CA COVID-19 Data - Bayesian Linear Models

Meltem Ozcan *

Abstract

In this paper we continue our exploration of the California COVID-19 mortality and infection dataset. We build and fit four linear regression models predicting log total number of cases in each county. The first model is an intercept only model, the second model has an additional term for log population density, the third model has an intercept and a term for the regional grouping, and the fourth model has an intercept as well as the grouping and log population density terms. We discuss model fit and validity and compare the models using AIC, BIC, DIC, and Bayes Factors resulting from the use of g-priors. We consider an additional model with Lasso and discuss the significance of the terms for grouping and population density.

Key Words: Bayesian linear regression, unequal variance, model comparison, g-priors, Lasso.

1. Introduction

As of June 2nd 2020, there have been 117,687 confirmed cases and 4,361 COVID-19 related deaths in the State of California. While counties such as Modoc and Trinity have had 0-1 confirmed cases, counties like Los Angeles have seen 58,160 positive cases. It is possible that county population and population density are factors associated with the spread of the virus. In this report, we explore four models and possible modifications to these models that incorporate the population information in the calculations of variance, and predict the total number of cases in each county using county-specific region and population density information.

*First author's affiliation, 1156 High St, Santa Cruz, CA 95064

2. Data

The dataset contains information on the number of COVID-19 cases and deaths in each of the 58 counties in California, as well as information on the population density and region of each county.

We created a new category, Group, that assigned each county to one of five groups: Group 1 (counties in Superior California and North Coast regions), Group 2 (counties in San Francisco Bay Area), Group 3 (counties in Central Coast), Group 4 (counties in Northern San Joaquin Valley and Southern San Joaquin Valley regions), and Group 5 (counties in Inland Empire, Los Angeles County, Orange County, and San Diego Imperial regions). The dataset contains data collected until April 13th 2020.

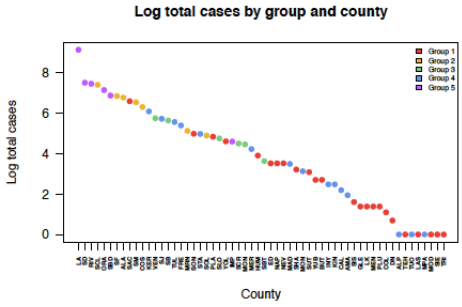

Figure 1: Log total number of cases for each county color coded by group.

We log-transformed the variables for the total number of cases and the population density as these variables are right-skewed. For the five counties that have no cases for the total number of cases, we set the log-transformed value to 0 as we do not have data to transform.

Figure 1 illustrates the log total number of cases for each of the 58 California counties. We can see that there were more cases observed in counties in Group 5 and 2 than counties in Group 1 and Group 4.

	G1	G2	G3	G4	G5
y_{ij}	2.34	6.28	4.80	3.18	7.13
	(1.87)	(0.93)	(0.80)	(2.14)	(1.47)
d_{ij}	3.69	7.00	4.92	3.75	6.06
	(1.64)	(0.83)	(0.93)	(1.77)	(1.71)
c_{ij}	11.19	13.64	12.63	11.62	14.59
	(1.39)	(0.71)	(0.89)	(1.63)	(1.33)

Table 1: Mean and standard deviations for log transformed total cases (y_{ij}), log population (c_{ij}), and log density (d_{ij}) for groups 1 through 5. Standard deviations are reported in parantheses.

Table 1 illustrates the mean and standard deviations of the log transformed total cases, population density and population variables. We see that Group 5 has the highest average log transformed total number of cases ($y_{ij} = logn_{ij}$), closely followed by Group 2. These two groups also have the highest average log transformed population and population density values, hinting at a positive relationship between these variables.

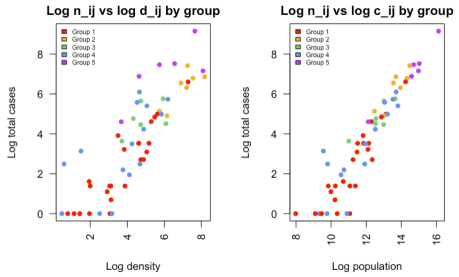

Figure 2: Plot on the left illustrates log total number of cases against log population density color coded by group. Plot on the right illustrates log total number of cases against log population color coded by group.

We also plotted log total cases against log population density and log population in Figure 2. There is a positive trend such that higher log population density is associated with higher log total number of cases. Similarly, higher log population values are associated with higher log total number of cases. This observation makes sense as COVID-19 is transmitted by particles in the air and individuals living in highly or densely populated counties might have a higher chance of coming in contact with other people who may have been infected or who are carriers. We see that the counties that have no reported COVID-19 cases are some of the least densely populated counties in California (belonging to Group 1 and Group 4).

3. Methods

We will take a Bayesian approach to building regression models and we will consider, fit and assess the following models:

$$\begin{aligned}
\text{Model 1} &: y_{ij} = \mu + \epsilon_{ij} \\
\text{Model 2} &: y_{ij} = \mu + \beta d_{ij} + \epsilon_{ij} \\
\text{Model 3} &: y_{ij} = \mu + \eta_j + \epsilon_{ij} \\
\text{Model 4} &: y_{ij} = \mu + \eta_j + d_{ij} + \epsilon_{ij}
\end{aligned}$$

where y_{ij} denotes the log transformed total number of cases, d_{ij} is the log population density of the ith county in the jth group, η_j is a group effect. We will assume that the error terms are normally distributed and centered around 0 with $\epsilon_{ij} \sim N(0, \sigma^2 10^3/c_{ij})$, where c_{ij} denotes the population for the ith county in the jth region. For all models, we consider a non-informative prior $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}) \sim (\sigma^2)^{-1}$.

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim N(0, \sigma^2 10^3/c_{ij})$$

This model can be written as

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 V)$$

where V is a diagonal matrix such that

$$V = diag\left(\frac{10^3}{c_1}, ..., \frac{10^3}{c_{58}}\right)$$

.

In order to transform our data and variables to a form that allows us to proceed with the equal variance formulas, we use the Cholesky decomposition of V ($V = LL^T$) to define $Y^* = L^{-1}Y$, $X^* = L^{-1}X$, and $\epsilon^* = L^{-1}\epsilon$ such that

$$Y^* = X^*\beta + \epsilon^*, \quad \epsilon^* \sim N(0, \sigma^2 I)$$

After this transformation, we can proceed our exploration with un-weighted data

and standard linear models that have known equal variance. The posterior distribution:

$$p(\boldsymbol{\beta^*}, \sigma^2|\boldsymbol{y^*}) = p(\boldsymbol{\beta^*}|\sigma^2, \boldsymbol{y^*})p(\sigma^2|\boldsymbol{y^*})$$

Conditional posterior of β^*:

$$\boldsymbol{\beta^*}|\sigma^2, \boldsymbol{y^*} \sim N\left((X^{*T}X^*)^{-1}X^{*T}y^*, (X^{*T}X^*)^{-1}\sigma^2\right.$$

Marginal posterior of σ^2:

$$\sigma^2|\boldsymbol{y^*} \sim IG\left((n-k)/2, (n-k)\hat{\sigma}^2/2\right)$$

Marginal posterior of β^*:

$$\boldsymbol{\beta^*}|\boldsymbol{y^*} \sim t_{n-k}(\hat{\beta}^*, \hat{\sigma}^2(X^{*T}X^*)^{-1})$$

After checking for posterior propriety, we will calculate the posterior hyperparameters, simulate 1000 draws of the posterior distribution of σ with these parameters, which we will use to simulate 1000 draws of from the posterior distribution of β. We will then simulate samples from the posterior predictive distributions for counties from each regional group using the four models.

This general process will be repeated for each model with different constructions of the design matrix X. For simplicity of notations, the rest of this report will refer to the un-weighted variables y^*, β^*, ϵ^* etc. without the asterisk i.e. y, β, ϵ. We proceed all computations with the transformed variables and transformed data.

4. Analyses and Results

4.1 Model 1: $y_{ij} = \mu + \epsilon_{ij}$

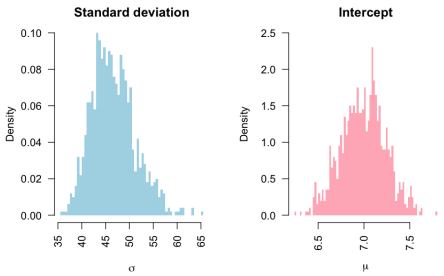

Figure 3: Simulations for σ and μ from Model 1.

For this model, we construct a 58x1 matrix containing 1s, and a β vector containing μ. We transform the data and variables as discussed above. As $n = 58 > k = 1$ and the rank of X equals 1, the posterior is proper.

Figure 3 illustrates the simulated draws from the posterior distributions of σ and μ. We calculate the mean posterior estimate for μ as

$$\hat{\mu} = 6.99$$

Then a regression equation for Model 1 is:

$$y_{ij} = 6.99 + \epsilon_{ij}$$

4.2 Model 2: $y_{ij} = \mu + \beta d_{ij} + \epsilon_{ij}$

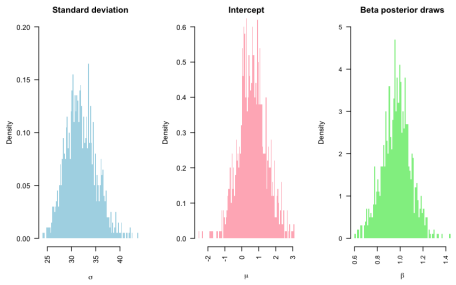

Figure 4: Simulations for σ, μ, and β from Model 2.

The design matrix for the second model contains 1s on the first column and the log density information for each county on the second column (58x2). The Beta vector contains μ and β. As $n = 58 > k = 2$ and the rank of X equals 2, the posterior is proper.

The mean $\boldsymbol{\beta}$ estimates are:

$$\hat{\mu} = 0.63$$
$$\hat{\beta} = 0.967$$

Then, we can plug in these values to get an equation for Model 2:

$$y_{ij} = 0.63 + 0.967d_{ij} + \epsilon_{ij}$$

4.3 Model 3: $y_{ij} = \mu + \eta_j + \epsilon_{ij}$

The Beta vector for Model 3 is a 5x1 vector that contains μ, η_1, η_2, η_3, and $\eta 4$. The

design matrix is a 58x5 matrix which contains 1s on the first column. η is a categorical variable designating membership to one of the five groups. We decided to add the constraint that $\eta_1 + \eta_2 + \eta_3 + \eta_4 = 0$. Therefore, columns 2 through 5 contain the information regarding η_1, η_2, η_3, and η_4 and the rows corresponding to observations belonging to Group 5 contain -1s in columns 2 through 5. We get η_5 by summing η_1, η_2, η_3, and η_4 and multiplying this value by -1.

As $n = 58 > k = 5$ and the rank of X equals 5, the posterior is proper.

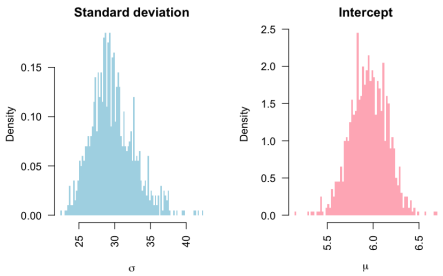

Figure 5: Simulations for σ and μ from Model 3.

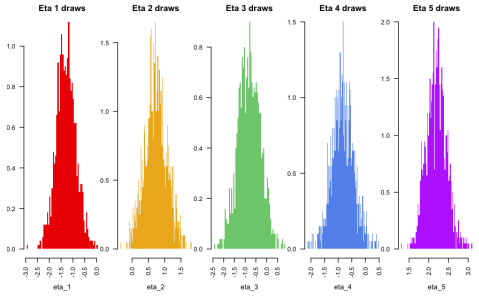

Figure 6: Simulations for η_1, η_2, η_3, η_4 and η_5 from Model 3.

The mean estimates for $\boldsymbol{\beta}$ are as follows:

$$\begin{aligned}\hat{\mu} &= \quad 5.95\\ \hat{\eta}_1 &= -1.288\\ \hat{\eta}_2 &= \quad 0.723\\ \hat{\eta}_3 &= -0.807\\ \hat{\eta}_4 &= -0.827\end{aligned}$$

Then, we can plug in these values to get an equation for Model 3:

$$\begin{aligned}y_{ij} =& 5.95 - 1.288\eta_1 + 0.723\eta_2 - 0.807\eta_3\\ &- 0.827\eta_4 + \epsilon_{ij}\end{aligned}$$

4.4 Model 4: $y_{ij} = \mu + \eta_j + d_{ij} + \epsilon_{ij}$

The Beta vector for Model 4 is a 6x1 vector that contains μ, $\eta_1, \eta_2, \eta_3, \eta 4$, and β. The design matrix is a 58x6 matrix which contains 1s on the first column. Model 4 also has the constraint that $\eta_1 + \eta_2 + \eta_3 + \eta_4 = 0$. Therefore, columns 2 through 5 contain the information regarding η_1, η_2, η_3, and η_4. As before, We get η_5 by summing η_1, η_2, η_3, and η_4 and multiplying this value by -1. Column 6 has β. As $n = 58 > k = 6$ and the rank of X equals 6, the posterior is proper.

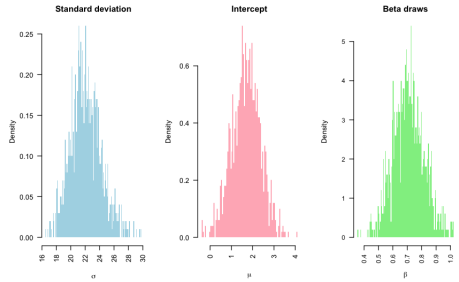

Figure 7: Simulations for μ, σ and β from Model 4.

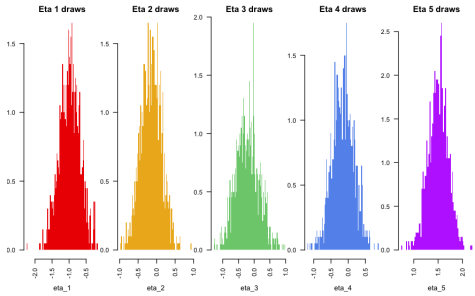

Figure 8: Simulations for η_1, η_2, η_3, η_4 and η_5 from Model 4.

Means for model 4 estimates for β are as follows:

$$\begin{aligned}\hat{\mu} &= \quad 1.69\\ \hat{\eta}_1 &= -0.982\\ \hat{\eta}_2 &= -0.152\\ \hat{\eta}_3 &= -0.214\\ \hat{\eta}_4 &= -0.5\\ \hat{\beta} &= \quad 0.708\end{aligned}$$

Then, we can plug in these values to get our

regression equation for Model 4:

$$y_{ij} = 1.69 - 0.982\eta_1 - 0.152\eta_2 - 0.214\eta_3 - 0.15\eta_4 + 0.708d_{ij} + \epsilon_{ij}$$

We notice that the coefficient estimates for Group 2 and Group 3 are close to being centered around 0 and smaller compared to the other coefficients. The influence of membership to Group 2 or Group 4 will be minimal in comparison to the other groups.

5. Model Comparison

5.1 Model validity

Figures 9,10, 11, 12, and 13 illustrate the samples of model predictions and vertical lines for the actual observation for Sacramento (Group 1), Santa Clara (Group 2), Ventura (Group 3), San Jose (Group 4), and Los Angeles (Group 5).

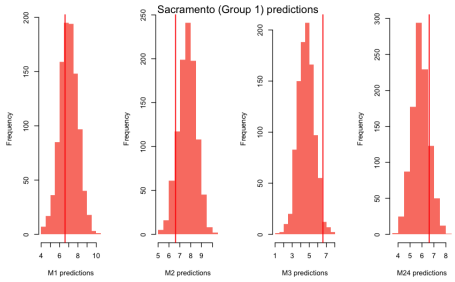

Figure 9: Posterior predictive samples of log total number of cases for Sacramento (Group 1), using the four models.

The observed log total number of cases for Sacramento, Santa Clara, San Jose and Ventura are captured in each of the posterior predictive samples generated using the four models. Model 3 slightly underestimates the log total number of cases for Sacramento, while Model 1 slightly overestimates the log total number of cases for Sacramento, Ventura and San Jose.

Posterior predictive samples for Los Angeles underestimate the log total number of cases for all four models, with Model 1 performing especially poorly. We had considered Los Angeles as a potential outlier in previous reports. We repeated this analysis

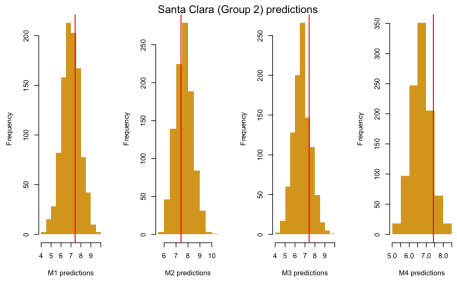

Figure 10: Posterior predictive samples of log total number of cases for Santa Clara (Group 2) using the four models.

with other counties in Group 5 such as San Diego. Samples obtained using Model 1 did not capture the observed log total number of cases, and the samples obtained using the remaining models contained the actual observed value for log total number of cases.

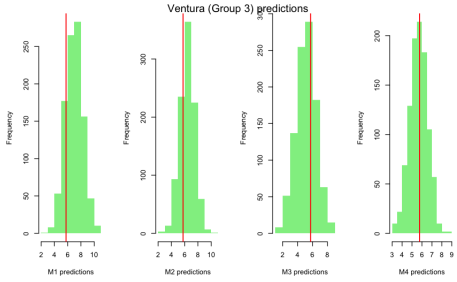

Figure 11: Posterior predictive samples of log total number of cases for Ventura (Group 3) using the four models.

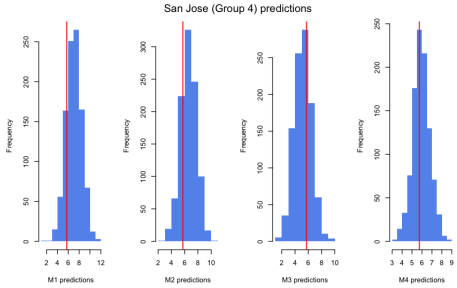

Figure 12: Posterior predictive samples of log total number of cases for San Jose (Group 4) using the four models.

Overall, Model 2, Model 3 and Model 4 appear to produce results consistent with the data as the actual observations for log total

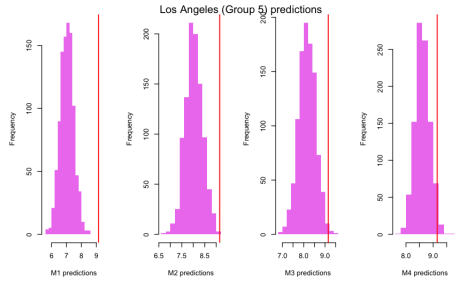

Figure 13: Posterior predictive samples of log total number of cases for Los Angeles (Group 5) using the four models.

	M1	M2	M3	M4
p	1	2	5	6
R^2	0	0.53	0.62	0.79
AIC	4	-38.3	-43.7	-77.4
BIC	8.1	-32.2	-31.4	-63
DIC	611.4	569.6	565.2	526.6

Table 2: Number of parameters and R-squared, AIC, BIC, and DIC values for models 1, 2, 3, and 4.

cases are captured in the posterior predictive samples for the counties considered. On the other hand, Model 1 produced predictions that are less then adequate, especially for counties in Group 5.

5.2 Information Criteria

Table 2 illustrates the AIC, BIC, and DIC scores for the four models as well as the number of parameters in the R-squared values for each model.

Model 1, the null model, has an R-squared value of 0 and as we increased the number of parameters in each model, the R-squared value increased as to be expected. Model 2 accounts for 53% of the variability in data, while the model with the most parameters, Model 4, has an R-squared value of .79, indicating that the model accounts for about 80% of the variability in the data. These somewhat low R-squared values indicate a room for improvement in the models.

After exploring the R-squared values, we computed the AIC, BIC, and DIC scores for each model. Individual scores are not meaningful on their own and become interpretable when compared to other scores calculated using the same information criterion. As such, we find that

$$AIC_{M1} > AIC_{M2} > AIC_{M3} > AIC_{M4},$$
$$BIC_{M1} > BIC_{M3} > BIC_{M2} > BIC_{M4}$$
$$DIC_{M1} > DIC_{M2} > DIC_{M3} > DIC_{M4}$$

Model 4 is the preferred model for AIC, BIC, and DIC, while Model 1 is the least preferred model by all three criteria. Model 3 and 4 have more parameters than the other two models. As BIC penalizes the number of parameters in the model, these models are more heavily penalized. Hence, BIC prefers Model 2 over Model 3 while AIC and DIC prefer Model 3 over Model 2.

	BF_{21}	BF_{31}	BF_{41}
$g = 0.1$	3.76	4.09	6.34
$g = 1$	$3.5 * 10^3$	$6.6 * 10^3$	$2.24 * 10^4$
$g = 10$	$1.56 * 10^7$	$3.96 * 10^7$	$4.81 * 10^{12}$
$g = 100$	$2.08 * 10^7$	$4.9 * 10^6$	$1.1 * 10^{13}$

Table 3: Bayes Factors comparing models 2, 3,and for against Model 1 for the various values of g.

5.3 Classical Approach

As an additional check, we take a classical approach and compare our nested models using the F-test. As we have determined Model 1 to be inadequate for modeling the data at hand, we draw comparisons between the remaining models. We conduct an F-test at level 95% to test the hypothesis that the reduced model (Model 2) is the best against the alternative hypothesis that the full model (Model 4) is preferred. As $16.27 > 2.55$ we reject the null hypothesis and conclude that Model 4 is preferred. Conducting an F-test comparing models 3 and 4, we reject the null hypothesis that Model 3 is preferred as $44.2 > 4.03$. Thus, the classical and Bayesian approaches agree that Model 4, $y_{ij} = \mu + \eta_j + d_{ij} + \epsilon_{ij}$, is the best model out of the four models considered.

	BF_{23}	BF_{24}	BF_{34}
$g = 0.1$	0.92	0.59	0.64
$g = 1$	0.54	0.02	0.03
$g = 10$	0.4	0.0	0.0
$g = 100$	4.28	0.0	0.0

Table 4: Bayes Factors comparing models 2, 3,and for against Model 1 for the various values of g.

5.4 Bayes Factors and G-priors

We next consider Bayes factors to compare the four models in pairs. It is possible to use g-priors given our current setup as we can scale the columns of our X matrix so that they have column means of 0. We utilize g-priors here in order to compute closed form Bayes factors. We consider fixed g values of 0.1, 1, 10 and 100 which result in shrinkage factors $\left(\frac{g}{1+g}\right)$ 0.091, 0.5, 0.91 and 0.99 respectively. The g-prior shrinks the posterior mean and predicted values to 0 by these factors.

Table 3 displays the Bayes factors for comparisons with the null model (Model 1), while Table 4 illustrates Bayes factors comparing the remaining three models on each of the g-values mentioned above.

The Bayes factors comparing the null model with the other three models all give support to the hypothesis that the larger models are better than the null model (since $BF > 1$). On a similar vein, the Bayes factors on Table 4 give support to the largest model, Model 4, when compared with models 2 and 3 (since $BF < 1$).

It is interesting to note that for g=100, Model 2 is preferred over Model 3.

6. Additional Models

6.1 Lasso

The various methods we used to compare our models point to Model 4 as the best model out of the four we have built. We can explore additional models such as the Lasso in order to induce sparsity and make our regression model more robust. The best model, Model 4, has 6 parameters. We can reduce some of the coefficients to 0 at the optimum.

$$min_\beta ||y - X\beta||^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

We estimate λ via a 10-fold cross validation and find that the minimum lambda is at 0.997. We use this value as the starting value and run 1000 simulations. The mean coefficient estimates from this setup is below:

$$\begin{aligned} \hat{\mu} &= \ \ 4.635 \\ \hat{\eta}_1 &= -0.634 \\ \hat{\eta}_2 &= \ \ 0 \\ \hat{\eta}_3 &= \ \ 0 \\ \hat{\eta}_4 &= \ \ 0 \\ \hat{\beta} &= \ \ 0.404 \end{aligned}$$

Then we can write a regression equation

$$y_{ij} = 4.635 - 0.634 + 0.404 d_{ij} + \epsilon_{ij}$$

This finding shows that Group 2, Group 3 and Group 4 are not meaningfully different from each other in a way that would warrant including specific terms for them in the model.

In other words, under this penalized setting, model fit is not improved by having separate terms for η_2, η_3, and η_4. Thus, we can reduce our number of parameters from 6 to 3.

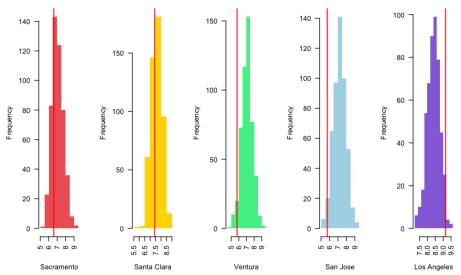

Figure 14: Predictions of log total number of cases for Sacramento (Group 1), Santa Clara (Group 2), Ventura (Group 3), San Jose (Group 4) and Los Angeles (Group 5) using the regression equation (lasso).

Figure 14 illustrates the posterior predictive samples for the total number of cases in

one country from each of the five regional groups, predicted using the Lasso model. The log total number of cases for Los Angeles is slightly underestimated and the log total number of cases for Ventura and San Jose are slightly overestimated, but the true values are still captured in the samples.

While out of the scope of this paper, one interesting extension to our exploration would be to consider a mixture model in light of our observation that counties in groups 1 and 5 are meaningfully different.

7. Conclusions

In this report we built, fit, assessed and compared four linear regression models with known unequal variance to predict the log transformed total number of COVID-19 cases in each of the 58 counties in California.

At the end of our exploration, we conclude that the grouping and population density variables are both significant variables. We found that Model 4, the model with an intercept and additional terms for log population density and regional group membership, is the preferred model out of the models considered. We also found that the number of parameters in the full model can be reduced to 3 as keeping the parameters pertaining to membership to groups 2, 3 and 4 is not justified by the data.

Future investigations can consider incorporating demographic information such as age intervals, gender, and race. In particular, it would be interesting to take a interval censoring approach and introduce latent variables regarding assignment to various age categories.

REFERENCES

Albert, Jim. (2009) Bayesian Computation with R. 2nd ed. Springer.

Gelman et al. (2013) Bayesian Data Analysis (3rd ed.). CRC Press.

California Department of Public Health (CDPH) Office of Public Affairs (https://www.cdph.ca.gov/Programs/OPA/Pages/NR20-111.aspx)