HW#1 Generalized Linear Models with TMB

1. Modeling Catch Rate Data for Alaska Pollock

In our data set we have survey catch rate data for Alaska Polluck (y) where $y \geq 0$. From this data we find that about 3% of the catches have a value of zero. Taking this into consideration we want to model both the probability of a zero catch as well as the distribution of catch rates when we do not have a zero catch. We take three seperate approaches for modeling this data

1. Delta Log Normal

$$\boldsymbol{\beta} = \beta_{intercept}$$
$$\lambda_i = \mathbf{x}_i\boldsymbol{\beta}$$
$$Pr(c_i) = \begin{cases} \theta, & \text{if } c_i = 0 \\ (1-\theta)Lognormal(\lambda_i, \sigma), & \text{otherwise} \end{cases}$$

2. Delta Gamma

$$\boldsymbol{\beta} = \beta_{intercept}$$
$$\lambda_i = \mathbf{x}_i\boldsymbol{\beta}$$
$$\zeta_i = exp(\lambda_i)$$
$$Pr(c_i) = \begin{cases} \theta, & \text{if } c_i = 0 \\ (1-\theta)gamma(\zeta_i, \sigma), & \text{otherwise} \end{cases}$$

3. Delta Log Normal with covariates

$$\boldsymbol{\beta} = \beta_{intercept}, \beta_{lat}, \beta_{long}$$
$$\lambda_i = \mathbf{x}_i\boldsymbol{\beta}$$
$$Pr(c_i) = \begin{cases} \theta, & \text{if } c_i = 0 \\ (1-\theta)Lognormal(\lambda_i, \sigma), & \text{otherwise} \end{cases}$$

Modeling Results

Below is the joint negative log likelihood for each model fitted to the data

	log_normal	gamma	log_normal_cov
jnll	66957.31	61922	61790.09

the parameter estimates for each model as specified above

	theta	sigma	beta_int	beta_lat	beta_long
log_normal	0.0347256	2.512687	2.8182012	NA	NA
gamma	0.0347257	272.164028	-0.9919551	NA	NA
log_normal_cov	0.0347257	2.338008	1.2465341	-0.6321875	-0.2274766

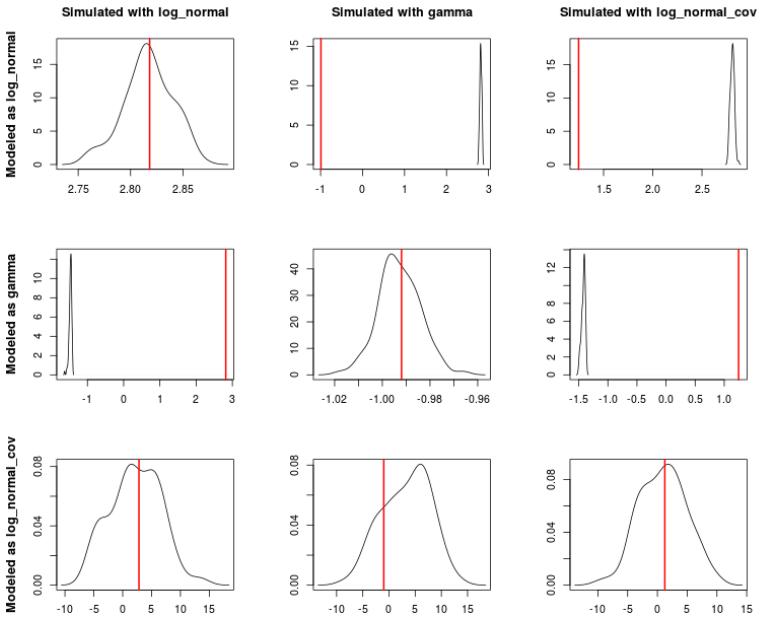
and the log predictive score

	log_normal	gamma	log_normal_cov
pred_score	6.033431	6.071572	6.06072

The models all visualize catch rate in different way yet there are similar out of sample predictive scores.

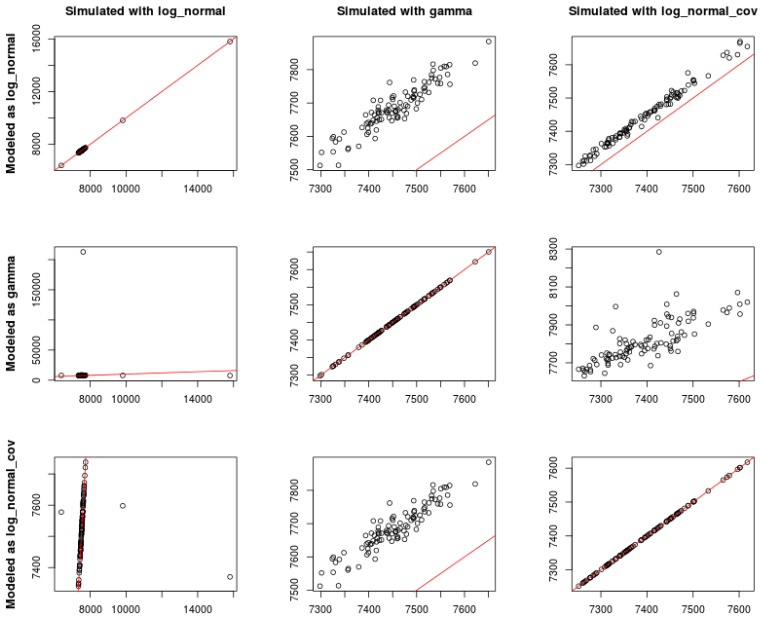
Simulation of catch rates

Using the parameters and model specifications shown above we can then simulate catch rate data to resemble the data that we have. In doing so we can examine the consequence of incorrectly specifying the model when we know the exact way that the data was generated. Below is the bias in the $\beta_{intercept}$ term when modeling the simulated data in various ways. The red bar shows the true $\beta_{intercept}$ term and the density plot shows the distribution of the estimated values of $\beta_{intercept}$ across 100 simulations.

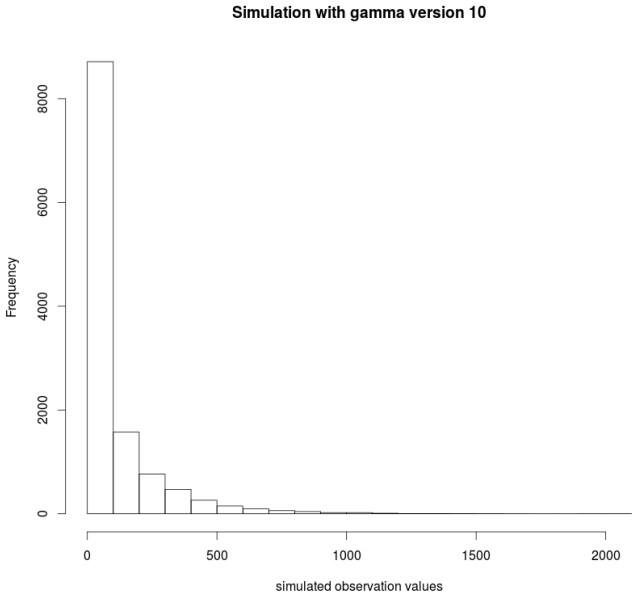


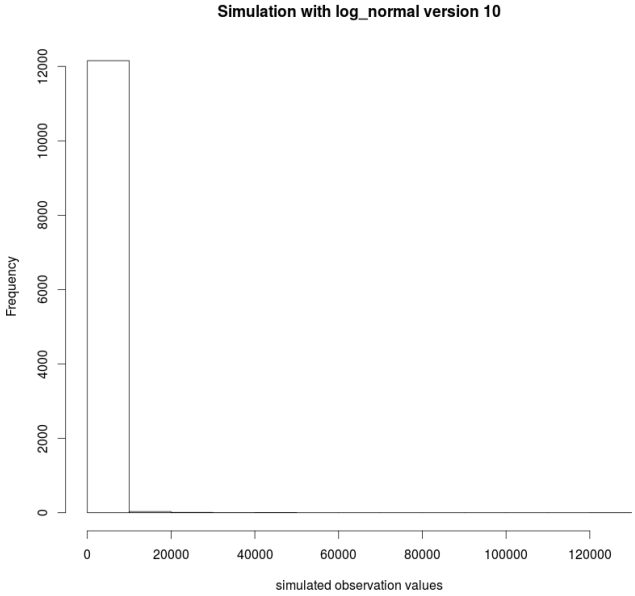
Using the correct model under which the data was simulated provoides a consistent unbiased estimator for the $\beta_{intercept}$ model.

Below is the nll for each model against a simulated set where the correct model specification is used for 100 simulations. The red line is $y = x$. If two models had the same nll then we would see there data points all fall on the red line. In this way it can be shown that using the lognormal model with covariates fits the data with less error as it it is less susceptible to outliers.



The two log normal models have similar data likelihoods which is to be expected as the model with covariates can decompose into the model without covariates. The gamma model sometimes struggles with identifying a good model when the simulations generated by the log normal values produce large numbers. Below are the histograms of two simulations one using the lognormal model and the other the gamma model to show the difference in the values generated.


Simulation with gamma version 10
Frequency
0
2000
4000
6000
8000
0
500
1000
1500
2000
simulated observation values


Simulation with log_normal version 10
Frequency
0
2000
4000
6000
8000
10000
12000
0
20000
40000
60000
80000
100000
120000
simulated observation values