Kyle Davis

Thesis Results and Discussion

Results:

Model 1:

Prior Distributions for Model 1:

$$General\ Subsidy \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim Normal(0{,}10000)$$
$$\beta \sim Normal(0{,}5)$$
$$\sigma \sim HalfCauchy(0{,}1)$$

Model 1 Summary Output:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: general_subsidy ~ 1 + investment_return
   Data: final_regdata_FTE (Number of observations: 12199)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                 Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept         -1824.61     91.27 -2001.71 -1646.73       3137 1.00
investment_return     0.16      0.00     0.15     0.16       4734 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma  9815.97     61.92  9697.80  9942.67       3762 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

In the table above we see the results for the first Bayesian model. This model estimates the posterior distribution of the parameters that generate our outcome variable *General Subsidy*. The prior on the intercept is normally distributed with mean 0 and standard deviation $10,000. This was selected to allow for the rather long tails that tend to be associated with this variable. While the majority of institutions hover around 0, there are some institutions that sit far out in the right tail (Yale University is one such example). The intercept had an estimated mean of -$1,824 with estimated error of $91.27. 95% of our distribution on the intercept parameter fall between -$2,001.71 and -$1,646.73. The coefficient on investment return is estimated to be 0.16 with an estimated standard error of 0. 95% of the distribution for the investment return

coefficient parameter falls between 0.15 and 0.16. Both intercept and investment return coefficient parameters had an Rhat of 1 indicating that the sampling chains converged.

Model 2:

Prior Distributions for Model 2:

$$General\ Subsidy \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 Investment\_Return_i + \beta_2 Non\_Instructional\_Expense_i$$
$$\alpha \sim Normal(0{,}10000)$$
$$\beta_1 \sim Normal(0{,}5)$$
$$\beta_2 \sim Normal(0{,}1)$$
$$\sigma \sim HalfCauchy(0{,}1)$$

Model 2 Summary Output:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: general_subsidy ~ 1 + investment_return + non_instructional_exp
   Data: final_regdata_FTE (Number of observations: 12199)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept            -3088.57     84.24 -3252.90 -2923.36       3442 1.00
investment_return        0.13      0.00     0.12     0.13       4013 1.00
non_instructional_exp    0.12      0.00     0.11     0.12       5669 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma  8814.92     55.63  8705.04  8922.25       3261 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

The table above specifies the results of the second Bayesian model. In this instance the variable non-instructional expense was also included. With the inclusion of this added variable we see that the intercept parameter estimate drops to -$3,088.57 and the standard error on this estimate drops as well to $84.24. 95% of the distribution for the intercept parameter falls between -$3252.90 and -$2923.36. The coefficient parameter on investment return dropped slightly to 0.13, while the standard error remained at 0. 95% of the distribution for the coefficient parameter on investment return falls between 0.12 and 0.13. The newly introduced variable non-instructional expense has a coefficient parameter estimate of 0.12 with a standard error of 0. 95% of the distribution for the coefficient parameter on non-instructional expense falls between 0.11 and 0.12. All Rhat values are equal to 1, showing that there is convergence in the simulated chains.

Model 3:

Prior Distributions for Model 3:

$$General\ Subsidy \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 Investment\_Return_i + \beta_2 Non\_Instructional\_Expense_i + \beta_3 Year_i$$
$$\alpha \sim Normal(0{,}10000)$$
$$\beta_1 \sim Normal(0{,}5)$$
$$\beta_2 \sim Normal(0{,}1)$$
$$\beta_3 \sim Normal(0{,}1)$$
$$\sigma \sim HalfCauchy(0{,}1)$$

Model 3 Summary Output:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: general_subsidy ~ 1 + investment_return + non_instructional_exp + year
   Data: final_regdata_FTE (Number of observations: 12199)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept            -3088.50     83.26 -3248.16 -2927.84       7009 1.00
investment_return        0.13      0.00     0.12     0.13       9339 1.00
non_instructional_exp    0.12      0.00     0.11     0.12       4748 1.00
year2005                 0.01      0.97    -1.91     1.89       7288 1.00
year2007                -0.00      1.00    -1.93     1.96       7074 1.00
year2008                 0.01      0.95    -1.88     1.88       7534 1.00
year2009                 0.04      0.99    -1.89     1.98       7391 1.00
year2010                -0.01      1.03    -2.01     2.04       7511 1.00
year2011                -0.01      1.00    -2.00     1.92       9762 1.00
year2012                -0.01      0.99    -1.92     1.92       5925 1.00
year2013                -0.03      1.03    -2.01     2.05       7768 1.00
year2014                -0.05      1.02    -2.06     1.95       8528 1.00
year2015                 0.06      1.00    -1.86     1.99       7829 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma  8815.70     56.72  8705.67  8929.79       8082 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

The table above shows the summary output for the third Bayesian model. This model introduces the factor variable year. With the inclusion of this added variable we see that the intercept parameter estimate drops to -$3,088.50 and the standard error on this estimate drops as well to $83.26. 95% of the distribution for the intercept parameter falls between -$3,248.16 and -$2,927.84. The newly introduced variable has coefficient parameter estimates

ranging from -0.05 to 0.06. Though the distributions for these variables are not exactly the same, all of the distributions have their upper and lower 95% confidence interval bands on opposite sides of 0. The importance of this will be discussed in the results section. All Rhat values are equal to 1, showing that there is convergence in the simulated chains.

Model 4:

Prior Distributions for Model 4:

$$General\ Subsidy \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 Investment\_Return_i + \beta_2 Non\_Instructional\_Expense_i + \beta_3 Year_i + \beta_4 Gifts_i + \beta_5 Grants\_Appropriations_i + \beta_6 Net\_Auxilary\_Revenue_i + \beta_7 Net\_Hospital\_Rev_i + \beta_8 Net\_Other\_Rev_i$$
$$\alpha \sim Normal(0,10000)$$
$$\beta_1 \sim Normal(0,5)$$
$$\beta_2 \sim Normal(0,1)$$
$$\beta_3 \sim Normal(0,1)$$
$$\beta_4 \sim Normal(0,5)$$
$$\beta_5 \sim Normal(0,5)$$
$$\beta_6 \sim Normal(0,1)$$
$$\beta_7 \sim Normal(0,10)$$
$$\beta_8 \sim Normal(0,0.5)$$
$$\sigma \sim HalfCauchy(0,1)$$

Model 4 Summary Output:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: general_subsidy ~ 1 + investment_return + non_instructional_exp + year + gifts_total + grants_appr + net_aux_rev +
net_hospital_rev + net_other_rev
   Data: final_regdata_FTE (Number of observations: 12199)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept           -3405.75     71.64 -3548.52 -3263.65       6685 1.00
investment_return       0.05      0.00     0.04     0.05       5813 1.00
non_instructional_exp  -0.02      0.00    -0.02    -0.01       6323 1.00
year2005                0.03      1.01    -1.93     2.03       8288 1.00
year2007               -0.00      1.02    -1.99     1.95       7088 1.00
year2008                0.02      0.98    -1.95     1.96       7249 1.00
year2009                0.04      0.98    -1.84     1.97       7391 1.00
year2010               -0.03      0.99    -2.02     1.88       8521 1.00
year2011               -0.07      1.00    -2.05     1.88       6397 1.00
year2012               -0.02      0.96    -1.87     1.85       7059 1.00
year2013               -0.06      1.01    -1.93     1.87       7240 1.00
year2014               -0.06      0.97    -1.94     1.87       7326 1.00
year2015                0.09      1.02    -1.86     2.06       7588 1.00
gifts_total             0.64      0.01     0.61     0.66       8328 1.00
grants_appr             0.99      0.01     0.97     1.02       6553 1.00
net_aux_rev            -0.74      0.04    -0.82    -0.67       7216 1.00
net_hospital_rev        0.60      0.01     0.57     0.63       7459 1.00
net_other_rev           0.43      0.01     0.41     0.45       5817 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma  6049.17     38.15  5974.71  6122.96       9485 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

In the fourth Bayesian model we add several net revenue sources including: gifts, grants/appropriations, auxiliary revenue, hospital revenue, and other revenue. All other variables were maintained. The estimate on the intercept parameter dropped to -$3,405.75 with a standard error of $71.64. 95% of the distribution on this estimate falls between -$3,548 and -$3,263. The estimate of the coefficient parameter on investment return has dropped to 0.05, with 95% of the distribution falling between 0.04 and 0.05. The estimate on the coefficient parameter of non-instructional expenses turned negative to -0.02, with 95% of the distribution falling between -0.02 and -0.01. The various levels of year all have similar properties associated with their coefficient parameter estimates. Most notably all of the distributions have a lower and upper band of the 95% confidence interval on opposite sides of zero. We will discuss this significance further in the discussion section. The coefficient parameter on gifts is estimated at 0.63 with standard error of 0.01. 95% of the distribution falls between 0.61 and 0.66. The coefficient parameter on grants and appropriations is estimated at 0.99 with standard error 0.01. 95% of the distribution on this parameter falls between 0.97 and 1.02. The coefficient parameter estimate for net auxiliary revenue is -0.74 with standard error of 0.04. 95% of the distribution on this parameter falls between -0.82 and -0.67. The coefficient parameter estimate for net hospital revenue is 0.60 with standard error 0.01. 95% of the distribution falls between 0.57 and 0.63. The coefficient parameter estimate on net other revenue is 0.43 with standard error 0.01. 95% of the distribution on this parameter fall between 0.41 and 0.45. All Rhat values are equal to 1, showing that there is convergence in the simulated chains.

Model 5:

Prior Distributions for Model 5:

$$General\ Subsidy \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 Investment\_Return_i + \beta_2 Non\_Instructional\_Expense_i + \beta_3 Gifts_i + \beta_4 Grants\_Appropriations_i + \beta_5 Net\_Auxilary\_Revenue_i + \beta_6 Net\_Hospital\_Rev_i + \beta_7 Net\_Other\_Rev_i$$
$$\alpha \sim Normal(0,10000)$$
$$\beta_1 \sim Normal(0,5)$$
$$\beta_2 \sim Normal(0,1)$$
$$\beta_3 \sim Normal(0,5)$$
$$\beta_4 \sim Normal(0,5)$$
$$\beta_5 \sim Normal(0,1)$$
$$\beta_6 \sim Normal(0,10)$$
$$\beta_7 \sim Normal(0,0.5)$$
$$\sigma \sim HalfCauchy(0,1)$$

Model 5 Summary Output:

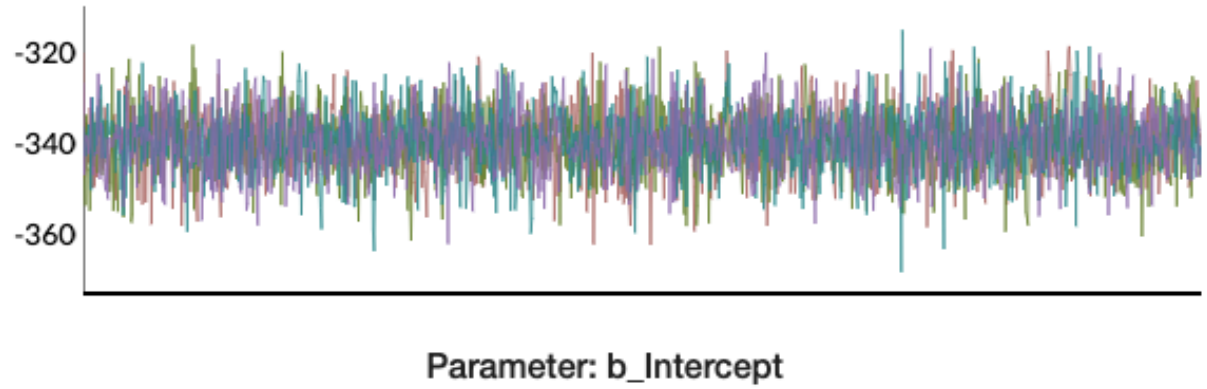
```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: general_subsidy ~ 1 + investment_return + non_instructional_exp + gifts_total + grants_appr + net_aux_rev +
net_hospital_rev + net_other_rev
   Data: final_regdata_FTE (Number of observations: 12199)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                   Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept          -3405.79     70.07 -3543.45 -3265.49       4535 1.00
investment_return      0.05      0.00     0.04     0.05       5484 1.00
non_instructional_exp -0.02      0.00    -0.02    -0.01       5808 1.00
gifts_total            0.64      0.01     0.62     0.66       4334 1.00
grants_appr            0.99      0.01     0.97     1.02       4016 1.00
net_aux_rev           -0.74      0.04    -0.82    -0.67       4415 1.00
net_hospital_rev       0.60      0.01     0.57     0.63       4156 1.00
net_other_rev          0.43      0.01     0.41     0.45       3554 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma  6049.56     38.53  5974.24  6126.87       5160 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

In the fifth Bayesian model we drop the categorical variable year, for reasons to be discussed later. The intercept parameter estimate shifted to -$3,405.79 with standard error $70.07. 95% of the distribution on this parameter falls between -$3,543.45 and -$3,265.49. The distributions for the coefficient parameters remain unchanged from model 4. All Rhat values are equal to 1, showing that there is convergence in the simulated chains.

Model Diagnostics:

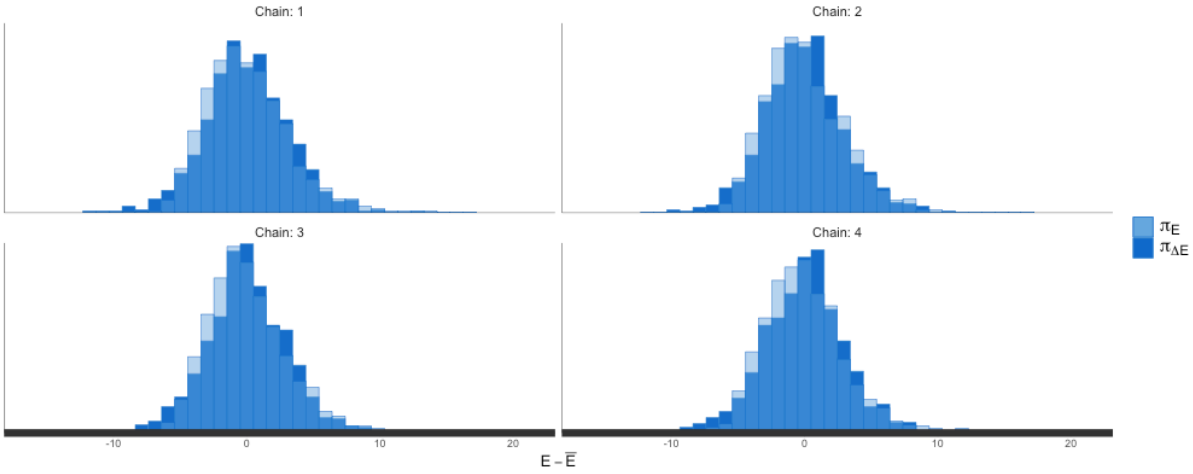
In order to discuss the results of the model, we must also explore the diagnostics affiliated with them. A true advantage of the Bayesian framework is the ability to predict well out of sample relative to other techniques (supervised learning, traditional regression analysis). To diagnose the model, we will use several measures including: chain consistency/convergence, energy information, MCMC chain autocorrelation, and review the posterior predictive distribution.

Checking the MCMC Chains:



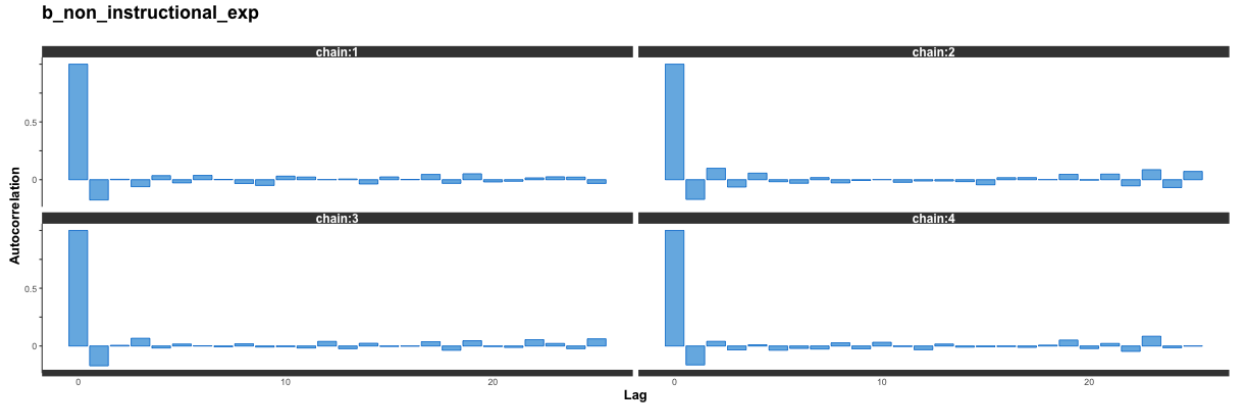
A markov process is a sequence of random variables with a specific dependence structure. The future is conditionally independent of the past given the present. However, nothing is marginally independent of anything else. Stan utilizes a specific method of markov chain monte carlo simulation called No U-Turn Sampling (NUTS). NUTS discretizes a continuous-time Hamiltonian process in order to solve a system of Ordinary Differential Equations. Above is a visual representation of the four MCMC chains and is a way of checking how NUTS performed during our simulation. From a visual perspective, NUTS seems to have performed well, though we will need to examine other parts of the posterior distribution to confirm this. The consistency of the random samples throughout the simulation process shows that the chains converged and worked well.

Energy Information:



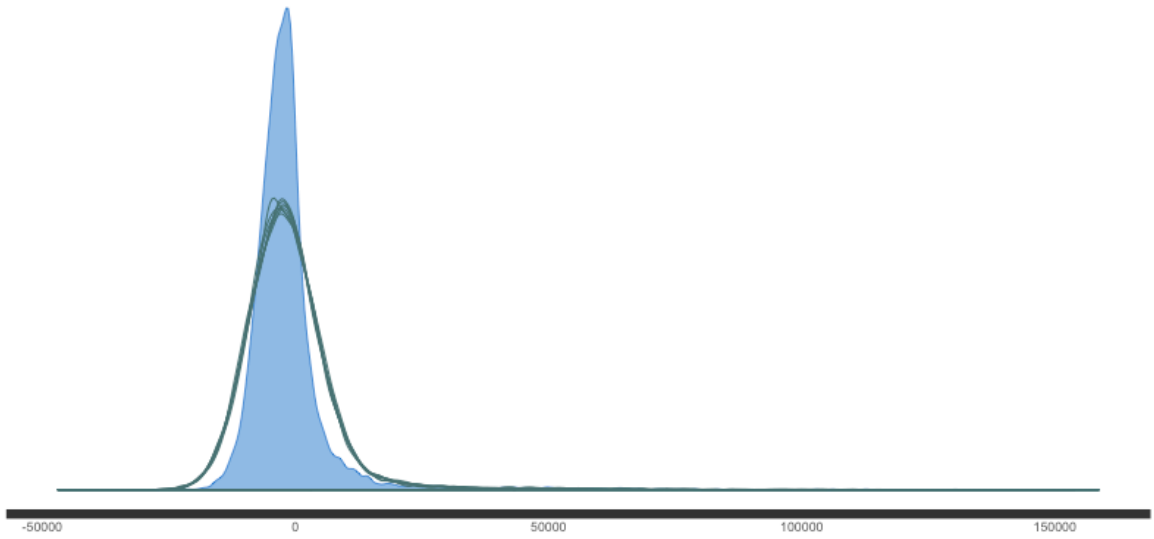
The energy information visual above represents a way to assess the low Bayesian fraction of missing information (LBFMI). When the tails of the posterior probability density function are very light, NUTS can have difficulty moving through Θ efficiently. This will result in an unreliable estimate of our effective sample size. This is not issue in our model as is indicated by the energy information plots. We want to see that the distribution of energy (π_{ε}) is the same as the distribution for the change in energy($\pi_{\varepsilon i}$). Visually we can see that this is the case for all four of our chains.

Autocorrelation:



In the plot above we examine the autocorrelation between draws in each of four chains. Specifically looking at the various draws for the coefficient parameter on non-instructional expense, though all the coefficient parameters experienced similar trends. We initially started out with some autocorrelation between draws, but this quickly went to zero. This is an indication that our model was well specified and that NUTS did not struggle with sampling our model.

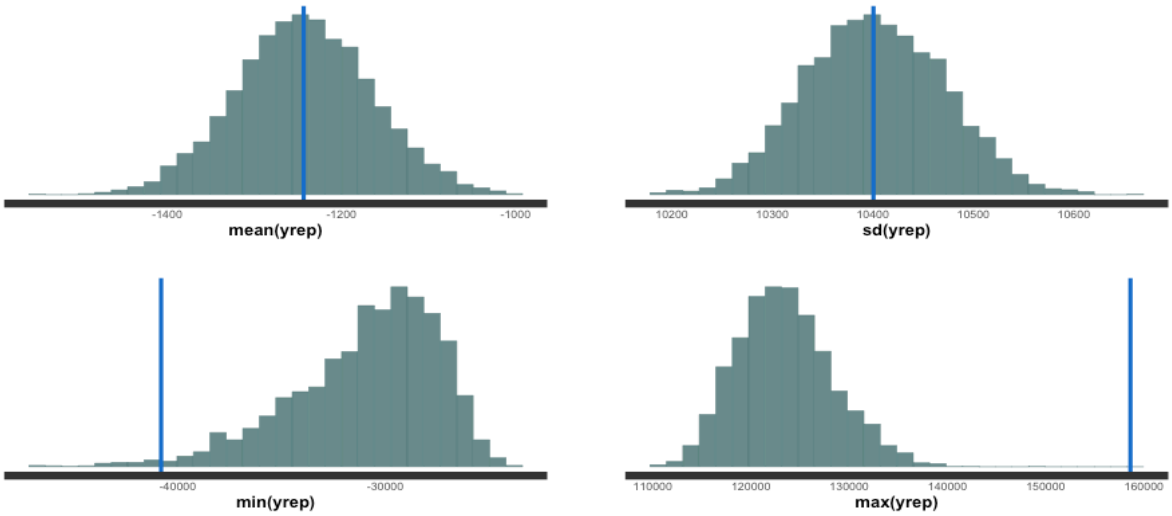
Posterior Distributions vs. Actual Distribution of Data:



The chart above is a way of comparing our posterior predictive distribution to the actual data. This allows us to get a sense of how well we are approximating the generating process, and which portions of the actual outcome distribution is our model missing. The light blue distribution is the actual distribution of general subsidy as seen in the dataset. The darker lines are the predicted posterior distributions. The posterior predictions seem to moderately capture the distribution of general subsidy. It has some trouble predicting the extreme outliers of the

distribution. The resulting predictive posterior distribution has more variance around the mean, but does not reach the extreme ends of the tails.

Distribution of Test Statistics:



Exploring the posterior predictive distribution further, the above plot shows how well the posterior distribution estimates the mean, standard deviation, minimum, and maximum of the actual general subsidy distribution. The posterior estimates the minimum and the maximum of general subsidy distribution well. The posterior struggles more with the tails of the distribution. The model consistently underestimates the maximum of general subsidy and consistently over estimates the minimum. This can be attributed to our normal priors that we have placed on our posterior predictive distribution. In further research a gamma distribution may be a more appropriate starting place to measure any non-linearities in the model.

Discussion:

Our fifth Bayesian model appears to have a solid fit of the data. We will now discussion the implications of the results as well as discuss the model diagnostics to check validity. The intercept parameter in model 5 is estimated to be centered around -$3,405.79. This starkly differs from previous research which found the average general subsidy to be positive. *Winston and Yen, 1995* estimated that in 1991 the average higher education institution produced a unit of education for $10,653, sold it for $3,101, creating an average general subsidy of $7,551. This paper went on to say that one of the fundamental parts of the economics of higher education is the permanent feature of general subsidies. Perhaps this difference is caused by the differences in time periods explored. Where *Winton and Yen, 1995* looked at data from the 1991 NCES IPEDS database. Though I source my data from the same organization, my data spans 2004-2015. Perhaps there were shifts in the way that higher education institutions were run during that time period versus the one relevant to this paper. The revelation is important however, as

the higher education market continues to increase in competition the effects of general subsidies on the institution should grow as well. We also saw that by adding new variables to the model we were able to tighten the distribution for the intercept parameter. By adding additional variables to our Bayesian model, we were able to reduce some of the uncertainty around this estimate.

This paper has theorized that increases in revenue sources not affiliated with Educational and General activities would have a relationship with the outcome variable general subsidy. However, this relationship would vary based on the perceived volatility of these sources. In creating this theory, we marry the concepts of background risk with that of the general subsidy. The decision to use net revenue variables in several cases was made in order to capture both the revenue and expenditure effects of these variables. Starting with net auxiliary revenue, we see that there is a negative estimated coefficient of -0.74. There is limited uncertainty around this coefficient estimate (standard error = 0.04) and lower and upper confidence intervals of -0.82 and -0.67 respectively. This result is expected as auxiliary expenses and revenues generally net out to zero. Auxiliary revenue is also tied to student consumption activity. A larger general subsidy would perhaps lead to less out of pocket spend on auxiliary services at the institution by students. Net hospital revenue shows a positive relationship with the outcome variable with an estimated coefficient parameter of 0.60 (standard error = 0.01). There is limited uncertainty around this parameter as well, giving us increased confidence in the fact that a relationship may exist. This is an expected result as hospital revenues tend to have stable revenue sources. I would imagine that the estimated coefficient parameter is not higher because so many institutions do not generate hospital revenues. The estimated coefficient parameter on net other revenue is 0.43 (standard error = 0.01). There is again limited uncertainty around this estimated parameter, giving us more confidence in the positive relationship. The magnitude of this revenue source parameter estimate is smaller because of the volatility in this measure. Net other revenue encompasses many different revenue sources on an annual basis, and therefore can bounce around quite a bit. One would expect that there is some positive correlation between net other revenue and general subsidy, however one would not expect this relationship to be stronger than is grants/appropriations. The coefficient parameter estimate on grants/appropriations has the largest magnitude. This is to be expected as appropriation and grant resources tend to have low volatility. Most grants and appropriations can span multiple years, and the organizations that issue the tend to only change their policies with sufficient warning. In another sense, this is one of the revenue sources that an institution should be able to forecast with some amount of certainty.

The results seem to match well with the proposed hypothesis. Generally speaking, increased revenue led to an increase in the general subsidy that a school offers. The source of that increased revenue does seem to have an impact on the magnitude of the relationship. As discussed previously, revenue sources that have less background risk associated with them (i.e. they are more stable and predictable) have a larger effect on general subsidy. This is highlighted the most by the two ends of the range of coefficient parameters. Auxiliary revenue and other revenue sources tend to be more volatile than other revenue sources. As a result, the magnitude of other revenue sources coefficient parameter is less than that of

grants/appropriations, hospital revenue, and gifts. Auxiliary revenue, which is dependent on more unknown parameters than the other predictors in our model, has a negative relationship with general subsidy. While we explored the specifics of this relationship already, the point remains the same: volatile revenue sources have less (to a negative) impact on general subsidy than more stable revenue sources.