Analyzing Treatment Costs in Clinical Trials

Henry Glick

Institutt for helseledelse og helseøkonomi

November 24, 2008

Outline

• Part 1. Univariate analysis
  – Policy relevant parameter for CEA
  – Cost data 101
  – T-tests
  – Response to the violation of normality
  – Primer on log cost
  – Why do different statistical tests lead to different inferences?
• Part 2. Multivariable analysis

Policy Relevant Parameter for CEA

• Policy relevant parameter: differences in the arithmetic, or sample, mean
  – In welfare economics, a project is cost-beneficial if the winners from any policy gain enough to be able to compensate the losers and still be better off themselves
  • Thus, we need a parameter that allows us to determine how much the losers lose, or cost, and how much the winners win, or benefit
  – From a budgetary perspective, decision makers can use the arithmetic mean to determine how much they will spend on a program
Policy Relevant Parameter for CEA (2)

- Other summary statistics such as median cost may be useful in describing the data, but do not provide information about the difference in cost that will be incurred or the cost saved by treating patients with one therapy versus another
  - They thus are not associated with social efficiency
- Lack of symmetry of cost distribution does not change fact that we are interested in the arithmetic mean
- Evaluating some other difference, be it in medians or geometric means, simply because the cost distribution satisfies the assumptions of the tests for these statistics, may be tempting, but does not answer the question we are asking

Are Sample Means Always the Best Estimator?

- In simulation, when cost data are sufficiently nonnormal, the relative bias (truth - observed)^2 for other parameters such as the median or adjusted geometric mean can sometimes be lower than the relative bias observed for the arithmetic mean
- HOWEVER,
  - Distribution required to be sufficiently nonnormal that ln(cost) is also substantially nonnormally distributed
  - In actual data, since we never know truth, it is difficult to determine whether other parameters will have lower relative bias than sample mean

Cost Data 101

- Common feature of cost data is right-skewness (i.e., long, heavy, right tails)
- Data tend to be skewed because:
  - Can not have negative costs
  - Most severe cases may require substantially more services than less severe cases
  - Certain events, which can be very expensive, occur in a relatively small number of patients
  - A minority of patients are responsible for a high proportion of health care costs
Typical Distributions Of Cost Data

• Heavy tails vs. "outliers"
  – Distributions with long, heavy, right tails will have means that differ from the median
  • Median is not a better measure of the costs on average than is the mean

Typical Distribution Of Cost Data (II)

Problem Not Related Solely to "Outliers"

• Distribution when 6 observations with cost > 7000 are eliminated
Mean, SD When 5 Observations with Cost > 7200 are Eliminated

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th>Trimmed (3*SD) *</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 0</td>
<td>Group 1</td>
</tr>
<tr>
<td>Mean</td>
<td>3015</td>
<td>3040</td>
</tr>
<tr>
<td>Median</td>
<td>2826</td>
<td>2901</td>
</tr>
</tbody>
</table>

* p = 0.003 and 0.0001 for nonnormality of groups 0 and 1, respectively

Univariate And Multivariable Analyses Of Economic Outcomes

- Analysis plans for economic assessments should routinely include univariate and multivariable methods for analyzing the trial data
- Univariate analyses are used for the predictors of economic outcomes
  - Demographic characteristics, clinical history, length of stay, and other resource use before entry of study subjects into the trial
- Univariate and multivariable analyses should be used for the economic outcome data
  - Total costs, hospital days, quality-adjusted life years

Univariate Analysis Of Costs

- Report:
  - Arithmetic means and their difference
    - Economic analysis is based on differences in arithmetic mean costs (because n x mean = total), not median costs; thus means are the statistic of interest
    - Measures of variability and precision, such as:
      - Standard deviation
      - Quantiles such as 5%, 10%, 50%,...
  - An indication of whether or not the difference in arithmetic means
    - Occurred by chance and is economically meaningful
Univariate Analysis: Parametric Tests Of Raw Means

- Usual starting point: T-tests and one way ANOVA
  - Used to test for differences in arithmetic means in total costs, QALYS, etc.
  - Makes assumption that the costs are normally distributed
  - Normality assumption is routinely violated for cost data, but in large samples t-tests have been shown to be robust to violations of this assumption when:
    - Samples moderately large
    - Samples are of similar size and skewness
    - Skewness is not too extreme

Steps in Performing a T-test

- Evaluate whether or not the outcome is normally distributed
  - sktest, joint test of skewness and kurtosis
  - Alternative tests:
    - swilk
    - sfrancia
- Evaluate whether or not the standard deviations of costs for the treatment groups are similar
- Perform the t-test and interpret it in relationship to the prior two tests

Results of Tests of Normality and Equivalence of S.D. of Costs

<table>
<thead>
<tr>
<th>Test</th>
<th>P-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sktest, group 0</td>
<td>0.0</td>
<td>Failed</td>
</tr>
<tr>
<td>sktest, group 1</td>
<td>0.0</td>
<td>Failed</td>
</tr>
<tr>
<td>Equality of standard deviations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdtest</td>
<td>0.0</td>
<td>Failed</td>
</tr>
</tbody>
</table>
Results of T-Test

Two-sample t test with unequal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>250</td>
<td>3015</td>
<td>100.1052</td>
<td>1582.802</td>
<td>2817.839 – 3212.161</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>3040</td>
<td>73.91742</td>
<td>1168.737</td>
<td>2894.417 – 3185.583</td>
</tr>
<tr>
<td>comb</td>
<td>500</td>
<td>3027.5</td>
<td>62.15917</td>
<td>1389.921</td>
<td>2905.374 – 3149.626</td>
</tr>
</tbody>
</table>

| diff | -25 | 124.4381 | -269.5399 | 219.5399 |

Ho: diff = 0  Satterthwaite’s degrees of freedom = 458.304
Ha: diff < 0  Ha: diff != 0  Ha: diff > 0
Pr(T < t) = 0.4204  Pr(|T| > |t|) = 0.8409  Pr(T > t) = 0.5796

Responses To Violation Of Normality Assumption

- Adopt nonparametric tests of other characteristics of the distribution that are not as affected by the nonnormality of the distribution (“biostatistical” approach)
- Transform the data so they approximate a normal distribution (“classic econometric” approach)
- Adopt tests of arithmetic means that avoid parametric assumptions (most recent development)
- OBSERVATION: If we abandon statistical testing of the arithmetic mean because distributional assumptions of the t-test are violated, does not imply that we are not interested in differences in the arithmetic mean

Response 1: Non-parametric Tests of Other Characteristics of the Distribution

- Rationale: Can analyze the characteristics that are not as affected by the nonnormality of the distribution
  - Wilcoxon rank-sum test
  - Kolmogorov-Smirnov test
Wilcoxon Rank-Sum

- Test of difference in medians
- Estimates the probability that a randomly selected patient from one treatment group has a higher cost than a randomly selected patient from another treatment group (Note: the area under the ROC curve is equivalent to the p-value of the Wilcoxon rank-sum test for a diagnostic test’s scores)

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

<table>
<thead>
<tr>
<th>treat</th>
<th>obs</th>
<th>rank sum</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>250</td>
<td>61183.5</td>
<td>62625</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>64066.5</td>
<td>62625</td>
</tr>
<tr>
<td>combined</td>
<td>500</td>
<td>125250</td>
<td>125250</td>
</tr>
</tbody>
</table>

unadjusted variance  2609375.00
adjustment for ties  -3.51
adjusted variance    2609371.49

Ho: cost(treat==0) = cost(treat==1)
z = -0.892
Prob > |z| =  0.3722

Kolmogorov-Smirnov

- Test of difference in the cumulative distribution function
- Estimates whether the maximum absolute difference between two cumulative distribution function estimates are significant
Kolmogorov-Smirnov Test

Two-sample Kolmogorov-Smirnov test for equality of distribution functions:

<table>
<thead>
<tr>
<th>Smaller group</th>
<th>D</th>
<th>P-value</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:</td>
<td>0.164</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>1:</td>
<td>-0.064</td>
<td>0.359</td>
<td></td>
</tr>
<tr>
<td>Combined K-S:</td>
<td>0.164</td>
<td>0.002</td>
<td>0.002</td>
</tr>
</tbody>
</table>

- Line 1 tests if group 0 has smaller values than group 1
- Line 2 tests if group 0 has larger values than group 1
- Line 3 provides a joint test

Potential Problem with Testing Other Characteristics of the Distribution

- Tests indicate that some measure of the cost distribution differs between the treatment groups, such as its shape or location, but not necessarily that the arithmetic means differ
- The resulting p-values need not be applicable to the arithmetic mean
- While one might decide to compare cost by use of tests like the Mann-Whitney U test, the numerator and denominator of the cost-effectiveness ratio should never be represented as a difference in median cost or effect
Response 2: Transform the Data

- Transform costs so they approximate a normal distribution
  - Common transformations
    - Log (arbitrary additional transformations required if any observation equals 0)
    - Square root
  - Estimate and draw inferences about differences in transformed costs

Estimates and Inferences Not Necessarily Applicable to Arithmetic Mean

- Goal is to use these estimates and inferences to estimate and draw inferences about differences in untransformed costs
  - Estimation: Simple exponentiation of mean of log costs results in geometric mean, which is a biased estimate of the arithmetic mean
  - Need to apply smearing factor
  - Inference: On the retransformed scale, inferences about the log of costs translate into inferences about differences in the geometric mean rather than the arithmetic mean

Results of T Test of Log

Two-sample t test with unequal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>250</td>
<td>7.863486</td>
<td>.0364313</td>
<td>.57603</td>
<td>7.791734 - 7.935239</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>7.95014</td>
<td>.023952</td>
<td>.3787148</td>
<td>7.902965 - 7.997314</td>
</tr>
<tr>
<td>comb</td>
<td>500</td>
<td>7.906813</td>
<td>.0218642</td>
<td>.4888993</td>
<td>7.863856 - 7.94977</td>
</tr>
<tr>
<td>diff</td>
<td></td>
<td>-.0866533</td>
<td>.0435998</td>
<td></td>
<td>-1.1723483 - .0009583</td>
</tr>
</tbody>
</table>

Ho: diff = 0  Satterthwaite's degrees of freedom = 430.373
Ha: diff < 0  Pr(|T| > |t|) = 0.0475
Ha: diff > 0  Pr(T > t) = 0.9762
Pr(T < t) = 0.0238
Nonparametric Tests of the Log

- Because the nonparametric tests are "order" statistics, and because the log transformation does not affect the order of costs, the results of the rank sum test and the Kolmogorov-Smirnov test on the log of cost are identical to the results of these tests when the log is not taken.

Primer on The Log Transformation Of Costs

Primer on the Log Transformation of Cost

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>75</td>
<td>55</td>
</tr>
<tr>
<td>Arithmetic mean</td>
<td>30</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Geometric mean</td>
<td>$\prod$ 24.662</td>
<td>36.993</td>
<td>44.247</td>
</tr>
<tr>
<td>Ratio, geometric / arithmetic mean</td>
<td>0.822</td>
<td>0.822</td>
<td>.983</td>
</tr>
<tr>
<td>Log, arithmetic mean cost</td>
<td>3.401197</td>
<td>3.806662</td>
<td>3.806662</td>
</tr>
<tr>
<td>Natural log</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.302585</td>
<td>2.70805</td>
<td>3.555348</td>
</tr>
<tr>
<td>2</td>
<td>3.401197</td>
<td>3.806662</td>
<td>3.806662</td>
</tr>
<tr>
<td>3</td>
<td>3.912023</td>
<td>4.317488</td>
<td>4.007333</td>
</tr>
<tr>
<td>Arithmetic mean, log cost</td>
<td>3.205269</td>
<td>3.610734</td>
<td>3.789781</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.8224</td>
<td>0.8224</td>
<td>.2265</td>
</tr>
<tr>
<td>$\exp^{\text{mea}}$</td>
<td>24.662</td>
<td>36.993</td>
<td>44.247</td>
</tr>
</tbody>
</table>
Primer On The Log Transformation Of Costs

• Observation: Simple exponentiation of the mean of the logs yields the geometric mean of costs, which in the presence of variability in costs (variance, skewness, kurtosis) is a biased estimate of the arithmetic mean
  - All else equal, the greater the variance, the skewness, or kurtosis, the greater the downward bias of the exponentiated mean of the logs
  - e.g., \((25 \times 30 \times 35)^{0.333} = 29.7196\)
  \((10 \times 30 \times 50)^{0.333} = 24.6621\)
  \((5 \times 30 \times 55)^{0.333} = 20.2062\)
  \((1 \times 30 \times 59)^{0.333} = 12.0664\)

• "Smearing" factor attempts to eliminate bias from simple exponentiation of the mean of the logs

Retransformation Of The Log Of Cost (I)

• Duan's common smearing factor:

\[ \Phi = \frac{1}{N} \sum_{i=1}^{N} \Phi (z_i - \bar{z_i}) \]

where in univariate analysis, \(\hat{z}_i\) = the group mean

• Most appropriate when treatment group variances are equivalent

Retransformation Of The Log Of Cost (II)

<table>
<thead>
<tr>
<th>Group</th>
<th>Observ</th>
<th>ln</th>
<th>(z_i)</th>
<th>(\Phi (z_i - \bar{z_i}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>2.708050</td>
<td></td>
<td>-0.9026834</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3.806663</td>
<td>0.1959289</td>
<td>1.216440</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4.317488</td>
<td>0.7067545</td>
<td>2.027401</td>
</tr>
<tr>
<td>Mean, 2</td>
<td></td>
<td>3.610734</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3.555348</td>
<td>-0.2344332</td>
<td>0.7910191</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.806663</td>
<td>0.0168812</td>
<td>1.017025</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4.007333</td>
<td>0.2175519</td>
<td>1.24303</td>
</tr>
<tr>
<td>Mean, 3</td>
<td></td>
<td>3.789781</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Smear</td>
<td></td>
<td></td>
<td>--</td>
<td>1.116732</td>
</tr>
</tbody>
</table>
Common Smearing Retransformation (I)

- Retransformation formula
  \[ E(Y) = \Phi \cdot e^{\phi^2} \]
  \[ E(Y) = \Phi \cdot e^{\phi^2} \]

- Retransformation

<table>
<thead>
<tr>
<th>Group</th>
<th>( \phi )</th>
<th>( e^{\text{predicted}} )</th>
<th>Predicted cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.116732 x 36.993</td>
<td>36.993</td>
<td>41.3</td>
</tr>
<tr>
<td>3</td>
<td>1.116732 x 44.247</td>
<td>44.247</td>
<td>49.4</td>
</tr>
</tbody>
</table>

Common Smearing Retransformation (II)

- Why are the retransformed subgroup-specific means -- 41.3 and 49.4 -- so different from the untransformed subgroup means of 45?
- Because the standard deviations of the subgroups' logs are substantially different
  \[ SD_2 = 0.8224; SD_3 = 0.2265 \]
- The larger standard deviation for group 2 implies that compared with the arithmetic mean, its geometric mean has greater downward bias than does the geometric mean for group 3
- Thus, multiplication of the 2 groups' geometric means by a common smearing factor cannot give accurate estimates for both groups' arithmetic means

Log Transformations and Normal Assumptions

- Log transformations and normal assumptions:
  - if met, t-test of the log may be more efficient than t-test of cost
  - if not met there are no efficiency gains
  - in either case, retransformation translates differences in variance, skewness, and kurtosis into differences in means
Subgroup-specific Smearing Factors (I)

- Manning has shown that in the face of heteroscedasticity – i.e., differences in variance – use of a common smearing factor in the retransformation of the predicted log of costs yields biased estimates of predicted costs.
- We obtain unbiased estimates by use of subgroup-specific smearing factors.
- Manning’s subgroup-specific smearing factor:
  \[ \phi_i = \frac{1}{N_i} \sum_{j=1}^{N_i} e^{(z_j - \bar{z}_j)} \]

Subgroup-specific Smearing Factors (II)

<table>
<thead>
<tr>
<th>Group</th>
<th>Observ</th>
<th>ln((z_i - \bar{z}_i))</th>
<th>(e^{(z_i - \bar{z}_i)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>2.708050</td>
<td>-0.9026834 0.4054801</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3.806663</td>
<td>0.1959289 1.216440</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4.317488</td>
<td>0.7067545 2.027401</td>
</tr>
<tr>
<td>Mean, 2</td>
<td>--</td>
<td>3.610734</td>
<td>-1.21644 Φ_2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3.553548</td>
<td>-0.2344332 0.7910191</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.806663</td>
<td>0.0168812 1.017025</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4.007333</td>
<td>0.2175519 1.24303</td>
</tr>
<tr>
<td>Mean, 3</td>
<td>--</td>
<td>3.789781</td>
<td>-1.0170245 Φ_3</td>
</tr>
</tbody>
</table>

Subgroup-specific Smearing Retransformation (I)

- Retransformation formulas:
  \[ E(\tilde{Y}_i) = \phi_i \ e^{z_i} \]
  \[ E(\bar{Y}_j) = \phi_j \ e^{z_j} \]
- Retransformation:

<table>
<thead>
<tr>
<th>Group</th>
<th>(\phi_i)</th>
<th>(e^{(z_i)})</th>
<th>Predicted cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.21644</td>
<td>x 36.993</td>
<td>45.00</td>
</tr>
<tr>
<td>3</td>
<td>1.0170245</td>
<td>x 44.247</td>
<td>45.00</td>
</tr>
</tbody>
</table>
Subgroup-specific Smearing Retransformation (II)

- All else equal, in the face of differences in variance (or skewness or kurtosis), use of subgroup-specific smearing factors yield unbiased estimates of subgroup means.
- Use of separate smearing factors eliminates efficiency gains from log transformation, because we cannot assume that p-value derived for the log of cost applies to the arithmetic mean of cost.

Potential Problems with Testing Transformation of the Data (I)

- Log transformation doesn’t always result in normality.

Potential Problems with Testing Transformation of the Data (II)

- When one uses a t-test to evaluate log cost, the resulting p-value has direct applicability to the difference in the log of cost.
- It generally also applies to the difference in the geometric mean of cost (i.e., we see similar p-values for the difference in the log and the difference in the geometric mean).
- The p-value for the log may or may not be directly applicable to the difference in arithmetic mean of untransformed cost.
Potential Problems with Testing Transformation of the Data (III)

• Whether the p-value for the log is applicable to the difference in the arithmetic mean of untransformed cost depends on whether the two distributions of the log are normal and whether they have equal variance and thus standard deviation
  – If log cost is normally distributed and if the variances are equal, inferences about the difference in log cost are generally applicable to the difference in arithmetic mean cost
  – If log cost is normally distributed and if the variances are unequal, inferences about the difference in log cost generally will not be applicable to the difference in arithmetic mean cost

Potential Problems with Testing Transformation of the Data (IV)

• For economic analysis, the outcome of interest is the difference in untransformed costs (e.g., “Congress does not appropriate log dollars. First Bank will not cash a check for log dollars”)
• Thus, the results on the transformed scale must be retransformed to the original scale
• “There is a very real danger that the log scale results may provide a very misleading, incomplete, and biased estimate...on the untransformed scale, which is usually the scale of ultimate interest” (Manning, 1998)
• “This issue of retransformation...is not unique to the case of a logged dependent variable. Any power transformation of y will raise this issue”

Response 3: Tests of Means that Avoid Parametric Assumptions

• Bootstrap estimates the distribution of the observed difference in arithmetic mean costs

  ![Histogram](image)

• Yields a test of how likely it is that 0 is included in this distribution (by evaluating the probability that the observed difference in means is significantly different from 0)
Implementation of Bootstrap

• Random draw with replacement from each treatment group (thus creating multiple samples)
• Calculate the difference in the mean for each sample
• For the percentile method: count the number of replicates for which the difference is above and below 0 (yielding a 1-tailed test of the hypothesis of a cost difference)
• For parametric tests:
  – Because each bootstrap replicate represents a mean difference, when we sum the replicates, the reported "standard deviation" is the standard error
  • Difference in means / SE = t statistic
  • Difference in means + 1.96 SE = 95% CI

Histogram of Bootstrap Results

Example: Distribution of Costs, Chapter 5

<table>
<thead>
<tr>
<th></th>
<th>Group 0</th>
<th>Group 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arith Mean</td>
<td>3615</td>
<td>3040</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1562.802</td>
<td>1168.737</td>
</tr>
<tr>
<td>Quantiles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>899</td>
<td>1426</td>
</tr>
<tr>
<td>25%</td>
<td>1819</td>
<td>2226</td>
</tr>
<tr>
<td>50%</td>
<td>2825.5</td>
<td>2900.5</td>
</tr>
<tr>
<td>75%</td>
<td>3752</td>
<td>3604</td>
</tr>
<tr>
<td>95%</td>
<td>6103</td>
<td>5085</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.03501</td>
<td>1.525386</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>4.910192</td>
<td>9.234913</td>
</tr>
<tr>
<td>Geom Mean</td>
<td>2608.571</td>
<td>2835.971</td>
</tr>
<tr>
<td>Mean ln</td>
<td>7.8634864</td>
<td>7.9501397</td>
</tr>
<tr>
<td>SD ln</td>
<td>.57602998</td>
<td>.37871479</td>
</tr>
<tr>
<td>Obs</td>
<td>250</td>
<td>250</td>
</tr>
</tbody>
</table>

Example: P Values from 6 Univariate Tests of the Difference in Cost

<table>
<thead>
<tr>
<th>SUMMARY TABLE</th>
<th>P-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIFFERENCE IN ARITHMETIC MEAN COST:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-test, difference in means:</td>
<td>0.8409</td>
<td>-220 to 270</td>
</tr>
<tr>
<td>nonparametric BS, diff in means:</td>
<td>0.8600</td>
<td>-218 to 275</td>
</tr>
<tr>
<td>Wilcoxon rank-sum:</td>
<td>0.3722</td>
<td></td>
</tr>
<tr>
<td>Kolmogorov-Smirnov:</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>t-test, difference in logs:</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>transformation to normal:</td>
<td>Sqrt</td>
<td></td>
</tr>
<tr>
<td>t-test, transformed variable:</td>
<td>0.2907</td>
<td></td>
</tr>
<tr>
<td>test for heteroscedasticity:</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

Why Do Different Statistical Tests Lead To Different Inferences?

- The tests are evaluating differences in different statistics
  - T-test of untransformed costs indicates we cannot infer that the arithmetic means are different
  - Bootstrap leads to same (lack of) inference and does not make the normality assumption
  - Wilcoxon rank-sum test also leads to the same inference, but its p-value relates more to the probability that the medians differ
  - T-test of log costs indicates one can infer that the mean of the logs are different, and thus the geometric means of cost are different
  - Kolmogorov-Smirnov test indicates one can infer that the distributions are different

(Repetition) Which Statistic Should Be Used To Summarize Cost Data? (I)

- Cost-effectiveness ratios (ΔC/ΔE) and NMB ([WTP ΔE] - ΔC) require an estimate of ΔC and ΔE, the differences in arithmetic means
- If arithmetic means are the most meaningful summary statistic of costs, we should test for significant differences in arithmetic mean costs
  - Parametric test of means
  - Non-parametric test of means (e.g., bootstrap methods)
Test For Differences In Arithmetic Means

• Because of distributional problems related to evaluating the arithmetic mean, there has been a growing use of nonparametric tests such as Wilcoxon and KS tests
  – Problem: Their use divorces hypothesis testing from estimation
    • i.e., we want to 1) estimate the magnitude of the difference in arithmetic means and 2) test whether that difference was observed by chance
    • Use of tests of medians or distributions to address the second task does not help with the first task
• Tests of transformed variables such as the log or square root pose similar problems

Outline (2)

• Part 1. Univariate analysis
• Part 2. Multivariable analysis
  – Ordinary least squares
    • Untransformed cost
    • Log of cost
  – General linear models (GLM)
  – Diagnostic tests
• General advice

Multivariable Analysis Of Economic Outcomes (I)

• Even if treatment is assigned in a randomized setting use of multivariable analysis may have added benefits:
  – Improves the power for tests of differences between groups (by explaining variation due to other causes)
  – Facilitates subgroup analyses for cost-effectiveness (e.g., more/less severe; different countries/centers)
  – Variations in economic conditions and practice pattern differences by provider, center, or country may have a large influence on costs and the randomization may not account for all differences
  – Added advantage: Helps explain what is observed (e.g., coefficients for other variables should make sense economically)
Multivariable Analysis Of Economic Outcomes (II)

• If treatment is not randomly assigned, multivariable analysis is necessary to adjust for observable imbalances between treatment groups, but it may NOT be sufficient

Common Multivariable Techniques Used for the Analysis of Cost (I)

• Ordinary least squares regression predicting costs after randomization
• Ordinary least squares regression predicting the log transformation of costs after randomization
• Generalized Linear Models
• On the horizon:
  – Generalized Gamma regression (Manning et al., NBER technical working paper 293)
  – Extended estimating equations (Basu and Rathouz, Biostatistics 2005)

Multivariable Analysis

• Different multivariable models make different assumptions
  – When assumptions are met, coefficient estimates will have many desirable properties
  – With cost analysis, assumptions are often violated, which may produce misleading or problematic coefficient estimates
    • Bias (consistency)
    • Efficiency (precision)
Ordinary Least Squares (OLS)

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon \]

- **Advantages**
  - Easy
  - No retransformation problem (faced with log OLS)
  - Marginal/Incremental effects easy to calculate
- **Disadvantages**
  - Not robust:
    - In small to medium sized data set
    - In large datasets with extreme observations
  - Can produce predictions with negative costs

Log Of Costs Ordinary Least Squares (log OLS)

\[ \ln Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon \]

- **Advantages**
  - Widely known transformation for costs
  - Common in the literature
  - Reduces robustness problem
  - Improves efficiency
- **Disadvantages**
  - Retransformation problem can lead to bias
  - Coefficients not directly interpretable
  - Not easy to implement
  - Residual may not be normally distributed even after log transformation

Generalized Linear Models (GLM)

- GLM models have the advantages of the log models, but
  - Don't require normality or homoscedasticity.
  - Evaluate a direct transformation of the difference in cost, and
  - Don't raise problems related to retransformation from the scale of estimation to the raw scale
- To build them, we must identify a "link function" and a "family" (based on the data)
Stata and SAS Code

• STATA code:
  \texttt{glm y x, link(linkname) family(familyname)}

• General SAS code (not appropriate for gamma family / log link):
  \texttt{proc genmod;}
  \texttt{model y=x/ link=linkname dist=familyname;}
  \texttt{run;}

SAS Code for a Gamma Family / Log Link

• When running gamma/log models, the general SAS code drops observations with an outcome of 0
• If you want to maintain these observations and are predicting y as a function of x (M Buntin):
  \texttt{proc genmod;}
  \texttt{a = _mean_;}
  \texttt{b = _resp_;}
  \texttt{d = b/a + log(a) variance var = a^2 deviance dev =d;}
  \texttt{model y = x / link = log;}
  \texttt{run;}

The Link Function

• Link function directly characterizes how the linear combination of the predictors is related to the prediction on the original scale
  – e.g., predictions from the identity link – which is used in OLS – equal:
  \[ \hat{Y} = \hat{\beta}_i X_i \]
The Log Link

- Log link is most commonly used in literature
- When we adopt the log link, we are assuming:
  \[ \ln(\mathbb{E}(y|x)) = X\beta \]
- GLM with a log link differs from log OLS in part because in log OLS, we are assuming:
  \[ \mathbb{E}(\ln(y|x)) = X\beta \]
- \( \ln(\mathbb{E}(y|x)) \neq \mathbb{E}(\ln(y|x)) \)
  i.e. log of the mean ≠ mean of the log costs

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>55</td>
</tr>
<tr>
<td>Arithmetic mean</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Log, arith mean cost</td>
<td>3.806662</td>
<td>3.806662 *</td>
</tr>
<tr>
<td>Natural log</td>
<td>1</td>
<td>2.70805</td>
</tr>
<tr>
<td>2</td>
<td>3.806662</td>
<td>3.806662</td>
</tr>
<tr>
<td>3</td>
<td>4.317488</td>
<td>4.007333</td>
</tr>
<tr>
<td>Arith mean, log cost</td>
<td>3.610734</td>
<td>3.789781 †</td>
</tr>
</tbody>
</table>

* Difference = 0; † Difference = 0.179047

Comparison of Results of GLM Gamma/Log and log OLS Regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE</th>
<th>z/T</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM, gamma family; log link</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>0.000000</td>
<td>0.405730</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>Constant</td>
<td>3.806662</td>
<td>0.286894</td>
<td>13.27</td>
<td>0.000</td>
</tr>
<tr>
<td>Log OLS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>0.179048</td>
<td>0.492494</td>
<td>0.36</td>
<td>0.74</td>
</tr>
<tr>
<td>Constant</td>
<td>3.610734</td>
<td>0.348246</td>
<td>10.32</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Passes Tests, But Can We Improve the Link?

- Stata’s power link provides a flexible link function
- It allows generation of a wide variety of named and unnamed links, e.g.,
  - power 2:  $\hat{u} = (B_iX_i)^{0.5}$
  - power 1 = Identity link;  $\hat{u} = B_iX_i$
  - power .5 = Square root link;  $\hat{u} = (B_iX_i)^{2}$
  - power .25:  $\hat{u} = (B_iX_i)^{4}$
  - power 0 = log link;  $\hat{u} = \exp(B_iX_i)$
  - power -1 = reciprocal link;  $\hat{u} = (B_iX_i)^{-1}$
  - power -2 = inverse quadratic;  $\hat{u} = (B_iX_i)^{-0.5}$

Negative Power Links

- When using a negative power link, negative coefficients yield larger estimates on the raw scale; positive coefficients yield smaller estimates

Selecting a Link Function

- There is no single test that identifies the appropriate link
- Instead can employ multiple tests of fit
  - Pregibon link test checks linearity of response on scale of estimation
  - Modified Hosmer and Lemeshow test checks for systematic bias in fit on raw scale
  - Pearson’s correlation test checks for systematic bias in fit on raw scale
- Ideally, all 3 tests will yield nonsignificant p-values
The Family

- Specifies the distribution that reflects the mean-variance relationship
  - Gaussian: Constant variance
  - Poisson: Variance is proportional to mean
  - Gamma: Variance is proportional to square of mean
  - Inverse Gaussian or Wald: Variance is proportional to cube of mean
- Use of the poisson, gamma, and inverse Gaussian families relax the assumption of homoscedasticity

Modified Park Test

- A “constructive” test that recommends a family given a particular link function
- Implemented after GLM regression that uses the particular link
- The test predicts the square of the residuals (res^2) as a function of the log of the predictions (lnyhat) by use of a GLM with a log link and gamma family to
  - Stata code
    - `glm res^2 lnyhat, link(log) family(gamma), robust`
- If weights or clustering are used in the original GLM, same weights and clustering should be used for modified Park test

Recommended Family, Modified Park Test

- Recommended family derived from the coefficient for lnyhat:
  - If coefficient =0, Gaussian
  - If coefficient =1, Poisson
  - If coefficient =2, Gamma
  - If coefficient =3, Inverse Gaussian or Wald
- Given the absence of families for negative coefficients:
  - If coefficient < -0.5, consider subtracting all observations from maximum-valued observation and rerunning analysis
Stata Commands: Modified Park Test

```stata
. gen res2 = ((cost-yhat)^2)
. gen lnyhat = ln(yhat)
. glm res2 lnyhat, link(log) family(gamma) robust nolog
Generalized linear models            No. of obs =      200
Optimization     : ML: Newton-Raphson Residual df =      198
Scale parameter =  5.37055
Deviance         =  556.0966603          (1/df) Deviance = 2.808563
Pearson          = 1063.368955          (1/df) Pearson =  5.37055
Variance function: V(u) = u^2            [Gamma]
Link function    : g(u) = ln(u)          [Log]
Standard errors  : Sandwich
Log pseudo-likelihood = -3667.728911     AIC =  36.6973
BIC =-492.9701783
-------------------------------------------------------------------
|             Robust
|     res2 |    Coef.   Std. Err.    z    P>|z|    [95% Conf. Interval]
|-------+-----------------------------------------------------------
|       |-----------------------------------------------------------
| lnyhat | .8059514   .6058605   1.33   0.183    -.3815133   1.993416
| _cons | 10.04718   5.417169   1.85   0.064    -.5702812   20.66463
-------------------------------------------------------------------
```

Stata Output, Modified Park Test

```stata
. test lnyhat==1
   ( 1) [res2]lnyhat = 1
   chi2(  1) =    0.10
   Prob > chi2 =    0.7488  → Implies poisson
. test lnyhat==2
   ( 1) [res2]lnyhat = 2
   chi2(  1) =    3.88
   Prob > chi2 =    0.0487  → Not gamma
. test lnyhat==3
   ( 1) [res2]lnyhat = 3
   chi2(  1) =   13.11
   Prob > chi2 =    0.0003  → Not inverse gaussian
```

GLM Comments (I)

- Advantages
  - Relaxes normality and homoscedasticity assumptions
  - Consistent even if not the correct family distribution
    - Choice of family only affects efficiency if link function and covariates are specified correctly
    - Gains in precision from estimator that matches data generating mechanism
  - Avoids retransformation problems of log OLS models
GLM Comments (II)

• Disadvantages
  − Can suffer substantial precision losses
    • If heavy-tailed (log) error term, i.e., log-scale residuals have high kurtosis (>3)
    • If family is misspecified

Retransformation (I)

• GLM avoids the problem that simple exponentiation of the results of log OLS yields biased estimates of predicted costs
• It does not avoid the other complexity of nonlinear retransformations (also seen in log OLS models):
  − On the transformed scale, the effect of the treatment group is estimated holding all else equal; however, retransformation (to estimate costs) reintroduces the covariate imbalances

Recycled Predictions

• Do not use the means of the covariates to avoid the reintroduction of covariate imbalance, because the mean of nonlinear retransformations does not equal the linear retransformation of the mean
• Rather, use the method of recycled predictions to create an identical covariate structure for the two groups by:
  − Coding everyone as if they were in treatment group 0 and predicting $\hat{Z}_0$
  − Coding everyone as if they were in treatment group 1 and predicting $\hat{Z}_1$
GLM Model Output

```
*****glm model (poisson/log)
  . glm cost treat $ivar, family(poisson) link(log)
Generalized linear models               No. of obs      =      200
Optimization: ML: Newton-Raphson        Residual df     =      193
Scale parameter =        1
Deviance        =   700567.946          (1/df) Deviance = 3629.886
Poisson         =  791555.8081          (1/df) Pearson  = 4101.325
Variance function: V(u) = u              [Poisson]
Link function    : g(u) = ln(u)          [Log]
Standard errors  : OIM
Log likelihood  = -351346.9719          AIC             =  3513.54
BIC             =  699545.3708
------------------------------------------------------------------
unit |    Coef.  Std. Err.     z   P>|z|  [95% Conf. Interval]
-<---------+--------------------------------------------------------
  treat | .4629637  .0015546  297.81  0.000   .4599168   .4660106
  age | .0082989  .0000756  109.72  0.000   .0081507   .0084472
  ejfract |-.0081781  .0001135  -72.07  0.000  -.0084006  -.0079557
  sex |-.0721448  .0016935  -42.60  0.000  -.0754639  -.0688256
  etiology | .2498528  .0015617  159.99  0.000   .2467919   .2529137
  race | .0462949  .0023699   19.53  0.000   .0416499   .0509398
  _cons | 8.359824   .005554 1505.18  0.000   8.348939    8.37071
------------------------------------------------------------------
```

Replaced Predictions (2)

```
replace treat=0
predict pois_0
replace treat=1
predict pois_1
gen pois_dif=pois_1-pois_0
replace treat=tmptreat
  . tabstat pois_1 pois_0 pois_dif
 stats |     pois_1     pois_0   pois_dif
--------+----------------------------------
 mean |    10843.55    6825.096    4018.451
```

Extended Estimating Equations

- Basu and Rathouz (2005) have proposed use of extended estimating equations (EEE) which estimate the link function and family along with the coefficients and standard errors
- Tends to need a large number of observations (thousands not hundreds) to converge
- Currently can’t take the results and use them with a simple GLM command (makes bootstrapping resulting models cumbersome)
Special Cases (I)

• A substantial proportion of observations have 0 costs
  – May pose problems to regression models
  – Commonly addressed by developing a “two-part” model in which the first part predicts the probability that the costs are zero or nonzero and the second part predicts the level of costs conditional on there being some costs
    • 1st part: Logit or probit model
    • 2nd part: log OLS or GLM model

Special Cases (II)

• Censored costs
  – Results derived from analyzing only the completed cases or observed costs are often biased
  – Need to evaluate the “mechanism” that led to the missing data and adopt a method that gives unbiased results in the face of missingness

Example Of ΔC Estimates & Its 95% CI

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>ΔC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>No covariates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS (untransformed costs)</td>
<td>4000</td>
<td>(2324 to 5706)</td>
</tr>
<tr>
<td>Log OLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homoskedastic retransform</td>
<td>6684</td>
<td>(449 to 5381)</td>
</tr>
<tr>
<td>Heteroskedastic retransform</td>
<td>4000</td>
<td>(2324 to 5706)</td>
</tr>
<tr>
<td>GLM (gamma/log link)</td>
<td>4000</td>
<td>(2324 to 5706)</td>
</tr>
<tr>
<td>GLM (poisson/log link)</td>
<td>4000</td>
<td>(2324 to 5706)</td>
</tr>
<tr>
<td>Multivariate model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS (untransformed costs)</td>
<td>4027</td>
<td>(2362 to 5792)</td>
</tr>
<tr>
<td>Log OLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homoskedastic retransform</td>
<td>6649</td>
<td>(4816 to 8479)</td>
</tr>
<tr>
<td>Heteroskedastic retransform</td>
<td>4268</td>
<td>(2566 to 6126)</td>
</tr>
<tr>
<td>GLM (gamma/log link)</td>
<td>4116</td>
<td>(2451 to 5984)</td>
</tr>
<tr>
<td>GLM (poisson/log link)</td>
<td>4018</td>
<td>(2351 to 5774)</td>
</tr>
</tbody>
</table>
Which Estimate Of $\Delta C$ Should We Use?

- While the $\Delta C$ estimates from the OLS and GLM models are quite similar, those from the log OLS models (particularly homoskedastic) are substantially different.
- Given that the different multivariable methods yield varying estimates of $\Delta C$, how do we judge which result is a better estimate of $\Delta C$?
- Series of diagnostic tests available to help compare performance of alternative multivariable models.

Current State Of The Art (I)

- Common characteristics of the distribution of costs can cause problems for a large number of the available multivariable techniques.
- It is most likely the case that no single model will always be most appropriate for estimating cost differences associated with medical therapies.

Current State Of The Art (II)

- Criteria for comparing alternative models may include a series of diagnostic tests discussed here.
  - Most tests detect problems but do not provide guidance on how to fix the problem.
  - Test may or may not help you identify a single best model.
  - Tests may help you clearly identify some models that perform very poorly and should be eliminated from consideration.
  - Use the combination of all test results to make inferences.
  - Use performance on tests to rank models.
General Advice (I)

- Use mean difference in costs between treatment groups estimated from a multivariable model as the numerator for a cost-effectiveness ratio
- No model is best in every situation
- Inference based on estimates of the % difference in means (estimated directly from log model) can differ from inferences based on estimates of $\Delta C$
  - For CEA, all inferences should be based on estimates of $\Delta C$

General Advice (II)

- Avoid the log of cost model unless you know that you are doing the retransformation correctly. Papers reporting results using the log model should be viewed with caution
  - When there is heteroskedasticity, the biases can be huge
  - Heteroskedasticity between treatment groups almost always exists
- Consider GLM models because they have the advantages of the log models without any transformation problems

General Advice (III)

- Establish criteria for adopting a particular multivariable model for analyzing the data prior to unblinding the data (i.e., the fact that one model gives a more favorable result should not be a reason for its adoption)
- Given that no method will be without problems, it may be helpful to report the sensitivity of our results to different specifications of the multivariable model
References

Measuring Treatment Costs


Alternative Multivariable Models


Alternative Multivariable Models (continued)

Alternative Multivariable Models (continued)


References

Non-parametric cost models (i.e. Cox)