ISYE 6420: Project

Spring 2022

Cale Williams

PIE à la Bayesian Logistic Regression

Application

The aim of this work is to model the relationship between an NBA team winning a game and a player's Player Impact Estimate (PIE) using Bayesian logistic regression. The NBA's PIE statistic "measures a player's overall statistical contribution against the total statistics in games they play in"[1]. It is given by:

$$\text{PIE} = \frac{\sum_i^l x_i - \sum_j^m x_j + \frac{1}{2}\sum_k^n x_k}{\sum_i^l X_i - \sum_j^m X_j + \frac{1}{2}\sum_k^n X_k}$$

where $x =$ player's statistic, $X =$ game's statistic (only during the player's minutes) and $i \in I$, $j \in J$, $k \in K$ for the following:

I	J	K
PTS	FGA	OREB
FGM	FTA	BLK
FTM	PF	
DREB	TO	
AST		
STL		

The statistics that make up I, J, and K are unimportant for the purposes of this investigation but can be found in the NBA glossary[1]. PIE includes both favorable and detrimental statistics. Note that $I, J, K \in \mathbb{Z}^{0+}$, i.e. the statistics are all integers ≥ 0. PIE $\in \mathbb{R}$, i.e. PIE can be a positive or negative decimal value or 0.

PIE is a ratio of an individual player's basic box score statistics and the game's box score statistics. Therefore, a reasonable hypothesis is that if a player owns a disproportionate share of the favorable statistics in a game relative to the other team's players on the court, his team is more likely to win.

For thoroughness, below is an example PIE calculation for former Yellow Jacket and current New Orleans Pelican, Jose Alvarado, for the Pelicans' game against the Golden State Warriors on November 5, 2021[2].

Set	Statistic	J. Alvarado	Game
I	PTS	5	25
I	FGM	2	10
I	FTM	0	3
I	DREB	0	11
I	AST	1	7

Set	Statistic	J. Alvarado	Game
I	STL	1	2
J	FGA	4	24
J	FTA	0	4
J	PF	0	5
J	TO	0	2
K	OREB	0	3
K	BLK	0	1

$$\begin{aligned}\text{PIE}_{J.Alvarado} &= \frac{(5+2+0+0+1+1)-(4+0+0+0)+\frac{1}{2}(0+0)}{(25+10+3+11+7+2)-(24+4+5+2)+\frac{1}{2}(3+1)} \\ &= \frac{5}{25} \\ &= 20\%\end{aligned}$$

With 9 other players on the court with Alvarado, a "fair share" of the PIE would be $\frac{1}{10} = 0.1$. Alvarado owns more, which is good for the Pelicans, per our hypothesis.

Logistic Regression

The game's outcome, Y, is the response variable:

$$Y = \begin{cases} 1 & \text{if Win} \\ 0 & \text{if Loss} \end{cases}$$

The probability of success, p, given the predictor variable, X_{PIE}, is defined using a link function:

$$p(x_{\text{PIE}}) = \Pr(Y = 1|x_{\text{PIE}}) = \frac{\exp(\beta_0 + \beta_1 x_{\text{PIE}})}{1 + \exp(\beta_0 + \beta_1 x_{\text{PIE}})}$$

where β_0 is the intercept of the log-odds and β_1 is the change in log-odds for an increase of one unit in PIE. The odds of a win ($Y = 1$) at a specified x_{PIE} is:

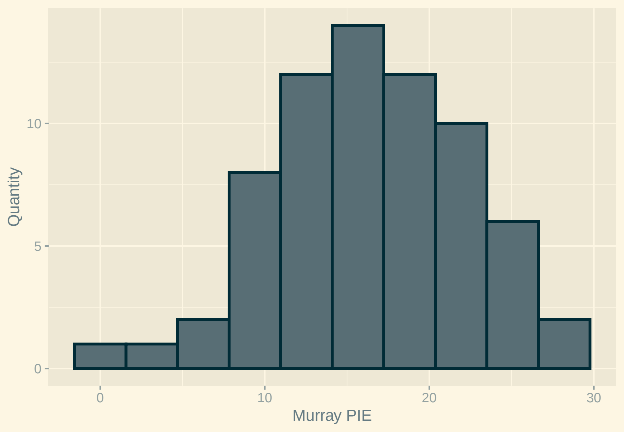
$$\frac{p(x_{\text{PIE}})}{1 - p(x_{\text{PIE}})} = e^{\beta_0 + \beta_1 x_{\text{PIE}}}$$

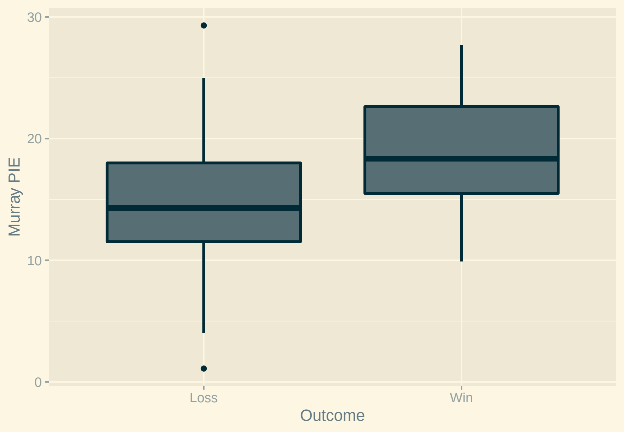
To build this regression model, this project uses the 2021-22 season of San Antonio Spurs guard, Dejounte Murray[3]. The goal is to relate a game's outcome with Murray's PIE in that game. Below are a few rows of the dataset used:

Game	Win/Loss	PIE [%]
1	W	15.8
2	L	9.8
3	L	9.5
⋮	⋮	⋮
80	L	Inactive
81	L	Inactive

Game	Win/Loss	PIE [%]
82	L	8.7

Removing the games in which Murray did not play yields 68 data points. Below is the distribution of Murray's PIE values along with his PIE grouped by game outcome. Note the potential outliers in the loss group.





In order to evaluate the regression model, 20% of the dataset is randomly selected as the testing set. The remaining 80% is used as the training set.

Bayesian

The Spurs won 4 of the 14 games Murray did not play in, which is a win percentage of $\frac{4}{14} = 0.286$. As the intercept, β_0 is the log-odds of winning for PIE $= 0$. Therefore, it is rational to set the mean of the prior on β_0 equal to the log-odds of the win percentage without Murray, since he neither added nor subtracted PIE in those games.

$$\log\left(\frac{0.286}{1-0.286}\right) = -0.916$$

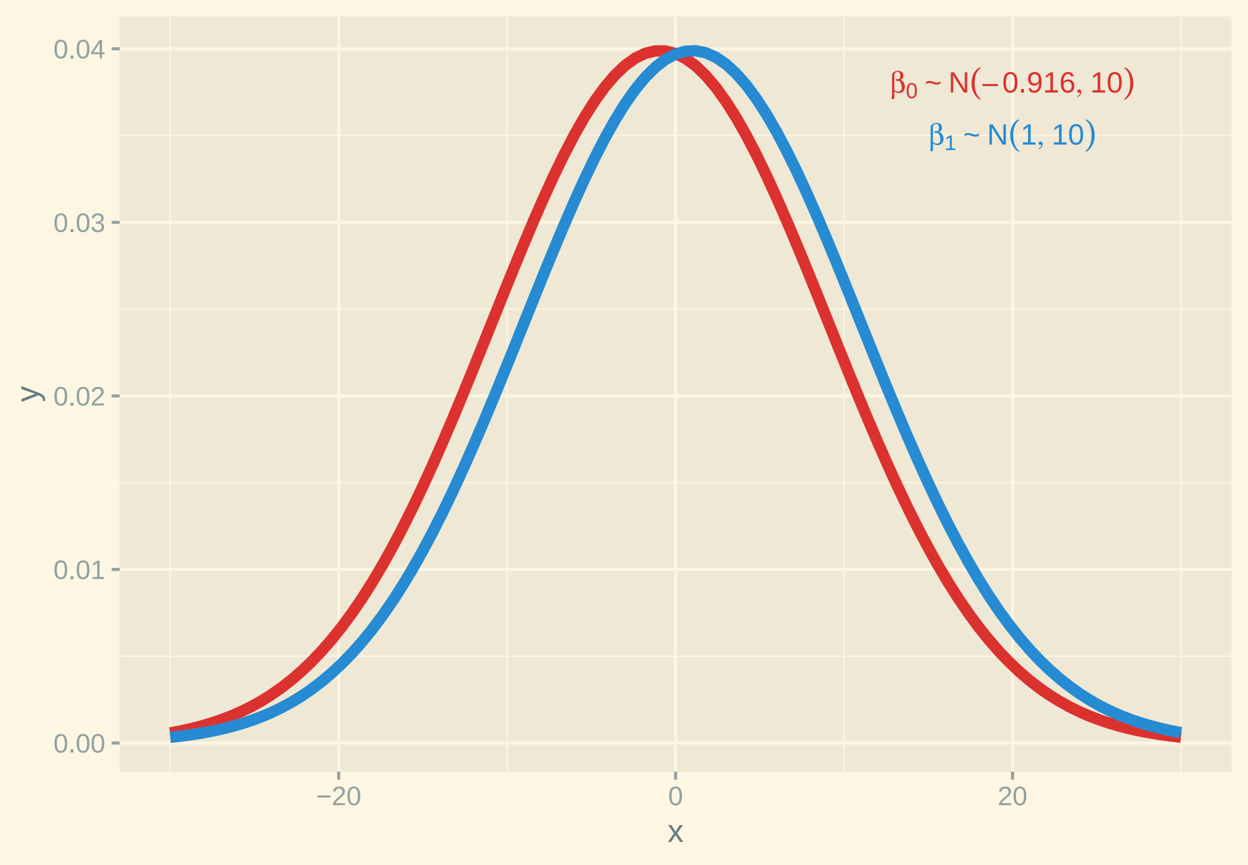
Trials were performed with varying standard deviations without significant changes in the results of β_0 or β_1, so the standard deviations are set to 10.

$$\beta_0 \sim \mathcal{N}(-0.916, 10)$$

Anticipating that the PIE for a player who plays the entire game is directly related to the probability of winning, the prior on β_1 is set as positive. However, because players don't play all 48 minutes (Murray averaged 34.8 minutes/game), the team's overall performance, and thus the outcome of the game, could be drastically different than would be reflected in Murray's PIE. So the prior on β_1 is not limited to positive values that a Gamma or Beta distribution would provide, but rather is given a Normal distribution with a slight shift positive:

$$\beta_1 \sim \mathcal{N}(1, 10)$$

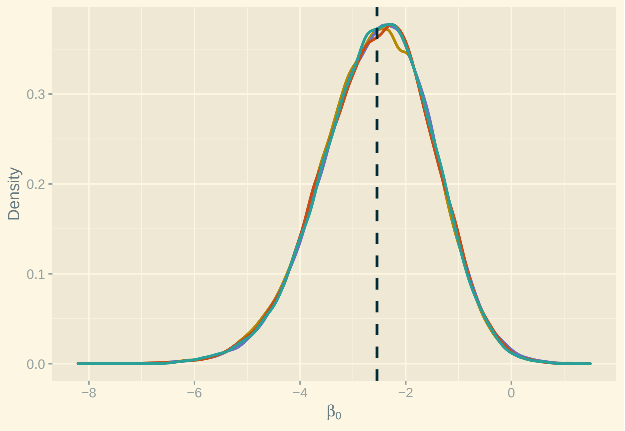
The probability density functions are below.

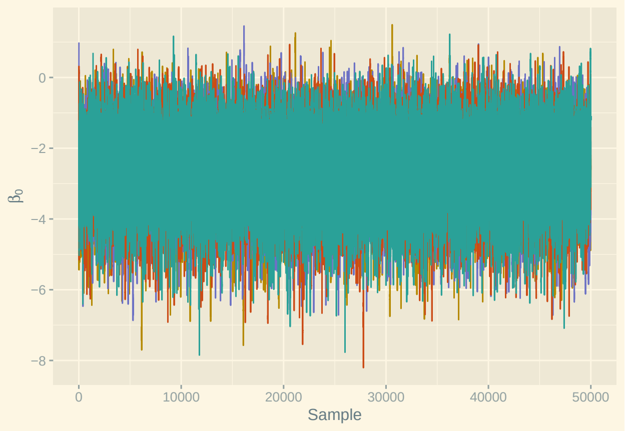


The likelihood is binomial with the link:

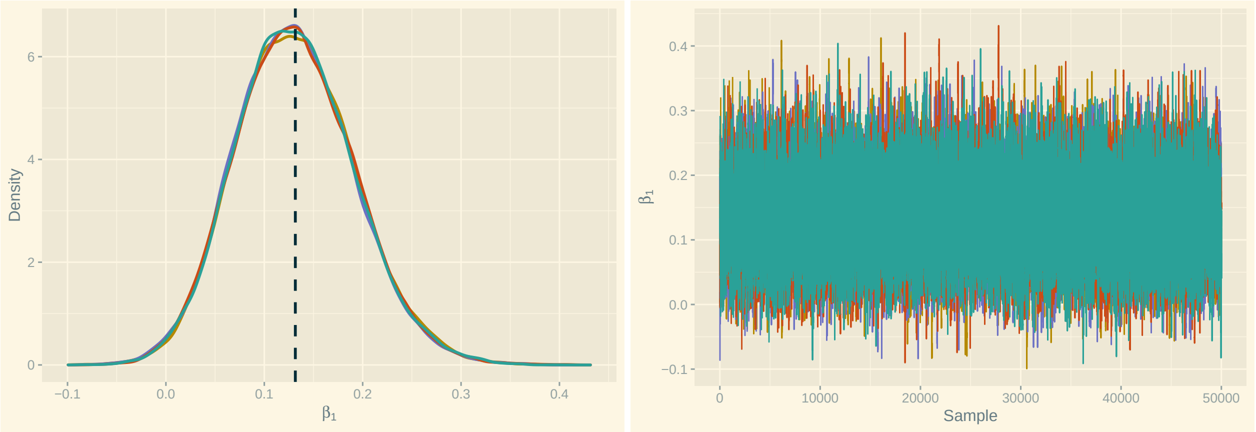
$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{\text{PIE}}$$

4 chains were run via Python's PyMC package with 50,000 samples each and 2,000 tuned. Density and trace plots are below.





For β_0, the chains converge as $\hat{R} = 1$. The posterior mean of β_0 is $\mu_{\beta_0} = -2.544$ with a 95% credible set of $(-4.628, -0.538)$. Because this set does not include 0, β_0 is assumed to have explanatory power on the outcome of games, Y.



For β_1, the chains also converge as $\hat{R} = 1$. The posterior mean of β_1 is $\mu_{\beta_1} = 0.132$, implying that for a 1 unit increase in PIE, the odds of winning increases by $e^{0.132} = 1.141 \Rightarrow 14\%$. The 95% credible set for β_1 is $(0.018, 0.253)$. Because this set does not include 0, β_1 is assumed to have explanatory power on the outcome of games, Y. Thus, Bayesian logistic regression returns a probability of winning equation of:

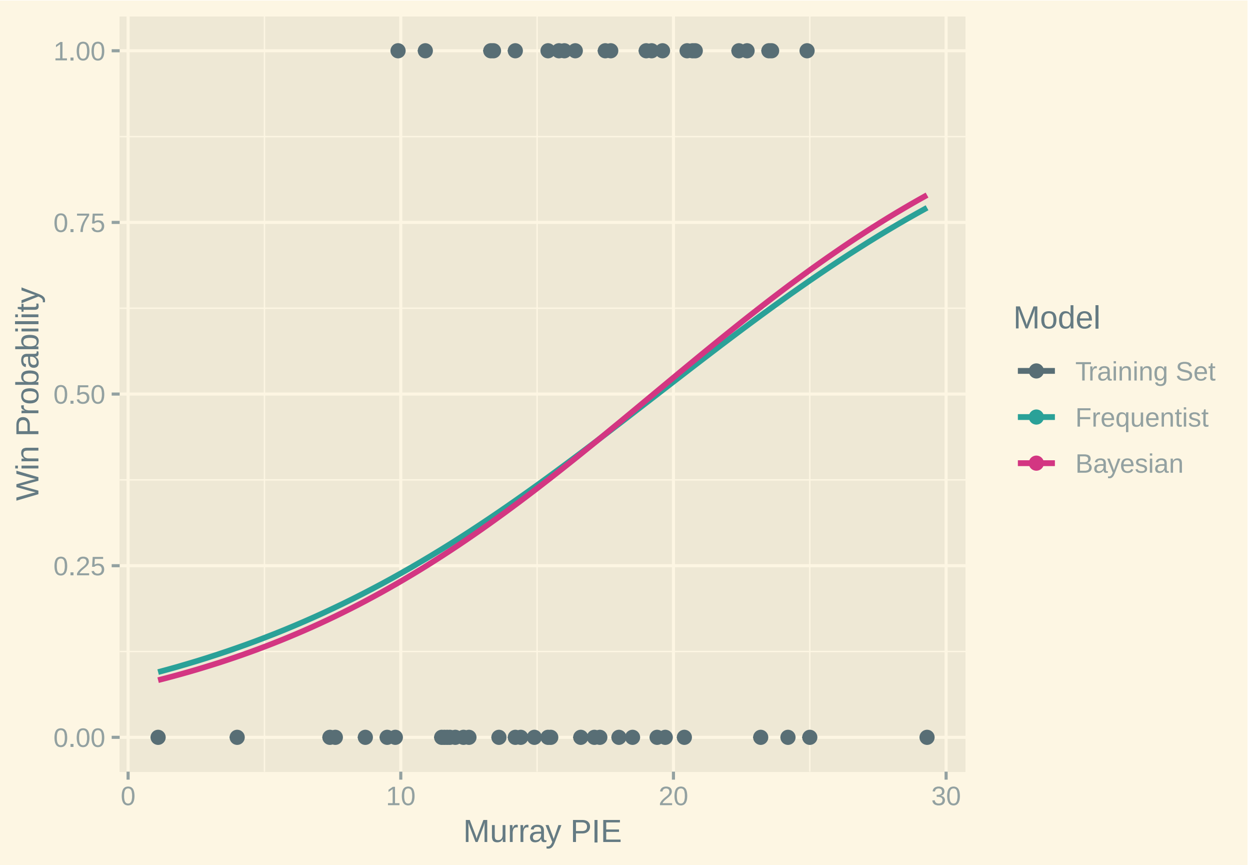
$$p = \frac{\exp(-2.544 + 0.132x_{\mathrm{PIE}})}{1 + \exp(-2.544 + 0.132x_{\mathrm{PIE}})}$$

Frequentist

As a comparison, a frequentist approach is also employed. A logistic regression model is fit on the training data via R's *glm* function. The resulting probability of success is given by:

$$p = \frac{\exp(-2.389 + 0.123x_{\mathrm{PIE}})}{1 + \exp(-2.389 + 0.123x_{\mathrm{PIE}})}$$

The coefficients are similar to the Bayesian model due to the relatively uninformative priors set in the previous analysis. The training set and resulting regression functions are shown below.

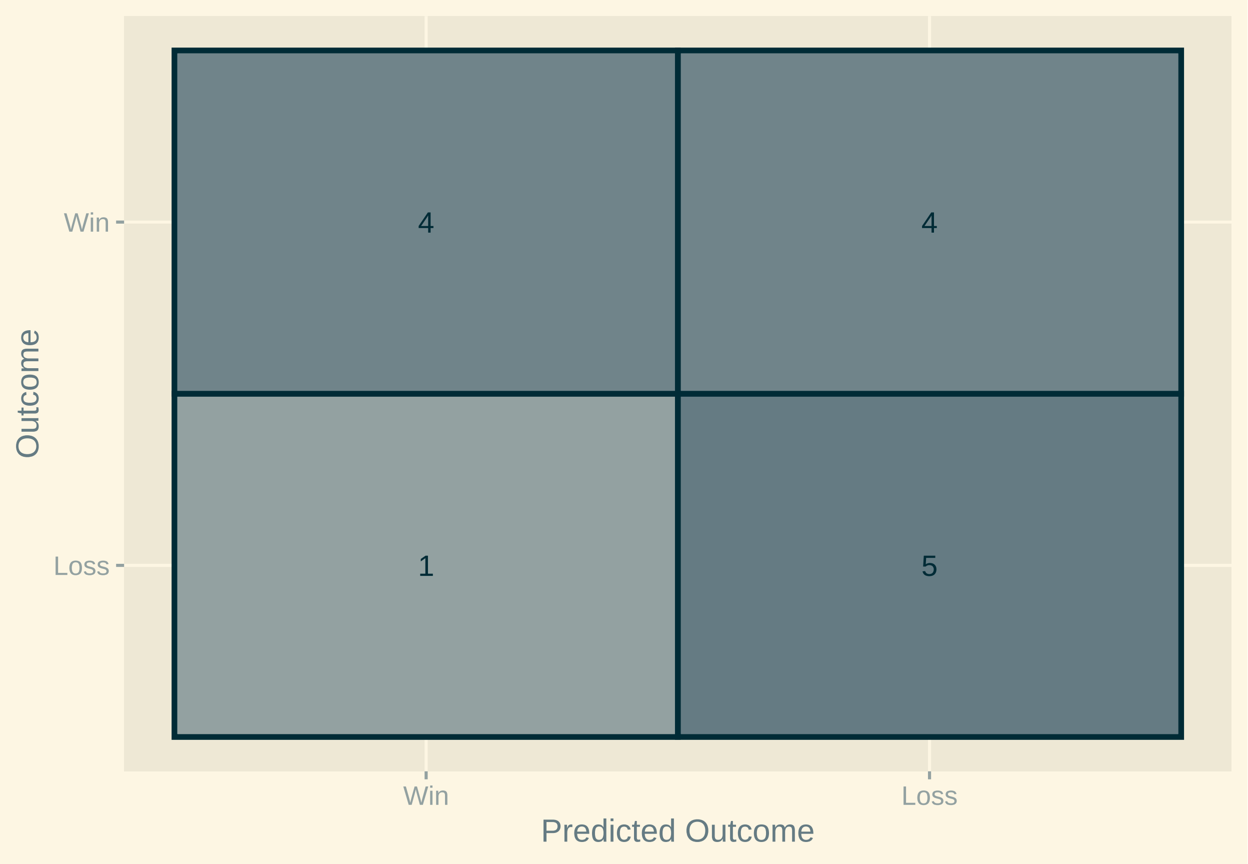


Prediction

Using the Bayesian model, predictions are made on the test set by averaging all of the outcomes (0s and 1s) and rounding to 0 or 1. In the plot above, the Bayesian regression curve crosses the $p = 0.5$ line at PIE $\approx$ 20. That inflection point is evident in the predictions below.

Game	Win/Loss	Outcome	Murray PIE	Mean Prediction	Prediction Std. Dev.	Predicted Outcome
27	L	0	9.7	0.231	0.422	0
28	W	1	11.8	0.280	0.449	0
30	W	1	25.2	0.671	0.470	1
31	W	1	10.6	0.250	0.433	0
34	L	0	14.1	0.339	0.473	0
36	L	0	18.0	0.457	0.498	0
38	W	1	16.8	0.420	0.494	0
40	W	1	25.8	0.686	0.464	1
47	W	1	17.2	0.433	0.495	0
49	W	1	27.7	0.729	0.445	1
55	L	0	21.8	0.579	0.494	1
56	L	0	17.3	0.434	0.496	0
62	L	0	10.7	0.252	0.434	0
64	W	1	22.9	0.610	0.488	1

The binary Bayesian predictions above do not differ from the frequentist predictions. The confusion matrix is:



The sensitivity and specificity are respectively:

$$\text{sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{4}{4+4} = 0.50$$

$$\text{specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} = \frac{5}{5+1} = 0.83$$

Conclusion

The goal of this report is to build a logistic regression model in a Bayesian manner regressing the outcome of an NBA game on a single player's PIE. This is achieved using Dejounte Murray's 2021-22 season. The analysis could be generalized for any player by simply updating the prior on β_0.

Future work could include more informative priors, hyperparameters on the priors, and more predictors to create a multiple logistic regression model.

All statistics are from nba.com/stats.

Appendix

[1] https://www.nba.com/stats/help/glossary/

[2] https://www.nba.com/game/nop-vs-gsw-0022100130/box-score?type=advanced

[3] https://go.nba.com/ddcfd

Accompanying Files

- *DMurray_PIE.xlsx*
  - Microsoft Excel spreadsheet.
  - Dataset of Dejounte Murray's PIE in the 2021-22 season.
- *Project-Williams.py*
  - Python script.
  - Defines priors, likelihood, and samples from posterior via MCMC.
- *post.csv*
  - Microsoft Excel Comma Separated Values spreadsheet.
  - Summary of posterior on regression coefficients.
- *predictions.csv*
  - Microsoft Excel Comma Separated Values spreadsheet.
  - Summary of predictions on test set.
- *Project-Williams.Rmd*
  - R Markdown file.
  - Report write-up with code for plots.