Predicting Buggy Freeroll Times from Multiple Parameters Using Multivariate Linear Regression

Isaiah Asah, Lameck Beni, Nate Klein, Lawrence Onyango

Overview

The goal of this project was to use multivariate linear regression (MVR) to predict freeroll times for buggies from a variety of easy to collect parameters. For the proof-of-concept demonstrated in this project, the parameters we used were air temperature, freeroll initial velocity, and buggy-driver combined weight. Using MVR, we were able to generate an equation relating each of our input variables to freeroll time with an R^2 value of 0.8.

Motivation

Buggy is one of the longest-standing traditions at CMU. In the modern era of buggy, times are close enough that teams are looking for ways to shave off every fraction of a second possible. Roughly half of every race is spent in the freeroll, so understanding what makes buggies roll faster is a focal point of buggy every year. Recently, data collection has become an important part of optimizing buggy and understanding ways to improve. We aim to create a proof of concept for a model that helps us understand the individual impacts of different factors on freeroll times. While the data and variables used in this project are not necessarily the most impactful or interesting, as data collection becomes more advanced and more data becomes available, the same methods will be applicable in finding more interesting results.

Data Collection

All of our data was collected from one buggy team's (CIA) rolls during the spring semester. Rolls in which there were outside factors significantly affecting freeroll time were excluded. These factors included bags attached to the push bar to reduce speed (required for the first few rolls each semester), mechanical failures which caused the buggies to slow down, or passing tests that significantly deviated from the normal driving line.

Freeroll times were measured using frame-by-frame analysis of footage as the buggy passed certain landmarks on the course: the crosswalk at the top of the hill for the start of the freeroll, and the construction tar line near Scaife Hall for the bottom of the hill.

Start of freeroll

End of freeroll

Initial velocity data was collected in a similar fashion using frame-by-frame analysis. In order to get accurate data, brightly colored lines were spray painted at known distances from the crosswalk: an orange line at the crosswalk, a blue line 20 ft from the first line, and another orange line at another 20 ft from there. Using the number of frames between the buggy crossing each line, the initial velocity was calculated via the following equation:

$$v_i = \frac{20\,ft}{frames\ between\ lines\ /\ 60\ frames\ per\ second}$$

Buggy weights were collected by weighing the buggies on a bathroom scale, and driver weights were self-reported. T air temperature was collected from timeanddate.com[1], which provides historical weather data.

Multivariate Linear Regression

Multivariate linear regression was used to find an equation relating our independent variables (total buggy weight, air temperature, and initial velocity) to our dependent variable (free roll time). This numerical method analyzes the relationship between our independent variables and dependent variable, assuming a linear relationship between these variables. Implementing this technique allows us to find the coefficients of the independent variables, which characterizes the best-fit line equation that represents the relationship between the independent and dependent variables.

Similarly to simple, one-dimensional linear regression, multivariate linear regression uses least squares regression to find these regression coefficients, though the process of finding these coefficients is more complicated due to having more than one independent variable. Least squares regression involves minimizing the sum of squared errors, with the error being the actual value of the dependent variable subtracted by the predicted value calculated using the regression coefficients. With multiple independent variables, this process of minimizing the sum of squared errors can be simplified by using matrices to perform our calculations. In particular, we would create an n length vector for our dependent variable y, where n is the number of data points. We would also need an n by $(k+1)$ matrix for our dependent variables x_k, where k is the number of dependent variables. We need an extra column where the value of every element is "1" as a part of this matrix in order to also find the intercept of our best-fit line. Lastly, we need an *n* length vector for our regression coefficients *b*. With these matrices, we can find the values of the elements in the *b* vector, which would be our regression coefficients, by following the general algorithm for minimizing the sum of least squares, though simplified by using some matrix operations.

The matrix math performed to arrive at the algorithm used for multivariate linear regression is as follows[2]:

$$\sum e_i^2 = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e'e$$

The sum of squared errors can be expressed simply due to matrix properties

$$\begin{aligned} e'e &= (y - Xb)'(y - Xb) \\ &= y'y - b'X'y - y'Xb + b'X'Xb \\ &= y'y - 2b'X'y + b'X'Xb \end{aligned}$$

$e = y - Xb$ so we can expand our previous equation to include the regression coefficients and the independent and dependent variable data

$$\frac{\partial e'e}{\partial b} = -2X'y + 2X'Xb = 0$$

As is typical for minimization problems, we take the derivative of the equation with respect to *b* and set the equation equal to zero to find the *b* that results in a minimum sum of squared errors.

$$b = (X'X)^{-1}X'y$$

We can then solve for *b*, our vector containing the regression coefficients, in terms of our independent and dependent variable data.

Assumptions

There are three assumptions that must be met in order for multivariate linear regression to be a valid tool for analysis. They are:

1. All of the independent variables are independent of one another
2. Each independent variable has a linear relationship with the dependent variable
3. The residuals of the model are normally distributed

The first assumption is satisfied mostly by common sense. Air temperature is not in any way related to initial velocity or to buggy weight. Buggy weight and initial velocity do have some correlation, but only when looking at data from an individual pusher. Heavier buggies are harder to accelerate, so some pushers may struggle to get them going as fast. However, since we have data from rolls done by multiple pushers, we have a spread of initial velocities for all of the different weights of buggies.

The linearity assumption is satisfied by the fact that the ranges of values for the independent variables we were looking at are small. In buggy, initial velocities only range from about 15 mph to 25 mph and buggy weights between 100 lbs and 140 lbs. Air temperature can vary more drastically, however due to the time of year the data was taken at, the temperatures only ranged from 25 F to 40 F. At these small ranges of values, we can approximate the relationships between each of them individually and freeroll times to be roughly linear.

Finally, the residuals of the regression model are assumed to be normally distributed in order for the model to be valid. After running the model, we plotted the residuals and found that they created a distribution that was approximately normal.

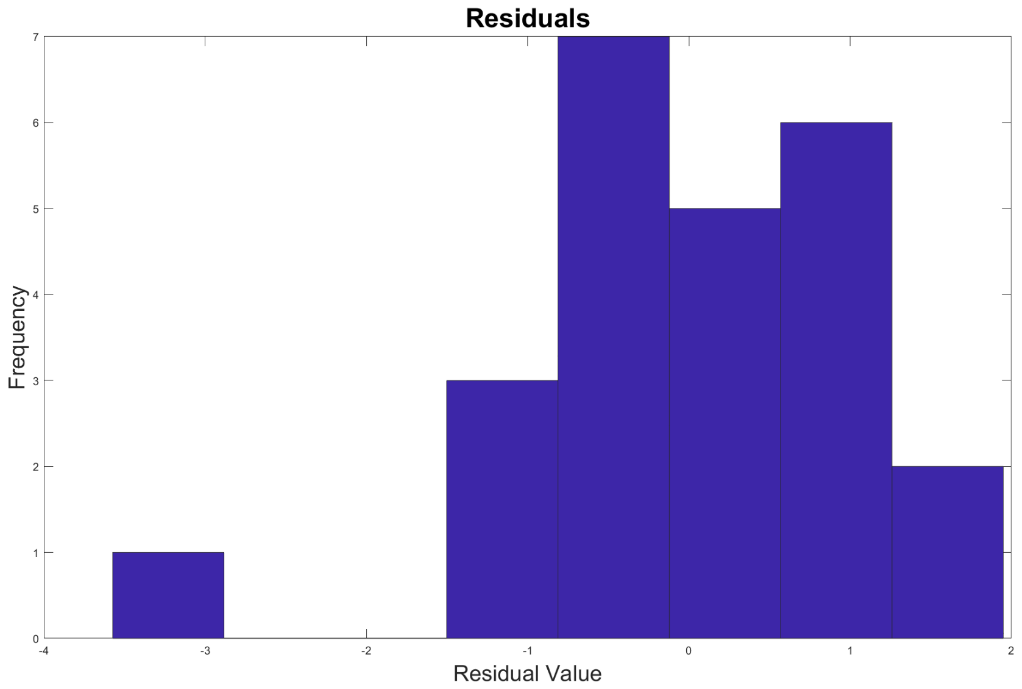
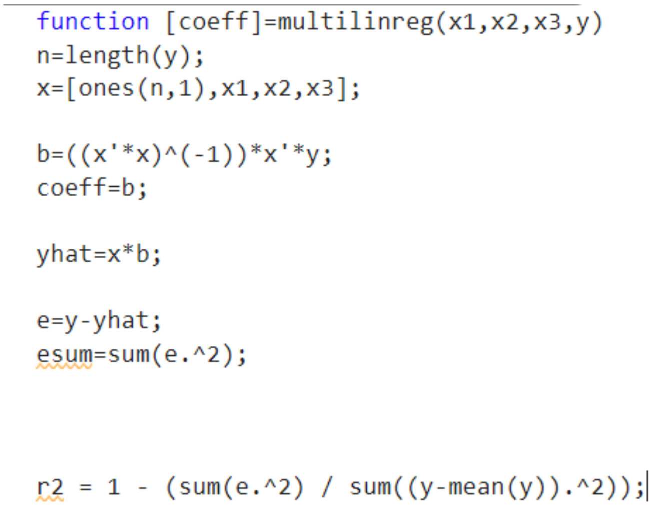


Figure 1: The residuals of the multiple regression model are mostly normally distributed with a center of 0.

Implementation

We implemented multivariate linear regression with our data by coding the algorithm as a function in MATLAB and passing our data in as inputs. The regression coefficients were the outputs of our function. The function also calculated the residuals, sum of residuals, and R-squared value of our best-fit equation.



```
function [coeff]=multilinreg(x1,x2,x3,y)
n=length(y);
x=[ones(n,1),x1,x2,x3];

b=((x'*x)^(-1))*x'*y;
coeff=b;

yhat=x*b;

e=y-yhat;
esum=sum(e.^2);



r2 = 1 - (sum(e.^2) / sum((y-mean(y)).^2));
```

We didn’t really run into any issues in implementing the numerical method with our data, as the algorithm is rather straightforward. One potential cause for alarm, however, is our relatively low number of data points (24 data points). With less data, our best-fit equation is likely less valid when used to predict free roll times using weights, air temperatures, or initial velocities outside of the range that our data captures. A lot of our data also did not span a large range, with only four actual different weight values, and air temperatures only ranging between 31 ℉ and 37 ℉. With an R-squared value of 0.8002, however, we can be fairly confident that our model can predict freeroll times when used with data that falls within our tested range, which is also fairly representative of typical buggy rolls.

Results

From our multivariate linear regression program, we received the following equation describing the data set:

$$Y = 92.0961 - 0.055X_1 + 0.619X_2 - 1.2423X_3$$

where X_1 is the combined weight of the buggy and person in pounds, X_2 is the air temperature of the run in Fahrenheit, and X_3 is the initial velocity that the buggy is pushed by in miles per hour. This would yield the Free Roll Time (Y), in seconds.

We decided to graph this four-dimensional plot to further analyze the correlations between the independent variables and the free roll time. Unfortunately, a pure 4d plot would be very hard to implement and analyze for our purposes, so we primarily focused on projecting different slices of the volume onto the various axes to find the correlations. To do this, we used the data to find the average of the two variables we were not analyzing and then plugged those into the equation so we only had to plot a 2d diagram. We used the following values for the averages of the data set:

Table 1: Average Values of Each Variable

Free Roll Time (s)	Weight (lbs)	Air Temperature (F)	Initial Velocity (mph)
62.03	124.38	33.75	20.38

The following graphs were obtained using this method:

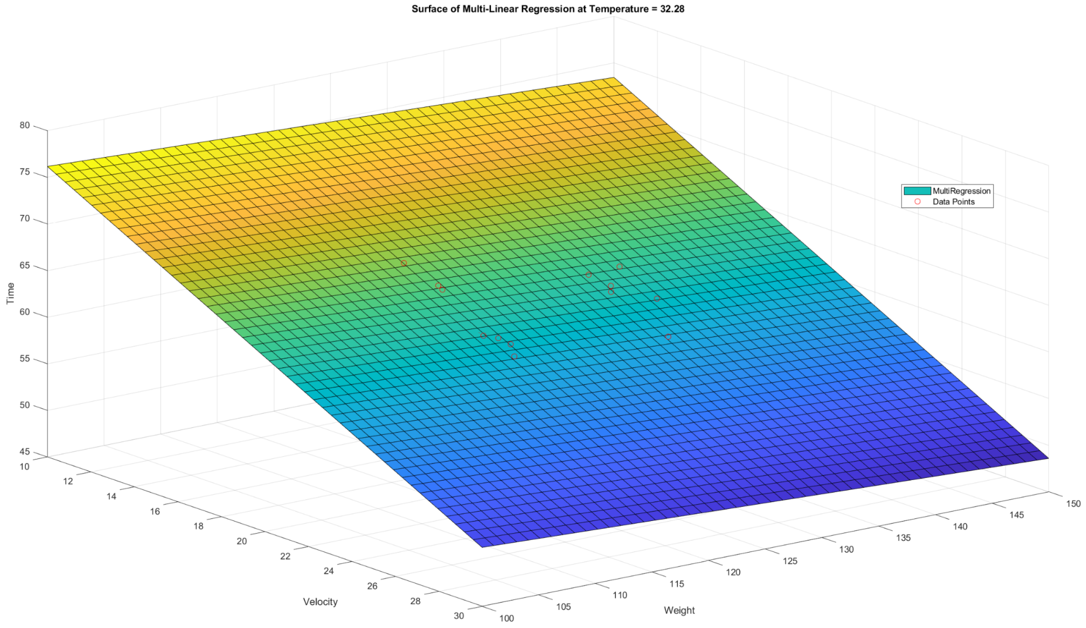


Figure 2: Freerolls times based on initial velocity and weight with air temperature held constant at the average value of 32.28 F

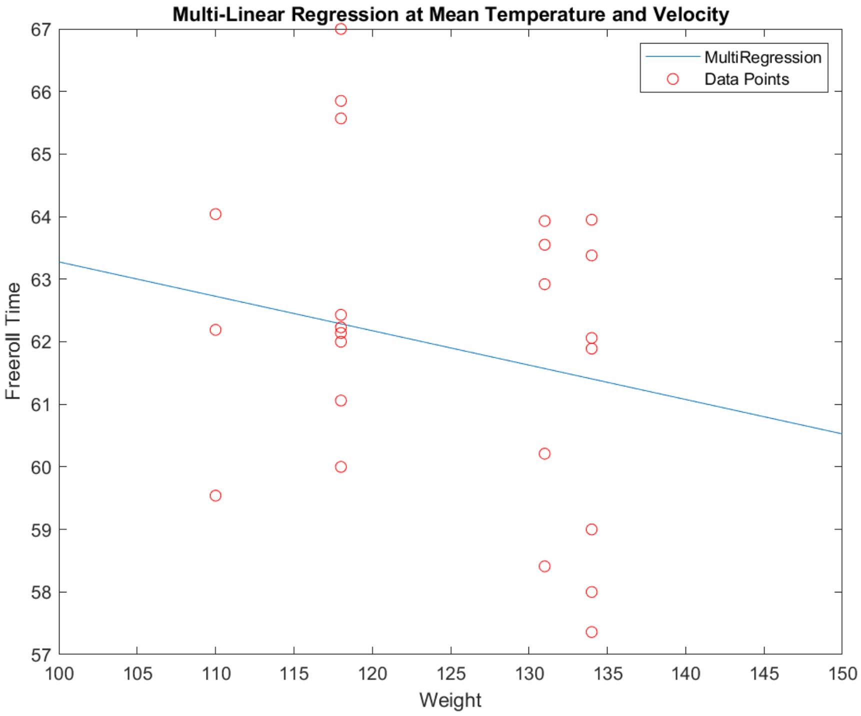


Figure 3: Graph of Freeroll time vs Weight with Temperature and Initial Velocity being held constant; Red is the Initial Data

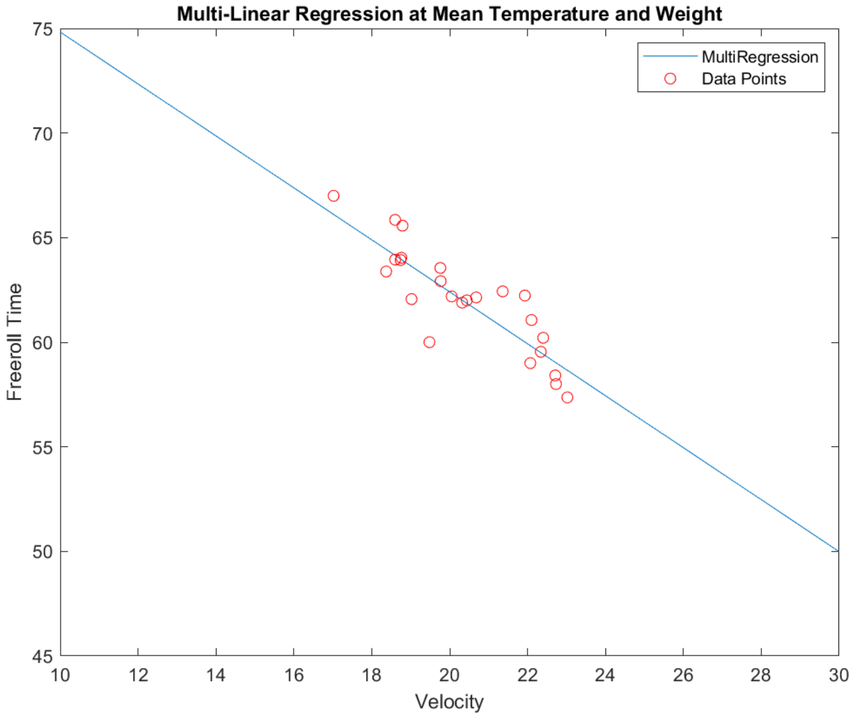


Figure 4: Graph of Freeroll time vs Initial Velocity with Temperature and Weight being held constant; Red is the Initial Data

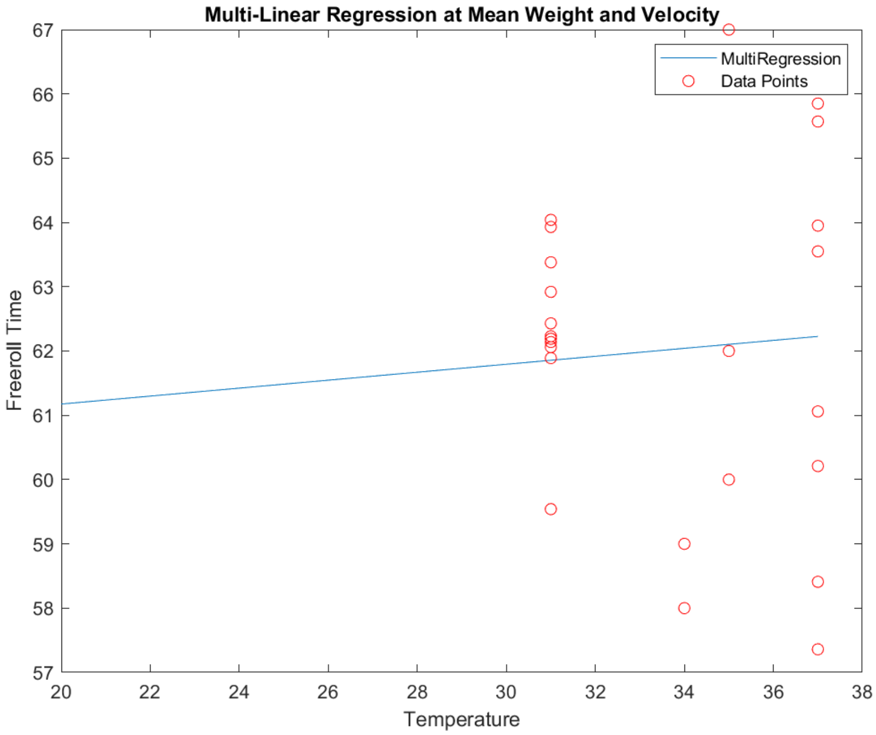


Figure 5: Graph of Freeroll time vs Temperature with Weight and Initial Velocity being held constant; Red is the Initial Data

From these graphs, it can be seen that freeroll time and the initial velocity are strongly correlated with a negative slope. The other two variables, air temperature, and weight, are weakly correlated to free roll time. This essentially means that the initial velocity has the largest impact in determining free roll time, followed by weight and then air temperature. Therefore, if someone wanted to optimize their buggy free roll time, they would push it as fast as possible while maintaining a fairly high weight. The air temperature would not contribute as much to the time, however, a lower one would be ideal assuming the positive linear correlation shown in Figure 3.

Conclusions

Our most important takeaways from this project are how each of the explanatory variables - initial freeroll velocity, air temperature, and buggy-driver combined weight - are correlated with the freeroll time. Through the use of MVR, we were able to generate an equation that shows that initial freeroll velocity has a negative correlation with freeroll time, air temperature has a positive correlation with freeroll time, and buggy-driver combined weight has a negative correlation with freeroll time.

Our resulting equation indicates that the air temperature and buggy-driver weight variables have impacts of similar magnitude on freeroll time, while initial freeroll velocity has a significantly larger impact on freeroll time. This means that, even if a roll has a high initial velocity but suboptimal air temperature and buggy-driver weight values, the freeroll time is still likely to be higher than that of a roll with a lower initial velocity, an optimal air temperature, and an optimal buggy-driver weight. These results confirmed our initial hypothesis that the velocity of the buggy as it is transitioning into the freeroll is the most important factor in achieving a fast freeroll time.

Improvements

In the future, we could improve the accuracy of our freeroll time predictions with more data. The data that we used was only for relatively small ranges of the variables. If we were to gather a greater range of experimental data that included more variation within each variable, we would be able to lower the R^2 value of our equation. In addition to collecting more data, we could include more explanatory variables to even further decrease the error within the equation. Some additional variables that we could consider are freeroll entry angle, chute turn radius, and transition angle.

To expand upon our project, we could use MVR with an additional set of data taken under a different condition - heated wheels. Heating the wheels of the buggy before it rolls can decrease the overall race time. Many of the competitive teams use this strategy on race day to minimize their total time. None of our data was collected in this condition, so if a team were to try to predict a race day freeroll time, with heated wheels, by using our equation, the predicted time would be inaccurate. We can account for this through expanding our project with heated wheel data to get a second freeroll time equation for this condition.

References

1. *Past weather in Pittsburgh, Pennsylvania, USA - Yesterday and Last 2 weeks*. Past Weather in Pittsburgh, Pennsylvania, USA - Yesterday or Further Back. (n.d.). Retrieved May 1, 2023, from https://www.timeanddate.com/weather/usa/pittsburgh/historic

2. Hank Jenkins-Smith, J. R. (n.d.). Quantitative research methods for political science, public policy and public administration: 4th edition with applications in R. 11 Introduction to Multiple Regression. Retrieved May 1, 2023, from https://bookdown.org/ripberjt/qrmbook/introduction-to-multiple-regression.html