

- Cost data can have observations with zero values, a truncated distribution
- Ln (0) is not defined
- Substituting small positive values for zero can result in biased parameters
- There are better methods

Next session- April 10

- Generalized Linear Model regressions
- Non-parametric statistical tests
- Two-part models
- How to determine which method is best?

Reading 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

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

$$\frac{dLn\,Y}{dx} = \beta, \; as \; dLnY = dY / Y$$

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

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