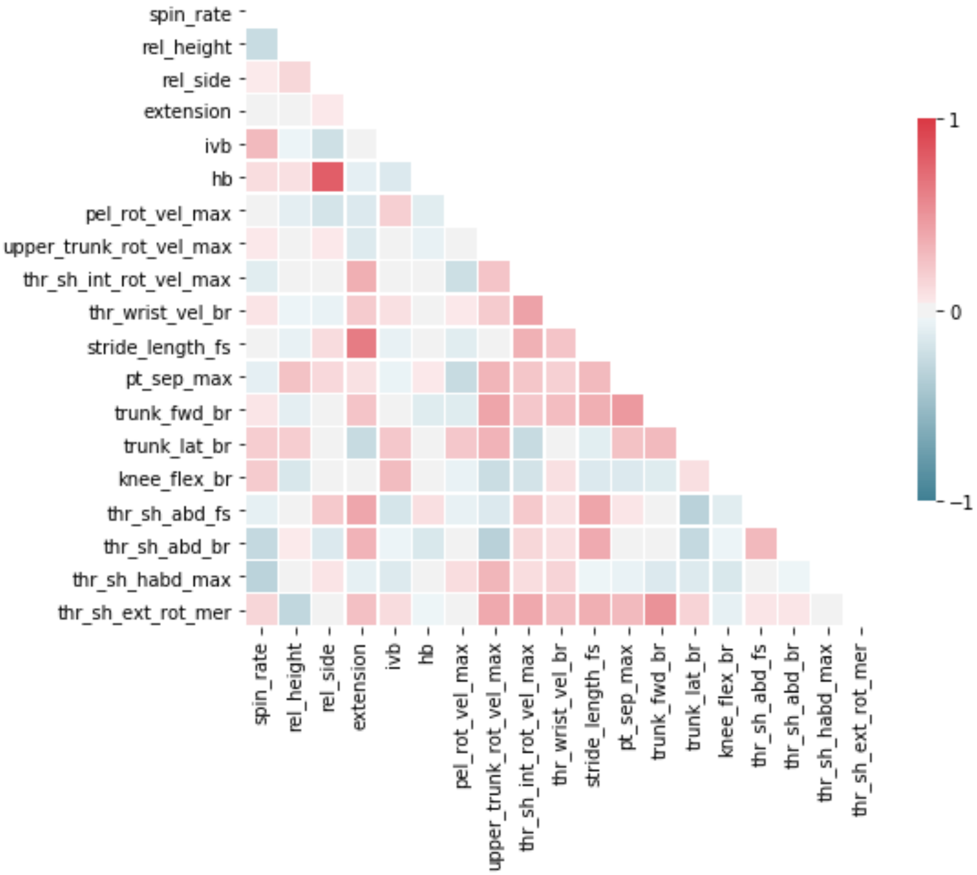
A Predictive Model of Fastball Velocity
Derek Bivona, PhD

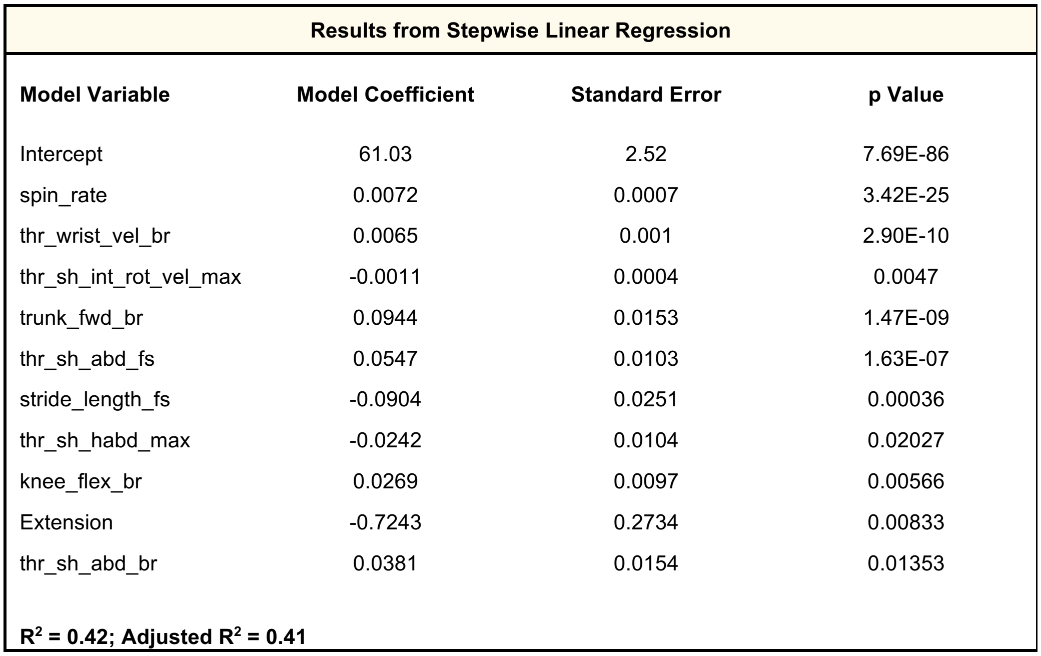
To build a model that predicts the velocity of a fastball, I followed the outline below:

1. **Impute Missing Data**: I imputed the missing data with the median value of the respective feature.

2. **Check Correlations**: I checked the correlation between all combinations of the features (in the heatmap below) before inputting them into a linear regression model to identify collinearity and eliminate unnecessary variables. Of all the variables, ‘hb’ and ‘rel_side’ were the most highly correlated pair and had a correlation coefficient of 0.801. I removed ‘hb’ since it has more missing entries than ‘rel_side.’ This correlation is reasonable as fastballs usually have little horizontal break, so the horizontal location of the ball will be correlated with where it is released. Of course, this is not true for cut fastballs or 2-seamers with run.



3. **Run Stepwise Linear Regression**: I ran a stepwise linear regression to elucidate the feature importance in predicting fastball velocity. I present my results in the table below:



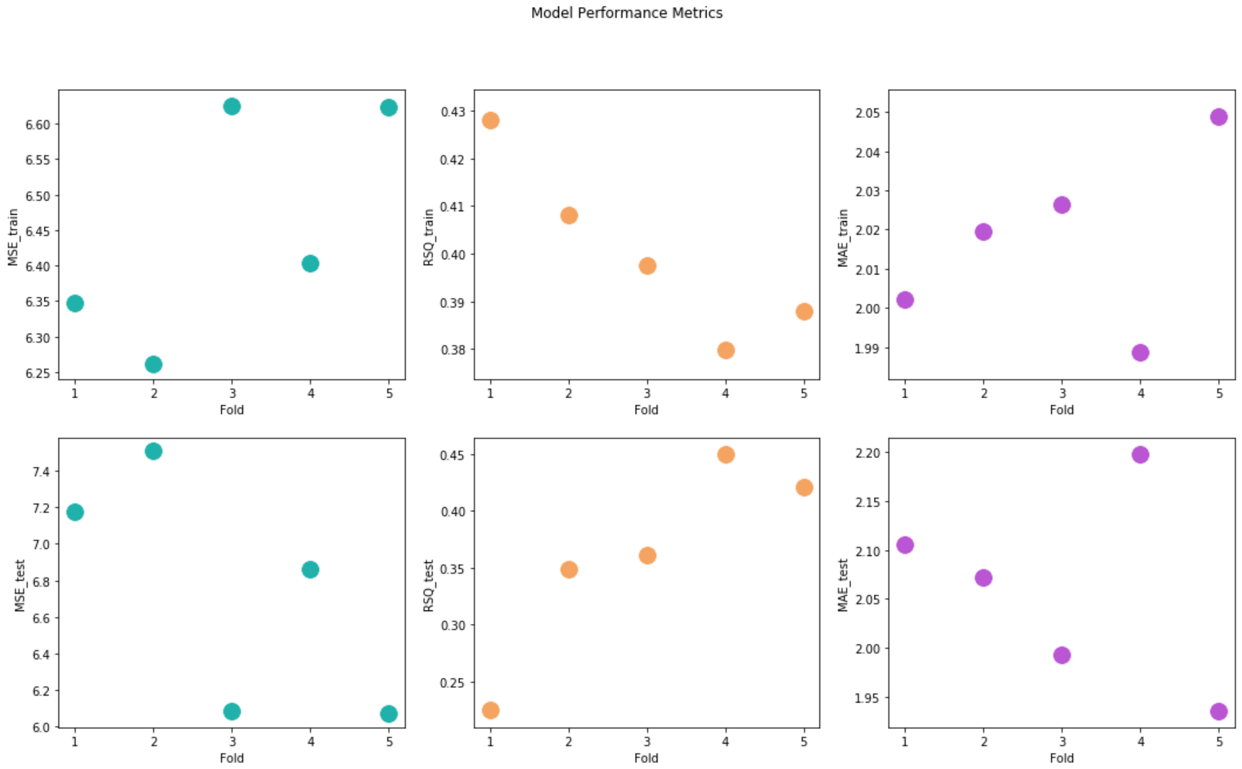
Results from Stepwise Linear Regression

Model Variable	Model Coefficient	Standard Error	p Value
Intercept	61.03	2.52	7.69E-86
spin_rate	0.0072	0.0007	3.42E-25
thr_wrist_vel_br	0.0065	0.001	2.90E-10
thr_sh_int_rot_vel_max	-0.0011	0.0004	0.0047
trunk_fwd_br	0.0944	0.0153	1.47E-09
thr_sh_abd_fs	0.0547	0.0103	1.63E-07
stride_length_fs	-0.0904	0.0251	0.00036
thr_sh_habd_max	-0.0242	0.0104	0.02027
knee_flex_br	0.0269	0.0097	0.00566
Extension	-0.7243	0.2734	0.00833
thr_sh_abd_br	0.0381	0.0154	0.01353

R^2 = 0.42; Adjusted R^2 = 0.41

Spin rate is the best predictor of fastball velocity, yet it alone cannot predict velocity.

4. **Run Linear Regression Models with Cross-Validation for Evaluation:** The best predictors from the previous step were ('spin_rate', 'thr_wrist_vel_br’, ’stride_length_fs’, ’trunk_fwd_br’, ’thr_sh_int_rot_vel_max', 'thr_sh_abd_fs’, ’thr_sh_habd_max'), and these were included in the final linear regression model. To evaluate the performance of the model with these features as the independent variables and the ‘velo’ as the dependent variable, I used 5-fold cross-validation while examining the mean squared error (MSE), r-squared value (RSQ), and mean absolute error (MAE) for both the training and testing cases for each fold. My results are shown in the figure below:



Even though the r-squared values are between ~ 0.3 and 0.5, the mean absolute errors range from 1.8 and 2.2, meaning (loosely) on average, the linear regression model can predict a fastball within ± 2 mph. These conclusions are true for both the training and testing cases.