

## Identifying an Appropriate Link and Family for Generalized Linear Models

ISPOR 20th Annual International Meeting

May 19, 2015

Henry Glick

[www.uphs.upenn.edu/dgimhsr](http://www.uphs.upenn.edu/dgimhsr)



---

---

---

---

---

---

---

---

### Generalized Linear Models (GLM)

- As Jalpa has indicated, to use generalized linear models, need to identify a link and a family
- OLS in GLM framework uses an identity link and a gauss family
- Log link/gamma family most commonly in literature
  - Log link: mean cost =  $\exp(\sum \beta_i X_i)$
  - Gamma family: variance increasing in magnitude as a function of the square of the mean
- No reason to believe that universal use of log/gamma combination is substantially better than universal use of any particular link/family combination



---

---

---

---

---

---

---

---

### Extended Estimating Equations

- One approach for identifying appropriate links and families is Basu and Rathouz's (2005) extended estimating equations (EEE) (implemented in Stata)
  - EEE estimates link function and family along with coefficients and standard errors
- Strongly recommend implementing EEE with your data; however:
  - Tends to need a large number of observations (thousands not hundreds) to converge
  - Can't identify a link and family with EEE and use the resulting link and family with a simple GLM command
- Our recommendations apply when can't use EEE or EEE won't converge



---

---

---

---

---

---

---

---

### Outline

- Introduce primary dataset used in examples
  - 2 other datasets used to make specific points
- Methods for identifying optimal family
  - Modified Park test
- Methods for identifying link function
  - Pregibon link test
  - Pearson correlation test
  - Modified Hosmer and Lemeshow test
  - AIC, BIC, Log likelihood
  - Informal summary measures (don't think we'll get to)



---

---

---

---

---

---

---

---

### Data Set

Explanatory variables	Rx0 (N=250)	Rx1 (N=250)	P-value
dissev	0.349 (0.112)	0.346 (0.113)	0.73
blcost	1630 (773)	1639 (770)	0.90
blqaly	0.784 (0.140)	0.787 (0.151)	0.85
race	0.516 (0.5)	0.496 (0.5)	0.72
Outcome			
cost1	3015 (1583)	3233 (1169)	--

Cost1: min=315; max=10692; skewness=1.13; kurtosis=6.27



---

---

---

---

---

---

---

---

### Family for GLM

- Specifies distribution that reflects mean-variance relationship
  - Gaussian: Constant variance (OLS/Log OLS)
  - Poisson: Variance proportional to mean
  - Gamma: Variance proportional to square of mean
  - Inverse Gaussian or Wald: Variance proportional to cube of mean
- Use of latter 3 families relaxes assumption of homoscedasticity



---

---

---

---

---

---

---

---

### Implications of Heteroscedasticity for OLS

- Coefficients remain linearly unbiased, but...
  - No longer have minimum variance
  - Resulting variance estimate is biased
    - “Only in some special cases...can it be determined whether the usual estimator...is biased upwards or downwards.”

Kennedy, A Guide to Econometrics



---

---

---

---

---

---

---

---

### Modified Park Test

- “Constructive” test that recommends a family conditional on a particular link function
- Implemented after GLM regression that uses particular link
- Test predicts square of residuals as a function of log of predictions by use of a GLM with a log link and gamma family



---

---

---

---

---

---

---

---

### Implementing Modified Park Test

- Run glm with a link you are interested in (e.g., identity) using some family
  - No rule about initial family used in MP test
    - Gauss or gamma probably least tempermental
- Predict  $\hat{y}$  and residuals
- Calculate log of  $\hat{y}$  ( $\ln\hat{y}$ ) and square of residuals ( $\text{res}^2$ )
- Estimate:  
`glm res2 lnhat,link(log) family(gamma) robust`
- If using weights, clustering, or “if” statement in original GLM, use same weights, clustering, and “if” statement for modified Park test



---

---

---

---

---

---

---

---

### Recommended Family, Modified Park Test

- Recommended family derived from coefficient for lnyhat:
  - If coefficient  $\sim 0$ , Gaussian
  - If coefficient  $\sim 1$ , Poisson
  - If coefficient  $\sim 2$ , Gamma
  - If coefficient  $\sim 3$ , Inverse Gaussian or Wald



---

---

---

---

---

---

---

---

### glm res2 lnyhat, link(log) family(gamma) robust

res2	Coef	Std Err	z	P> z	[95% Conf Int]
lnyhat	<b>1.3459</b>	0.3354	4.01	0.000	0.6886 to 2.0032
_cons	3.3234	366.11	1.25	0.212	-1.8960 to 8.5428

test lnyhat = 0  
chi<sup>2</sup> (1) = 16.11; p = 0.0001  
test lnyhat==1  
chi<sup>2</sup> (1) = 1.06; **p = 0.30**  
test lnyhat==2  
chi<sup>2</sup> (1) = 3.80; p=0.05



eeicd2011r.dta

---

---

---

---

---

---

---

---

### Issues

- Coefficients <0
  - If coefficient  $\leq -0.5$ , consider subtracting all observations from maximum-valued observation and rerunning analysis
    - Works sometimes, but not always
- Coefficient > 3.5
  - Continue to use inverse Gaussian for larger coefficients ???



---

---

---

---

---

---

---

---

### SEs for Poisson Family

- When using poisson family, for both Stata and SAS, SEs for coefficients can be improbably small
  - E.g., 0.0000 for all variables
- In Stata, correct by use of variance-covariance matrix of the estimators (VCE) option:
 

```
glm depvar indepvars, link[xxx] family(poisson)
vce(bootstrap, [strata(treat)] reps(200) nodots)
```




---

---

---

---

---

---

---

---

---

---

### Modified Park Test, 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
- Recommended family may not run
  - 1.6 won't run for (recommended) poisson family, but will for gauss
- May be no recommended family
  - 1.7 won't run for any family



eeicd2011r.dta

---

---

---

---

---

---

---

---

---

---

### Link Function

- Link function directly characterizes how linear combination of predictors is related to prediction on original scale
- While log link is most commonly used in literature, need not be best fitting link
- SAS and Stata power links allow generation of wide variety of named and unnamed links, e.g.,
  - power 1 = Identity link  $\hat{y} = \beta_i X_i$
  - power .5 = Square root link  $\hat{y} = (\beta_i X_i)^2$
  - power 0 = log link  $\hat{y} = \exp(\beta_i X_i)$
  - power -1 = reciprocal link  $\hat{y} = 1/(\beta_i X_i)$




---

---

---

---

---

---

---

---

---

---

### Selecting a Link

- Literature is mixed on whether there is a single statistic that can be used to identify an optimal link
- Manning et al. proposed selection should be based on a combination of at least 3 tests: Pregibon link test, Pearson's correlation test, and modified Hosmer and Lemshow test
- Hardin and Hilbe have suggested use of links with (smaller) AIC or BIC statistics or links with (larger) log likelihood statistics
- In what follows, discuss Manning's suggestion, but return to AIC and BIC



---

---

---

---

---

---

---

---

### Link Tests

- Pregibon link test evaluates linearity of response on scale of estimation
  - e.g., if log or square root link is used, evaluates response on log and square root of cost scales, not cost scale
- Pearson's correlation test evaluates presence of systematic bias in fit on raw scale
  - e.g., on cost scale
- Modified Hosmer–Lemeshow test also evaluates systematic bias in fit on raw scale (write for details about implementation)



---

---

---

---

---

---

---

---

### Implementing Pregibon Link Test

- Run glm with a link and family
- Predict  $(\sum_i \beta_i X_i)$  and  $(\sum_i \beta_i X_i)^2$  on scale of estimation
- Estimate:  
`glm depvar  $(\sum_i \beta_i X_i)$   $(\sum_i \beta_i X_i)^2$ ,link([xxx]) family[xxx] robust`  
(family[xxx] and link[xxx] represent link and family used in initial glm)
- P-value on coefficient for  $(\sum_i \beta_i X_i)^2$  <0.05 indicates lack of linearity
- If using weights, clustering, or "if" statement in original GLM, use same weights, clustering, and "if" statement for modified Park test



---

---

---

---

---


---

---

---

glm cost1 xb xb2, link(log) family(gamma) robust

res2	Coef	Std Err	z	P> z	[95% Conf Int]
xb	9.9140	3.9930	2.48	0.013	2.088 to 17.740
xb2	-0.5546	0.2476	-2.24	<b>0.025</b>	-1.040 to -0.069
_cons	-35.787	16.0917	-2.22	0.026	-67.326 to -4.248

eeid2011r.dta 

---

---

---

---

---

---

---


---

---

---

Implementing Pearson Correlation Test

- Run glm with a link and family
- Predict cost ( $\hat{y}$ ) and cost residuals (res)
- Estimate:
 
$$\text{corr } \hat{y} \text{ res}$$
- In stata: `pwcorr  $\hat{y}$  res,sig`
- P-value for correlation <0.05 indicates lack of fit
- If using weights, clustering, or "if" statement in original GLM, use same weights, clustering, and "if" statement for modified Park test



---

---

---

---

---

---

---

---


---

---

pwcorr pcost res,sig

	pcost	res
pcost	1.0000	
res	-0.0665	1.0000

**0.1378**

eeid2011r.dta 

---

---

---

---

---

---

---

---

---

---

### Diagnosing a Link

Link	Pears	Pregib	mHM
0.4	.6842	.1422	.6426
0.5	.7091	.2040	.6434
0.6	.7399	.2850	.4615
0.7	.7772	.3872	.701
0.8	.8213	.5111	.8777
0.9	.8729	.6556	.5906
1.0	.9323	.8168	.7636
1.1	<b>.9999</b>	<b>.9885</b>	<b>.9193</b>
1.2	.9239	.8375	.9298
1.3	.8391	.6703	<b>.9725</b>
1.4	.7455	.5186	.785
1.5	.6433	.3888	.7608

- No link is least significant for all 3 tests (i.e., dominant)
- 1.1 link dominates all links except 1.2 and 1.3 links




---

---

---

---

---

---

---

---

---

---

### AIC, BIC, Log Likelihood

Link	AIC	BIC	LL
0.4	445.449	214752	-111356
0.5	444.854	214455	-111208
0.6	444.354	214205	-111083
0.7	443.951	214004	-110982
0.8	443.648	213852	-110906
0.9	443.445	213751	-110855
1.0	<b>443.348</b>	<b>213702</b>	<b>-110831</b>
1.1	443.359	213707	-110834
1.2	443.481	213769	-110864
1.3	443.721	213889	-110924
1.4	444.085	214070	-111015
1.5	444.581	214318	-111139

- AIC, BIC, LL yield a similar, but not identical solution

#### Issues

- Unstable across recommended families
- AIC and BIC don't always agree




---

---

---

---

---

---

---

---

---

---

### AIC, BIC, Log Likelihood Unstable Across Links

Link	Family	AIC	BIC	LL
-0.7	Gamma	18.0677	-2990.45	-4510.78
-0.6	Gamma	18.0666	-2990.71	-4510.65
-0.5	Gamma	18.0661	-2990.96	-4510.53
.	.	.	.	.
-0.1	Gamma	18.0645	-2991.78	-4510.12
0.0	P/G	448.760	216408	-111942
0.1	P/G	447.793	215924	-111724
0.2	Poisson	446.92	215487	-111723
.	.	.	.	.
0.9	Poisson	443.45	213751	-110855
1.0	<b>Poisson</b>	<b>443.35</b>	<b>213702</b>	<b>-110830</b>
1.1	Poisson	443.36	213707	-110833



eeicd2011r.dta

---

---

---

---

---

---

---

---

---

---



### Continuous Families

- EEE already uses continuous families
- Once this feature becomes part of glm software, we won't be able to distinguish changes in log likelihood, AIC, BIC, and deviance statistics that were:
  - Due to better fit OR
  - Due to changes in family




---

---

---

---

---

---

---

---

### AIC, BIC, and Log Likelihood Needn't Agree

Link	LL	AIC	BIC
-.64	-9	-9	-9
-.63	-5931.4072	17.299584	<b>-3378.90</b>
-.62	-5931.2616	17.299160	-3378.66
-.61	-5931.1228	17.298756	-3378.41
-.6	-5930.9913	17.298373	-3378.17
-.59	-5930.8676	17.298013	-3377.92
.	.	.	.
-.5	-5930.2011	17.296073	-3375.7923
-.49	-5930.1868	17.296031	-3375.5701
-.48	<b>-5930.1864</b>	<b>17.296030</b>	-3375.3522
-.47	-5930.2004	17.296071	-3375.1392
-.46	-5930.2291	17.296155	-3374.9312
-.45	-5930.2731	17.296283	-3374.7287




---

---

---

---

---

---

---

---

### Zeros

- Since log of 0 is undefined, in log OLS either:
  - Use 2-part model (prediction of any cost (yes/no) followed by prediction of log cost among those with any cost), or
  - Add arbitrary small quantity to all observations
  - Exclude observations of 0 from analysis
- No problem including observations of 0 when using glm with log link
  - However, presence of large fractions of zeros can make it hard to identify stable link/family combinations
    - As with OLS, two-part models can avoid problems posed by large fractions of zeros




---

---

---

---

---


---

---

---

```
glm cost1 treat dissev bl* race,link(power 1.1)
family(poisson) vce(bootstrap,strata(treat)
reps(200))
```

Cost1	Coef	Std Err	z	P> z	[95% Conf Int]
treat	759	243	3.13	0.002	284 to 1235
dissev	9842	1017	9.68	0.000	7849 to 11,835
blcost	.973	.224	4.35	0.000	0.534 to 1.411
blqaly	-1804	996	-1.81	0.070	-3756 to 148
race	-1812	274	6.61	0.000	-2349 to -1274
_cons	3950	936	4.22	0.000	2156 to 5784

eeid2011r.dta 

---

---

---

---

---

---

---


---

---

---

### How Not to Calculate Between-Group Predicted Cost Differences

- For all multivariable models, **INAPPROPRIATE** to calculate predicted between-group differences in means by:
  - Running regression (OLS, GLM, Logit, etc.)
  - Making prediction for each observation
  - Calculating mean of predictions for groups 0 and 1
  - Subtracting group 0's mean of predictions from group 1's
- Reintroduces between group differences in covariates that were controlled for in multivariable model



---

---

---

---

---

---

---


---

---

---

### Calculating Between-Group Predicted Cost Differences

- For OLS, can use sample means for covariates and treatment group indicator to estimate adjusted mean for each group
  - NOTE: sample means are same for each treatment group
- For multiplicative models (e.g., log, power 1.1, logit), **CAN'T** use this approach
  - Mean of retransformations  $\neq$  retransformation of mean



---

---

---

---

---

---

---

---

---

---

### Recycled Predictions

- Should instead use method of recycled predictions to create an identical covariate structure for each group by:
    - Generating a temporary 0/1 variable that equals the treatment status variable and including it in model
    - After running model, assigning 0s to temporary variable for all observations independent of actual treatment status
    - Predicting  $pcost_0$ , predicted cost had everyone been in treatment group 0
    - Assigning 1s to temporary variable for all observations independent of actual treatment status
    - Predicting  $pcost_1$ , predicted cost had everyone been in treatment group 1
- Stata "margins" syntax: `margins r.ib(last).treat`




---

---

---

---

---

---

---

---

### Results From Various Model Specifications

Link	Family	$\Delta C$	SE	P-value
T-test	--	218	124	0.08
Identity	Gauss	215	108	0.046
Identity	Poisson	304	103	0.003
Log	Gamma	337	109	0.002
power 1.1	Poisson	310	101	0.002



eeicd2011r.dta

---

---

---

---

---

---

---

---

### Link Fit Statistics

Link	Pregibon	Pearson	M-H&L
Identity/Gauss	0.702	1.00 *	0.375
Identity/Poisson	0.817	0.932	0.764
Log/Gamma	0.025	0.138	0.416
power 1.1/Poisson	0.989	0.999	0.919

\* For identity/gauss, Pearson statistic=1.0 by definition



eeicd2011r.dta

---

---

---

---

---

---

---

---

### Links Used in Bootstrap Draws

Link	%
-1.4 to -0.1	4.1
0 to 0.7	20.9
0.8	5.83
0.9	6.83
1.0	7.66
1.1	8.85
1.2	11.76
1.3	9.55
1.4	8.29
1.5	5.91
1.6+	10.33

eeid2011r.dta



---

---

---

---

---

---

---

---

---

---

### Summary

- Log/gamma not always preferred link/family
- Need to conduct diagnostic tests to identify appropriate link/family
- Establish criteria for choice of preferred link/family prior to unblinding data
  - Fact that one model gives a more favorable result should not be a reason for its adoption
- Report sensitivity of results to different link/family specifications



---

---

---

---

---

---

---

---

---

---