STATISTICAL HORIZONS

Analysis of Cost Data

Henry Glick, Ph.D.

Upcoming Seminar:

February 24-26, 2022, Remote Seminar

Multivariable Analysis of Patient-Level Treatment Cost

Henry A. Glick, Ph.D.

Cost Data 101

- Commonly right (positively) skewed (i.e., long, heavy, right tails)
- Tend to be skewed because:
  - Can have 0 costs, but not negative costs
  - Most severe cases may require substantially more services than less severe cases
  - Certain very expensive events occur in a relatively small number of patients
  - A minority of patients are responsible for a high proportion of health care costs
- Tends to complicate analysis

Multivariable Analysis Strategy

- Analysis of cost
  - Start with everyone's "old" favorite: OLS
  - Briefly review log OLS
  - Transform OLS into GLM and check fit of gauss family (with diagnostic)
    - Revise family if necessary
  - Start with everyone's "new" favorite: GLM gamma/log
  - Check fit of gamma family
    - Revise family if necessary
  - “Tune” link (with diagnostics)
- QALY Appendix

GLM Relax OLS and Log OLS Assumptions

- Ability to choose among different families relaxes Gauss family assumption of constant variance
  - Gauss: constant variance
  - Poisson: variance proportional to mean
  - Gamma: variance proportional to square of mean
  - Inverse gauss: variance proportional to cube of mean
- Ability to choose among different links relaxes assumption that:
  - $E(y/x) =$ (OLS)
  - $E(\ln(y)/x) =$ (log OLS)

Rerun OLS as GLM With Identity Link and Gauss Family

```
glm cost i.treat dissev blc blq race,
link(identity) family(gauss)
```

General syntax:

```
glm [depvar] [indepvars] [if xxx],link(xxx) family(xxx)
```

But is Gauss Right Family

- Modified Parks test is a “constructive” test that recommends a family given a particular link function
- Implemented after GLM regression that uses particular link
- Test predicts square of residuals (res^2) as a function of log of predictions (lnyhat) by use of a GLM with a log link and gamma family
  - Stata code

    glm res^2 lnyhat,link(log) family(gamma), robust
- **If weights or clustering are used in original GLM, same weights and clustering should be used for modified Park test**

Modified Parks Test of Family For Different Links

Link	Family	Coef	P-value
-0.7	Gamma	1.6777	0.24
-0.6	Gamma	1.6469	0.20
-0.5	Gamma	1.6175	0.17
.	.		
-0.1	Gamma	1.5150	0.09
0.0	P/G	1.5378	0.15
0.1	P/G	1.5163	0.13
0.2	Poisson	1.4954	0.12
.	.	.	.
1.4	Poisson	1.3039	0.38
1.5	Poisson	1.2997	0.39
1.6	Poisson	1.1528	0.63
1.7	--	--	--

- Power links of 0.0 and 0.1 demonstrate toss-ups (poisson/gamma)
- Recommended family may not run
  - 1.6 won't run for (recommended) poisson family, but will for gauss
- May be no recom-mended family
  - 1.7 won't run for any family

eeict2011r.dta

GLM Diagnostics, Identity/Gaussian

FITTED MODEL: Link = Identity ; Family = Gaussian

Results, Modified Park Test (for Family)

Coefficient: 1.391784

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

Family	Chi2	P-value
Poisson:	**1.4021**	**0.2364**
Gamma:	3.3790	0.0660
Gaussian NLLS:	17.6936	0.0000
Inverse Gaussian or Wald	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

cost, eeict2011.dta

```
glm cost i.treat dissev bl*race, link(identity)
    family(poisson) vce(bootstrap, reps(200)
        strata(treat) seed(1234))
```

Variance function: V(u) = u [Poisson]
Link function: g(u) = u [Identity]

Log likelihood = -113575.9606

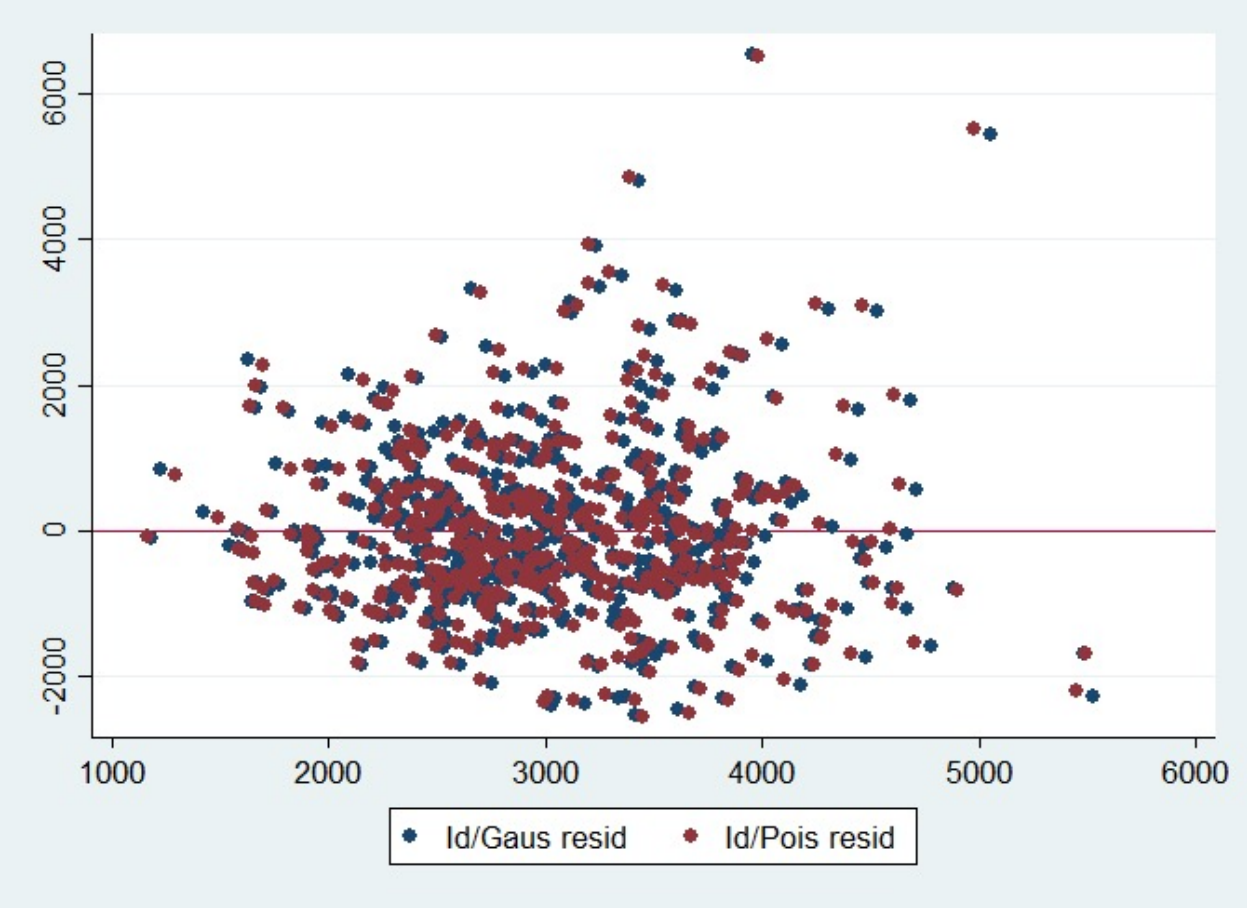
AIC 454.3278
BIC 219209.5

cost	Coef	BS St Err	z	P>|z|	95% CI	
1.treat	113.1149	103.0793	1.10	**0.272**	-88.91687	315.1466
dissev	4008.434	429.8734	9.32	**0.000**	3165.898	4850.97
blcost	.3861271	.0781542	4.94	**0.000**	.2329476	.5393066
blqaly	-765.3726	366.625	-2.09	**0.037**	-1483.944	-46.80076
race	-746.574	111.639	-6.69	**0.000**	-965.3823	-527.7657
_cons	1925.985	343.2156	5.61	**0.000**	1253.295	2598.676

cost, eeict2011.dta

Effect of Changing Families

- Residuals plotted against predicted costs for gauss and poisson families demonstrating heteroscedasticity



cost, eeict2011.dta

```
glm cost i.treat dissev blcost blqaly
    race, link(log) family(gamma)
```

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

Log likelihood = -4494.155729

AIC 18.00062
BIC -2988.518

cost	Coef	Std Err	z	P>|z|	95% CI	
1.treat	.0446782	.0356359	1.25	0.210	-.0251669	.1145232
dissev	1.409376	.1739606	8.10	0.000	1.06842	1.750333
blcost	.0001227	.0000257	4.78	0.000	.0000724	.0017300
blqaly	-.2579657	.1223431	-2.11	0.035	-.4977537	-.0183796
race	-.2613111	.0395492	-6.61	0.000	-.3388262	-.1837961
_cons	7.610573	.1220851	62.34	0.000	7.371291	7.849856

Interpretation of 0.0447 Coefficient

- As with log OLS, it is sometimes assumed that log/gamma GLM coefficients for dichotomous variables have a % difference interpretation
- Not exactly true, but unlike log OLS, whether or not variances are equal (homoscedasticity), transformation of coefficient has a percentage difference (in predicted costs) interpretation
- glm cost i.treat dissev blcost blqaly race, link(log) family(gamma)

```
------------------------------------------------------------------------------
        cost | Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.treat |   .0446782   .0356359     1.25   0.210    -.0251669    11.45232
```

Is Gamma Correct Family for Log Link?

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

Family	Chi2	P-value
Gamma:	**1.9560**	**0.1619**
Poisson:	4.0897	0.0431
Inverse Gaussian or Wald	23.2272	0.0000
Gaussian NLLS:	29.6281	0.0000

Results of tests of GLM Log link

Pearson Correlation Test:	.2460
Pregibon Link Test:	.1273
Modified Hosmer and Lemeshow:	.6199

- Certainly a reasonable family for log link

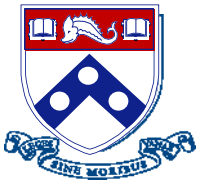
cost, eeict2011.dta

Is Log Best Link Available?

- So far evaluated identity link (with an “optimized” poisson family) and log link (with an “optimized” gamma family)
- While log link is most commonly used in literature, need not be the best fitting link
- What link should we use?

Selecting a Link Function

- Literature mixed on whether there's a single statistic that can be used to identify optimal link
  - Compare model performance of all permutations of candidate link and variance functions???
- Manning proposed selection based on at least 3 tests:
  - Pearson's correlation test evaluates systematic bias in fit on raw scale
  - Pregibon link test evaluates linearity of response on scale of estimation
  - Modified Hosmer and Lemeshow test evaluates systematic bias in fit on raw scale
- Ideally, all 3 tests yield nonsignificant p-values



Can We Improve Link?

- Iteratively evaluate power links (in 0.1 intervals) between -2 and 2
  - Use modified Park test to select a family
  - Rerun GLM with preferred power link / family
  - Evaluate fit statistics
  - Don't show you results here, but then fine tune power link in 0.01 intervals within candidate regions of power link

Power 0.65 Link / Poisson Family

```
glm cost i.treat dissev bl* race, link(power
.65) family(poisson) vce(bootstrap, reps(200)
      strata(treat) seed(1234) nodots)
```

Variance function: V(u) = u [Poisson]
Link function: g(u) = u^(.65) [Power]

Log likelihood = -113515.3

AIC = 454.0853
BIC = 219088.2

Cost	Coef	Std Err	z	P>|z|	95% CI	
i.treat	3.493932	4.188398	0.83	0.404	-4.715177	11.70304
dissev	161.4855	17.74034	9.10	0.000	126.715	196.2559
blcost	.0150344	.0030678	4.90	0.000	.0009215	.0210473
blqaly	-30.51369	14.51284	-2.04	0.042	-59.86632	-1.161064
race	-30.27001	4.51284	-6.71	0.000	-39.11501	-21.425
_cons	138.8326	13.95714	9.95	0.000	111.4771	166.1881

cost, eeict1.dta

Run GLM DIAGNOSTICS, .65/Poisson

FITTED MODEL: Link = Power .65 ; Family = Poisson

Results, Modified Park Test (for Family)

Coefficient: 1.495248

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

Family	Chi2	P-value
Poisson:	2.3212	0.1276
Gamma:	2.4111	0.1205
Gaussian NLLS:	21.1587	0.0000
Inverse Gaussian or Wald:	21.4285	0.0000

Results of tests of GLM Log link

Pearson Correlation Test:	**.9027**
Pregibon Link Test:	**.7469**
Modified Hosmer and Lemeshow:	**.5870**

cost, eeict1.dta

Summary Link Fit Measures

P-Value Based

AIC/BIC/Log Likelihood

Improbable Predictions

- In some datasets, some link/family combinations (including log/gamma) can yield improbable predictions
- Example below is from a bootstrap predicting group 1's hospital costs from a substance abuse clinical trial

Link	Family	$\hat{y}$	SE	Min $\hat{y}$	Max $\hat{y}$
Observed	NA	5103	1064	2081	10041
Identity	Gauss	4934	2185	-3880	15601
Log	Gamma	13,263	21301	1544	426,526
Fitted *	Fitted	5815	5153	-33	174,816

* Link and family for each draw determined using link and family tests

Group 1 cost: mean=5089; min=145; max=40246; skewness=2.11; kurtosis=6.68

heroin2.dta

What to Do If/When Model Has Bad Fit Statistics

- Fit statistics for links and families are data dependent
- If no link/family pair yields good fit, consider changing:
  - Variables included in model
    - Add or subtract as needed
  - How variables are specified
    - e.g., continuous vs quadratic vs log vs square root vs 2 categories vs N categories