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.

Generalized Linear Models (GLM)

- These models have the advantages of the log models, but (a direct transformation of) ΔC 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:
  
  \[ \text{glm } y \ x, \text{ link(linkname) family (familyname)} \]

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

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:

\[ \begin{align*}
  \text{a} &= _\text{mean}_; \\
  \text{b} &= _\text{resp}_; \\
  \text{d} &= \text{b/a + log(a)} \\
  \text{variance var} &= \text{a}^2 \\
  \text{deviance dev} &= \text{d;} \\
  \text{model } y &= x / \text{ link = log;} \\
  \text{run; }
\end{align*} \]
The Link Function (I)

- Link function directly characterizes how the linear combination of the predictors is related to the mean on the original scale
  - e.g. identity, log, power # (square root, etc.)
- For example, if one uses a log link, one is assuming:
  \[ \ln(E(y/x)) = X\beta \]
- GLM with a log link differs from log OLS in part because in log OLS, one is assuming:
  \[ E(\ln(y/x)) = X\beta \]
- \[ \ln(E(y/x)) \neq E(\ln(y)/x) \]
  i.e. log of the mean ≠ mean of the log costs

\[
\begin{array}{llll}
\text{Variable} & \text{Group 2} & \text{Group 3} \\
\hline
\text{Observations} & 15 & 36 \\
2 & 45 & 45 \\
3 & 75 & 55 \\
\text{Arithmetic mean} & 45 & 45 \\
\text{Log, arith mean cost} & 3.806662 & 3.806662 * \\
\text{Natural log} & 2.70805 & 3.555348 \\
2 & 3.806662 & 3.806662 \\
3 & 4.317488 & 4.007333 \\
\text{Arithmetic mean} & 3.610734 & 3.789781 † \\
\end{array}
\]

* Difference = 0; † Difference = 0.179047

Comparison of Results of GLM 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 3</td>
<td>0.0000000</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 3</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>
The Link Function (II)

- Log link is most commonly used in literature but may not be the best in all cases
- No single test identifies the appropriate link
- Can employ multiple tests of fit:
  - Pregibon link test checks linearity of response on scale of estimation
  - Modified Hosmer-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: variance is proportional to cube of mean

Modified Park Test

- A “constructive” test that recommends a family given a particular link function
- Implemented after GLM regression assuming a family & link
- Example: glm cost treat $ivar, family(gamma) link(log)
  1. Predict value of y and log transform it
     predict yhat
     gen lnyhat=ln(yhat)
  2. Save raw scale residuals and square them
     gen res=cost-yhat
     gen r2=((res)^2)
(3) Regress \( \ln(r^2) \) on \( \ln(\text{yhat}) \) and a constant using GLM with log link and gamma distribution

\[ \text{glm } r^2 \ln(\text{yhat}), \text{link(}\log\text{)} \text{ family(}\text{gamma}\text{)} \text{ robust nolog} \]

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

\[ \text{test ln(\text{yhat})==0} \]
\[ \text{test ln(\text{yhat})==1} \]
\[ \text{test ln(\text{yhat})==2} \]
\[ \text{test ln(\text{yhat})==3} \]

GLM Comments (I)
- Advantages
  - No retransformation problems of log OLS models
  - Because the link function allows modeling of the log of mean costs [i.e. \( \ln(E(y|x)=X\beta) \)] unlike the log OLS that models the mean of log costs [i.e. \( E(\ln(y)|x)=X\beta) \)]
  - 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

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

Retransformation (II)

• 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 $\hat{Z}_1$

Multivariate Analysis: Outline

• Analysis of cost
  – Start with everyone’s "old" favorite: OLS
  – Check the fit of the gauss family used in OLS
    • Revise family if necessary
  – Start with everyone’s "new" favorite: GLM gamma/log
  – Check the fit of the gamma family
    • Revise family if necessary
  – Tune the link
• QALYs
  – You evaluate OLS and GLM gamma/log
  – We review OLS and GLM gamma/log and tune link
• Summary
Sample Dataset

.cclear
.set more off
.use rchapter5
.sum

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>500</td>
<td>250.5</td>
<td>144.4818</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>treat</td>
<td>500</td>
<td>0.5</td>
<td>0.5005008</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>cost</td>
<td>500</td>
<td>1389.92</td>
<td>315</td>
<td>10499</td>
<td></td>
</tr>
<tr>
<td>qaly</td>
<td>500</td>
<td>0.212114</td>
<td>0.4798</td>
<td>0.95119</td>
<td></td>
</tr>
<tr>
<td>dissev</td>
<td>500</td>
<td>0.1124773</td>
<td>0.025</td>
<td>0.729</td>
<td></td>
</tr>
<tr>
<td>race</td>
<td>500</td>
<td>0.506</td>
<td>0.5004647</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>blcost</td>
<td>500</td>
<td>770.5504</td>
<td>111.0891</td>
<td>4926.931</td>
<td></td>
</tr>
<tr>
<td>blqaly</td>
<td>500</td>
<td>0.4895464</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>500</td>
<td>0.484</td>
<td>0.5002444</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

Inspect the Cost Data (I)

.sum cost if treat==0,detail

Total cost

Percentiles       Smallest
1%  622           315
5%  899           589
10% 1093           622              Obs 250
25% 1170           640              Sum of Wgt. 250
50% 2825.5        Mean 3015
75% 3752          Largest 1582.802
90% 5085          Std. Dev. 2505262
95% 6470          Variance 9.234913
99% 7540          Skewness 4.910192

Inspect the Cost Data (II)

.sum cost if treat==1,detail

Total cost

Percentiles       Smallest
1%  1093           681
5%  1426           899
10% 1832          1093              Obs 250
25% 2226           1170              Sum of Wgt. 250
50% 2900.5        Mean 3040
75% 3604          Largest 1168.737
90% 4404          Std. Dev. 1365946
95% 5085          Variance 1.525386
99% 6470          Skewness 9.234913
Inspect the QALY Data (I)

```
.drop_all
.use rchapter5
.sum qaly if treat==0,detail
```

<table>
<thead>
<tr>
<th>QALYs</th>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
<td>.0861</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>.16518</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>.268535</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>.41286</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>.63118</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>.74512</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>.834055</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>.86954</td>
</tr>
<tr>
<td></td>
<td>Largest</td>
<td>.95119</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>.2171849</td>
</tr>
</tbody>
</table>

```
.sum qaly if treat==1,detail
```

<table>
<thead>
<tr>
<th>QALYs</th>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
<td>.15776</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>.2283</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>.333105</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>.46083</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>.65595</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>.73928</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>.856215</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>.948104</td>
</tr>
<tr>
<td></td>
<td>Largest</td>
<td>.95119</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>.2051629</td>
</tr>
</tbody>
</table>

Common Starting Point: OLS Regression

```
regress cost treat dissev blcost blqaly race
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>Df</th>
<th>MS</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2473e+5</td>
<td>5</td>
<td>416e+5</td>
<td>0.000</td>
</tr>
<tr>
<td>Resid</td>
<td>9640e+5</td>
<td>499</td>
<td>193e+4</td>
<td>0.2490</td>
</tr>
<tr>
<td>Total</td>
<td>9840e+5</td>
<td>494</td>
<td>193e+4</td>
<td>0.2565</td>
</tr>
</tbody>
</table>

| Cost  | Coef  | Std Err | T     | P>|t| | [95% Conf Int] |
|-------|-------|---------|-------|-----|----------------|
| Treat | 21.993 | 107.77  | 0.20 | 0.838 | -189.74 - 233.74 |
| dissev| 453.41 | 516.34  | 0.81 | 0.000 | 3018.92 - 5047.91 |
| blcost| .3945 | .0758  | 5.20 | 0.000 | 0.2455 - 0.5435 |
| blqaly| -.7733 | .37198 | -2.08 | 0.038 | -1.504.16 - -2.42.45 |
| race  | -768.02 | 118.75 | -6.47 | 0.000 | -1001.35 - -534.69 |
| _cons | 1966.32 | 366.11 | 5.37 | 0.000 | 1247.00 - 2685.64 |
Rerun OLS as GLM (gauss/identity)

- Variance function: $V(u) = 1$ (Gaussian)
- Link function: $g(u) = u$ (Identity)
- Log likelihood = -4253

| Variable | Coef     | Std Err | t     | P>|t| | 95% Cl |
|----------|----------|---------|-------|-----|--------|
| treat    | 21.99324 | 107.766 | 0.20  | 0.838 | -189.2247 | 233.2112 |
| dissev   | 4033.414 | 516.340 | 7.81  | 0.000 | 3021.406 | 5045.423 |
| blcost   | 0.944632 | 0.75840 | 5.20  | 0.000 | 0.2458189 | 0.5431576 |
| bIqaly   | -773.301 | 371.978 | -2.08 | 0.038 | -1502366 | -44.23705 |
| race     | -768.020 | 118.754 | -6.47 | 0.000 | -1000.775 | -535.2645 |
| cons     | 1966.319 | 366.106 | 5.37  | 0.000 | 1248.765 | 2683.874 |

- Create temp=treat and Construct Recycled Predictions
- gen temp=treat

```
replace temp=0
predict polsc0
replace temp=1
predict polsc1
sum polsc*
```

```
Variable | Obs  | Mean  | Std. Dev. | Min  | Max
----------|------|-------|-----------|------|------
polsc0    | 500  | 3016.503 | 703.866   | 1184.116 | 5527.065
polsc1    | 500  | 3038.497 | 703.866   | 1206.109 | 5549.059
```

```
replace temp=treat
quietly do glmdiagnostic
glmdiag
```

```
Fitted model: Link = Identity; Family = Gaussian
```

Replace temp=treat; Run GLM DIAGNOSTICS (glmdiag)

```
Results, Modified Park Test (for Family)

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>1.391784</th>
</tr>
</thead>
</table>

Family, Chi2, and p-value in descending order of likelihood

<table>
<thead>
<tr>
<th>Family</th>
<th>Chi2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>1.4021</td>
<td>0.2364</td>
</tr>
<tr>
<td>Gamma</td>
<td>3.3790</td>
<td>0.0660</td>
</tr>
<tr>
<td>Gaussian NLLS</td>
<td>17.6936</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Inverse Gaussian or Wald

| Coefficient | 23.6244 | 0.0000 |

Results of tests of GLM Identity link

| Pearson Correlation Test | 1.0000 |
| Pregibon Link Test       | 0.8913 |
| Modified Hosmer and Lemeshow | 0.3487 |
GLMDIAG Saved Results

```
. return list
scalars:
   r(ln_coef) =  1.591162
   r(p_gaus) =  1.00000000000e-07
   r(p_pois) =  .043146
   r(p_gam) =  .1619374
   r(p_igaus) =  1.40000000000e-06
   r(p_pearson) =  .246
   r(p_pregibon) =  .1273
   r(p_h_m) =  .6199
macros:
   r(family) : "gamma"
```

Change Family to Poisson and Rerun Model

Variance function:  V(u) = u
Link function:         g(u) = u
Log likelihood = -113576

```
|    | Coef  | Std Err |   t | P>|t| | 95% C.I     |
|----|-------|---------|-----|-----|----------------|
| 1  | 113.11| 4.798526|23.57|0.000|103.71|122.52 |
| 2  | 4008.43|22.67209|176.80|0.000|3964.00|4052.67 |
| 3  | 3861272 |0.036013 |107.22|0.000 |.791 .9392 |
| 4  | -765.3726 |16.58928 | -46.14 |0.000 | -797.89 | -732.86 |
| 5  | -746.5739 | 5.324134|140.22|0.000 | -757.01 | -736.14 |
| 6  | 1925.985 |16.49097 | 116.79|0.000 |1893.664 |1958.307 |
```

• Problem with p-values?

Change in Family Leads to Fairly Big Differences in Point Estimate (Not Sure About SE)

```
| Cost     | Coef  | Std Err |    z |   P>|z| | [95% Conf Interval] |
|----------|-------|---------|------|-----|---------------------|
| Gaussian / Identity |
| temp     | 21.99 | 107.77  | 0.20 |0.838| -189.2247 | 233.2112 |
| Poisson / Identity |
| temp     | 113.11| 4.80    | 23.57|0.000| 103.71 | 122.5198 |
```
Suppose We Started with GLM Gamma/Log?

Variance function:  \( V(u) = u^2 \)  
Link function:         \( g(u) = \ln(u) \)

Log likelihood = -4494

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>Std Err</th>
<th>t</th>
<th>P&gt;abs(t)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temp</td>
<td>.0446782</td>
<td>.0356359</td>
<td>1.25</td>
<td>.210</td>
<td>-.0251669 to .1145232</td>
</tr>
<tr>
<td>dsev</td>
<td>1.409376</td>
<td>.1739606</td>
<td>8.10</td>
<td>.000</td>
<td>1.06842 to  1.750333</td>
</tr>
<tr>
<td>blcost</td>
<td>.000122</td>
<td>.0000257</td>
<td>4.78</td>
<td>.000</td>
<td>.0000724 to  .0017300</td>
</tr>
<tr>
<td>blqaly</td>
<td>-.2579657</td>
<td>.1223431</td>
<td>-.21</td>
<td>.035</td>
<td>-.4977537 to  -.0183796</td>
</tr>
<tr>
<td>race</td>
<td>.2601311</td>
<td>.0395492</td>
<td>6.61</td>
<td>.000</td>
<td>-.3388262 to   1.837961</td>
</tr>
<tr>
<td>cons</td>
<td>7.61057</td>
<td>.1220851</td>
<td>62.34</td>
<td>.000</td>
<td>7.371291 to   7.849856</td>
</tr>
</tbody>
</table>

replace temp=0
predict pgmglc0
replace temp=1
predict pgmglc1
sum pgmglc*

Variable |     |     |     |   |     |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pgmglc0</td>
<td>500</td>
<td>2964.034</td>
<td>733.726</td>
<td>1542.916</td>
<td>6767.186</td>
</tr>
<tr>
<td>pgmglc1</td>
<td>500</td>
<td>3099.465</td>
<td>767.251</td>
<td>1613.414</td>
<td>7076.388</td>
</tr>
</tbody>
</table>

replace temp=treat
glmdiag
Recycled vs Treatment-Specific Predictions

```stata
. quietly glm cost temp dissev bl* race, link(log) family(gamma)
. predict pcost
   (option mu assumed; predicted mean cost)
. replace temp=0
   (250 real changes made)
. predict pcost0
   (option mu assumed; predicted mean cost)
. replace temp=1
   (500 real changes made)
. predict pcost1
   (option mu assumed; predicted mean cost)
```

Recycled vs Treatment-Specific Predictions (II)

```stata
. tab treat, sum(pcost)
Treatment | Summary of predicted mean cost
| Mean   | Std. Dev. | Freq. |
----------|----------|--------|
0         | 2973.8331| 789.66446| 250      |
1         | 3089.2184| 705.44167| 250      |
----------|----------|----------|--------|
Total     | 3031.5257| 750.21371| 500      |

. sum pcost?
```

Variable | Obs | Mean     | Std. Dev | Min    | Max    |
---------|-----|----------|----------|--------|--------|
pcost0   | 500 | 2964.034 | 733.7266 | 1542.916| 6767.186 |
pcest1   | 500 | 3099.465 | 767.2515 | 1613.414| 7076.388 |

$115 vs $135 estimated difference in cost

Run GLM DIAGNOSTICS, Gamma/Log

Fitted Model: Link = Log; Family = Gamma

Results, Modified Park Test (for Family)

Coefficient: 1.5912

Family, Chi2, and p-value in descending order of likelihood

<table>
<thead>
<tr>
<th>Family</th>
<th>Chi2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>1.9560</td>
<td>0.1619</td>
</tr>
<tr>
<td>Poisson</td>
<td>4.0897</td>
<td>0.0431</td>
</tr>
<tr>
<td>Inverse Gaussian or Wald</td>
<td>23.2272</td>
<td>0.0000</td>
</tr>
<tr>
<td>Gaussian NLLS:</td>
<td>29.6281</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Results of tests of GLM Identity link

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation Test</td>
<td>0.246</td>
<td></td>
</tr>
<tr>
<td>Pregibon Link Test</td>
<td>0.1274</td>
<td></td>
</tr>
<tr>
<td>Modified Hosmer and Lemeshow</td>
<td>0.6199</td>
<td></td>
</tr>
</tbody>
</table>
Passes Tests, But Can We Improve the Link?

- We can generate a wide variety of links by use of the power link
  - power 1 = identity link; \( g(u) = B \sum X_i \)
  - power 0.5 = square root link; \( g(u) = (B \sum X_i)^2 \)
  - power 0 = log link; \( g(u) = \exp(B \sum X_i) \)
  - power -1 = reciprocal link; \( g(u) = 1 / (B \sum X_i) \)
- More generally for negative power links:
  \[
  g(u) = \frac{1}{\exp(B \sum X_i)}
  \]

Passes Tests, But Can We Improve the Link?

- We iteratively evaluate power links (in 0.1 intervals) between 1 (identity) and 0 (log)
  - Fine tune with 0.05 intervals in range of optimum

GLM Diagnostics for Power Links Between 0 and 1

<table>
<thead>
<tr>
<th>power</th>
<th>mod park</th>
<th>pearson</th>
<th>pregibon</th>
<th>h&amp;m</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>gamma</td>
<td>0.246</td>
<td>0.127</td>
<td>0.620</td>
</tr>
<tr>
<td>0.1</td>
<td>gamma</td>
<td>0.303</td>
<td>0.185</td>
<td>0.699</td>
</tr>
<tr>
<td>0.2</td>
<td>gamma</td>
<td>0.368</td>
<td>0.260</td>
<td>0.704</td>
</tr>
<tr>
<td>0.3</td>
<td>gamma</td>
<td>0.443</td>
<td>0.356</td>
<td>0.714</td>
</tr>
<tr>
<td>0.4</td>
<td>gamma</td>
<td>0.526</td>
<td>0.472</td>
<td>0.578</td>
</tr>
<tr>
<td>0.5</td>
<td>gamma</td>
<td>0.618</td>
<td>0.606</td>
<td>0.661</td>
</tr>
<tr>
<td>0.65</td>
<td>gamm/pois</td>
<td>0.719</td>
<td>0.757</td>
<td>0.695 *</td>
</tr>
<tr>
<td>0.7</td>
<td>poisson</td>
<td>0.903</td>
<td>0.747</td>
<td>0.587 **</td>
</tr>
<tr>
<td>0.75</td>
<td>poisson</td>
<td>0.929</td>
<td>0.823</td>
<td>0.405 *</td>
</tr>
<tr>
<td>0.8</td>
<td>poisson</td>
<td>0.956</td>
<td>0.910</td>
<td>0.323 *</td>
</tr>
<tr>
<td>0.9</td>
<td>poisson</td>
<td>0.985</td>
<td>0.981</td>
<td>0.319</td>
</tr>
<tr>
<td>1.0</td>
<td>poisson</td>
<td>0.951</td>
<td>0.858</td>
<td>0.421</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.888</td>
<td>0.702</td>
<td>0.513</td>
</tr>
</tbody>
</table>
QALY Evaluation

- Start with GLM gauss/identity
  - Evaluate GLM diagnostics
  - If necessary, reestimate GLM with better fitting family
- Also assess GLM gamma/log
  - Evaluate GLM diagnostics
  - If necessary, reestimate GLM with better fitting family
Common Starting Point: GLM with Gauss/Identity (OLS)

Variance function: V(u) = 1
Link function: g(u) = u
Log likelihood = 81.96004

| Cost  | Coef  | Std Err | t     | P>|t| | 95% CI |
|-------|-------|---------|-------|-----|--------|
| Temp  | .0417238 | .0184664 | 2.28  | 0.034 | .0055304 | .0779172 |
| Dissev| -.1528245 | .0836938 | -1.83 | 0.068 | -.3168613 | .0112124 |
| blqaly| -.0000361 | .0000122 | -2.96 | 0.003 | -.00006 | -.0000122 |
| blqal| 2089839 | .0837204 | 3.28  | 0.001 | .0840943 | .338735 |
| cons  | .5212725 | .0824228 | 6.35  | 0.000 | .3989260 | .6436189 |

Run GLM DIAGNOSTICS, Gauss/Identity

Fitted Model: Link = Identity; Family = Gaussian
Results, Modified Park Test (for Family)
Coefficient: -1.3983
Family, Chi², and p-value in descending order of likelihood
Family | Chi² | P-value
Gaussian NLLS: 8.0884 | 0.0045
Poisson: 23.7938 | 0.0000
Gamma: 47.7724 | 0.0000
Inverse Gaussian or Wald: 80.0234 | 0.0000

Results of tests of GLM Identity link
Pearson Correlation Test: 1
Pregibon Link Test: .7332
Modified Hosmer and Lemeshow: .8933

replace temp=0
predict pglmig0
replace temp=1
predict pglmig1
sum pglmig*

Variable | Obs | Mean | Std. Dev. | Min | Max
----------|-----|------|-----------|-----|-----
pglmig0 | 500 | .5733035 | .0476744 | .4297185 | .6776147 |
p glmig1 | 500 | .8150273 | .0476744 | .4714423 | .7193385 |
Can We Improve the Link?

<table>
<thead>
<tr>
<th>power</th>
<th>family</th>
<th>pearson</th>
<th>pregibon</th>
<th>h&amp;m</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.5</td>
<td>gauss</td>
<td>0.990</td>
<td>0.988</td>
<td>0.980</td>
</tr>
<tr>
<td>-0.4</td>
<td>gauss</td>
<td>0.993</td>
<td>0.909</td>
<td>0.980</td>
</tr>
<tr>
<td>-0.3</td>
<td>gauss</td>
<td>0.996</td>
<td>0.937</td>
<td>0.984</td>
</tr>
<tr>
<td>-0.2</td>
<td>gauss</td>
<td>0.999</td>
<td>0.964</td>
<td>0.984</td>
</tr>
<tr>
<td>-0.1</td>
<td>gauss</td>
<td>0.999</td>
<td>0.992</td>
<td>0.985</td>
</tr>
<tr>
<td>0.0</td>
<td>gauss</td>
<td>0.997</td>
<td>0.982</td>
<td>0.992</td>
</tr>
<tr>
<td>0.1</td>
<td>gauss</td>
<td>0.996</td>
<td>0.955</td>
<td>0.993</td>
</tr>
<tr>
<td>0.2</td>
<td>gauss</td>
<td>0.995</td>
<td>0.928</td>
<td>0.994</td>
</tr>
<tr>
<td>0.3</td>
<td>gauss</td>
<td>0.994</td>
<td>0.902</td>
<td>0.996</td>
</tr>
<tr>
<td>0.4</td>
<td>gauss</td>
<td>0.994</td>
<td>0.876</td>
<td>0.996</td>
</tr>
<tr>
<td>0.5</td>
<td>gauss</td>
<td>0.994</td>
<td>0.850</td>
<td>0.996</td>
</tr>
</tbody>
</table>

GLM with Gauss/Power -.1

Variance function: V(u) = 1  [Gaussian]
Link function: g(u) = u^{0.1}  [Power]
Log likelihood = 82.03257
AIC = -30813
BIC = -3055.145

| Cost | Coef   | Std Err | t     | P>|t| | 95%  | Cl      |
|------|--------|---------|-------|------|------|---------|
| Temp | -.0074317 | .0032673 | -.27 | .023 | -.0138355 | -.0010278 |
| Disser | .0287873 | .0150143 | 1.92 | .055 | -.0006401 | .0582147 |
| disco | 6.71e-06 | 2.26e-06 | 2.97 | .003 | 2.28e-06 | .0000111 |
| biqt | -.0359387 | .0111592 | -3.22 | .001 | -.0570046 | -.0140729 |
| cons | 1.064875 | .0110311 | 96.53 | .000 | 1.043254 | 1.086495 |

replace temp=0
predict pglmmpgq0
replace temp=1
predict pglmmpgq1
sum pglmmpgq*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>pglmmpgq0</td>
<td>500</td>
<td>.5732042</td>
<td>.0461443</td>
<td>.4478679</td>
<td>.6842101</td>
</tr>
<tr>
<td>pglmmpgq1</td>
<td>500</td>
<td>.6151152</td>
<td>.0498658</td>
<td>.4797741</td>
<td>.7351478</td>
</tr>
</tbody>
</table>
STATA Bootstrap

- We bootstrap these models primarily to estimate standard errors and p-values on the cost scale
- We use Stata’s most basic bootstrap command, bsample
  - A number of its other more powerful bootstrap commands build some parametrics into the analysis, whereas we are in part interested in nonparametrics
- We also need to create a dataset to save our results
  - We use the “append” command to add the results from a particular bootstrap sample to this dataset

Bootstrap Program: Creating bsmvpred

* drop _all is similar to clear, but maintains local variables, scalars, and matrices
* drop _all
  gen pglmig0=.
  gen pglmig1=.
  gen pglmig2=.
  gen pglmig3=.
  gen pglmig4=.
  gen pglmpg0=.
  gen pglmpg1=.
  gen pglmpg2=.
  save bsmvpred,replace

Bootstrap Program: Starting the Bootstrap

set more off
* If you want to be able to replicate your results, set seed
*set seed 2345

* Major loop: runs N times: forvalues i=1/N {
  forvalues i=1/1000 {
    * Displays a count every 50 iterations (to make sure something is happening)
    if (`i'/50)==round((`i'/50),1) {
      display `i'
    }
  }
  drop _all
  use rchapter5
  strata(treat): maintains sample size per group
  * cluster(id): if participants had multiple observations, sample all of them
  bsample,treat}
Bootstrap Program: Cost Estimation (1)

gen temp=treat
quietly glm cost temp dissev bl* race,link(identity) family(gauss)
quietly replace temp=0
quietly predict pglmigc0
quietly replace temp=1
quietly predict pglmigc1
quietly replace temp=treat
quietly glm cost temp dissev bl* race,link(identity) family(poisson)
quietly replace temp=0
quietly predict pglmipc0
quietly replace temp=1
quietly predict pglmipc1

Bootstrap Program: Cost Estimation (2)

quietly replace temp=treat
quietly glm cost temp dissev bl* race,link(log) family(gamma)
quietly replace temp=0
quietly predict pglmlgc0
quietly replace temp=1
quietly predict pglmlgc1
quietly replace temp=treat
quietly glm cost temp dis race blc blq,link(power .65) family(poisson)
quietly replace temp=0
quietly predict pglmppc0
quietly replace temp=1
quietly predict pglmppc1

Bootstrap Program: QALY Estimation

quietly replace temp=treat
quietly regress qaly temp dissev bl*
quietly replace temp=0
quietly predict pglmigq0
quietly replace temp=1
quietly predict pglmigq1
quietly replace temp=treat
quietly glm qaly temp dissev blcost blqaly, family(gauss) link(power -0.1)
quietly replace temp=0
quietly predict pglmpgq0
quietly replace temp=1
quietly predict pglmpgq1
Bootstrap Program: Estimate Treatment Group Mean Costs

```
sum pglmigc0, meanonly
local pglmigc0=r(mean)
sum pglmigc1, meanonly
local pglmigc1=r(mean)
sum pglmipc0, meanonly
local pglmipc0=r(mean)
sum pglmipc1, meanonly
local pglmipc1=r(mean)
sum pglmigq0, meanonly
local pglmigq0=r(mean)
sum pglmigq1, meanonly
local pglmigq1=r(mean)
sum pglmigc0, meanonly
local pglmigc0=r(mean)
sum pglmigc1, meanonly
local pglmigc1=r(mean)
sum pglmipc0, meanonly
local pglmipc0=r(mean)
sum pglmipc1, meanonly
local pglmipc1=r(mean)
sum pglmigq0, meanonly
local pglmigq0=r(mean)
sum pglmigq1, meanonly
local pglmigq1=r(mean)
```

Bootstrap Program: Estimate Treatment Group Mean QALYS

```
sum pglmigq0, meanonly
local pglmigq0=r(mean)
sum pglmigq1, meanonly
local pglmigq1=r(mean)
sum pglmigc0, meanonly
local pglmigc0=r(mean)
sum pglmigc1, meanonly
local pglmigc1=r(mean)
sum pglmigq0, meanonly
local pglmigq0=r(mean)
sum pglmigq1, meanonly
local pglmigq1=r(mean)
```

Bootstrap Program: Keep 1 Row of Data, Substitute Group Means, Append and Save

```
quietly keep if _n==1
quietly replace pglmigc0=`pglmigc0'
quietly replace pglmigc1=`pglmigc1'
quietly replace pglmipc0=`pglmipc0'
quietly replace pglmipc1=`pglmipc1'
quietly replace pglmlgc0=`pglmlgc0'
quietly replace pglmlgc1=`pglmlgc1'
quietly replace pglmppc0=`pglmsppc0'
quietly replace pglmppc1=`pglmsppc1'
quietly replace pglmigq0=`pglmigq0'
quietly replace pglmigq1=`pglmigq1'
quietly replace pglmigq0=`pglmigq0'
quietly replace pglmigq1=`pglmigq1'
quietly append using bsmvpred
quietly save bsmvpred, replace
```
Bootstrap Program: Use bsmvpred; Clean Up Empty Row; Calculate Mean Differences

```
drop _all
use bsmvpred
drop if pglmigc0==.
capture drop pglmigd
capture drop pglmipd
capture drop pglmlgd
capture drop pglmppd
capture drop pglmigqd
capture drop pglmpgd
gen pglmigd=pglmigc1-pglmigc0
ngen polsipd=pglmipc1-pglmipc0
ngen pglmigq=pglmigq1-pglmigq0
gen pglmpgd=pglmipq1-pglmipq0
save,replace
```

**Sum (3000 Draws)**

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>pglmig0</td>
<td>3000</td>
<td>3014.932</td>
<td>91.51177</td>
<td>2692.253</td>
<td>3388.113</td>
</tr>
<tr>
<td>pglmig1</td>
<td>3000</td>
<td>3037.588</td>
<td>72.58048</td>
<td>2728.398</td>
<td>3316.496</td>
</tr>
<tr>
<td>pglmip0</td>
<td>3000</td>
<td>2969.405</td>
<td>88.46574</td>
<td>2684.362</td>
<td>3324.534</td>
</tr>
<tr>
<td>pglmip1</td>
<td>3000</td>
<td>3083.115</td>
<td>89.30655</td>
<td>2781.27</td>
<td>3329.921</td>
</tr>
<tr>
<td>pglmlg0</td>
<td>3000</td>
<td>2961.086</td>
<td>74.90699</td>
<td>2786.943</td>
<td>3355.442</td>
</tr>
<tr>
<td>pglmlg1</td>
<td>3000</td>
<td>3099.429</td>
<td>74.90699</td>
<td>2786.943</td>
<td>3355.442</td>
</tr>
<tr>
<td>pglmpg0</td>
<td>3000</td>
<td>3070.295</td>
<td>89.80905</td>
<td>2682.656</td>
<td>3342.144</td>
</tr>
<tr>
<td>pglmpg1</td>
<td>3000</td>
<td>3070.9</td>
<td>72.17738</td>
<td>2765.581</td>
<td>3331.024</td>
</tr>
<tr>
<td>pglmig0</td>
<td>3000</td>
<td>.5732414</td>
<td>.0130528</td>
<td>.5237578</td>
<td>.6158496</td>
</tr>
<tr>
<td>pglmig1</td>
<td>3000</td>
<td>.6150669</td>
<td>.0128176</td>
<td>.5727083</td>
<td>.6582406</td>
</tr>
<tr>
<td>pglmpg0</td>
<td>3000</td>
<td>.5732115</td>
<td>.0130297</td>
<td>.5243673</td>
<td>.6158617</td>
</tr>
<tr>
<td>pglmpg1</td>
<td>3000</td>
<td>.6150899</td>
<td>.0127998</td>
<td>.5717807</td>
<td>.6579012</td>
</tr>
</tbody>
</table>

**Sum Differences**

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>pglmigd</td>
<td>3000</td>
<td>22.65688</td>
<td><strong>107.9732</strong></td>
<td>-407.6902</td>
<td>415.9185</td>
</tr>
<tr>
<td>polsipd</td>
<td>3000</td>
<td>113.7097</td>
<td><strong>101.4317</strong></td>
<td>-280.5315</td>
<td>499.9663</td>
</tr>
<tr>
<td>pglmipd</td>
<td>3000</td>
<td>138.3427</td>
<td><strong>107.8001</strong></td>
<td>-261.5327</td>
<td>568.0652</td>
</tr>
<tr>
<td>pglmppd</td>
<td>3000</td>
<td>89.30476</td>
<td><strong>104.0829</strong></td>
<td>-315.1501</td>
<td>479.8337</td>
</tr>
<tr>
<td>pglmigd</td>
<td>3000</td>
<td>.0418255</td>
<td><strong>.0181843</strong></td>
<td>-.0305826</td>
<td>.1084224</td>
</tr>
<tr>
<td>pglmpgd</td>
<td>3000</td>
<td>.0418844</td>
<td><strong>.0179765</strong></td>
<td>-.0277929</td>
<td>.1076626</td>
</tr>
</tbody>
</table>
Calculating Nonparametric P-Value

```
use bsmvpred
sum pglmigd
Variable | Obs Mean Std. Dev. Min Max
---------+----------------------------------------------
pglmigd | 3000 22.65688 107.9732 -407.6902 415.9185
local den=r(N)
sum pglmigd if pglmigd<0
Variable | Obs Mean Std. Dev. Min Max
---------+----------------------------------------------
pglmigd | 1257 -77.55559 62.92353 -407.6902 -.0107422
local num=r(N)
local p1=num/den
if p1>.5 { local p1=1-p1 }
local p2=2*p1
display p2
.838
```

Calculating Nonparametric CI

```
sort pglmigd
list pglmigd if _n==round((.025*3000),1)+1
| _n==round((.975*3000),1)
  +-----------+
  | pglmigd |
  +-----------+
  76. | -190.7117 |
  2925. | 232.584 |
  +-----------+
```

Calculating Parametric P-Value

```
Required data:
  point estimate for difference (OLS):  22
  DOF:  498
  Bootstrapped SE:
sum pglmigd
Variable | Obs Mean Std. Dev Min Max
---------+------------------------------------------
pglmigd | 3000 22.65688 107.9732 -407.6902 415.9185
local se=r(sd)
display 2*ttail(498,(22/se))
.83862872
```
Calculating Parametric CI

Required data:
point estimate for difference (GLS): 22
T-statistic for DOF = 498
Bootstrapped SE: 107.9732

. local tstat=invttail(498,.025)
. display `tstat'
1.964739
. display 22-(`tstat''*`se')
-190.13916
. display 22+(`tstat''*`se')
234.13916

Results (Principal Analysis and Bootstrapped SE)

<table>
<thead>
<tr>
<th>GLM</th>
<th>Link/fam</th>
<th>PE</th>
<th>BS</th>
<th>Nonpar 95% CI</th>
<th>Par 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GLM</td>
<td>SE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P-val</td>
<td></td>
<td>Nonpar</td>
<td>Par</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95% CI *</td>
<td>BS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P-val</td>
<td>BS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GLM</td>
<td>PE</td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
<td>BS</td>
<td>SE</td>
</tr>
<tr>
<td>id/gau</td>
<td>22</td>
<td>.838</td>
<td>108</td>
<td>.839</td>
<td>-191 to 233</td>
</tr>
<tr>
<td>id/pois</td>
<td>113</td>
<td>.000</td>
<td>101</td>
<td>.266</td>
<td>-91 to 308</td>
</tr>
<tr>
<td>log/gam</td>
<td>135</td>
<td>.210</td>
<td>108</td>
<td>.211</td>
<td>-78 to 346</td>
</tr>
<tr>
<td>pow/pois</td>
<td>88</td>
<td>.100</td>
<td>104</td>
<td>.398</td>
<td>-116 to 287</td>
</tr>
<tr>
<td>id/gau</td>
<td>.0417</td>
<td>.024</td>
<td>.100</td>
<td>.022</td>
<td>.0059 to .0780</td>
</tr>
<tr>
<td>pow/pois</td>
<td>.0419</td>
<td>.023</td>
<td>.179</td>
<td>.020</td>
<td>.0061 to .0775</td>
</tr>
</tbody>
</table>

* sort pglmigcd; list pglmigcd if _n==76|_n==2925
† PE ± (1.964739 * SE)

Summary

- We've provided you with programs to conduct univariate analysis (t-tests and bootstraps) as well as multivariate analysis of cost data using GLMs (including diagnostic tests to select the more appropriate family and link functions and bootstrapping to obtain confidence intervals and p-values for the cost estimates)
- No single model will always be the best in every situation of estimating cost differences associated with medical therapies
- 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