Econometrics Course: Cost as the Dependent Variable (I)

Paul G. Barnett, PhD
April 12, 2017
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)

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5,768</td>
</tr>
<tr>
<td>Median</td>
<td>1,750</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>18,874</td>
</tr>
<tr>
<td>Skewness</td>
<td>13.98</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>336.3</td>
</tr>
</tbody>
</table>
Skewness and kurtosis

- **Skewness (3\textsuperscript{rd} moment)**
  - Degree of symmetry
  - Skewness of normal distribution $= 0$
  - Positive skew: more observations in right tail

- **Kurtosis (4\textsuperscript{th} 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

- **Medical Only**
- **Medical+Rx**
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 is the goal of your cost analysis?

- Find difference in mean cost between two or more groups of patients
- Simulate the cost of a particular group of patients
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
Expected value of a random variable (cont)

- Expected value from a roll of the die
  \[ E(W) = \sum W_i \cdot p(W_i) \]
  \[ = 1 \left( \frac{1}{6} \right) + 2 \left( \frac{1}{6} \right) + 3 \left( \frac{1}{6} \right) + 4 \left( \frac{1}{6} \right) + 5 \left( \frac{1}{6} \right) + 6 \left( \frac{1}{6} \right) \]
  \[ = 3.5 \]
Review of OLS assumptions

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

\[ \hat{Y} \mid (X = 0) = \alpha \]

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

\[ \hat{Y} \mid (X = 1) = \alpha + \beta \]
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)

\[
\hat{Y} \mid (X = 1) = \alpha + \beta_1 + \beta_2 \bar{Z}
\]

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

- 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

\[ Y = 0.72 + 0.88 \times X \]
The influence of a single outlier observation

\[ Y = 22.9 + 0.42 \, X \]
Log Transformation of Cost

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

<table>
<thead>
<tr>
<th>COST</th>
<th>LN(COST)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10</td>
<td>2.30</td>
</tr>
<tr>
<td>$1,000</td>
<td>6.91</td>
</tr>
<tr>
<td>$100,000</td>
<td>11.51</td>
</tr>
</tbody>
</table>
Same data - outlier is less influential

$\text{Ln } Y = 2.87 + 0.011 \text{ X}$
Same data - outlier is less influential

\[ \ln Y = 2.99 + 0.008 X \]
Annual per person VHA costs FY10

- Medical Only
- Medical+Rx
Effect of log transformation

Annual per person VHA costs FY10

- Medical Only
- Medical+Rx
### Descriptive statistics: VHA costs FY10

(5% sample, includes outpatient pharmacy)

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>Ln Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5,768</td>
<td>7.68</td>
</tr>
<tr>
<td>Median</td>
<td>1,750</td>
<td>7.67</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>18,874</td>
<td>1.50</td>
</tr>
<tr>
<td>Skewness</td>
<td>13.98</td>
<td>-0.18</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>336.3</td>
<td>1.12</td>
</tr>
</tbody>
</table>
Log linear model

Regression with log dependent variable

\[ \ln Y = \alpha + \beta X + \mu \]
Log linear model

- \( \ln (Y) = \alpha + \beta X + \mu \)

- 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

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

What is $\hat{Y}$?
Can we retransform by taking antilog of fitted values?

With the model:

\[ \ln (Y) = \alpha + \beta_1 X + \beta_2 Z + \mu \]

Does

\[ \hat{Y} = e^{\alpha + \beta_1 X + \beta_2 Z} \]
What is fitted value of $Y$?

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

$$= e^{\alpha + \beta_1 X + \beta_2 Z} E(e^{\mu_i})$$

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

*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

\[ E(Y) = E(e^{\alpha + \beta X_1 + \beta Z_2 + \mu_i}) \]

\[ = \left( e^{\alpha + \beta X_1 + \beta Z_2} \right) E(e^{\mu_i}) \]

\[ = \left( e^{\alpha + \beta X_1 + \beta Z_2} \right) \frac{1}{n} \sum_{i=1}^{n} (e^{\mu_i}) \]
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

- Save mean of my_smear
  - as &smear_est (a single value)
- Predict the log cost (PREDICTED)
- Transform back into natural units (dollars of cost)
  
  Predicted_Cost = EXP(Predicted_lncost)*&smear_ext;
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 predicted 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?
The simple approach to comparing groups controlling for case mix

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

\[ \text{LNCost} | (X = 0) = \alpha + \beta_2 \bar{Z} \]

- 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 $\ln(Y)$ 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

\[ \ln Y = -0.40 + 0.12X \]
Substitute $0.10 for Zero Cost Records

\[ \ln Y = 2.47 + 0.15 X \]

0 20 40 60 80 100
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
  - 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 $a$ with linear dependent variable
- Find the linear combination of parameters and variables, e.g.

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

\[ E(Y \mid X = 1, Z = \bar{Z}) = \left( e^{\mu - \beta - \beta Z} \right)^{1/n} \sum_{i=1}^{n} (e^{U_i}) \]
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- April 26

- 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


HERC@va.gov
Supplemental reading on Log Models

- Smearing estimator for retransformation of log models

- Alternatives to smearing estimator
Appendix: Derivation of the meaning of the parameter in log model

\[ \ln Y = \alpha + \beta X + \mu \]

\[ \frac{d\ln Y}{dx} = \beta, \text{ as } d\ln Y = \frac{dY}{Y} \]

\[ \frac{dY}{Y} = \beta \]

\[ \beta \] is the proportional change in \( Y \) for a small change in \( X \)