A Study of Generalized and Vector Generalized Linear Models

Sam Lorenz

January 11, 2023

Abstract

In many real-life applications, modeling using the classical linear modeling paradigm (CLM) will not suffice due to violations of its requirements, namely that the response variable has a conditionally Normal distribution. When this requirement is not met, one usually opts for using a generalized linear model (GLM). GLMs require that the response variable has a conditional distribution in the exponential family of distributions. Should this requirement also be violated, we can perform an extension of the GLM— a vector generalized linear model (VGLM).

In this paper, we will discuss some of the key similarities and differences between these modeling paradigms by comparing the models in R with a data set and studying the mathematical foundation of CLM, GLM, and VGLM.

Introduction

This shall be written at the end. (As will the final abstract.)

1 All Things Statistical Modeling

In this section, we will discuss three types of regression modeling: Classical Linear Models, Generalized Linear Models, and Vector Generalized Linear Models. I will discuss how they differ from one another, the requirements of each type of modeling, how they are used, and how the parameters are estimated.

1.1 The Classical Linear Model

I love to read fantasy novels. Specifically, I like young authors' fantasy the best, because it is an easy plot to get lost in without much romance, violence, and other things found in adult fantasy books. I especially like to read fantasy books where a kingdom is built, an army is trained, a lot of magic is used, and the ending is happy. Let's pretend I am Queen in one of these fantasy novels. I am concerned with building an army. Let's say I have information about how many enemies the witches can take down and information about how many enemies my Normal soldiers can take down. I want to know if I should try to find more witches, or regular soldiers to optimize the number of enemies I take down. How should I do this?

One solution is to use the classical linear modeling paradigm. A classical linear model has the form

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

Here, the Y_i are my enemies (dependent variable), and x_1 is the force of the

witches (independent variable) and the x_2 is the force of the regular soldiers (another independent variable). The ε_i are just the residuals, or errors, in my model. So, in order to see how my army would fare, I run two linear models and evaluate the relationship between my number of enemies and the force of the witches and then the force of the regular soldiers. This being said, whichever slope (the x) is closer to zero (or more negative) corresponds to a decrease in the number of enemies, and therefore, the group I would choose to select to fight in my army.

The classical linear model, especially as fit using ordinary least squares (OLS), is a fantastic way to test relationships between variables. Although it is simple and powerful — and well-established — several requirements must be met in order to properly use OLS estimation:

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

All of these requirements can be combined into one distributional statement as

$$\mathbf{Y} \mid \mathbf{XB} \sim \mathcal{N}\Big(\mathbf{XB};\ \sigma^2\mathbf{I}\Big) \tag{2}$$

Violations of these requirements are easy to find. Many can be mitigated using a transformation (or two) of the dependent variable. Such transformations include the logarithmic function and the logit function. However, transformations may not be sufficient to fix all of the violations, and they always increase the level of complexity of the model. In the case that the requirements for OLS are violated, there are other types of modeling we can use in order to analyze relationships.

1.2 Generalized Linear Models (GLMs)

Let's pretend I test my CLM model for violations of the requirements and find that the residuals do not follow a Normal distribution, or there is non-linearity in my model. Maybe the forces of soldiers and witches correspond to an exponential rise in the number of enemies taken down. Instead of trying to fix this by doing several transformations and failing, I could instead use a *generalized* linear model (GLM). A GLM can be used to model different distributions in the exponential family; our OLS requirements could be violated because the distribution is binomial rather than normal. The GLM can model a linear relationship between the dependent and independent variables regardless of linearity since it uses a special link function.[Woo06] The form of a generalized linear model is very similar to the form of a classical linear model. It is still

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{3}$$

The major difference is that the conditional distribution of $\mathbf{Y}$ is no longer restricted to just the Normal distribution as in (2). Any member of the exponential class of families will work (see Section 1.2.4). With that being said, this generalization introduces a few complications. Technically, these are not complications; they are clarifications of things assumed, but never examined, in classical linear models.

1.2.1 The Linear Predictor

The first thing is the linear predictor. The linear predictor, $\eta = \mathbf{XB}$, can be written in scalar form as

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{4}$$

This is identical to the linear predictor in classical linear models. We are

still trying to estimate the values of the individual β_j using the independent variables x_i.

1.2.2 The Conditional Distribution

The second item to specify is the conditional distribution of the dependent variable. In CLMs, this was the Normal distribution (and *only* the Normal distribution). In GLMs, this can be *any* member of the exponential family of distributions (see Section 1.2.4).

Examples of important distributions in the exponential class include the Normal, the Poisson, the Binomial, and the Bernoulli distribution. All four are quite useful in modeling different types of dependent variables.

1.2.3 The Link Function

Each distribution has its own mean (expected value). The third, and final, "new" item is the function that links the linear predictor, η with the mean of the conditional distribution, μ. Since the linear predictor is unbounded and the means of the conditional distributions *may be* bounded. The link function effectively unbounds the expected value function.[Yee15]

In the case of the classical linear model, the conditional distribution was the Normal distribution. Its mean is unbounded. Thus, the typical link function used was the "identity" function. That is, $\mu_i = \eta_i$. In other words, generalized linear models contain the CLM and extend it.

Returning to my example above, my residuals do not follow a Normal distribution. Instead, they follow an exponential. The mean of an exponential distribution is bounded below by 0 (and unbounded above). Thus, a link function can transform the μ so it is no longer bounded and now I can make predictions about who I want in my army.

Specifically, Nelder and Wedderburn[NW72, Woo06] defined $g(\cdot)$ to be

a link function if

$$g(\mu_i) = \eta_i \tag{5}$$

Here, μ_i is the expected value of the distribution, conditioned on the values of the independent variables, and η_i are the values calculated from the linear predictor.

1.2.4 The Exponential Family

A requirement of generalized linear models is that the conditional distribution is a member of the exponential family of distributions. The exponential family refers to a set of distributions in which the probability function *can be* written in the form

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{6}$$

Here, a, b, and c are arbitrary functions, ϕ is the 'scale' parameter, and θ is the canonical parameter.[Woo06] Before investigating the meaning of these five elements, let us look at an example of a probability mass function written in the form in (6)— the Poisson.

$$f(y) := \frac{\lambda^y \ \mathrm{e}^{-\lambda}}{y!} \tag{7}$$

$$= \exp\left[\log\left[\ \frac{\lambda^y \ \mathrm{e}^{-\lambda}}{y!}\ \right]\ \right] \tag{8}$$

$$= \exp\left[\log \lambda^y \ + \log \mathrm{e}^{-\lambda} - \log y!\ \right] \tag{9}$$

$$= \exp\left[\ y\ \ \log(\lambda) - \lambda - \log(y!)\ \right] \tag{10}$$

$$= \exp\left[\ \frac{y\ \ \log(\lambda) - \lambda}{1} - \log(y!)\ \right] \tag{11}$$

And so, if we define

- $\theta = \log(\lambda)$, that is $\lambda = e^{\theta}$,
- $b(\theta) = \lambda$, that is $b(\theta) = e^{\theta}$,
- $a(\phi) = 1$, and
- $c(y, \phi) = -\log(y!)$,

then we see that equation (7) can be written in the form of (6); thus, I have shown that the Poisson is a member of the exponential class of distributions.

Note that we can now interpret the five elements. The scale parameter $a(\phi)$ refers to the level of dispersion of the distribution. In this case, that level is fixed at 1. It can be shown that

$$\mathbb{V}[Y_i \mid \mathbf{XB}] = a(\phi)\mathbb{V}[\mu] \tag{12}$$

Thus, the dispersion is the ratio of the variances. If the data are conditionally Poisson, then the ratio of the variances will be 1. If they are not, then the ratio of variances will be greater than 1 (overdispersion) or less than 1 (underdispersion). Cases of over- and underdispersion frequently arise when modeling aggregated count data.

The $b(\theta)$ function is used to calculate the expected value and variance of the distribution. It can be shown that $b'(\theta) = \mu$ and that $b''(\theta) = \mathbb{V}[\mu]$. This fact is central to the simplifications in the maximum likelihood estimation method typically used for GLMs (see Section 1.3.1).

The canonical link function is $\theta = \log(\lambda)$. While canonical links have some useful statistical properties, there is no real reason to limit ourselves to using them. Any function that is able to transform the bounded μ function to an unbounded η function will work.

The $c(y, \phi)$ function has no meaning beyond ensuring the function is a probability function; that is, it sums (or integrates) to 1 over its domain (sample space).

The GLM can only model using a distribution in the exponential family. So, if the witch and regular soldier data followed an exponential distribution (which *is* a member of the exponential family), then I could use a GLM to make predictions.

From the point of a practitioner, the CLM model really is not too different than a GLM. We could still use CLM to predict if I want witches or regular soldiers in my army, but this would be a tedious amount of work. We already know we would transform the dependent variable, check our assumptions again (probably realize our residuals are overdispersed), then back-transform the variable again to do the analysis. With a GLM, we just run the model with the appropriate distribution and link function, back transform for our predictions and plots, and be good to go. It is faster, cleaner, and easier to do.

1.2.5 Generalized Additive Models

An interesting extension of a GLM is a generalized *additive* model (GAM). This differs from the GLM in a couple of ways. Let's first look at the form:

$$g(\mu(x_i)) = \eta_i = \beta_1 + f_2(x_{i2}) + \cdots + f_p(x_{ip}) \tag{13}$$

Notice how this is different from the GLM. We are now requiring an additive effect between the covariates.

More importantly, however, GAMs use **smoothing functions** to transform the independent variables. There are a multitude of smoothing functions that are used for different purposes, it is up to the researcher using GAMs to choose which one best fits their goal. This allows a researcher to fit data that

is not a nice, smooth curve; a GAM can fit data with multiple local minima and maxima, so if we have some wavy data, we can apply a smoothing function and still make a model to estimate beta. For example, let's say my data of witches and how many enemies they take down grows, then hits a peak, then falls (they get tired after about 50 enemies, or so). A GLM could still fit this, but it may be better to try a GAM to smooth out that peak so I can model it better.

1.3 Estimation Methods

Perhaps more interesting than the GLM and the extensions themselves is how the model is actually estimated. How is it possible we type a few simple commands and we have a model ready at our disposal? It is an interesting, but complex, algorithm.

This section discusses two types of estimation methods used to fit GLMs and VGLMs. Note there are other ways to do so, but these are two of the most used and easiest in terms of understanding.

1.3.1 Maximum Likelihood Estimation

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.[LL22]

This idea is a little complex, so let's think about an example. Let's say I have won the war and now my kingdom is safe and happy. As Queen of a magical kingdom, I have a lot of information on old battles and when a mystic, ancient foe will decide it wants my kingdom. Suppose I have ten data points on when previous battles were, each point representing the time there was before that battle. I want to predict when the next battle will be so I can

ensure I have the proper armada in place.

First, I need to decide what model I think best represents my data. Which distribution do my data follow best? Well, we are concerned with the time between events. I want to best predict when the next war will happen, so let my data follow an exponential distribution. We are aiming to estimate λ; in other words, we are trying to find a curve where λ best fits our data so we can get an accurate prediction of when the next war will be. If we have a graph of multiple curves with different rates, we are trying to find the curve that was most likely responsible for creating the data I already have. Maximum likelihood estimation is a method that will find the value of λ that will result in the curve that best fits the data.

Example. The above was what maximum likelihood estimation is, and what the goal of the algorithm is. How do we actually calculate the parameter values, or the **maximum likelihood estimates** (MLEs)? So, using the example above, how am I actually going to find the MLE of λ? Well, we need to calculate the total probability of observing all the data, so the joint probability distribution of the observed points.[LL22] This can be very difficult, considering we will need to use the conditional probabilities. Let us make our first assumption that each data point is generated independently of the others. Using this, the total probability of observing all of the data is just the product of observing each individual point (the marginal probabilities).

So, now all we have to do is find the joint probability of observing my data. This is done by calculating the product of the individual probabilities, also known as the likelihood:

$$\mathcal{L}(\lambda; \mathbf{x}) = \prod_{i=1}^{n} \lambda \, e^{-\lambda x_i} \tag{14}$$

Notice that this is explicit that the likelihood is a function of the parameter

of interest, λ, given the data.

Calculus is the typical method for calculating the value of λ that produces the maximum likelihood. However, it is easier to take the derivative of a sum instead of a product. Thus, one maximizes the *logarithm* of the likelihood function:

$$\begin{aligned} l(\lambda; \mathbf{x}) &= \log\Big(\mathcal{L}(\lambda; \mathbf{x})\Big) \\ &= \log\left(\prod_{i=1}^{n} \lambda\ e^{-\lambda x_i}\right) \\ &= \sum_{i=1}^{n} \log\left(\lambda\ e^{-\lambda x_i}\right) \\ &= \sum_{i=1}^{n} \left(\log\lambda\ -\lambda\ x_i\right) \end{aligned}$$

In this case, solving for the parameter, λ, is a straightforward application of differential calculus.[LL22]

$$\begin{aligned} \frac{d}{d\lambda}\sum_{i=1}^{n} \left(\log\lambda\ -\lambda\ x_i\right) &= \sum_{i=1}^{n} \frac{d}{d\lambda}\left(\log\lambda\ -\lambda\ x_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{1}{\lambda} - x_i\right) \\ &= \left(\frac{n}{\lambda} - n\overline{x}\right) \end{aligned}$$

Setting this to zero and solving for the estimator gives us

$$\begin{aligned} 0 &\overset{\text{set}}{=} \frac{n}{\hat{\lambda}} - n\overline{x} \\ n\overline{x} &= \frac{n}{\hat{\lambda}} \end{aligned}$$

$$\hat{\lambda} = \frac{1}{\overline{x}}$$

In other words, the maximum likelihood estimator of λ is $\frac{1}{\overline{x}}$.

» I will add more. «

1.3.2 Iteratively Reweighted Least Squares (IRLS)

Cries in IRLS

1.4 Vector Generalized Linear Models (VGLMs)

(KNOW)

Things VGLM wise, also about differences between these and GLMs, and IRLS for VGLM.

1.4.1 Vector Generalized Additive Models (VGAMs)

(KNOW)

Things VGAM wise

2 Putting Models to Use in R

Now that we know what models are available, what their requirements are, and how they differ, we are ready to see a data example. Let us use our fantasy example, but this time, apply it to electoral forensics.

I am trying to rule over more than one kingdom. Instead of going to battle and putting people in danger, I am going to run against King Darrow. He is known as a brutal King, ruling over Vetraheim (the Winter realm) with an iron fist. He has been the King since the election ten years ago. Many have run against him, but somehow, King Darrow always comes out on top. So, I began my campaign, built alliances, and became popular quite fast. It was going smoothly, until I found out King Darrow doubled his military and was punishing those that were going against him. I had to back off a little, out of concern for the people.

When it was election time, to my surprise, King Darrow won by over 90% of the votes counted. I was floored. I went to my council, and they believed King Darrow was guilty of electoral fraud, specifically differential invalidation. This means the more votes he received, the lower his invalidation rate. My council worked on collecting the number of ballots cast for King Darrow and me, the number of valid votes, and the number of invalid votes. Our independent variable here is the proportion of valid votes King Darrow had and our dependent variable is the proportion of invalid votes, since we want to see how his invalidation rate is affected by the more valid votes he has. We are going to look at the three models we discussed to see if there is evidence for differential invalidation against King Darrow.

2.1 The CLM Model

The first step is to check our four assumptions of the residuals. So, we make a model for the dependent versus the independent variable, then find the residuals for that model and apply visual as well as numerical tests to the residuals.

2.1.1 The Model

I am using R for the modeling. After loading the data, I then created my independent and dependent variables. Our independent variable is the number of invalid votes over the total number of votes, since we are trying to determine if there is a relationship between the invalidation rate and King Darrow's votes. In R, this is written like so:

```
pInv <- Invalid / (Valid + Invalid)
```

Our dependent variable is the proportion of votes for King Darrow over the total number of votes since we are trying to see changes based on King Darrow's votes and the invalidation rate. This is a similar notation in R:

```
pDarrow <- King_Darrow_Votes / Valid
```

Now, we are ready to make our linear model, get our residuals, and begin our assumption tests. Below is the model, and how to access the residuals:

```
modA <- lm(pInv ~ pDarrow)

eA <- residuals(modA)
```

Now, we can perform our assumption tests.

2.1.2 Constant Expected Value

The first requirement tested is the constant expected value and the independence of the residuals.

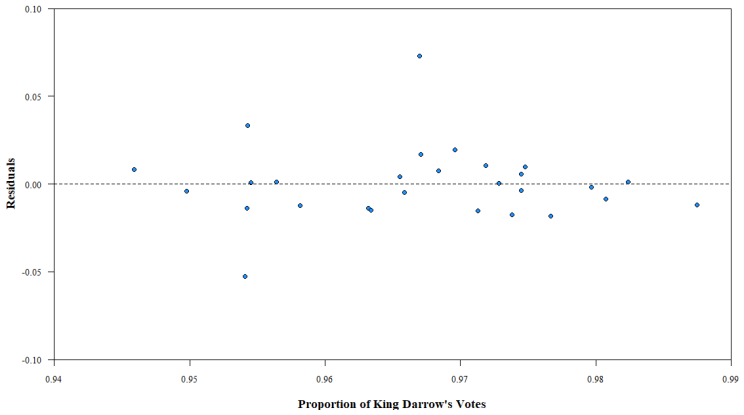

Figure 1: *A residuals plot of the CLM election model.*

The numeric test for seeing if there is a constant expected value is the Wald–Wolfowitz runs test, created by Abraham Wald and Jacob Wolfowitz. This test aims to check elements are mutually independent. The null hypothesis is that the expected value of the residuals is constant and zero. When we do the runs test, our p-value is 0.700, meaning we have sufficient evidence that the expected value of the residuals is constant and zero and that the residuals are independent.

Had the null hypothesis been rejected, we would need to examine the residual plot to determine if the violation is significant enough to affect the estimates. Figure 1 shows the relationship between the residuals (vertical axis) and the independent variable (horizontal axis).

There isn't a pattern here; it slopes down a little, but if there was an obvious quadratic shape, it would be a bit more worrisome.

2.1.3 Constant Variance (Homoskedasticity)

The next requirement we will test for is constant variance (also known as homoscedasticity). The typical numeric test for homoskedasticity is the Breusch–Pagan test, created by Trevor Breusch and Adrian Pagan in 1979. It tests whether or not the residuals of a model is dependent on the independent

variable.[BP79] The null hypothesis is that the residuals do not depend on the independent variable, and are therefore homoskedastic.

When performing the Breusch-Pagan test on the model, we get a p-value of 0.450, meaning we cannot reject the null hypothesis that the residuals have constant variance. The model passes this test.

Had we rejected the null hypothesis of homoskedasticity, and found a violation of one of the OLS requirements, we would examine the residuals plot to determine if the violation is significant enough to affect our estimates of the standard errors. This allows one to see if there are any patterns, which means the data are heteroskedastic. These patterns are pretty obvious; the shape of the graph will have a funnel-shape, trumpet-shape, or balloon-shape, where the data essentially puffs out along the y-axis in the center.

Looking at Figure 1, one sees no significant pattern. There are a few points that dip above where $y = 0$, but things fall into place as we would expect. It would be pretty clear and obvious if there was a pattern.

2.1.4 Normality of Residuals

The final requirement that needs to be checked is the Normality of the residuals. The typical numeric test for normality is called the Shapiro-Wilk test, named after Samuel Sanford Shapiro and Martin Wilk in 1965. [For20, SW65] The null hypothesis is that the variable is Normally distributed, that the residuals are generated from a Normal distribution. If the p-value is below 0.05, then we reject the null hypothesis and conclude the variables are *not* from a normal distribution.

For this model, the Shapiro-Wilk test indicated that the residuals were not generated from a Normal distribution (p-value = 0.003). Thus, we have a violation of the requirements of OLS.

To determine if the violation is severe enough to significantly affect the

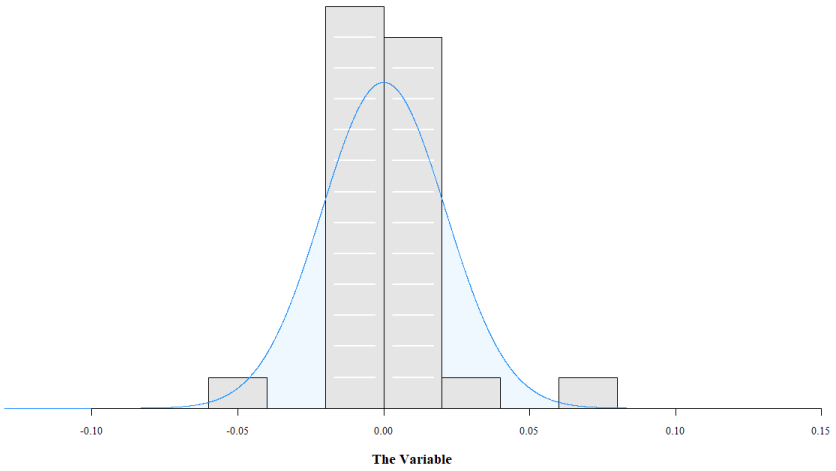

Figure 2: *A histogram of the residuals from the model, with a Normal distribution provided to illustrate the deviation from Normality.*

estimated confidence intervals, one should examine a histogram with an overlay to show what the Normally distributed residuals should look like. Figure 2 is a histogram with the overlay of the residuals.

The assumptions for the residuals seem to preform well, although there are some violated assumptions. It is possible for some residuals to be outliers due to the distribution of the data.

2.1.5 A Second Model

Overall, the OLS model without any transformations seems to perform well; the only issue we had was with the normality of the residuals. When the sample size is large, perhaps around $n = 100$, then we can use the Central Limit Theorem and assume that our residuals are normal.

In this case, the sample size is $n = 28$, since there are 28 kingdoms in the land I am trying to rule. So, let us apply a transformation to our independent variable and see about fixing the Normality violation.

To select an appropriate transform, we must consider the boundedness of the dependent variable. The invalidation rate, being a proportion, is

bounded between 0 and 1. Thus, a logit transform is appropriate.

Let's see what happens with normality. Let us perform another Shapiro-Wilk test on this transformed model and see if our violations are fixed. We get a p-value of 0.251, therefore, we have sufficient evidence our residuals are now normal. Let's check our other requirements. Since we have done this above, below is a table of the other requirements.

Test	p-value
Constant Expected Value	0.4411
Constant Variance	0.9368

It looks like all of our requirements are met, and there are no violations. We can now use the logit-transformed model to preform our testing. With this new model, let us finally see if King Darrow is guilty of differential invalidation. Let us look at a summary of the model:

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

Since we used a logit transform, the slope should be interpreted as the logarithm of the increase in the *odds* of a ballot being invalidated. In general, we can interpret the effect as for every 1 increase in the independent variable, the odds of the dependent variable increase by a factor of $\mathbb{e}^{\beta}$. So, with our model, that would be an odds change to 1.47×10^{-14} of its previous value.

More meaningfully, if we increase the support for King Darrow in the kingdom by just 10%, the odds of a ballot being invalidated decreases by 96% ($= 1 - \mathbb{e}^{-31.848 \times 0.10} = 1 - 0.04$). This is a rather large effect. Figure 3 illustrates this effect (icky green curve).

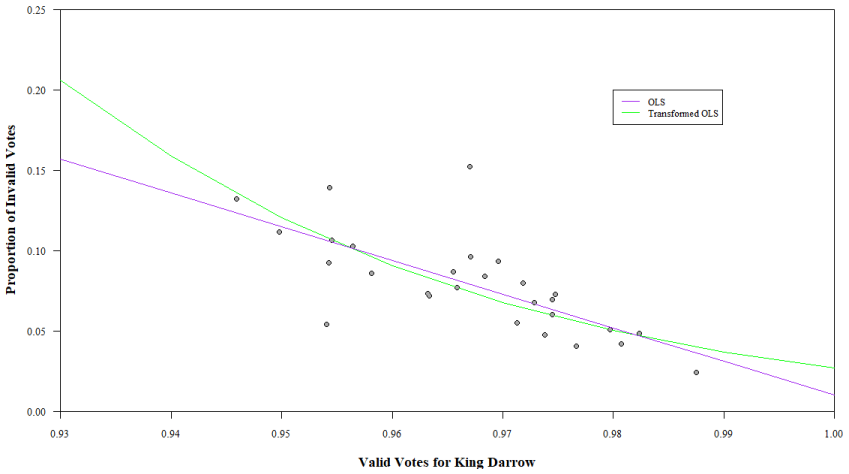

Figure 3: *A graphic of both linear models. This shows the difference in the transformed model and non-transformed model.*

Most importantly, notice that the slope is negative and the p-value is less than 0.05. These two things indicate that there is sufficient evidence that there is differential invalidation that *favors* King Darrow in this election.

Notice the purple line is our original OLS model without any transformations; this is quite obliviously linear. Although it fits the data, it is not really an accurate model due to the assumption violations. The green curve is the OLS model with a logit transformation; notice how much nicer this fits the behavior of the data. Furthermore, the assumptions are met, so we can actually use this model for predictions. The transformed OLS model does a great job, but it is quite complex and monotonous to transform, check assumptions, then back-transform.

Let us now look at the generalized linear model and the differences it has from the classical linear model.

2.2 GLM Model

Recall the GLM is more flexible than the CLM. It allows us to select certain features to better reflect what we know about the response variable; we just

need to know the link function and be modeling data within the exponential family.[Woo06]

Let us first think about the distribution of our data, and if it actually is in the exponential family. Our dependent variable is the proportion of invalid votes, and we are aiming to see if there is a relationship between the proportion of invalid votes (invalidation rate) and the proportion of votes for King Darrow. Thus, we are looking for the probability of obtaining a given number of successes (invalid votes, k) over the total number of trials (valid ballots cast, n). Written in this way, it is clear that the number of invalid ballots may follow a Binomial distribution. Since there are many distributions that can model the number of successes out of the number of trials, we cannot be absolutely certain that the conditional distribution is the Binomial. However, it is a good place to start.

One benefit to using the Binomial is that it is a member of the exponential family. Here, in much the same way as I did for the Poisson in Section 1.2.4, I show that the Binomial distribution is also a member of the exponential family. Recall the exponential form is:

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{15}$$

The PMF for the Binomial distribution is:

$$f(y) = \binom{n}{x} \pi^x \ (1-\pi)^{n-x} \tag{16}$$

Let us continue as in the Poisson example (§ 1.2.4).

$$f(y) = \binom{n}{x} \pi^x \ (1-\pi)^{n-x} \tag{17}$$

$$= \exp\left[\ \log\left(\binom{n}{x} \pi^x \ (1-\pi)^{n-x}\right)\ \right] \tag{18}$$

$$= \exp\left[\log \binom{n}{x} + x \ \log \pi + (n-x) \log (1-\pi) \right] \tag{19}$$

This simplifies to

$$f(y) = \exp\left[x\big(\log(\pi) - \log(1-\pi)\big) + n \log(1-\pi) + \log \binom{n}{x} \right] \tag{20}$$

Next, let's apply our logarithmic rules and rewrite this as:

$$f(y) = \exp\left[\frac{x \log\left(\frac{\pi}{(1-\pi)}\right)}{1} + n \log(1-\pi) + \log \binom{n}{x} \right] \tag{21}$$

Defining the following shows that this is in exponential form (6):

- $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, that is $\pi = \frac{e^\theta}{1+e^\theta}$,
- $b(\theta) = n \log\left(1 + e^\theta\right)$,
- $a(\phi) = 1$, and
- $c(y, \phi) = \log \binom{n}{x}$,

Here, I have shown that the Binomial is a member of the exponential class of distributions. I have also shown that the canonical link is the logit function,

$$\operatorname{logit} \pi := \log\left(\frac{\pi}{1-\pi}\right) \tag{22}$$

As an aside, the inverse of the logit function is the logistic,

$$\operatorname{logistic} x := \frac{e^x}{1+e^x} \tag{23}$$

For GLMs, the second thing to specify is a link function. From the proof that the Binomial is exponential class, we see that the canonical link for the Binomial distribution is the logit function (22).

Now, we have all of our requirements covered and we are ready to model. In R, the function to perform generalized linear modeling is `glm`. The entire line of code is

```
modC <- glm(pInv ~ PDarrow, family=binomial(link = "logit"))
```

The resulting output looks like

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

Once again, we have a negative slope. This indicates that the more votes King Darrow receives, the lower the invalidation rate. Note that the p-value is much less than out usual $\alpha = 0.05$, meaning that there is evidence that the differential invalidation aided King Darrow.

One thing to notice here is how much shorter this section is as compared to the OLS section; we did not need to do nearly as many steps or transformations. All we needed to do was figure out the likely distribution of our dependent variable, decide on an appropriate link function, and run the model. Figure 4 illustrates the effect of support for the King on the invalidation rate according to the GLM model.

While we are here, let us notice some similarities and differences between the CLM models and our GLM model (Figure 5). The regular OLS line is linear, so it does not fit the curvature of the data as well. Furthermore, that has no transformations to fix the requirements of OLS. Usually, we can get away with the violation of Normality when we have a large enough sample size (this is just the Central Limit Theorem). The logit-transform of our CLM model fits pretty well, but it does start a bit steep. The GLM does a bit better since the curve starts lower and looks like it fits the data slightly more precisely.

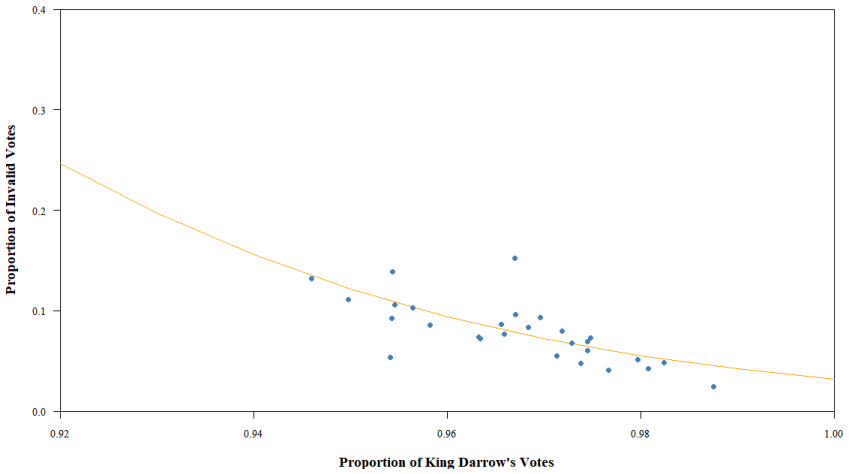

Figure 4: *Give a short description of this graphic. The first part should be a noun clause. If you add to it, the rest should be complete sentences.*

All of this aside, the transformed CLM and the GLM curves look almost the same. The most notable difference is time because with the GLM we did not have to worry about transformations, checking requirements, and then back-transforming. We just ran our model and did one back-transform to get our curve. That is the one advantage to using the GLM over the transformed-CLM is simply time.

Plus, we do end up getting a slightly most accurate model since we can take advantage of the actual distribution of the data as well as the link function. To really drive this home, below is a graphic of the CLM, the transformed CLM, and the GLM model.

So, to recap: the CLM was okay, but it was too linear for our data. The transformed-CLM was a lot better, but was a pain in the neck to do and took a long time. The GLM was even better, but we have one small issue.

Note that we are aggregating many ballots over multiple kingdoms. Because we are working with aggregate data, the model is most likely overdispersed. This is an important observation because the dispersion parameter for the Binomial distribution, $a(\phi) = 1$. If that parameter were a variable (as in

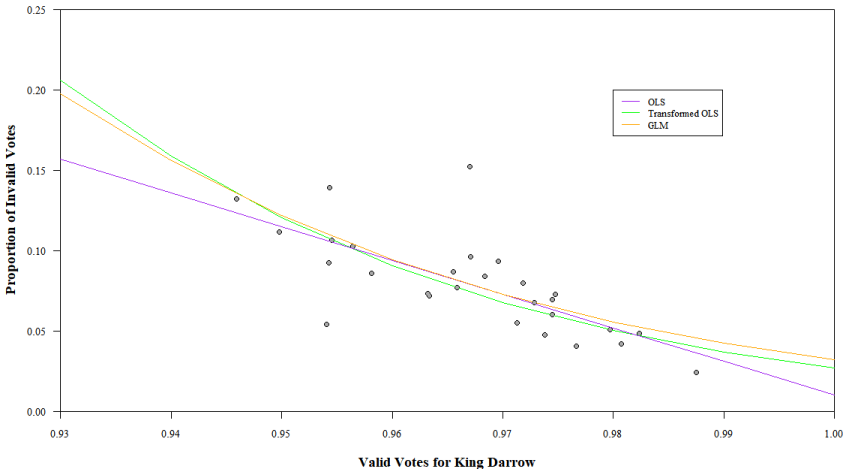

Figure 5: *A plot of the GLM model. Notice how well it fits the data, and the similarities and differences from the OLS models.*

the case of the Normal distribution), then overdispersion would be modeled by that parameter. However, this is not the case for the Binomial (or for the Poisson).

One way to check for overdispersion is to calculate the ratio of the residual deviance to the degrees of freedom. If it is much greater than one, then there is overdispersion. One can also obtain a confidence interval for the residual deviance under the assumption that there is no dispersion.[For20, §12.3.2]

Using the Chi-squared distribution, we are 95% confident that the residual deviance would be between 13.8 and 41.9. Because the observed residual deviance is much greater than this interval, we should conclude that there is evidence of overdispersion in the model.

Now that we know we have overdispersion, how do we fix this so we can make a better model and be sure King Darrow is actually guilty of electoral fraud before we bring evidence in front of the Court?

2.3 VGLM Model

In order to fix overdispersion in our data, we can use a distribution that is similar to the Binomial, but does not have the dispersion parameter being $a(\phi) = 1$. There are several, but a natural one is the Beta-binomial distribution. This distribution is very similar to the Binomial. The main difference between the Binomial and the Beta-Binomial is that the Beta-Binomial allows the success probability to vary according to a Beta distribution.

Its probability mass function is

$$\binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha,\beta)} \tag{24}$$

In this function, like the Binomial, x is the number of successes and n is the number of trials. In addition, α and β are parameters describing the variability of the success probability (π in the Binomial model). Finally, $B(\cdot,\cdot)$ is the beta function.

Using the beta distribution changes the scale of the overdispersed data, and therefore, makes a more reliable model.

The beta-binomial distribution is not a member of the exponential family. Thus, we cannot use the GLM framework. We will have to use a VGLM model instead.[Yee15]

Our coding syntax will not look much different than that of the GLM:

```
modD <- vglm(pInv ~ PDarrow, family=betabinomial)
```

The link function for the beta-binomial defaults to the logit function, so there is no need to specify it. The summary of our model is:

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

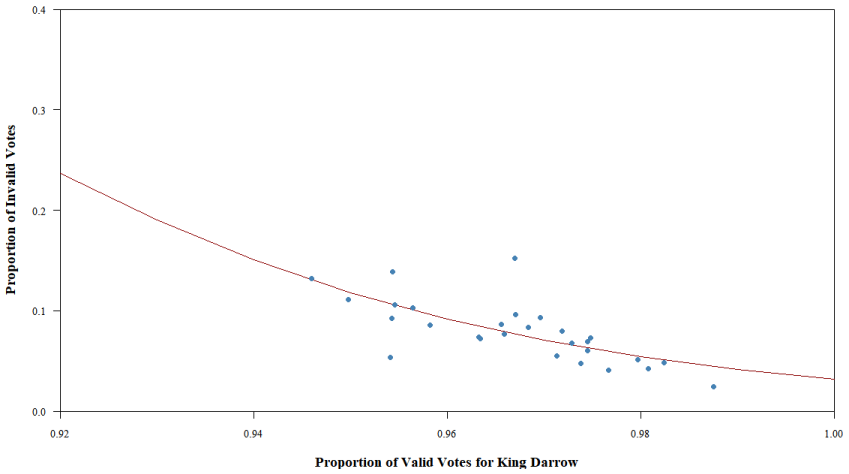

Figure 6: *Give a short description of this graphic. The first part should be a noun clause. If you add to it, the rest should be complete sentences.*

Again, note that the slope is negative — and the p-value is significant — so we do have evidence of differential invalidation that aided King Darrow.

Notice these results are quite similar to our GLM model. The p-values are even smaller, but the curve matches rather well. The difference is we were able to clean up the possibility of overdispersion, which makes for a more accurate model.

Now, we can be confident that everything is perfect, or at least more perfect, before we present the results to the Court. The downfall of the VGLM is time. Because there are more parameters to be fit, the IRLS iterations tend to take a bit longer. But, it is worth it to have a more precise model. Figure 6 is a graphic of the VGLM model.

2.4 Reflection

Let's think about what we did and why. We started with the classical linear model, as fitted using OLS. After modeling, we noticed violations of the requirements. If our sample size were larger, I would have felt more comfortable relying on the Central Limit Theorem and using OLS. However, our sample

size was only 28. Plus, regular OLS was too linear to really fit our data well.

So, we took some time to transform the dependent variable using a logit function. Then, we checked our assumptions again, saw everything was good, and then back-transformed the estimates. This made a rather precise model, but it was not perfect and took a lot of time and brain power since we had to keep checking the requirements.

So, we took advantage of what we knew about our data and fit everything using a GLM, which gave us a really accurate model and took hardly any time. This was great, until we realized we could have overdispersion because we were aggregating so many ballots over the kingdom level.

To compensate for this, we then fit a new type of distribution to help regulate overdispersion and used a VGLM instead since the distribution we decided on was outside of the exponential family. This gave us an even more precise model but took quite a bit of computer power.

Overall, it is important to realize when to use each type of model. Really, this is left up to the practitioner. We can use transformed OLS if we know a lot about it and not so much about other distributions and GLMS; or, we can use a GLM to make our lives easier if we are confident about distributions in the exponential family and link functions. Then, we can use a VGLM for modeling what we cannot use OLS or a GLM for and to make a really precise model. It is hard to go wrong either way. We talked a lot about differences, so just to show how similar these can be, Figure 7 is a graphic of all the models plotted together. Note that all models, except for the original line, make similar predictions.

Notice how close the GLM and the VGLM model are to each other. They both start at a slightly different place and end up a little different, which could be due to the VGLM correcting for overdispersion and using a different family. The first OLS without any transformations was pretty

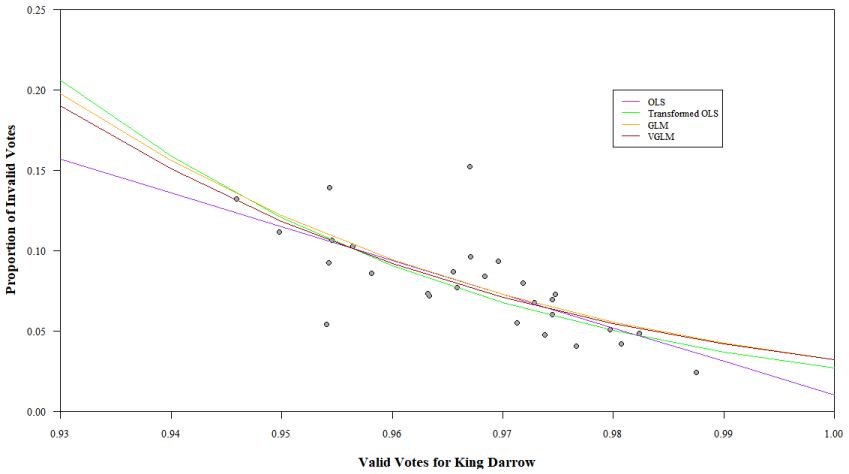

Figure 7: *Give a short description of this graphic. The first part should be a noun clause. If you add to it, the rest should be complete sentences.*

out of the park. The transformed OLS was okay, but still a bit of and took a lot of physical time. The GLM worked really well, minus the chance of overdispersion. Computationally, the VGLM was time-consuming, but worth it so we can have a precise model to bring thorough evidence to the Court.

We have just explored one of the two differences a VGLM has from a GLM. A VGLM can model data outside of the exponential family, like the beta-binomial, the Cauchy, and many more. This is really helpful for overdispersed data. Now, we are ready to look at the other difference. The VGLM can model more than one linear predictor. Recall the form of the VGLM, and the i subscript connected to the η. This means I can have a dependent variable that is modeled by a distribution with two parameters, and one linear predictor corresponds to the first parameter, and the other can correspond to the second. It is important to mention there can be more than two linear predictors for a two parameter distribution. Let us delve into an example of this.

3 Lengthy Application to the Gamma

The purpose of this section is to illustrate the second key difference between the VGLM compared to the GLM; the VGLM can handle more than one linear predictor, whereas the GLM can just predict one. An example of a distribution with more than one parameter is the Gamma distribution. In this section, we will use κ and θ to illustrate differences in the GLM and VGLM.

3.1 Types of data that may be gamma-distributed

Use your witch data to illustrate what is happening here.

3.2 Gamma Distribution

fact that it is two-;parameters

3.3 Gamma and OLS

3.4 Gamma and GLM

implementation is just one parameter

uses ========== parameterization

3.5 Gamma and VGLM

H MATRIX HERE I WORKED HARD ASF FOR THAT

4 Real Data example

5 Conclusion

References

[BP79] Trevor S. Breusch and Adrian R. Pagan. A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, 47:1287–1294, 1979.

[For20] Ole J. Forsberg. *Linear Models and Ř uritá Království : Using the Kingdom for Greater Insight*. https://rur.kvasaheim.com/, 0.704442d edition, 2020.

[LL22] Yang Liu and Baoding Liu. Estimating unknown parameters in uncertain differential equation by maximum likelihood estimation. *Soft Computing*, 26:2773–2780, 2022.

[NW72] John A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.

[SW65] Samuel Sanford Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[Woo06] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science. Chapman and Hall/CRC, 2006.

[Yee15] Thomas W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Series in Statistics. Springer, 2015.