Comparison of Log-Linear and Generalized Linear modeling Frameworks

A study of direct demand models for non-motorized modes (19-03044)

Aditya Medury[1], Robert Schneider[2], Julia Griswold[1] and Offer Grembek[1]
[1]UC Berkeley Safe Transportation Research & Education Center
[2]University of Wisconsin-Milwaukee

Log-linear Models: Estimating *E*(log(y)) vs *E*(y)

Pedestrian and bicycle activity at intersections and segments is routinely modeled as direct demand models in traffic safety literature. These models estimate the expected number of trips at a given location for a given time period as a function of socio-economic and built environment characteristics. As the dependent variables are typically positive, many studies utilize log-linear regression models, which use a logarithmic transformation to ensure that estimates derived from the model are positive once back-transformed. While commonly used, this back-transformation approach does not estimate the mean of the depedent variable.

Let *y* be the dependent variable (e.g., annual pedestrian volumes, peak hour bicycle counts), and the **X** be the explanatory variables (e.g., population, employment, network density, slope). If *y* > 0, the log-linear regression framework, solved using ordinary least squarses (OLSLT), is applicable as follows:

$$log(y|\mathbf{X}) = \mathbf{X}\beta + \epsilon; \epsilon \sim N(0, \sigma^2(\mathbf{X}))$$

The back-transformed estimate that is typically used as the model output is:

$$\mathbf{E}(log(y|\mathbf{X})) = \mathbf{X}\beta \Rightarrow \hat{y}_{OLSLT} = exp(\mathbf{X}\beta)$$

However the mean estimate of *y* for the log-linear model is given by:

$$\mu_{OLSLT} = \mathbf{E}(y|\mathbf{X}) = \mathbf{E}(exp(\mathbf{X}\beta + \epsilon)|\mathbf{X})$$
$$= exp(\mathbf{X}\beta)\ exp(0.5\sigma^2(\mathbf{X}))$$

Thus, the backtransformed estimate of a log-linear model, which represents the median of *y*, is biased. However, under heteroscecadisticty, even the mean OLSLT estimate can be biased, since the functional form of the variance is unknown.

Generalized Linear Model Alternatives

For generalized linear models (GLM), it is possible to use a log link to estimate **E**(*y*|**X**) as follows:

$$\mathbf{E}(y|\mathbf{X}) = exp(\mathbf{X}\beta)$$

However, for a given link function, different assumptions on the mean-variance relationship lead to different types of GLM specifications:

GLM Type	Variance Specification
Gaussian (μ_{GLMGLL})	σ^2 (constant)
Negative Binomial (μ_{GLMNB})	$\mu_{GLMNB} + \alpha\mu^2_{GLMNB}$
Quasi-Poisson (μ_{GLMQP})	$\theta\mu_{GLMQP}$

Evaluation Criteria

We use both simulated and empirical datasets to evaluate log-linear and other GLM specifications using the following criteria:

- Stability of coefficients: variation in sign/magnitude across data samples
- Overfitting: compare root mean squared errors (RMSE) for training and test data

Simulated Datasets

Deterministic Component

$\mathbf{X} = [1\ x_1\ x_2]$, where:
$x_1 \sim unif(0,1)$
$x_2 \sim bernoulli(0.4)$ $(x_2 \in \{0,1\})$
$\beta = [10\ 1\ 1]$

Error Component

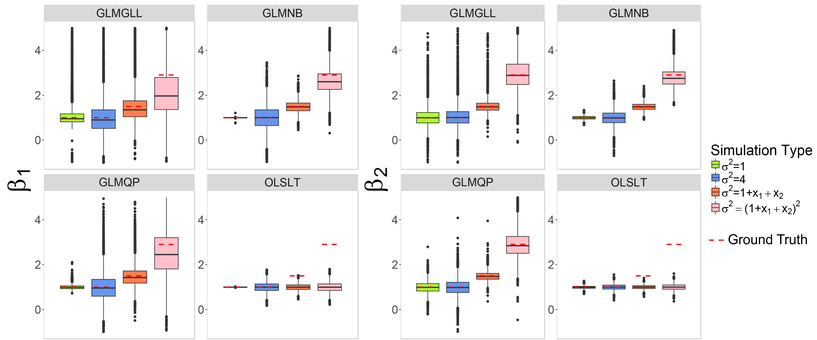
ϵ is homoscedastic (i.e. $\sigma^2(\mathbf{X}) = \sigma^2$):
$\sigma = \{1,2\}$
ϵ is heteroscedastic (i.e. $\sigma(\mathbf{X}) = \sigma f(\mathbf{X})$):
$\sigma^2 = 1; f(\mathbf{X}) = (1 + x_1 + x_2)^{0.5}$
$\sigma^2 = 1; f(\mathbf{X}) = 1 + x_1 + x_2$

Training/test data: 10,000 observations; number of simulations: 10,000 runs

Stability of Coefficients Coefficients

Ground truth for GLM coefficients was estimated as follows:

$$\beta_{GLM,i} = \frac{\partial log(\mathbf{E}(y|\mathbf{X}))}{\partial x_i} = \beta_i + 0.5\frac{\partial(f(\mathbf{X})^2)}{\partial x_i}$$



Once the constant variance assumption is violated, the coefficients under OLSLT cannot be estimated accurately.

Training vs Test Data (Under Heteroscedasticity)

Percentage of simulations when "Method A" yields a lower/equal RMSE than "Method B"

$\sigma^2(\mathbf{X}) = (1 + x_1 + x_2)$

	Training Data					Test Data				
	Method B					Method B				
Method A	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB
OLSLT (Median)	100%	0%	0%	0%	0%	100%	0%	7%	5%	3%
OLSLT (Homoscedastic Mean)	100%	100%	0%	0%	7%	100%	100%	31%	28%	27%
GLMGLL	100%	100%	100%	100%	100%	93%	69%	100%	33%	36%
GLMQP	100%	100%	0%	100%	96%	95%	72%	67%	100%	39%
GLMNB	100%	93%	0%	4%	100%	97%	73%	64%	61%	100%

$\sigma^2(\mathbf{X}) = (1 + x_1 + x_2)^2$

Method A	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB
OLSLT (Median)	100%	0%	1%	0%	1%	100%	0%	28%	25%	16%
OLSLT (Homoscedastic Mean)	100%	100%	1%	0%	5%	100%	100%	35%	32%	24%
GLMGLL	99%	99%	100%	99%	94%	72%	65%	100%	39%	34%
GLMQP	100%	100%	1%	100%	85%	75%	68%	61%	100%	37%
GLMNB	99%	95%	6%	15%	100%	84%	76%	66%	63%	100%

- Under heteroscedasticity, OLSLT is inferior to other GLM alternative.
- GLMGLL is most prone to overfitting: outperforms in training (over test) data.

Empirical Dataset

Our empirical dataset is comprised of annual pedestrian counts for 1,268 intersections in California. The Breusch-Pagan Test for heteroscedasticity yielded a p-value < 0.001.

Stability of Coefficients

1000 simulations run with 20% of data used as test data.

Coefficient	OLSLT			GLMGLL			GLMQP			GLMNB		
	Mean	1st Percentile	99th Percentile	Mean	1st Percentile	99th Percentile	Mean	1st Percentile	99th Percentile	Mean	1st Percentile	99th Percentile
(Intercept)	5.66	5.19	6.12	2.21	-1.45	8.13	6.01	4.84	7.69	6.93	6.28	7.43
logEmpQw	0.37	0.34	0.41	0.49	0.35	0.68	0.44	0.39	0.50	0.32	0.28	0.39
PopH_20k	1.42E-04	1.33E-04	1.51E-04	9.42E-05	5.53E-05	1.45E-04	1.04E-04	9.00E-05	1.18E-04	1.19E-04	1.07E-04	1.32E-04
logStSegHw	0.30	0.20	0.39	0.93	-0.04	1.47	0.34	0.06	0.55	0.27	0.16	0.41
WalkComPctH	2.87	2.28	3.49	2.21	0.82	3.85	2.67	2.03	3.31	3.03	2.13	3.73
logSchoolsH	0.05	0.03	0.06	-0.09	-0.13	0.04	-0.03	-0.05	0.03	0.04	0.02	0.05
PrincArt	0.43	0.36	0.50	0.31	0.01	0.70	0.28	0.14	0.42	0.39	0.31	0.47
MinorArt	0.36	0.30	0.44	0.22	-0.09	0.63	0.17	0.03	0.32	0.36	0.27	0.45
Int4way	0.38	0.29	0.46	0.02	-0.32	0.57	0.25	0.11	0.47	0.28	0.17	0.39
Signal	0.21	0.13	0.30	-0.03	-0.24	0.23	0.11	0.01	0.22	0.26	0.16	0.35

GLMGLL shows counterintuitive signs for multiple variables; GLMQP in one case.

Training vs Test Data

RMSE Comparisons for All Intersections (1268 observations)
Breusch-Pagan Test for Heteroskedasticity (p-value < 0.001)

	Training Data					Test Data				
	Method B					Method B				
Method A	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB
OLSLT (Median)	100%	100%	0%	0%	12%	100%	93%	49%	33%	40%
OLSLT (Homoscedastic Mean)	0%	100%	0%	0%	0%	7%	100%	7%	3%	5%
GLMGLL	100%	100%	100%	100%	100%	51%	93%	100%	36%	51%
GLMQP	100%	100%	0%	100%	100%	67%	97%	64%	100%	66%
GLMNB	88%	100%	0%	0%	100%	60%	95%	49%	34%	100%

RMSE Comparisons for Intersections with PopH < 5,000 (~ 51% of original data)
Breusch-Pagan Test for Heteroskedasticity (p-value = 0.02)

Method A	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB	OLSLT (Median)	OLSLT (Homoscedastic Mean)	GLMGLL	GLMQP	GLMNB
OLSLT (Median)	100%	0%	3%	0%	2%	100%	26%	63%	17%	19%
OLSLT (Homoscedastic Mean)	100%	100%	3%	0%	100%	74%	100%	64%	8%	66%
GLMGLL	97%	97%	100%	97%	97%	37%	36%	100%	30%	35%
GLMQP	100%	100%	3%	100%	100%	83%	92%	70%	100%	84%
GLMNB	98%	0%	3%	0%	100%	81%	34%	65%	16%	100%

Homoscedastic mean for OLSLT performs the worst among all GLM options for entire dataset, but is superior to all but GLMQP when applied to a smaller subset of intersections with less population (and lower concern of heteroscedasticity)

Discussion

- Tests for heteroscedasticity are essential when considering log-linear models.
- GLM models, in particular GLMQP and GLMNB, are more robust alternatives, especially when the dependent variable can be zero (e.g., bicycle volumes).
- Trade-off between coefficient stability and predictive quality of the final model depends on the desired use of the direct demand model.
- Random parameter specifications will be considered in future research.

This research was funded by the California Department of Transportation.