# Project Luther Summary
# Predicting High School Track Results

Dana Lindquist
October 12, 2018

## Project Design

This project was motivated by the large quantity of data available on track and field (and cross country) meets on the web site www.athletic.net. The web site has lists of athletes competing in all events as well as detailed information for each athlete. For this project the focus was on Washington State high school athletes. The web site has full high school data for athletes graduating between 2006 and 2017.

The goal was to predict future performance of an athlete based on past performance. Data was pulled from the web site for girls and boys competing in the 400 meter and 1600 meter events. For each of the 4 high school years (9th, 10th, 11th and 12th grades) the personal record (PR) for these events were compiled. Only athletes who had competition results for all four years were included in the data set. More data points were collected for the 1600 race because for most competitions any number of athletes can enter this event. For the 400m races the entrants are limited due to a limited number of lanes on the track.

The data set included:

| | # girls | # boys | Total | # Schools |
|---|---|---|---|---|
| 400 meters | 382 | 444 | 826 | 186 |
| 1600 meters | 852 | 1204 | 2056 | 198 |

## Tools

- Pandas
- Numpy
- Statsmodels
- Patsy
- Scikit Learn
- Beautiful Soup
- Yellowbrick
- Matplotlib

# Data

For each athlete the following data was collected.

| Variable | Type | Description | Use in Model |
|---|---|---|---|
| Athlete's Name | str | Record descriptor | N |
| Athlete's ID number | int | Record descriptor from athletic.net | N |
| Athlete's School | Int | Each school will be given an integer number | N |
| Athletic District | int | 9 Districts in the state of Washington | Y |
| 9th Grade PR (Personal Record) | float | Best competition time for this school year | Y |
| 10th Grade PR (Personal Record) | float | | Y |
| 11th Grade PR (Personal Record) | float | | Y |
| 12th Grade PR (Personal Record) | float | | Target |

# Models

Race times from 9th, 10th and 11th grade were used to predict the 12th grade times.

An investigation of model complexity was performed to determine what level of complexity was required to

<u>Algebraic</u>

A simple examination of the data shows that plotting the ratio of growth from 10th grade to 11th grade over the growth from 11th grade to 12th grade is almost 1

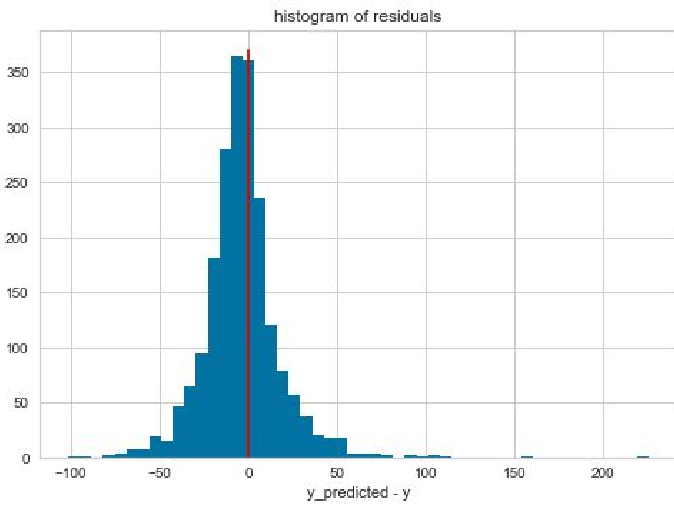
$$\frac{t_{12}}{t_{11}} / \frac{t_{11}}{t_{10}} \approx 1$$

From this knowledge a simple algebraic model can be tested to predict the 12th grade times. This model assumes this ratio of times is equal to 1. Or in other words:

$$t_{12}(predicted) = \frac{t_{11}^2}{t_{10}}$$

This model produces the following results

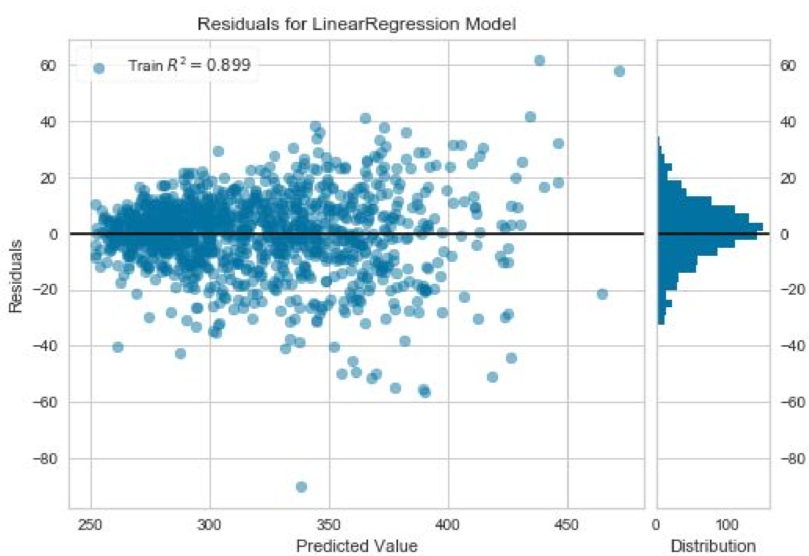
| | 400 meters RMSE (sec) | 1600 meters RMSE (sec) |
|---|---|---|
| Algebraic model | 3.70 (6.3% of ave) | 21.64 (6.8% of ave) |




This is pretty good but from the above plot of the residuals we are predicting lower times with this model then the data shows.

Linear Regression

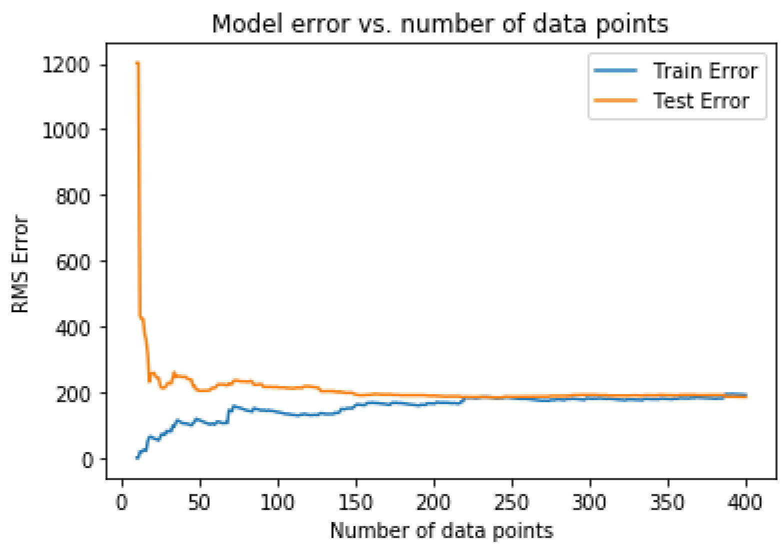
A simple linear regression model was fit to the data. One-hot encoding was used to account for the different districts.




From the above plot you can see that the residuals are more evenly distributed and the model is producing better results.

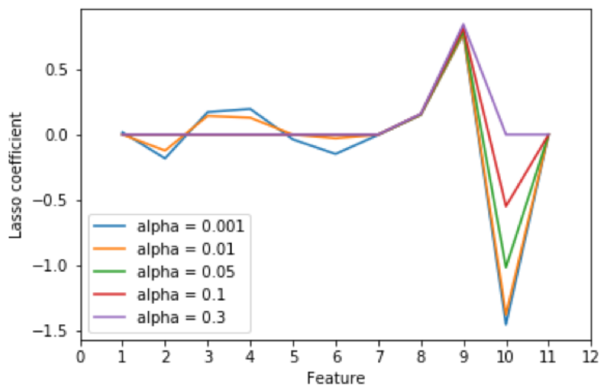
| | 400 meters RMSE (sec) | 1600 meters RMSE (sec) |
|---|---|---|
| Algebraic model | 3.70 (6.3% of ave) | 21.64 (6.8% of ave) |
| Linear Regression | 2.18 (3.7% of ave) | 13.54 (4.3% of ave) |

Further examination with this linear regression model shows that significantly less data then is available is required to train this model. For the 1600 meter race the effect of data set size is shown in the following plot. Of the 2000+ data points that are available, only 1/10 of this data is required to fit a model. The results are similar for the 400 meter race.

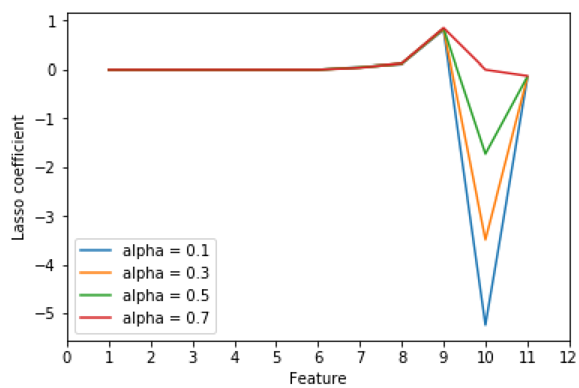



<u>Lasso</u>

The LASSO regularization model can be used to examine the importance of the features in the linear regression model.




400 meter feature coefficients




1600 meter feature coefficients

For the 400 meter race:

1. 11th grade PR
2. 10th grade PR
3. Sex
4. Athletic district

For the 1600 meter race:

1. 11th grade PR
2. Sex
3. Graduation year

This information indicates that the number of features can be reduced for the model to have similar accuracy.

<u>Other models</u>

Two other models were tried.  In the hope of picking up more of the taper off of times during the 12th grade year, higher order terms were added to the model.  Two features were added, one with the square of the 12th grade PR and the other with the square of the 11th grade PR.  This addition to the model did not improve the accuracy of the prediction.  The other model was a mixed effects regression which showed similar results as the linear regression.

| | 400 meters RMSE (sec) | 1600 meters RMSE (sec) |
|---|---|---|
| Algebraic model | 3.70 (6.3% of ave) | 21.64 (6.8% of ave) |
| Linear Regression | 2.18 (3.7% of ave) | 13.54 (4.3% of ave) |
| Adding Higher Order terms to Linear Regression | 2.26 | 13.42 |
| Mixed Effects Model | 2.14 | 13.59 |

## Conclusions and Future Extensions

The accuracy of these linear models was quite good and a definite improvement over a simple algebraic model.  But there was a limit to what the models could predict.  This should be improved by adding more features to the model.  The full event list for each athlete is available on www.athletic.net.  It would be interesting to add more details about the athlete's competitions within each school year.  It would also be interesting to add information on which other events in which the athlete competed.

This approach could easily be extended to an investigation of field events as well as running events.

A very relevant use of this type of predictive model of future performance is for athletic recruiting, for example colleges recruiting from high schools.  The work here shows merit and indicates the potential of extending this work to the prediction of college performance based on high school performance.