OUTLINE

● Univariate Analysis
  - Statistical Tests
  - General Advice

● Multivariable Analysis
  - Multivariable Techniques
  - Diagnostic Tests
  - General Advice

● APPENDICES
  - Technical Notes on Diagnostic Tests
  - Stata Programs

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
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, a relatively small number of patients
  
  A minority of patients are responsible for a high proportion of health care costs

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

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>23,019</td>
<td>16,052</td>
</tr>
<tr>
<td>&lt;75,000</td>
<td>20,430</td>
<td>15,960</td>
</tr>
</tbody>
</table>
UNIVARIATE ANALYSIS OF COSTS

- Report:
  - Arithmetic means and their difference
    * Economic analysis is based on differences in arithmetic mean costs (because \( n \times \text{mean} = \text{total} \)), 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%,... (particularly if data are skewed)
  - An indication of whether or not the difference in arithmetic means:
    * Occurred by chance
    * Is economically meaningfully

HOSPITAL COSTS, TIRILAZAD MESYLATE FOR SUBARACHNOID HEMORRHAGE *

<table>
<thead>
<tr>
<th></th>
<th>Vehicle</th>
<th>6 mg / kg / day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ($)</td>
<td>20,287</td>
<td>25,185</td>
</tr>
<tr>
<td>SD</td>
<td>22,542</td>
<td>22,619</td>
</tr>
<tr>
<td>Distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>4,506</td>
<td>10,490</td>
</tr>
<tr>
<td>25%</td>
<td>9,691</td>
<td>13,765</td>
</tr>
<tr>
<td>50%</td>
<td>13,773</td>
<td>18,834</td>
</tr>
<tr>
<td>75%</td>
<td>23,044</td>
<td>31,069</td>
</tr>
<tr>
<td>95%</td>
<td>53,728</td>
<td>51,771</td>
</tr>
</tbody>
</table>

PARAMETRIC TESTS OF MEANS

- Usual starting point: T-tests and one way ANOVA
  - Used to test for differences in means in total costs, QALYS, etc.
  - Makes assumption that the costs are normally distributed
  - While the normality assumption is routinely violated for cost data, in large samples these tests have been shown to be robust to violations of this assumption
RESPONSES TO VIOLATION OF ASSUMPTIONS (I)

- Adopt nonparametric tests of other characteristics of the distribution that are not as affected by the nonnormality of the distribution ("biostatistical" approach)
  - Wilcoxon rank-sum test for difference in medians
  - Kolmogorov-Smirnov test for difference in cumulative distribution function

RESPONSES TO VIOLATION OF ASSUMPTIONS (II)

- Transform costs to approximate a normal distribution ("classic econometric" approach)
  - Log transformation or square root transformation
    * For the log transformation, one is making estimates and inferences about the ratio of the treatment group means
    * For economic analysis, the outcome of interest is the difference in untransformed costs (e.g., "Congress does not appropriate log dollars")
    * Need to retransform log costs to original scale
    * Retransformation issues: Simple exponentiation of log costs results in geometric mean (not arithmetic mean). Need to apply smearing factor to obtain unbiased estimates

LOG OF COSTS DISTRIBUTION. SAH EXAMPLE

- "There is a very real danger that the log scale results may provide a very misleading, incomplete, and biased estimate of the impact of covariates 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"
RESPONSES TO VIOLATION OF ASSUMPTIONS (III)

- Adopt tests of means that avoid parametric assumptions (most recent development)
  - Non-parametric Bootstrap (Efron)
    * Estimates the distribution of the observed difference in mean costs
    * 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)

HOSPITAL COSTS, TIRILAZAD MESYLATE FOR SUBARACHNOID HEMORRHAGE *

- Statistical Comparison of Differences

<table>
<thead>
<tr>
<th>Test</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-test of mean difference</td>
<td>0.16</td>
</tr>
<tr>
<td>Wilcoxon rank-sum (Mann-Whitney)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>0.001</td>
</tr>
<tr>
<td>T-test, log of cost difference</td>
<td>0.001</td>
</tr>
<tr>
<td>T-test, cost - retransformed log</td>
<td>0.16</td>
</tr>
<tr>
<td>Bootstrap (non parametric, 1-tail)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

WHY WOULD DIFFERENT STATISTICAL TESTS LEAD TO DIFFERENT INFERENCES?

- The tests are evaluating differences in different statistics and may have different degrees of efficiency
  - T-test of untransformed costs indicates one cannot infer that the arithmetic means are different
  - Bootstrap leads to same (lack of) inference and does not make the normality assumption
  - 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
  - Wilcoxon rank-sum test indicates one can infer that the medians are different
  - Kolmogorov-Smirnov test indicates one can infer that the distributions are different
- When distributions are skewed, means and medians can be measuring very different things

WHICH STATISTIC SHOULD BE USED TO SUMMARIZE COST DATA?

- What statistical formulation best characterizes the policy or decision problem of interest?
- For cost-effectiveness analysis: C (arithmetic mean)
  - Social perspective: In economic theory, arithmetic mean costs and differences in arithmetic mean costs yield social efficiency (Kaldor-Hicks)
  - Budgetary perspective: arithmetic mean costs are a better summary of budgetary impact than median costs or log of costs
- Cost-effectiveness ratios (C/E) and NMB ([C - E] - C) require an estimate of C where:
  - C = \bar{C}_t - \bar{C}_s
  - E = \bar{E}_t - \bar{E}_s

TEST FOR DIFFERENCES IN MEANS

- If arithmetic means are the most meaningful summary statistic of costs, one should test for significant differences in arithmetic mean costs
  - Parametric test of means
  - Non-parametric test of means
    * Bootstrap methods
- 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
MULTIVARIABLE ANALYSIS OF ECONOMIC OUTCOMES

- Possibly more than in epidemiologic analysis, multivariable estimates of costs generally thought to be better than univariate estimates.

- Even if treatment is assigned in a randomized setting use multivariable analysis because:
  - Improves the power for tests of differences between groups (by explaining variation due to other causes).
  - Practice pattern differences by provider, center, or country may have a large influence on costs and the randomization may not account for all imbalances between groups.
  - Variations in economic conditions often not controlled for in a randomized trial, therefore multivariable analysis takes on added importance.
  - Additional advantage: Helps explain what is observed (e.g., coefficients for other variables should make sense economically)

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

MULTIVARIABLE TECHNIQUES USED FOR THE ANALYSIS OF COSTS

- More common techniques
  - Ordinary least squares regression predicting costs after randomization.
  - Ordinary least squares regression predicting the log transformation of costs after randomization (retransformation/smearing [Duan, 1983, Manning, 1998]).
  - Generalized Linear Models (retransformation without problems posed by smearing).

- Other techniques
  - Generalized Gamma regression (Manning et al., NBER technical working paper 293).

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 REGRESSION (OLS)

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon, \]

- Coefficient on treatment indicator produces an estimate of \( \beta \) (C)
- Assumptions generally violated with cost data:
  - Variance of error term is constant.
    * Heteroskedasticity - Unbiased coefficients, but potentially inefficient
  - Error term is normally distributed
    * Violation has no effect on coefficients
    * Necessary only for test of statistical significance
    * In large samples we can rely on the central limit theorem
    * Has been shown that regression analysis is robust against violations of normality
    * Difficult to defend in practice
    * See tests for normality (pp. xx - yy)

COMMENTS ON OLS

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

See Computer Output, p. 67
## LOG OF COSTS: ORDINARY LEAST SQUARES REGRESSION

- Coefficient on treatment indicator produces an estimate of the percentage difference in mean costs between treatment groups.
- For cost effectiveness analysis we are interested in the predicted mean costs of treatment. With a log of cost regression we are predicting log of costs, not costs.
  - Is it possible to reliably transform predicted log of costs into an unbiased prediction of mean costs?
  - Not trivial because the mean of the log of costs is not equal to the log of the mean of costs
    - i.e. $E(\ln(y)/x)$ is not equal to $\ln(E(y/x))$
- Residual may not be normally distributed even after log transformation.

<table>
<thead>
<tr>
<th>Cost</th>
<th>S.E.</th>
<th>t</th>
<th>p</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>4027</td>
<td>853.8</td>
<td>4.72</td>
<td>0.000</td>
<td>2343 to 5711</td>
</tr>
</tbody>
</table>

```stata
replace treat=0
predict ols_0
replace treat=1
predict ols_1
replace treat=tmptreat
gen ols_dif=ols_1-ols_0

tabstat ols_1 ols_0 ols_dif
```

### Univariate Result:

```stata
. tabstat cost if treat==1
variable | mean
----------|--------
cost | 10832.22

. tabstat cost if treat==0
variable | mean
----------|--------
cost | 6832.225
```
COMMENTS ON LOG OLS MODEL

- Advantages
  - Widely known transformation for costs
  - Common in the literature
  - Reduces robustness problem
  - Improves efficiency

- Disadvantages
  - Retransformation problem could lead to bias
  - Coefficients not directly interpretable
  - Not easy to implement

RETRANSFORMATION AFTER THE LOG OF COST REGRESSION (I)

- The result of the regression is the predicted log of costs ($\hat{Z}_i$) where:

$$\hat{Z}_i = \hat{\alpha} + X_i \hat{\beta}_1 + T_i \hat{\beta}_T$$

- Estimation of the effect of treatment on predicted costs is a nonlinear retransformation of the regression coefficients

- Nonlinear retransformations have to account for two complexities
  - In the logged scale, the (multiplicative) effect of the treatment group is estimated holding all else equal; however, exponentiation of predicted log costs (to estimate costs) reintroduces the covariate imbalances
  - Simple exponentiation of the predicted log costs leads to biased estimates (Duan, 1983)
RETRANSFORMATION AFTER THE LOG OF COST REGRESSION (II)

- Avoiding reintroduction of covariate imbalance (method of recycled predictions)
  - Code everyone as if they were in treatment group A and predict $\bar{Z}_{iA}$
  - Code everyone as if they were in treatment group B and predict $\bar{Z}_{iB}$

- Avoiding the bias of simple exponentiation of the predicted log (smearing retransformation)
  - Retransform the predicted log of costs (both $\bar{Z}_{iA}$ and $\bar{Z}_{iB}$) into the original scale of costs by use of a smearing factor $N$ as follows:

$$E(Y_{iA}/X_i) = \Phi \cdot e^{\bar{Z}_{iA}}$$
$$E(Y_{iB}/X_i) = \Phi \cdot e^{\bar{Z}_{iB}}$$

where

$$\Phi = \frac{1}{N} \sum_{i=1}^{N} e^{(Z_i - Z_i)}$$

SMEARING AND HETEROSKEDASTICITY (I)

- Use of a single smearing factor introduces bias when there is heteroskedasticity in the data

- For example:
  - If the variance of the residuals is greater for the treatment group than for the comparison group, the smearing correction parameter will underestimate the costs for the treatment group and overestimate the costs for the standard

- Solution:
  - Estimate the smearing correction parameter separately for the two treatment groups
    * For those actually in group A
      $$\Phi_A = \frac{1}{N_A} \sum_{i=1}^{N_A} e^{(Z_{iA} - Z_{iA})}$$
    * For those actually in group B
      $$\Phi_B = \frac{1}{N_B} \sum_{i=1}^{N_B} e^{(Z_{iB} - Z_{iB})}$$
SMEARING AND HETEROSEDASTICITY (II)

- Retransformation in the face of heteroskedasticity
  - When everyone is coded as if they are in treatment group A
    \[ E(Y_A/X_i) = \Phi_A \ e^{\gamma_A} \]
  - When everyone is coded as if they are in treatment group B
    \[ E(Y_B/X_i) = \Phi_B \ e^{\gamma_B} \]

- See Breusch-Pagan test for heteroskedasticity
- See Manning (1998) for unbiased estimation when heteroskedasticity exists in the data
- See Ai and Norton (2000) for estimates of standard errors for the retransformation problem with heteroskedasticity

---

LOG RESULT - SMEARING FACTOR:

ESTIMATE THETA

**smearing factor comes from the regression**

* residual
predict res, r
replace res=exp(res)

***HOMOSKEDASTIC SMEARING FACTOR
sum res
scalar sm=r(mean)

***HETEROSKEDASTIC SMEARING FACTORS
sum res if treat==0
scalar sm0=r(mean)
sum res if treat==1
scalar sm1=r(mean)

. display sm
1.3056459

. display sm0
1.4702791

. display sm1
1.1410127

---

See Computer Output, p. 67
LOG RESULT - RETRANFORMATION:

***ESTIMATE Z’S FOR EACH TREATMENT GROUP
replace treat=0
predict lc0
replace treat=1
predict lc1
replace treat=tmp_treat

*****HETEROSEDASTIC PREDICTION
***RETRANSFORM AND APPLY SMEARING FACTOR
gen lols_0=exp(lc0)*sm0
gen lols_1=exp(lc1)*sm1
gen lols_t=lols_0 if treat==0
replace lols_t=lols_1 if treat==1
gen lols_dif=lols_1-lols_0

.tabstat lols_1 lols_0 lols_dif
    stats |    lols_1    lols_0  lols_dif
---------+------------------------------    mean |  11156.22  6888.652  4267.573

*****HOMOSKEDASTIC PREDICTION
***RETRANSFORM AND APPLY SMEARING FACTOR
gen lolsho_0=exp(lc0)*sm
ngen lolsho_1=exp(lc1)*sm
gen lolsho_dif=lolsho_1-lolsho_0
ngen lolsho_t=lolsho_0 if treat==0
replace lolsho_t=lolsho_1 if treat==1

tabstat lolsho_1 lolsho_0 lolsho_dif
    stats |    lolsho_1  lolsho_0  lolsho_dif
---------+-------------------------------------    mean |  12765.92  6117.301  6648.621

GENERALIZED LINEAR MODELS (GLM)

- These models have the advantages of the log models, but C is estimated directly so it does not require any smearing correction
- To build them, one must identify a "link function" and a "family" (based on the data)

STATA code:
```
glm y x, link(linkname) family (familyname)
```

SAS code:
```
proc genmod;
model y=x/ link=linkname dist=familyname;
run;
```
THE LINK FUNCTION

- Specifies the relationship between the covariates and the mean
  - e.g. identity, log, power # (square root, etc.)
- GLMs are attractive because the link function directly characterizes how the mean on the raw untransformed scale is related to the predictors
  - e.g. ln(E(y/x)=X$
- Log link has been most commonly used in literature but may not necessarily be the best in all cases
- Little guidance in current literature for applied researcher on how to identify correct link function – most studies on health care expenditures use (assume) log link!
  - Compare model performance of all permutations of candidate link and variance function
- Basu and Rathouz (2005) propose extended estimating equations (EEE) which estimate the link function and variance structure along with other components of the model based on the data

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 - variance is proportional to cube of mean
- Modified Park test used to determine family (Appendix 1)
  - If $\delta=0$ Gaussian NLLS
  - If $\delta=1$ Poisson
  - If $\delta=2$ Gamma
  - If $\delta=3$ Inverse Gaussian or Wald
GLM COMMENTS

- Advantages
  - No retransformation problems of log OLS models
    * Because the link function allows modeling of the log of mean costs [i.e. \( \ln(\text{E}(y/x)) = X \)] unlike the log OLS that models the mean of log costs [i.e. \( \text{E}(\ln(y)/x) = X \)]
  - Gains in precision from estimator that matches data generating mechanism
  - Consistent even if not the correct family distribution
    * Choice of family only affects efficiency if link function and covariates are specified correctly

- Disadvantages
  - Can suffer substantial precision losses
    * If heavy-tailed (log) error term [log-scale residuals have high kurtosis (>3)]
    * If variance function (i.e. family) is misspecified

---

See Computer Output, p. 68

GLM (GAMMA/LOG LINK) RESULT

```
replace treat=0
predict gam_0
replace treat=1
predict gam_1
gen gam_dif=gam_1-gam_0
replace treat=tmptreat
```

```
.tabstat gam_1 gam_0 gam_dif   
  stats |     gam_1     gam_0   gam_dif
---------+------------------------------    mean |  10905.44  6789.416  4116.023
```

See Computer Output, p. 68

GLM (POISSON/LOG LINK) RESULT

```
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
```
SPECIAL CASES

- 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

- 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 ESTIMATES OF $C$ AND ITS 95% CI

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$C$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>No covariates</td>
<td>4000</td>
<td>(2324 to 5706)</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 $C$ SHOULD ONE USE?

- While the $C$ 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 $C$, how does one judge which result is a better estimate of $C$?
- Series of diagnostic tests available to help compare performance of alternative multivariable models.

DIAGNOSTIC TESTS

1. Skewness/Kurtosis
   - Tests for normality of residuals
2. Heteroskedasticity test (Breusch-Pagan test)
   - Tests whether residuals are heteroskedastic
3. Modified Park test (GLM family test)
   - Used to determine the family distribution in GLM
4. Pregibon Link test
   - Checks linearity of response on scale of estimation
5. Modified Hosmer Lemeshow test
   - Checks for systematic bias in fit on raw scale
6. Pearson’s Correlation test
   - Checks for systematic bias in fit on raw scale
7. Copas test
   - Tests for overfitting and cross-validation

NOTE: Details on tests are provided in APPENDIX 1.

BREUSCH-PAGAN TEST FOR HETEROSKEDASTICITY

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test whether residuals are heteroskedastic</td>
<td>If reject null then residuals are heteroskedastic</td>
</tr>
<tr>
<td>If log-scale residuals are heteroskedastic, Log OLS will be biased if appropriate smearing correction not applied</td>
<td></td>
</tr>
</tbody>
</table>

- Test result:

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breusch-Pagan</td>
<td>32.3</td>
</tr>
</tbody>
</table>

- Log OLS residuals are heteroskedastic in treatment variable
- Hence, results from log OLS with homoskedastic retransformation are biased!
CURRENT STATE OF THE ART

- 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.
- 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 $\) C
  - For CEA, all inferences should be based on estimates of $\) 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 one’s results to different specifications of the multivariable model.
REFERENCES

Measuring Treatment Costs


Alternative Multivariable Models


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


APPENDIX 1. Technical Notes on Diagnostic Tests

This section provides details on the statistical tests discussed in this lecture.

1. Skewness/Kurtosis
2. Heteroskedasticity test (Breusch-Pagan test)
3. Modified Park test (GLM family test)
4. Pregibon Link test
5. Modified Hosmer Lemeshow test
6. Pearson's Correlation test
7. Copas test

Tests for Normality and Heteroskedasticity

<table>
<thead>
<tr>
<th>Test</th>
<th>Purpose</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Skewness/</td>
<td>Test normality of residuals</td>
<td>If skewness $\neq 0$ or kurtosis $&gt;3$ then residuals not normally distributed (OLS assumption violated)</td>
</tr>
<tr>
<td>kurtosis</td>
<td></td>
<td>If kurtosis $&gt;3$ on log-scale residuals GLM may suffer precision losses</td>
</tr>
<tr>
<td>2. Breusch-</td>
<td>Test whether residuals are heteroskedastic</td>
<td>If reject null then residuals are heteroskedastic</td>
</tr>
<tr>
<td>Pagan</td>
<td></td>
<td>If log-scale residuals are heteroskedastic, Log OLS will be biased if appropriate smearing correction not applied</td>
</tr>
</tbody>
</table>

Test for Heteroskedasticity (Breusch-Pagan test)

Is the treatment variable heteroskedastic?

After OLS cost regression:

e.g. regress cost treat $ivar

hettest treat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance
Variables: treat

<table>
<thead>
<tr>
<th>chi2(1)</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.03</td>
<td>0.1546</td>
</tr>
</tbody>
</table>

After log OLS cost regression:

e.g. regress lcost treat $ivar

hettest treat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance
Variables: treat

<table>
<thead>
<tr>
<th>chi2(1)</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>32.26</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

YES in log model! (Also heteroskedastic in race & etiology)

See Computer Output, pp. 69 and 70
Summary of Results: Tests for Normality and Heteroskedasticity

After OLS cost regression:

Normality of Residuals
Coefficient of Skewness: 1.96
Coefficient of Kurtosis: 8.85
Joint test statistic of Skewness/Kurtosis: 72.1 \( (p<0.0000) \)

Heteroskedasticity in treatment variable
Breusch-Pagan test statistic: 2.0 \( (p=0.15) \)

After Log-OLS regression:

Normality of Residuals
Coefficient of Skewness: -0.12
Coefficient of Kurtosis: 2.9
Joint test statistic of Skewness/Kurtosis: 0.52 \( (p=0.77) \)

Heteroskedasticity in treatment variable
Breusch-Pagan test statistic: 32.3 \( (p=0.0000) \)

- OLS residuals are homoskedastic in treatment variable BUT they are not normally distributed!
- Log-OLS residuals are normal BUT heteroskedastic in treatment variable!

Tests for Determining Family Distribution for GLM

<table>
<thead>
<tr>
<th>Test</th>
<th>Purpose</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Modified Park</td>
<td>Determine family distribution for GLM</td>
<td>If 8=0 Gaussian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If 8=1 Poisson</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If 8=2 Gamma</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If 8=3 Inverse Gaussian</td>
</tr>
</tbody>
</table>

Modified Park Test (GLM family test)

Purpose: To determine family distribution for GLM model (i.e. relationship between mean and variance)

AFTER GLM COST REGRESSION:
e.g. glm cost treat $ivar, family(gamma) link(log)

(1) Predict value of y and log transform it
    gen lnyhat=ln(yhat)

(2) Save raw scale residuals and square them
    gen res=cost-yhat
    gen r2=((res)^2)

(3) Regress \( \ln(r2) \) on \( \ln(yhat) \) and a constant using GLM with gamma distribution
    glm r2 lnyhat , link(log) family(gamma) robust nolog

(4) Coefficient on \( \ln(yhat) \) gives the family
    If 8=0 Gaussian NLLS
    If 8=1 Poisson
    If 8=2 Gamma
    If 8=3 Inverse Gaussian or Wald

    test lnyhat==0
    test lnyhat==1
    test lnyhat==2
    test lnyhat==3

See Computer Output, p. 71
Summary of Results: Modified Park Test

Modified Park Test found $\theta = 0.81$ (Poisson!)

- Test Ho $\theta=1$ (p=0.74) Ũ Poisson
- Test Ho $\theta=2$ (p=0.04) Ũ Not Gamma
- Test Ho $\theta=3$ (p=0.000) Ũ Not Inverse Gaussian

(Gaussian) $0 < \theta < 1$ (Poisson) Compromise!

Within-sample Diagnostics for Model Fit

<table>
<thead>
<tr>
<th>Test</th>
<th>Purpose</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Pregibon Link</td>
<td>Test linearity of response on scale of estimation</td>
<td>If reject null then problem with the functional form of X</td>
</tr>
<tr>
<td>5. Modified Hosmer Lemeshow</td>
<td>Determine systematic bias in fit on raw scale</td>
<td>If reject null and observe systematic pattern then problem with either the link function or functional form of X</td>
</tr>
<tr>
<td>6. Pearson Correlation</td>
<td>Determine systematic bias in fit on raw scale</td>
<td>If reject null and high correlation co-efficient then problem with either the link function or functional form of X</td>
</tr>
</tbody>
</table>

Pregibon’s Link Test

- Purpose: To determine linearity of response on scale of estimation

After COST REGRESSION:
e.g. glm cost treat $\$ivar, family(gamma) link(log)

(1) Create two new variables
(a) Create prediction of (xb)
   predict xb, xb

(b) Create variable of squared prediction
   gen xbsq=xb^2

(2) Refit model with the two new variables as the only predictors
   glm cost xb xbsq, family(gamma) link(log) robust

(3) Co-efficient on square of the prediction should not be significantly different from zero. i.e. test Ho: co-efficient on xbsq=0
   lincom xbsq (alternatively, test xbsq)

(4) If fail to reject null (i.e. xbsq not significant predictor) then keep model the same; if reject null then problem with functional form of x

See Computer Output, p. 72
Modified Hosmer Lemeshow Test

Purpose: To determine systematic pattern of bias in model fit on raw scale

After COST REGRESSION:
e.g. glm cost treat $ivar, family(gamma) link(log)

(1) Obtain predicted value of y on raw scale
   predict yhat

(2) Compute residual on raw scale
   gen res=cost-yhat

(3) Create 10 groups, sorted by x or xb
   xtile xbtile=xb, nq(10)
   tab xbtile, gen(xbt)

(4) Conduct an F-test of whether the mean of the raw scale residuals
    across all groups of the deciles are not significantly different from
    zero
   reg res xbt1 xbt2 xbt3 xbt4 xbt5 xbt6 xbt7 xbt8 xbt9 xbt10, nocons robust
   test xbt1 xbt2 xbt3 xbt4 xbt5 xbt6 xbt7 xbt8 xbt9 xbt10, nocons robust

(5) Look for systematic patterns by plotting mean residuals by deciles
    (e.g. U-shape)

(6) If reject null and find systematic pattern then there is a problem with
    either the left hand side (wrong power function) or right hand side
    (wrong functional form of x)

See Computer Output, p. 73

Pearson’s Correlation Test

Purpose: To determine systematic bias in the prediction of E(y|x)

After COST REGRESSION:
e.g. glm cost treat $ivar, family(gamma) link(log)

(1) Obtain predicted value of y on raw scale
   predict yhat

(2) Compute residual on raw scale
   gen res=cost-yhat

(3) Check correlation between res and yhat
   pwcorr yhat res, sig

<table>
<thead>
<tr>
<th>yhat</th>
<th>res</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>-0.0335</td>
</tr>
</tbody>
</table>
|  1.0000     |  0.6372 => p-value

(4) If statistic is significantly different from zero then model is providing a
    biased prediction (i.e. estimated impact of the x on y (slope) is either too
    high or too low) and suggests that either the link function or linear
    specification is incorrect
### Summary of Results: Within-sample Diagnostics for Model Fit

<table>
<thead>
<tr>
<th>Method</th>
<th>Stat</th>
<th>p</th>
<th>Stat</th>
<th>p</th>
<th>Stat</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>0.17</td>
<td>0.68</td>
<td>0.69</td>
<td>0.73</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td>Log OLS (Homo)</td>
<td>11.78</td>
<td>0.001</td>
<td>1.70</td>
<td>0.08</td>
<td>-0.309</td>
<td>0.000</td>
</tr>
<tr>
<td>Log OLS (Hetero)</td>
<td>11.78</td>
<td>0.001</td>
<td>0.70</td>
<td>0.73</td>
<td>-0.154</td>
<td>0.03</td>
</tr>
<tr>
<td>GLM gamma/log</td>
<td>1.39</td>
<td>0.24</td>
<td>0.65</td>
<td>0.77</td>
<td>-0.034</td>
<td>0.64</td>
</tr>
<tr>
<td>GLM poisson/log</td>
<td>0.09</td>
<td>0.77</td>
<td>0.58</td>
<td>0.83</td>
<td>-0.002</td>
<td>0.98</td>
</tr>
</tbody>
</table>

- Both Log OLS models (with homoskedastic as well as heteroskedastic retransformation) fail the Pregibon Link test and Pearson Correlation test!

### Copas Test

**Purpose:** To test for overfitting using split sample cross validation

1. Randomly split sample into two equal groups A & B
2. Estimate model on sample A and retain its co-efficients
3. Forecast to sample B
   \[ \hat{y}_B = X_B \hat{\beta}_A \]
4. Regression model for sample B
   \[ y_B = \delta_0 + \delta_1 \hat{y}_B + \eta \]
   Test \( \hat{\delta}_1 = 1 \)
   Alternatively,
   \[ \text{res}_B = \delta_0 + \delta_1 \hat{y}_B + \eta \]
   Test \( \hat{\delta}_1 = 0 \)
5. If reject null hypothesis, then overfitting may be a problem - need to prune the model and check for outliers
6. Repeat split sample experiment (Steps 1 to 4) for 1000 times to get distribution and report the % of times the null is rejected – use to rank models, since test of \( \star \_1 = 1 \) often fails
Summary of Results: Copas Test

<table>
<thead>
<tr>
<th></th>
<th>Using Own-Sample Predictions</th>
<th>Using Cross-Sample Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>P *</td>
</tr>
<tr>
<td>OLS</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Log OLS (Homo)</td>
<td>7.65</td>
<td>0.707</td>
</tr>
<tr>
<td>Log OLS (Hetero)</td>
<td>2.84</td>
<td>0.284</td>
</tr>
<tr>
<td>GLM gamma/log</td>
<td>0.84</td>
<td>0.044</td>
</tr>
<tr>
<td>GLM poisson/log</td>
<td>0.70</td>
<td>0.029</td>
</tr>
</tbody>
</table>

* Represents proportion of 1000 replicates for which the F-statistic is significant at 5% level

- OLS model shows significant degree of overfitting (34.5% of the replicates fail Copas test using out-of-sample predictions)
- GLM poisson/log link model performs best (2.9% & 3.8% replicates fail Copas test within-sample & out-of-sample, respectively)

See Summary Table, p. 74

APPENDIX 2. STATA PROGRAMS

This section provides sample programs for the topics discussed in this lecture.

1. Multivariable.do
2. Insideboot.do
3. Bootmulti.do
4. Bootresult.do
5. Multivariable_tests.do
6. Mse_ape_copastest.do

1. Multivariable.DO

******************************THIS PROGRAM DOES THE MULTIVARIABLE ANALYSES***************************
clear
set memory 10000
capture log close
log using multivariable.log, replace
use mdmcea

******control variables
global ivar "age ejfract sex etiology race"

**create log of costs
gen lcost=log(cost)
gen tmptreat=treat

*****OLS
regress cost treat $ivar
predict ols_t
replace treat=0
predict ols_0
replace treat=1
predict ols_1
replace treat=tmptreat
gen ols_dif=ols_1-ols_0
tabstat cost if treat==1
tabstat cost if treat==0
tabstat ols_1 ols_0 ols_dif

***log costs
regress lcost treat $ivar

***smearing factor
capture drop res
predict res, r
replace res=exp(res)
sum res
scalar sm=r(mean)
sum res if treat==0
scalar sm0=r(mean)
sum res if treat==1
scalar sm1=r(mean)
display sm
display sm0
display sm1

replace treat=0
predict lc0
replace treat=1
predict lc1
replace treat=tmptreat
gen lols_0=exp(lc0)*sm0
gen lols_1=exp(lc1)*sm1
gen lols_t=lols_0 if treat==0
replace lols_t=lols_1 if treat==1
gen lols_dif=lols_1-lols_0
tabstat lols_1 lols_0 lols_dif

****assume homoskedastic!
gen lolsho_0=exp(lc0)*sm
gen lolsho_1=exp(lc1)*sm
ngen lolsho_dif=lolsho_1-lolsho_0

replace lolsho_t=lolsho_0 if treat==0
replace lolsho_t=lolsho_1 if treat==1
tabstat lolsho_1 lolsho_0 lolsho_dif

*****glm model (gamma/log link)
glm cost treat $ivar, family(gamma) link(log)
predict gam_t
replace treat=0
predict gam_0
replace treat=1
predict gam_1
gen gamdif=gam_1-gam_0
replace treat=tmptreat
gen gam_t=gam_0 if treat==0
replace gam_t=gam_1 if treat==1
tabstat gam_1 gam_0 gamdif

*****glm model (poisson/log link)
glm cost treat $ivar, family(poisson) link(log)
predict pois_t
replace treat=0
predict pois_0
replace treat=1
predict pois_1
gen poisdif=pois_1-pois_0
replace treat=tmptreat
gen pois_t=pois_0 if treat==0
replace pois_t=pois_1 if treat==1
tabstat pois_1 pois_0 poisdif

sum ols_1 lols_1 lolsho_1 gam_1 pois_1
sum ols_0 lols_0 lolsho_0 gam_0 pois_0
sum ols_dif lols_dif lolsho_dif gam_dif pois_dif
log close
2. Insideboot.do
**********************
***THIS PROGRAM IS JUST LIKE multivariable.do
***BUT IT IS A VERSION THAT IS CALLED BY bootmulti.do
***IN ORDER TO GET THE CONFIDENCE INTERVALS
**********************

******control variables
global ivar "age ejfract sex etiology race"

**create log of costs
gen lcost=log(cost)

***
gen tmptreat=treat

******OLS
regress cost treat $ivar
predict ols_t
replace treat=0
predict ols_0
replace treat=1
predict ols_1
replace treat=tmptreat
gen ols_dif=ols_1-ols_0
tabstat cost if treat==1
tabstat cost if treat==0
tabstat ols_1 ols_0 ols_dif

***log costs
regress lcost treat $ivar

***smearing factor
capture drop res
predict res, r
replace res=exp(res)
sum res
scalar sm=r(mean)
sum res if treat==1
scalar sm1=r(mean)
display sm
display sm0
display sm1
replace treat=0
predict lc0
replace treat=1
predict lc1
replace treat=tmptreat
gen lols_0=exp(lc0)*sm0
gen lols_1=exp(lc1)*sm1
gen lols_t=lols_0 if treat==0
replace lols_t=lols_1 if treat==1
ngen lols_dif=lols_1-lols_0
tabstat lols_1 lols_0 lols_dif

****assume homoskedastic!
gen lolsho_0=exp(lc0)*sm
gen lolsho_1=exp(lc1)*sm
gen lolsho_dif=lolsho_1-lolsho_0
ngen lolsho_t=lolsho_0 if treat==0
replace lolsho_t=lolsho_1 if treat==1
gen lolsho_dif=lolsho_1-lolsho_0

******glm model (gamma/log link)
glm cost treat $ivar, family(gamma) link(log)
predict gam_t
replace treat=0
predict gam_0
replace treat=1
predict gam_1
gen gam_dif=gam_1-gam_0
replace treat=tmptreat
gen gam_t=gam_0 if treat==0
replace gam_t=gam_1 if treat==1
tabstat gam_1 gam_0 gam_dif
****glm model (poisson/log link)

```stata
glm cost treat tvar, family(poisson) link(log)
*predict pois_t
replace treat=0
predict pois_0
replace treat=1
predict pois_1
gen pois_dif=pois_1-pois_0
replace treat=tmptreat
gen pois_t=pois_0 if treat==0
replace pois_t=pois_1 if treat==1

tabstat pois_1 pois_0 pois_dif
sum ols_1 lols_1 lolsho_1 gam_1 pois_1
sum ols_0 lols_0 lolsho_0 gam_0 pois_0
sum ols_dif lols_dif lolsho_dif gam_dif pois_dif

sum ols_1, meanonly
gen ols1=r(mean)
sum ols_0, meanonly
gen ols0=r(mean)

sum lols_1, meanonly
gen lols1=r(mean)
sum lols_0, meanonly
gen lols0=r(mean)

sum lolsho_1, meanonly
gen lolsho1=r(mean)
sum lolsho_0, meanonly
gen lolsho0=r(mean)

sum gam_1, meanonly
gen gam1=r(mean)
sum gam_0, meanonly
gen gam0=r(mean)

sum pois_1, meanonly
gen pois1=r(mean)
sum pois_0, meanonly
gen pois0=r(mean)
```

keep ols1 ols0 lols1 lols0 lolsho1 lolsho0 gam1 gam0 pois1 pois0
keep if _n==1

3. Bootmulti.do

**** THIS PROGRAM RUNS insideboot.do TO GET bootresult.dta
**** AFTER RUNNING THIS PROGRAM, RUN bootresult.do
**** TO GET CONFIDENCE INTERVALS

version 7
* BOOTSTRAP METHOD

capture erase temp1.dta
capture program drop loopm

program define loopm
local i=`1'
while `i'>0 {
    local i = `i' - 1
    use mdmcea
    preserve
    keep if treat==0
    bsample
    save temp1,replace
    restore
    keep if treat==1
    bsample
    append using temp1
    save temp1,replace

    **** calling in the insideboot.do file to run
    do insideboot
    append using bootm
    save bootm, replace
}
end

clear
set more 1
set seed 123456789
set obs 1
gen temp = 1
save bootm, replace

quietly loopm 2000
drop if temp==1
drop temp
capture erase bootresult.dta
save bootresult, replace
capture erase templ.dta
capture erase bootm.dta

4. Bootresult.do

************************************************************
****THIS PROGRAM GIVES THE CI'S FOR THE MULTIVARIABLE**REGRESSIONS** RUN THIS PROGRAM AFTER GETTING bootresult.dta
************************************************************
clear
set memory 10000
capture log closelog using bootresult.log, replace
use bootresult

gen olsdif=ols1-ols0
gen lolsdif=lols1-lols0
gen lhodif=lolsho1-lolsho0
gen gamdif=gam1-gam0
gen poisdif=pois1-pois0
capture program drop get95ci
program define get95ci sort `1' list `1' if (_n==1+round((_N*0.025),0))|(_n==_N-round((_N*0.025),0))
quietly sum `1'
display "mean   " r(mean)
display "st err   " r(sd)
display "t-stat   " abs(r(mean)/r(sd))

end
get95ci olsdif
get95ci lolsdif
get95ci lhodif
get95ci gamdif
get95ci poisdif
log close

5. Multivariable_tests.do

************************************************************
*********THIS PROGRAM CONDUCTS THE STATISTICAL TESTS**FOLLOWING MULTIVARIABLE ANALYSES
************************************************************
clear
set memory 10000
capture log closelog using multivariable_test.log, replace
use mdmcea, clear
******control variables*******************
global ivar "age ejfract sex etiology race"

******create log of costs
global lst1 "xbt1 xbt2 xbt3 xbt4 xbt5 xbt6 xbt7 xbt8 xbt9 xbt10"

******OLS
regress cost treat $ivar
predict xb, xb
gen mu=xb
gen res= cost-mu
summ res

/* PEARSON CORR TEST */
pwcorr mu res, sig

/* PREGIBON LINK TEST */
qui gen xbsq=xb^2
reg cost xb xbsq, robust
test xbsq

/* MODIFIED HOSMER LEMESHOW TEST */
qui xtile xbtile=xb, nq(10)
qui tab xbtile, gen(xbt)
reg res $lst1, nocons robust
test $lst1
drop xb mu res xbtile xbt1-xtile10 xbsq

***log costs
regress lcost treat $ivar
predict xb, xb

***smearing factor
capture drop res
predict res, r
replace res=exp(res)
summ res
scalar sm=r(mean)
summ res if treat==0
scalar sm0=r(mean)
summ res if treat==1
scalar sm1=r(mean)
drop res

display sm
display sm0
display sm1

replace treat=0
predict lc0
replace treat=1
predict lc1

replace treat=tmp/treat

gen lols_0=exp(lc0)*sm0
gen lols_1=exp(lc1)*sm1
gen lols_t=lols_0 if treat==0
   replace lols_t=lols_0 if treat==1
gen lols_res=cost-lols_t
gen res=lols_res
gen mu=lols_t

/* PEARSON CORR TEST */
pwcorr mu res, sig

/* PREGIBON LINK TEST */
gen xbsq=xb^2
reg lcost xb xbsq, robust
test xbsq

/* MODIFIED HOSMER LEMESHOW TEST */
xtile xbtile=xb, nq(10)
qui tab xbtile, gen(xbt)
reg res $lst1, nocons robust
test $lst1
drop xb mu res xbtile xbt1-xtile10 xbsq r2

****assume homoskedastic!
regress lcost treat $ivar
predict xb, xb

gen lolsho_0=exp(lc0)*sm
ngen lolsho_1=exp(lc1)*sm
ngen lolsho_dif=lolsho_1-lolsho_0
ngen lolsho_t=lolsho_0 if treat==0
   replace lolsho_t=lolsho_1 if treat==1
gen lolsho_res=cost-lolsho_t
gen res=lolsho_res
gen mu=lolsho_t

/* PEARSON CORR TEST */
pwcorr mu res, sig

/* PREGIBON LINK TEST */
gen xbsq=xb^2
gen mu=xbsq, robust
test xbsq

/* MODIFIED HOSMER LEMESHOW TEST */
xtile xbtile=xb, nq(10)
qui tab xbtile, gen(xbt)
reg res $lst1, nocons robust
test $lst1

drop xb mu res xbtile xbt1-xbt10 xbsq r2

*****glm model (gamma/log link)
glm cost treat $ivar, family(gamma) link(log)
predict xb, xb
replace treat=0
predict gam_0
replace treat=1
predict gam_1
gen gam_dif=gam_1-gam_0
replace treat=tmp_treat
gen gam_t=gam_0 if treat==0
replace gam_t=gam_1 if treat==1
gen gam_res=cost-gam_t

/* PEARSON CORR TEST */
pwcorr mu res, sig

/* PREGIBON LINK TEST */
gen r2=(cost-mu)^2
glm cost xb xbsq, family(gamma) link(log) robust nolog
test xbsq

/* MODIFIED HOSMER LEMESHOW TEST */
xtile xbtile=xb, nq(10)
tab xbtile, gen(xbt)
reg res $lst1, nocons robust
test $lst1

/* Modified Park Test */
gen r2=((cost-mu)^2)
gen lnmu=ln(mu)

61 62
clear all
capture log close
set memory 500000
set more 1
global seed=5735766

clear all
set obs 1
gen simul=. save mdmcea_splithalf, replace
log using mdmcea_splithalf.log, replace

gui local k=1
gui while `k' <= 1000 {
tempfile tmp1 tmp2
use mdmcea, clear
gen tmptreat=treat
gen lcost=ln(cost)
global seed = $seed +10000
set seed $seed
save `tmp1', replace
keep patid
keep if patid != patid[_n-1]
sample 50
}
save 'tmp2', replace
merge patid using 'tmp1'
drop _merge
replace split=0 if split==.
/* Split data into two equal parts */

****** control variables ********************
global ivar "age ejfract sex etiology race"

** For demonstration purposes Copas test for only OLS regression shown in this sample program - Can similarly include other multivariable models

****** OLS
regress cost treat $ivar if split==1
predict ols1_t replace treat=0
predict ols1_0 replace treat=1
predict ols1_1 replace treat=tmptreat
/* In-sample Prediction */
gen own_ols = ols1_t if split==1
replace own_ols=. if split==0
/* Out-sample Prediction */
gen cross_ols = ols1_t if split==0
replace cross_ols=. if split==1
/* Efficiency Measures */
gen own_ols_mse = (cost-own_ols)^2
gen own_ols_ape = abs(cost-own_ols)
gen cross_ols_mse = (cost-cross_ols)^2
gen cross_ols_ape = abs(cost-cross_ols)

summ own_ols_mse
gen own_olsmse = r(mean)
summ own_olsape
gen own_olsape = r(mean)
summ cross_ols_mse
gen cross_olsmse = r(mean)
summ cross_ols_ape
gen cross_olsape = r(mean)

/* COPAS TEST */

**OLS**
gen own_res = cost-own_ols /* Create Residuals */
gen cross_res = cost-cross_ols /* Create Residuals */

reg own_res own_ols, robust
gen own_b0_ols = _b[_cons]
gen own_b1_ols = _b[own_ols]
gen own_ols_copusmse = e(rmse)^2

test _b[_cons] = 0, notest
test _b[own_ols] = 0, accumulate
gen own_fstat_ols = r(F)
gen own_pval_ols = (r(p)<=0.05)

reg cross_res cross_ols, robust
gen cross_b0_ols = _b[_cons]
gen cross_b1_ols = _b[cross_ols]
gen cross_ols_copusmse = e(rmse)^2

test _b[_cons] = 0, notest
test _b[cross_ols] = 0, accumulate
gen cross_fstat_ols = r(F)
gen cross_pval_ols = (r(p)<=0.05)

drop own_res cross_res

gen simul = `k'
keep if _n == 1
keep simul own_b0_ols own_b1_ols own_fstat_ols own_pval_ols own_ols_copusmse /* */
cross_b0_ols cross_b1_ols 
cross_fstat_ols cross_pval_ols cross_ols_copusmse

append using mdmcea_splitwhalf
save mdmcea_splitwhalf, replace

noi di "Simul: " `k'
local k = `k' +1

/* */

ci own_b0_ols own_b1_ols own_fstat_ols own_pval_ols /*
From pp. 16 and 23

*** OLS (untransformed)

```
. regress cost treat $ivar
```

```
Source |      SS      df       MS        Number of obs =     200
---------+----------------------------     F(  6,   193) =    5.45
Model | 1.1771e+09    6  196177149     Prob > F      =  0.0000
Residual | 6.9525e+09  193  36023463.2     R-squared     =  0.1448
---------+----------------------------     Adj R-squared =  0.1182
Total | 8.1296e+09  199  40852217.5     Root MSE      =    6002
---------+----------------------------
```

```
|       | Coef. | Std. Err. | t   | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|-----|------|-----------------------|
| treat | 4026.585 | 853.8104 | 4.72 | 0.000 | 2342.587 - 5710.582 |
| age   | 75.1648 | 42.17426 | 1.78 | 0.076 | -8.016836 - 158.3464 |
| ejfract | -76.18302 | 64.84257 | -1.17 | 0.241 | -204.0741 - 51.70804 |
| sex   | -652.0594 | 975.3994 | -0.67 | 0.505 | -2575.87 - 1271.752 |
| etiology | 2264.755 | 887.3851 | 2.55 | 0.011 | 514.5371 - 4014.972 |
| race  | 481.5117 | 1284.211 | 0.37 | 0.708 | -2051.37 - 3014.401 |
| _cons | 2676.532 | 3087.71 | 0.87 | 0.387 | -3413.456 - 8766.52 |
```

. ***Log costs

```
. regress lcost treat $ivar
```

```
Source |       SS       df       MS       Number of obs =     200
---------+------------------------------    F(  6,   193) =   11.41
Model |  39.2814791     6  6.54691318    Prob > F      =  0.0000
Residual | 110.692964   193  .573538672    R-squared     =  0.2619
---------+------------------------------    Adj R-squared =  0.2390
Total | 149.974443   199  .753640416    Root MSE      =  .75732
---------+------------------------------
```

```
|       | Coef. | Std. Err. | t   | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|-----|------|-----------------------|
| treat | .7356582 | .107733 | 6.83 | 0.000 | .5231724 - .9481441 |
| age   | .0072738 | .0053215 | 1.37 | 0.173 | -.003222 - .0177696 |
| ejfract | -.016828 | .0081818 | -2.06 | 0.041 | -.0329652 - .0006908 |
| sex   | -.112238 | .1230754 | -0.91 | 0.363 | -.3549834 - .1305074 |
| etiology | .3244058 | .1191698 | 2.90 | 0.004 | .1035643 - .5452473 |
| race  | .3329655 | .162041 | 2.05 | 0.041 | .0133669 - .6525641 |
| _cons | 7.998734 | .3896055 | 20.53 | 0.000 | 7.230302 - 8.767165 |
```

67
### From p. 30

#### *** glm model (gamma/log)

```
glm cost treat $ivar, family(gamma) link(log)
```

<table>
<thead>
<tr>
<th>Generalized linear models</th>
<th>No. of obs</th>
<th>Residual df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization</td>
<td>ML: Newton-Raphson</td>
<td>193</td>
</tr>
<tr>
<td>Scale parameter</td>
<td>0.522547</td>
<td></td>
</tr>
<tr>
<td>Deviance</td>
<td>101.9865419</td>
<td>(1/df) Deviance = 0.5284277</td>
</tr>
<tr>
<td>Pearson</td>
<td>100.8515681</td>
<td>(1/df) Pearson = 0.522547</td>
</tr>
</tbody>
</table>

- Variance function: $V(u) = u^2$  
- Link function: $g(u) = \ln(u)$
- Standard errors: OIM
- Log likelihood = -2009.40744  
- AIC = 20.16407
- BIC = -920.5887098

| treat | Coef. | Std. Err. | z    | P>|z| | 95% Conf. Interval |
|-------|-------|-----------|------|-----|-------------------|
| 0.4738967 | 0.1026917 | 4.61 | 0.000 | 0.2726247 | 0.6751687 |
| age | 0.0064758 | 0.0048601 | 1.33 | 0.183 | -0.0030498 | 0.0160014 |
| ejfract | -0.0077655 | 0.0081786 | -0.95 | 0.342 | -0.0237954 | 0.0082643 |
| sex | -0.0722942 | 0.1178488 | -0.61 | 0.540 | -0.3032737 | 0.1586853 |
| etiology | 0.2685697 | 0.1103177 | 2.43 | 0.015 | 0.052351 | 0.4847884 |
| race | 0.1668937 | 0.1587082 | 1.05 | 0.293 | -0.1441686 | 0.4779561 |
| _cons | 8.354721 | 0.3801616 | 21.98 | 0.000 | 7.609618 | 9.099825 |

#### ***** glm model (poisson/log)

```
glm cost treat $ivar, family(poisson) link(log)
```

<table>
<thead>
<tr>
<th>Generalized linear models</th>
<th>No. of obs</th>
<th>Residual df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization</td>
<td>ML: Newton-Raphson</td>
<td>193</td>
</tr>
<tr>
<td>Scale parameter</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Deviance</td>
<td>700567.946</td>
<td>(1/df) Deviance = 3629.886</td>
</tr>
<tr>
<td>Pearson</td>
<td>791555.8081</td>
<td>(1/df) Pearson = 4101.325</td>
</tr>
</tbody>
</table>

- Variance function: $V(u) = u$  
- Link function: $g(u) = \ln(u)$
- Standard errors: OIM
- Log likelihood = -351346.9719  
- AIC = 3513.54
- BIC = 699545.3708

| cost | Coef. | Std. Err. | z    | P>|z| | 95% Conf. Interval |
|------|-------|-----------|------|-----|-------------------|
| 0.4629637 | 0.0015546 | 297.81 | 0.000 | 0.4599168 | 0.4660106 |
| age | 0.0082989 | 0.0000756 | 109.72 | 0.000 | 0.0081507 | 0.0084472 |
| ejfract | -0.0081781 | 0.0001135 | -72.07 | 0.000 | -0.0084006 | -0.0079557 |
| sex | -0.0721448 | 0.0016935 | -42.60 | 0.000 | -0.0754639 | -0.0688256 |
| etiology | 0.2498528 | 0.0001135 | 159.99 | 0.000 | 0.2467919 | 0.2529137 |
| race | 0.0462949 | 0.0023699 | 19.53 | 0.000 | 0.0416499 | 0.0509398 |
| _cons | 8.359824 | 0.005554 | 1505.18 | 0.000 | 8.348939 | 8.37071 |

---

68
Tests for Normality of Residuals

Are the residuals normally distributed?

(1) Skewness for a normal distribution is zero
   Negative values – left skewed data
   Positive values – right skewed data

(2) Kurtosis for a standard normal distribution is 3

After OLS cost regression:

```
. sum res_ols, detail
```

```
Residuals
Percentiles   Smallest
1%   -6865.266     -6954.6
5%   -6285.974   -6933.082
10%  -5321.376    -6797.45       Obs                 200
25%  -3998.92   -6778.228       Sum of Wgt.         200
50%  -1597.021                   Mean           2.78e-06
75%   2502.182    17393.61       Largest Std. Dev.  5910.781
90%   7335.55    17773.35       Variance       3.49e+07
95%  10781.04    25293.73       Skewness       1.960199
99%  21533.54    32942.74       Kurtosis        8.85317
```

```
. sktest res_ols
```

```
Skewness/Kurtosis tests for Normality
-------- joint -------
Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
----------|----------------------------------------
res_ols   |   0.000         0.000         72.09       0.0000
```

NO FOR OLS
After log OLS cost regression:

. sum res_log, detail

Residuals

------------------------------------------
<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>-1.879498</td>
<td>-2.256186</td>
</tr>
<tr>
<td>5%</td>
<td>-1.2353</td>
<td>-2.033567</td>
</tr>
<tr>
<td>10%</td>
<td>-1.025368</td>
<td>-1.725428</td>
</tr>
<tr>
<td>25%</td>
<td>-.4435903</td>
<td>-1.623668</td>
</tr>
<tr>
<td>50%</td>
<td>-.055706</td>
<td>Mean</td>
</tr>
<tr>
<td>75%</td>
<td>.5797378</td>
<td>1.407953</td>
</tr>
<tr>
<td>90%</td>
<td>.9839813</td>
<td>1.572954</td>
</tr>
<tr>
<td>95%</td>
<td>1.207211</td>
<td>1.787229</td>
</tr>
<tr>
<td>99%</td>
<td>1.680091</td>
<td>1.860877</td>
</tr>
</tbody>
</table>

---------- Largest Std. Dev.  .745819

Obs 200          Sum of Wgt.  200

5%     .055706

. sktest res_log

Skewness/Kurtosis tests for Normality

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pr(Skewness)</th>
<th>Pr(Kurtosis)</th>
<th>adj chi2(2)</th>
<th>Prob&gt;chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>res_log</td>
<td>0.475</td>
<td>0.945</td>
<td>0.52</td>
<td>0.7707</td>
</tr>
</tbody>
</table>

YES FOR LOG
/* Modified Park Test */
. gen r2 = ((cost-yhat)^2)
. gen lnyhat = ln(yhat)

. glm r2 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 = 2.808569
 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 = -3667.729811               AIC = 36.6973
BIC =-492.9701783

------------------------------------------------------------------------------
|               Robust               
|       r2       | 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
------------------------------------------------------------------------------

. test lnyhat==1
   ( 1)  [r2]lnyhat = 1
          chi2(  1) =    0.10
          Prob > chi2 =    0.7488

. test lnyhat==2
   ( 1)  [r2]lnyhat = 2
          chi2(  1) =    3.88
          Prob > chi2 =    0.0487

. test lnyhat==3
   ( 1)  [r2]lnyhat = 3
          chi2(  1) =   13.11
          Prob > chi2 =    0.0003
From pp. 44

```stata
/* PREGIBON LINK TEST */
gen xbsq=xb^2
glm cost xb xbsq, family(gamma) link(log) robust nolog
```

Generalized linear models
Optimization : ML: Newton-Raphson
Scale parameter = .5366878
Deviance = 101.0412175 (1/df) Deviance = .5128996
Pearson = 105.727504 (1/df) Pearson = .5366878

Variance function: V(u) = u^2  [Gamma]
Link function : g(u) = ln(u)  [Log]
Standard errors : Sandwich

Log pseudo-likelihood = -2008.934778  AIC = 20.11935
BIC = -942.7273038

|    | Coef.  | Robust Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----|--------|------------------|-----|-----|---------------------|
| xb | 16.67377 | 13.38723         | 1.25| 0.213| -9.564707           | 42.91225 |
| xbsq | -0.863838 | 0.7334817      | -1.18| 0.239| -2.301436           | 0.5737597 |
| _cons | -71.02776 | 61.05194       | -1.16| 0.245| -190.6874           | 48.63184 |

.test xbsq

( 1)  [cost]xbsq = 0

    chi2( 1) =    1.39
    Prob > chi2 =    0.2389
/* MODIFIED HOSMER LEMESHOW TEST */
.xtile xbtile=xb, nq(10)
.tab xbtile, gen(xbt)
10 quantiles of xb

<table>
<thead>
<tr>
<th>of xb</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>10.00</td>
<td>10.0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>10.00</td>
<td>20.0</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>10.00</td>
<td>30.0</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>10.00</td>
<td>40.0</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>10.00</td>
<td>50.0</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>10.00</td>
<td>60.0</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>10.00</td>
<td>70.0</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>10.00</td>
<td>80.0</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>10.00</td>
<td>90.0</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>10.00</td>
<td>100.0</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

.reg res xbt1 xbt2 xbt3 xbt4 xbt5 xbt6 xbt7 xbt8 xbt9 xbt10, nocons robust

Regression with robust standard errors

Number of obs = 200
F( 10,   190) = 0.65
Prob > F = 0.7652
R-squared = 0.0293
Root MSE = 5983.8

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
</tr>
<tr>
<td></td>
<td>-------</td>
</tr>
<tr>
<td>xbt1</td>
<td>-858.0814</td>
</tr>
<tr>
<td>xbt2</td>
<td>-1082.153</td>
</tr>
<tr>
<td>xbt3</td>
<td>2533.632</td>
</tr>
<tr>
<td>xbt4</td>
<td>-390.9682</td>
</tr>
<tr>
<td>xbt5</td>
<td>-10.61213</td>
</tr>
<tr>
<td>xbt6</td>
<td>960.5702</td>
</tr>
<tr>
<td>xbt7</td>
<td>170.9792</td>
</tr>
<tr>
<td>xbt8</td>
<td>-330.7193</td>
</tr>
<tr>
<td>xbt9</td>
<td>-519.5159</td>
</tr>
<tr>
<td>xbt10</td>
<td>-686.1769</td>
</tr>
</tbody>
</table>

.test xbt1 xbt2 xbt3 xbt4 xbt5 xbt6 xbt7 xbt8 xbt9 xbt10
( 1)  xbt1 = 0
( 2)  xbt2 = 0
( 3)  xbt3 = 0
( 4)  xbt4 = 0
( 5)  xbt5 = 0
( 6)  xbt6 = 0
( 7)  xbt7 = 0
( 8)  xbt8 = 0
( 9)  xbt9 = 0
(10)  xbt10 = 0
(1)  xbt1 = 0
(2)  xbt2 = 0
(3)  xbt3 = 0
(4)  xbt4 = 0
(5)  xbt5 = 0
(6)  xbt6 = 0
(7)  xbt7 = 0
(8)  xbt8 = 0
(9)  xbt9 = 0
(10) xbt10 = 0
F( 10,   190) = 0.65
Prob > F = 0.7652
<table>
<thead>
<tr>
<th>Model</th>
<th>Pregibon Link</th>
<th>Modified Hosmer-Lemeshow</th>
<th>Pearson correlation</th>
<th>Modified Park</th>
<th>Copas (Rank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>Passed</td>
<td>Passed</td>
<td>Passed</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Log OLS (Homo)</td>
<td>Failed</td>
<td>Passed</td>
<td>Failed</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>Log OLS (Hetero)</td>
<td>Failed</td>
<td>Passed</td>
<td>Failed</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>GLM gamma/log</td>
<td>Passed</td>
<td>Passed</td>
<td>Passed</td>
<td>Failed</td>
<td>2</td>
</tr>
<tr>
<td>GLM poisson/log</td>
<td>Passed</td>
<td>Passed</td>
<td>Passed</td>
<td>Passed</td>
<td>1</td>
</tr>
</tbody>
</table>

- GLM poisson/log link seems to have performed well on all tests as compared to other models!
- However, in this example dataset, the incremental costs estimated from the GLM poisson ($4,018) and untransformed OLS ($4,027) were not very different (although Log OLS (homoskedastic) was quite different ($6649))
- In other examples, we have seen substantial differences across models!