Honors for Spring of 2023

Sam Lorenz

Conclusions of the Project

Outline of the Presentation

We will walk through the entirety of the project. Specifically:

- Recap on important information
- Application of data belonging and not belonging to the exponential family:
  - Recap of the example in the paper
- Application of using data from a two-parameter distribution
  - Gamma example from the text
  - Real data example using Professor Solomon's data
- Ending conclusions

Recall the Modeling Types

The main goals of this project were to:

- Discover similarities and differences between the CLM, GLM, and VGLM
- Learn the requirements of each type of modeling
- Applications of these modeling types
- Understand how and when to use each type of model, as well as the why

Information on CLM

A CLM model has the form:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

In order to use the CLM scheme fit by Ordinary Least Squares, there are four requirements that must be met:

1. The residuals follow a Normal distribution ($\varepsilon \sim \mathcal{N}(\mu, \sigma)$)
2. The residuals are homoskedastic ($\mathbb{V}\left[\varepsilon\right] = \sigma^2 < \infty$)
3. The expected value of the residuals is a constant zero ($\mathbb{E}\left[\varepsilon\right] = 0$)
4. The residuals are independent, $\mathbb{C}ov\left[\varepsilon_i, \varepsilon_j\right] = 0$ (unless $i = j$)

What happens when these requirements are violated?

Information on the GLM

A GLM has a form similar to the CLM:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

The generalization introduces a few factors; these are clarifications of things assumed, but never examined, in classical linear models. These are:

- The linear predictor
- The conditional distribution
- The link function

VGLM Information

The differences between a VGLM and GLM are slight but powerful. The form of a VGLM is written as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

The vector generalized model still requires the linear predictor, the conditional distribution, and the link function.

- Now, we can model more than one linear predictor as well as distributions outside of the exponential family.

When do we use each type of model? What are the differences?

Example from Paper

We will investigate the three models using an example of an election where there is a possibility for differential invalidation.

- What is differential invalidation?
- Graphically, what does it look like?
- How do we test for it?

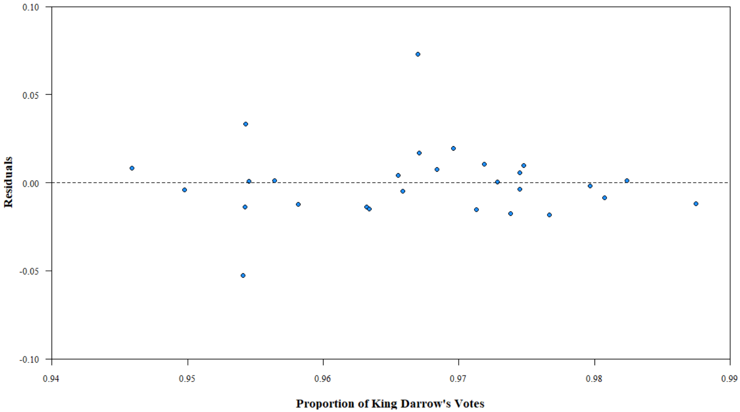
Requirements First: Constant Expected Value

The numeric test is the Wald–Wolfowitz runs test, created by Abraham Wald and Jacob Wolfowitz.

- What are the null and alternative hypothesis?
- The p-value is 0.7001, therefore, we have sufficient evidence that the expected value of the residuals is constant and zero and that the residuals are independent.

Requirements: Constant Expected Value

Let us look at a graphical example.



Requirements: Constant Variance

We can use the same graphic for the constant expected value to check for homoskedasticity. Let us go into the numeric test, which is the Breusch–Pagan test, created by Trevor Breusch and Adrian Pagan.

- What are the null and alternative hypothesis?
- The p-value is 0.4501, therefore, we have sufficient evidence that the variance of the residuals is constant.

Requirements: Normality

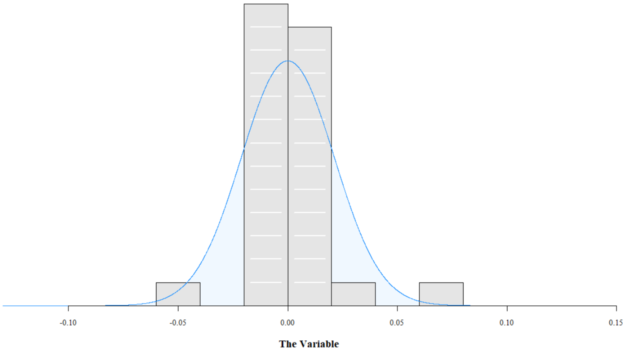
Next, we can do a numeric test using the Shapiro-Wilk test.

- What are the null and alternative hypotheses?
- The p-value for this is 0.0030, therefore, we do not have sufficient evidence that our residuals are Normally distributed.
- Note on Central Limit Theorem

Requirements: Normality

Let us graphically test the normality assumption.

- Graphical tests



Transformed OLS Model

Let's think about our data and how it behaves, specifically, the boundedness of the dependent variable.

- A logit transform is appropriate.

Now, we need to re-check our assumptions. Let's save some time and just look at the results of the required tests.

Test	p-value
Normality	0.2514
Constant Expected Value	0.4411
Constant Variance	0.9368

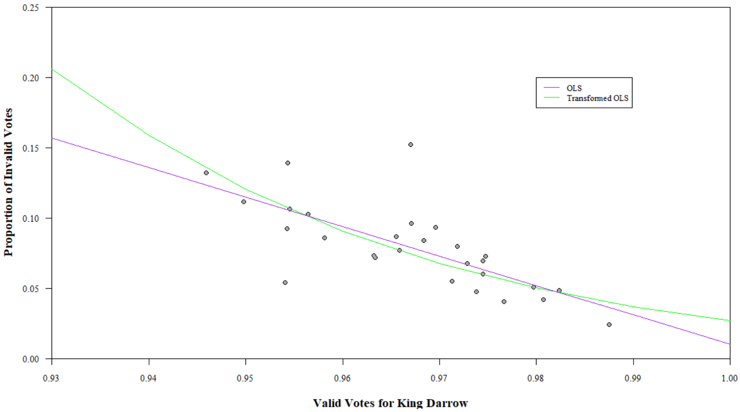
The Actual Regression

Now that we have our model, we are ready to see if we have evidence of differential invalidation. Below is the regression table:

	Estimate	p-value
Intercept	28.629	1.130×10^{-5}
King Darrow Support	-31.848	3.200×10^{-6}

Thus, there is evidence of differential invalidation.

Graphic of the OLS Models



GLM Requirements

Let's try modeling with the GLM. First, we need to specify a few things: the distribution as well as the link function.

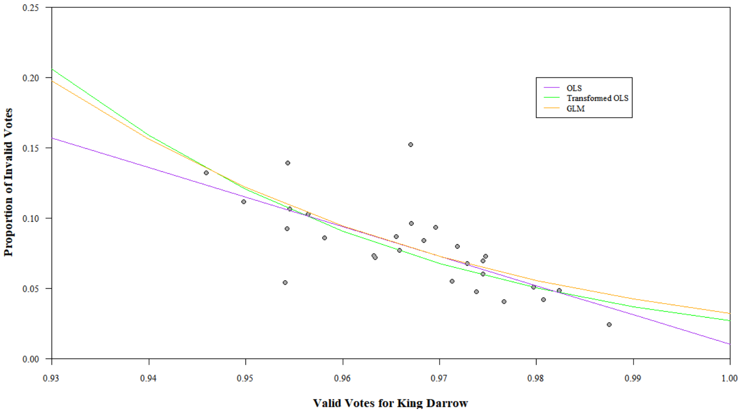
- What do we know about the dependent variable?
- What is the link function?

Now, using this information, we can make our model. The output is:

	Estimate	P-value
Intercept	25.226	2.00×10^{-16}
King Darrow Support	-28.634	2.00×10^{-16}

We do have evidence of differential invalidation.

GLM Graphic


0.25
0.20
0.15
0.10
0.05
0.00
Proportion of Invalid Votes
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1.00
Valid Votes for King Darrow
OLS
Transformed OLS
GLM

VGLM Specifications

Why would we need to use a VGLM here?

- Overdispersion

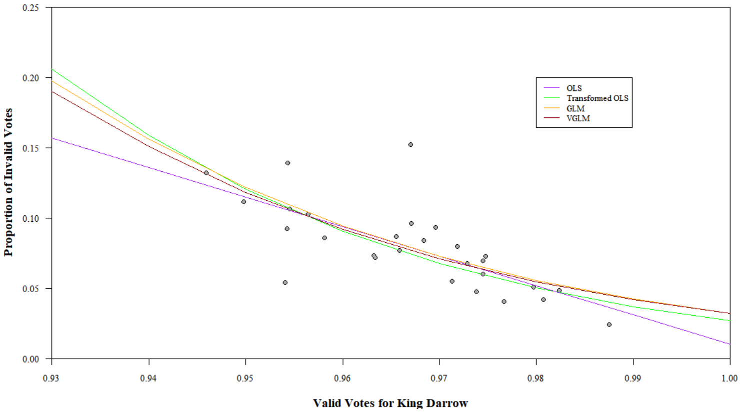
We still need to define the distribution as well as the link function.

- Distribution: Beta-Binomial (cannot use GLM since this is outside the exponential family).
- Link function: Logit

Our output for the model is then:

	Estimate	P-value
Intercept	24.654	4.300×10^{-7}
King Darrow Support (effect)	-28.059	2.830×10^{-8}

VGLM Graphic



Discussion

We just saw how to use our three models for testing an election for differential invalidation.

- OLS was nice, but took time.
- GLM worked better, but we have issues with overdispersion.
- a VGLM was pretty much perfect... other than the model itself taking a lot of time to run.
- The first big difference between the GLM and VGLM is the VGLM can model distributions that are outside the exponential family.

When do we use these models?

Gamma Distributed Data

The Gamma Distribution is a two-parameter distribution in the exponential family.

- It is frequently used to model "time until" something happens.
- There are options for the parameters, but most of Yee's packages use κ as the shape and θ as the scale.
- The Gamma distribution also can model the time between events

How the Models Fit Gamma-Distributed Data

CLM with OLS and the GLM:

- Hold one parameter constant and makes predictions using the other
- Uses an additive effect rather than fully makes the distinction of κ and θ
- May not make good predictions

With the VGLM:

- Remember, the VGLM can model more than one linear predictor
- We can specify what variables are for κ and which are for θ using contraint matrices.

Actual Values from Data Generating Process

Note that the true values are:

	Kappa	Theta
Kappa Intercept	0.693	0.000
Theta Intercept	0.000	1.099
Witch Level	0.000	-0.051
Soldier Level	0.000	-0.073
Number of Witches	-0.010	0.000
Number of Soldiers	-0.030	0.000

Thus, the true formulas are

$$\kappa = 0.693 - 0.010 \text{ Number of Witches} - 0.030 \text{ Number of Soldiers}$$

$$\theta = 1.099 - 0.051 \text{ Witch Level} - 0.073 \text{ Soldier Level}$$

OLS and Gamma

When the model is done in `R`, we get coefficients of:

Intercept	3.17×10^{-7}
Witch Level	0.229
Soldier Level	5.864
Number of Witches	2.267
Number of Soldiers	1.252

We cannot separate these into κ and θ due to how the OLS model fits the Gamma distribution. Thus, the prediction formula is estimated to be

$$\begin{aligned} y = 3.17 \times 10^{-7} &+ 2.267 \text{ Number of Witches} + 1.252 \text{ Number of Soldiers} \\ &+ 0.229 \text{ Witch Level} + 5.864 \text{ Soldier Level} \end{aligned}$$

GLM and Gamma

When the model is done in `R`, we get coefficients of:

Intercept	-0.396
Witch Level	8.547
Soldier Level	6.225
Number of Witches	17.241
Number of Soldiers	19.231

We cannot separate these into κ and θ due to how the GLM model fits the Gamma distribution. Thus, the prediction formula is estimated to be

$$
\begin{aligned}
y = -0.396 &+ 17.241 \text{ Number of Witches} + 19.231 \text{ Number of Soldiers} \\
&+ 8.547 \text{ Witch Level} + 6.225 \text{ Soldier Level}
\end{aligned}
$$

VGLM and Gamma

We will use constraint matrices to specify our prior knowledge that some variables belong to κ and some belong to θ.

- The distribution will be "`GammaR`," which is just a specific parameterization of the Gamma distribution. The parameters are θ for scale and κ for shape. The pdf is

$$f(x;\ \kappa, \theta) = \frac{1}{\theta^{\kappa}\,\Gamma(\kappa)}\ x^{\kappa-1}\ \mathrm{e}^{-x/\theta}$$

- This will slightly change the output since θ is first, so keep this in mind for constructing the constraint matrices.
- Furthermore, since $\theta = 1/\text{rate}$, then $\log(\theta)$ is the same as -$\log(\text{rate})$. This will be important for transformations.

What are Constraint Matrices and How to Specify Them?

This is a pretty hefty slide, so I am going to use the board.

- All constraint matrices do is help constrain the independent variables so they are specified for certain parameters or dependent variables.
- In this case, we are specifying η_1 as θ and η_2 as κ, and both are predicting the dependent variable, time.

VGLM and Gamma

When the constrained model is done in `R`, we get coefficients of:

	Kappa	Theta
Kappa Intercept	0.817	0.000
Theta Intercept	0.000	1.052
Witch Level	0.000	-0.046
Soldier Level	0.000	-0.070
Number of Witches	-0.015	0.000
Number of Soldiers	-0.032	0.000

Thus, the two prediction formulas are estimated to be

$$\kappa = 0.817 - 0.015 \text{ Number of Witches} - 0.032 \text{ Number of Soldiers}$$
$$\theta = 1.052 - 0.046 \text{ Witch Level} - 0.070 \text{ Soldier Level}$$

VGLM and Gamma

Comparing the estimates with the true values:

	True		Estimated	
	Kappa	Theta	Kappa	Theta
Kappa Intercept	0.693	0.000	0.817	0.000
Theta Intercept	0.000	1.099	0.000	1.052
Witch Level	0.000	-0.051	0.000	-0.046
Soldier Level	0.000	-0.073	0.000	-0.070
Number of Witches	-0.010	0.000	-0.015	0.000
Number of Soldiers	-0.030	0.000	-0.032	0.000

Conclusions of Gamma Example

- When we want to show how both parameters in a distribution are being estimated and how they affect the variable, VGLM is the best.
- A lot of information can be lost since the CLM or GLM holds a parameter constant.
- Using two linear predictors is the second big difference between the GLM and VGLM.
- VGLM can model parameters of interest.

Overview of the Data

These data were collected by Professor Solomon (thanks!). Variables included are:

- Student identifier
- Timing of the test
- Number of words read correctly per minute
- Teacher
- Grade level

In this example, the dependent variable is words read correctly per minute and the research variable will be the time of the test.

Setting up the VGLM

We are going to use a distribution family called "`Gamma2`." The difference is that it gives the expected value estimates ($\mu = \kappa\theta$) as well as shape estimates (θ). Its pdf is:

$$f(x;\ \mu, \theta) = \frac{\exp\left[-\frac{\theta\, y}{\mu}\right] \left(\frac{\theta\, x}{\mu}\right)^{\theta - 1} \theta}{\mu\, \Gamma(\theta)}$$

- Let us do the constraint matrices on the board

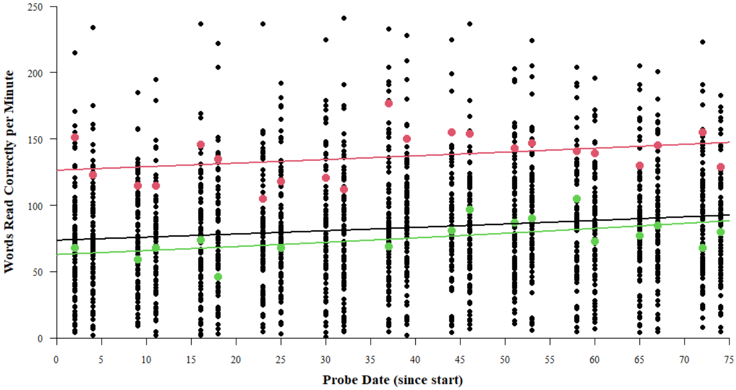
Output of Regression

The results of this model are

	Mean (μ)	Theta (θ)
Kappa Intercept	4.297	0.000
Theta Intercept	0.000	1.040
Probe Timing	0.003	0.000

Graphic

This is a graphic of all observations, with Student 1 and Student 2 singled out:



Discussion

This could not be done by GLM or CLM using OLS.

- The constraint matrices are very versatile.
- In real life, many data will be better modeled using two separate parameters.
- The linear predictors are practically unlimited as compared to that of the GLM or CLM.

The Entire Project in One Slide

This year, I accomplished:

- Learning the requirements and quirks of the CLM, GLM, and VGLM
- Understanding similarities and differences of each type of modeling
- When to use which type and why

Ideas for further research are:

- Types of VGLMs, such as Reduced-Rank VGLMs
- Generalized Additive Models and Vector-Generalized Additive Models
- More real-life applications of these models.