1 Introduction

In this project, we are attempting to determine how to predict which players will perform best in a fantasy football season based on certain underlying statistics. The dataset we will use is an amalgamation of two sets of statistics: "Next Gen Stats" powered by AWS and Pro Football Reference. The Next Gen Stats has underlying statistics about football players from the 2019 season, while Pro Football Reference has Fantasy Points statistics.

The goal of this analysis is to determine which and to what degree underlying variables contribute to total fantasy points scored during the 2019 season. The deviance information criterion (DIC) will guide which models best explain the relationship between these underlying variables and total fantasy points.

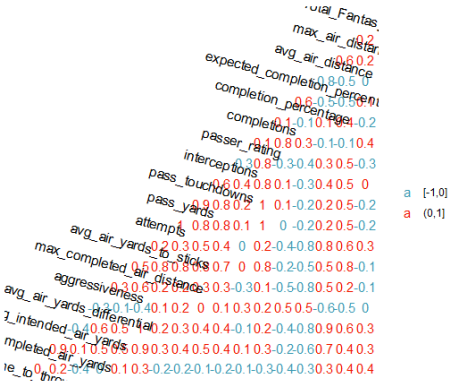
Since the response variable (total fantasy points) is a continuous variable, and we want to see how these explanatory variables affect the response variable, Bayesian Linear Regression is a sensible model to use. Because football statistics differ between different player positions, our analysis will be repeated for three different player classes: Quarterbacks, Running Backs, and Receivers (Wide Receivers and Tight Ends).

Due to the 5 page constraint, this document gives the highlights, while skipping over many details. The detailed analysis is done in the RMarkdown files (and corresponding knitted PDF).

2 Quarterbacks

2.1 Data Exploration

We first want to get an understanding of how different Next Gen Stat variables are correlated to Total Fantasy Points. After some data cleanup, we plot the correlations, where the far right column shows how each variable is correlated with Total Fantasy Points (our response variable):



2.2 Model Fitting

2.2.1 Approach 1: Bayesian Linear Regression on Positively-Correlated Variables

Based on this, we create a model that includes only the variables that are positively correlated with Total Fantasy Points (i.e. where the far right column is red):

$$Total.Fantasy.Points_i \sim (\mu_i, \tau)$$

$$\begin{aligned}\mu_i = \beta_1 + \beta_2 * avg.time.to.throw + \beta_3 * avg.completed.air.yds + \beta_4 \\ * avg.intended.air.yds + \beta_5 * avg.air.yds.to.sticks + \beta_6 * pass.rating \\ + \beta_7 * completion.percent + \beta_8 * avg.air.dist + \beta_9 * max.air.dist\end{aligned}$$

$$\beta_i \sim N(0, 0.001)$$

$$\tau \sim Ga(0.01, 0.01)$$

When sampling from this posterior 50,000 times, we get the below credible sets:

	2.50%	25%	50%	75%	97.50%
b[1]	-60.592	-19.3447	1.712	22.978	63.41
b[2]	-44.627	-3.8577	16.421	36.946	76.98
b[3]	-35.788	-7.9053	6.57	20.91	47.66
b[4]	-38.739	-8.2659	9.021	27	61.58
b[5]	-50.398	-20.4081	-4.498	11.087	41.32
b[6]	1.49	5.013	6.585	8.21	11.21
b[7]	-19.413	-12.7455	-9.006	-5.266	2.05
b[8]	-36.824	-15.3911	-3.957	7.59	29.76
b[9]	-6.567	-0.8384	2.072	5.083	10.94

From this, we see b[6] (passer rating) is the only variable whose 95% credible set does not contain 0. So we are confident of its positive impact on total fantasy points.

2.2.2 Approach 2: Bayesian Linear Regression with LASSO-selected Variables

In our next approach, we'll again perform linear regression, but this time the scaled variables will be selected via LASSO regression (elastic net).

After scaling and selecting variables with a non-zero impact, we get the below variables as being significant. There is a lot of overlap with the first model, with interceptions replacing completion percentage:

$$Total.Fantasy.Points_i \sim (\mu_i, \tau)$$

$$\mu_i = \beta_1 + \beta_2 * avg.time.to.throw + \beta_3 * avg.completed.air.yds + \beta_4 * avg.intended.air.yds + \beta_5 * avg.air.yds.to.sticks + \beta_6 * pass.rating + \beta_7 * interceptions + \beta_8 * avg.air.dist + \beta_9 * max.air.dist$$

$$\beta_i \sim N(0, 0.001)$$
$$\tau \sim Ga(0.01, 0.01)$$

When sampling from this posterior 50,000 times, we get the below parameter credible sets:

	2.50%	25%	50%	75%	97.50%
b[1]	-60.651	-18.9636	2.52575	23.746	64.313
b[2]	-42.864	-2.6034	17.73639	38.182	76.458
b[3]	-27.799	-0.7969	13.01755	26.714	52.569
b[4]	-46.202	-16.1804	-0.07492	15.929	46.817
b[5]	-38.469	-5.6421	11.89112	29.587	60.519
b[6]	0.175	3.4911	5.2453	7.054	10.658
b[7]	-18.827	-11.2237	-7.26163	-3.477	3.285
b[8]	-24	-8.2853	0.60783	9.639	26.386
b[9]	-46.438	-29.4829	-20.7131	-11.779	6.679

From this, we see b[6] (passer rating) is the only variable whose 95% credible set does not contain 0. So we are confident of its positive impact on total fantasy points. We could look at the 50% Credible Set to see that b[3] is almost always positive and b[7] and b[9] is always negative, suggesting the positive impact of average completed air yards and the negative impact of interceptions and max air distance.

2.2.3 Approach 3: Bayesian Linear Regression with Stochastic Search-selected Variables

Finally, we'll use Stochastic Search Variable Selection (SSVS) to find which model is most frequently visited by the Gibbs sampler. Because of the large number of covariates (and thus possible models), I had to reduce this sampling to 1000 iterations to prevent timing out.

Using the same scaled data as model 2, we get the below model as being the most likely:

$$Total.Fantasy.Points_i \sim (\mu_i, \tau)$$

$$\mu_i = \beta_1 + \beta_2 * passer.rating + \beta_3 * completion.percentage$$

$$\beta_i \sim N(0, 0.001)$$
$$\tau \sim Ga(0.01, 0.01)$$

When sampling from this posterior 50,000 times, we get the below parameter estimates and credible sets:

	2.50%	25%	50%	75%	97.50%
b[1]	-52.848	-12.578	8.858	30.03	70.3774
b[2]	2.174	5.283	6.693	8.106	10.9033
b[3]	-12.558	-8.466	-6.39	-4.293	0.3241

2.3 Model Comparison

We use the DIC to evaluate which model is best. When doing this, we see model 2 has the lowest mean and penalized deviance, so is considered the best model among the three.

3 Running Backs

3.1 Model Fitting

Due to page limit constraints, we'll just explain the model with the lowest DIC (model 1):

3.1.1 Bayesian Linear Regression on Positively-Correlated Variables

$$Total.Fantasy.Points_i \sim (\mu_i, \tau)$$

$$\mu_i = \beta_1 + \beta_2 * exp.rush.yds + \beta_3 * rush.pct.over.exp + \beta_4 * rush.att + \beta_5 * rush.yds.over.exp + \beta_6 * avg.rush.yds + \beta_7 * rush.yds.over.exp.per.att$$

$$\beta_i \sim N(0, 0.001)$$

$$\tau \sim Ga(0.01, 0.01)$$

These are the quantiles for each parameter:

	2.50%	25%	50%	75%	97.50%
b[1]	-66.8742	-25.963	-4.906	15.821	55.652
b[2]	-0.3102	3.845	5.741	7.581	11.136
b[3]	-59.5748	-19.404	1.764	22.883	62.924
b[4]	-30.4903	-18.482	-12.602	-6.242	7.081
b[5]	-0.1309	2.687	4.152	5.62	8.426
b[6]	-37.5387	-13.768	-1.516	10.996	35.786
b[7]	-53.1314	-15.321	4.595	24.591	63.395

While none of the 95% CS's exclude 0, b[2] (expected rush yards) and b[5] (rush yards over expected) are close to being positive, suggesting a positive effect of those variables on total running back fantasy points.

4 Receivers

4.1 Model Fitting

Due to page limit constraints, we'll just explain the model with the lower DIC (model 3):

4.1.1 Bayesian Linear Regression with Stochastic Search-selected Variables

$$Total.Fantasy.Points_i \sim (\mu_i, \tau)$$

$$\mu_i = \beta_1 + \beta_2 * pct.share.intended.air.yards + \beta_3 * receptions$$

$$\beta_i \sim N(0, 0.001)$$
$$\tau \sim Ga(0.01, 0.01)$$

When sampling from this posterior 50,000 times, we get the below parameter estimates and credible sets:

	2.50%	25%	50%	75%	97.50%
b[1]	-62.5684	-40.49	-28.724	-16.748	6.34
b[2]	0.9042	1.466	1.753	2.036	2.575
b[3]	12.5527	16.258	18.104	19.94	23.471

From this, b[2] (pct.share.intended.air.yards) and b[3] (receptions) both have positive 95% Credible Sets, suggesting these having a clear positive impact on total fantasy points.

5 Conclusion

For Quarterbacks, passer rating is a strong positive predictor of total fantasy points, while average completed air yards is a weak positive predictor and interceptions and max air distance are weak negative predictors.

For Running Backs, expected rushing yards and rush yards over expected are weak positive predictors of total fantasy points.

For Receivers, receptions and percent share of intended air yards are strong positive predictors of total fantasy points.

Cross-validation was out of scope of this analysis, so these results possibly include overfitting. More analysis would need to be done to see the results on validation datasets to confirm the best model.