LINEAR REGRESSION APPLIED TO RACE CAR PERFORMANCE PREDICTIONS

Dwight Southerland

University of Central Arkansas
College of Business
3404 London Road
Conway, AR 72034
(501) 472-6640
dsoutherland@conwaycorp.net

ABSTRACT

In the arena of championship sportsman drag racing, software tools have been developed based on linear regression formulae and theory much like what has been used for prediction in business analysis for growth. This paper describes the history of the development of these software tools, how they are used to improve a competitor's performance and how the application is very similar to business analysis to predict future performance. The weighted least squares formulation that is used in business marketing research combined with biological, behavioral and social sciences data to help predict future success of business growth is also used for predicting drag race car performance for hundreds of competitors in week after week of intense competition.

Linear Regression Applied to Race Car Performance Predictions

When most people hear of drag racing, the first image that comes to their mind is the long, loud fast dragsters that are the top of the sport. These cars termed “Top Fuel” dragsters are one of the professional categories of vehicles that are publicized the most and the focus of most marketing from the racing associations. However, the majority of participants in the sport of drag racing are in much more accessible and mundane vehicles. In fact, if you were to go to a local drag race event, the majority of the participants would be cars that would look more like you see on your morning commute, albeit louder and behaving worse than a commuter eating his breakfast, shaving and texting at the same time. The most common venue for drag racing in all the non-professional categories is handicapped racing where each participant will predict the exact elapsed time for his vehicle before the race begins. His goal to win the race is to perform better than his opponent in reaction time at the start and then not be quicker than his predicted elapsed time, called the “dial in”. Being able to set this “dial in”, or performance prediction, is one of the two major components that determine the winner of the contest. This sets up a prediction problem that can be addressed with the same linear regression methods used in business for forecasting. When a single contest can be won by thousandths or even ten-thousandths of a
second, the ability to control any component of that contest becomes more and more valuable when competing for prizes in the tens of thousands of dollars.

**History of the Development of the Application**

There are several products on the market that serious sportsman racers use for performance forecasting today. Most all are based on some work that was done during the 1980s by a company called Performance Trends. The variables for the regression formula are several weather environmental factors that can affect an engine’s ability to make power. This company originally bundled a measuring instruments package to measure and enter the variables into a specially programmed hand-held calculator for the linear regression calculations. Since there are several variables related to changing weather conditions and each one has a differing effect, the resulting problem is a rather involved multiple factor linear regression calculation. The initial program relied on air temperature, humidity and altitude as the variables to forecast the performance of the next race and the packaged product included a laboratory quality thermometer, hygrometer and an FAA certified altimeter along with the specially programmed calculator. (Other atmospheric conditions also affect vehicle performance such as wind velocity and direction, and we will come to those later.)

Of course the racers who benefitted the most from these initial products were those racers who had sufficient and accurate enough data to input for the linear regression calculation to be meaningful, since the calculation program depended on an accurate database. A small contingency of serious racers had adopted a regimen of data gathering at events and the rest of the pool of competitors were beginning to take note that there was a disproportionate number of these “scientific” racers who were doing well in competition. More and more racers had begun incorporating records of weather conditions into their record keeping because they knew there were causal relationships between the variables and their performance, but lacked the scientific calculations for objective results. By the time the Performance Trends calculator hit the market, “weather stations”, as they were referred to, were common in the pits of the sportsman racers. The Performance Trends calculator, however, quickly became the “had to have” tool because the forecasting calculations were much more accurate than the mental estimating that had become the norm.

Several other companies followed with their similar products during the next few years due to the success of Performance Trends, but all used the same basic calculations. By the mid-1990s, this application of linear regression forecasting had become standard operations for all but the most novice racers. However, some racers had begun noticing that although the altitude factor was a meaningful variable, the actual condition that affected performance was more than just a standard calculated differential of atmospheric pressure due to changes in altitude. The actual measured barometric pressure was also an indicator of weather conditions that affected performance since high pressure areas and low pressure areas created much the same effect as change in altitude. Soon, barometers began being common and racers were struggling to integrate this new data into their forecasting. The programs were designed to input altitude and how to add barometric pressure was not part of the calculation. However, smart racers began using standard procedures from pilots and the instrumentation already available to calculate *density altitude*, which is pressure altitude corrected for non-standard temperature. Such
calculations and formulae were readily available from the Aircraft Owners and Pilots Association training. Pressure altitude is the actual barometric pressure read from the now common barometers and the density altitude is calculated using standard formulas used by pilots of small aircraft. Using density altitude rather than measured physical altitude for input into the calculations would allow for a more accurate referential standardization for the data. So these smart racers began creating new databases using the density altitude as input to the linear regression calculators. Soon this quest for more accurate data began showing up in another new level of forecasting accuracy as another wave of consistent winners. As this new procedure was becoming known, the guys at Performance Trends upgraded their product to include the new programming and the new set of instrumentation and once again were in the forefront of the drag racing prediction product market.

The Current State of the Technology

Such established the standard approach to drag race forecasting for the next twenty years. During that period, packaging on the market has changed so that completely self-contained units with the measuring instruments and calculator in one box have become common. And, stationary weather stations with transmitter capability were developed so that the driver could see a current calculation via his cell phone the moment before his race and make any necessary adjustments to his dial-in. With the additional capabilities of processors for storage and input, products have been offered that would input the readings from digital instrumentation directly into the dataset of the calculation and then automatically store the data at intervals into the weather sample database. Through all this development, the calculations remained the same multiple linear regressions formulas.

The recognition that the relevancy of the data used in the calculation was necessary (moving from purely physical altitude to density altitude) follows rules of discipline that have been recognized since Carl Friedrich Gauss first published his works on statistics prediction that is recognized as one of history’s first presentations of linear regression. Another rule that Gauss recognized was the importance of volume of data and in the sport of drag racing, accumulating enough accurate data to validate the “weather station” tools as advantageous took some time. Even the most ardent competitors in the sportsman ranks would be hard-pressed to accumulate 100 runs of data without some major component failure or parts change that would again become a variable that was outside the arena of applicability for the weather station calculators. That common phenomenon introduced another layer of complexity into the development of the weather station predictor. Since the linear regression model adapted for weather conditions of temperature, humidity and density altitude was already a multiple linear regression process, considering the addition of these other variables pushed the prediction process into the domain of multivariate linear progression. Also, it was common for racers to “tune” their engine combination to improve efficiency due to changing weather conditions. For example, higher air density allowed the engine to ingest more oxygen per cylinder filling which meant that more fuel could be burned and produce more power. Lower air density created the opposite effect. The racer would change the fuel delivery in the carburetor or fuel injection system to account for these changes in atmospheric conditions to improve performance. Or, as another example, high humidity levels inhibit fuel combustion so the racers often will change ignition timing to initiate the combustion in the cylinders a bit sooner to promote complete fuel burning. Of course, this
meant that additional variables were present to affect the accuracy of the performance predictions, so the development of the regression calculations for predictions expanded.

In the analysis of the complexity due to the many possible variables, additional layers of linear regression territories were identified and subsequent formulation was developed. First, a set of variables were identified that affected engine power output along the lines of the tuning examples cited above. These variables included for example, fuel delivery, ignition timing, camshaft timing, engine operating temperature, fuel combustion properties, etc. And most of the identified variables had relative effects conditioned by the already identified atmospheric conditions. A second set of variables were identified that affected the application of that power to the vehicle, including gear ratios, vehicle weight, power transmission characteristics whether it was through fluid drives (automatic transmissions with torque converters) or mechanical drives (manual transmissions with clutches), wind direction and velocity, and so on. So now the multiple linear regression prediction process that started out to only account for atmospheric temperature, humidity and density altitude became a series of multivariate linear progression calculations that fed from one to the other to eventually calculate a single number! The “weather station” products became very sophisticated and it is common to see hand held wind anemometers being carried around at any major drag race event to add that data to the new level of more complete and accurate calculations.

All this complication demanded more of what is necessary for any successful linear regression analysis – data. Few if any racers had enough experience with the effects of the variables for engine performance and vehicle performance for their particular situation to provide the data needed for input to these calculations. Fortunately, a different approach was available. There was a bountiful source of data from the automotive industry that would give reasonably accurate estimates of the changes in power output of an engine when variables were changed. There was also an abundance of research information and theory that was readily available about the physical effects wind and mass on the vehicle. There were common calculations that could accurately plot an acceleration curve for a vehicle based on power applied, mechanical efficiencies and gear ratios. All this data was integrated into the weather station calculators to become a very sophisticated process. Thank goodness this technology was developing in the trails of powerful laptops and hand held processors.

The savvy and drag racer today then has available a tool that will account for changes in the race environment and the vehicle to be able to predict what his next performance will be. For the spectator, the racing is much closer, and for the racer the potential for winning is also closer. All this is possible through another application of the predictive power of linear regression that is so widely used in business.

**Expansion of the Technology**

Looking at the development of this software from a different direction, it is easy to see that along with the development of race performance prediction was the development of formulas that predicted engine horsepower output. Another software market product that sprang up used linear progression calculations combined with automotive industry resource data that allowed racers to predict what affect changes in an engine design would have on power output. A racer could
effectively create a database that identified all the characteristics of his engine, input that data into software that used a variety of linear regression processes and the output would be the theoretical horsepower of that engine. Of course the software would only estimate the output, but the weight of the relationship of the variables in the linear regression calculations is backed by massive volumes of data resulting from the years of automotive production and development. Having this level of accuracy of describing the relationship of variables adds credibility to the process since the majority of the calculations are basically weighted least squares linear regressions. Needless to say, the application of this type of software extends beyond sportsman level drag racing to all levels of automotive racing of all kinds. NASCAR, Indianapolis type cars, Formula 1 cars, motorcycle racing, airplane racing, boat racing, if it uses an internal combustion engine, there is software to help predict the power output and the effect of changes in the component variables, all based on Guass’ linear regression formulas. Professional automotive race teams in every venue around the globe now depend on the predictive analysis of such types of software.

Then there is the additional extension of prediction to the overall performance of the vehicle. Once you can predict the power output and characteristics of an engine, the definition of a vehicle and the performance of that vehicle is another set of linear regressions based on industry and scientific data that gives accurate values of the effects of changes in variables. Software was developed by many of the same companies that started the race car performance prediction that isolated the calculations using those variables to create output that would allow a racer to predict the performance of his vehicle, and there are applications for all types of racing. In the case of the sportsman drag racer, this software also becomes a tool for development. By selectively changing variables for traction coefficients, gear ratios, aerodynamics, transmission characteristics, weight distribution, and so on, the racer has a tool to predict performance improvements without the cost of the changes and the expense of the testing process that would be necessary. (See the Appendix for illustrations of the output from common inexpensive software used to predict horsepower output and vehicle performance.) Bundling these packages for engine output prediction, vehicle development prediction and the “weather station” at the racetrack for elapsed time prediction has an added advantage that the original three atmospheric variables – temperature, humidity and density altitude – are key input to all three areas of calculations.

These tools have become so effective that racers can create imaginary combinations of vehicles, engines and drivelines and have some reasonably accurate prediction of the performance without ever building the car. This obviously saves untold amounts of time and money for anyone wanting to participate in racing. The popularity of these software products is great in the auto racing industry and wide varieties are available. Versions are available that are inexpensive enough so that the most budget-minded racing enthusiast can afford to purchase up to “professional” versions that would make sense only for those whose livelihood is dependent on having the best of tools. As with many products, you get what you pay for and in this case the difference between the cheap and the expensive versions is the volume and accuracy of the built in data, the number and variety of variables for input and the number and complexity of the calculations that are part of the programming. As an example, consider that a simple linear regression calculation model that determines the linear relationship between a single dependent variable \( y \) and the \( p \)-vector of regressors \( x_i \) takes the form the model takes the form
\[ y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \]

where \( T \) denotes the transpose, so that \( x_i^T \beta \) is the inner product between vectors \( x_i \) and \( \beta \). Of course this eventually stacks to the well-known \( Y = \beta X + \varepsilon \), but the complexity of the calculation is obvious. Then consider that each subsequent regressor (independent variable) that is not related to any other previously calculated regressor compounds the calculations. So, adding vehicle weight to the calculation along with air temperature, as an example, requires calculating each possible weight value for each possible temperature value. When the variables considered for engine power output and vehicle dynamics may each be forty to fifty values for even the simple sportsman category race car, and each additional value potentially compounds the number of calculations, the processing ability of the software becomes a huge requirement.

The original product developers mentioned in the beginning of the report, Performance Trends, is now one of the leaders in professional level race car development software. The cost of their full suite of products for professional top fuel dragster racing companies would rival the cost of software used in banks. Yet the core of all their analysis and prediction tools is least squares linear regression.

**Effects of the Technology on Engine Development**

As a side note, one of the purposes for this research was to address a known concern about the current capabilities and performance of common commercial products that many sportsman drag racers use. It seems that the accuracy of the “weather station” products decline when the vehicle definition deviates from what is considered a “mainstream” drag race car model. Most sportsman drag race cars utilize V8 engines that have good cylinder air flow by intake system and carburetors or fuel injection systems that flow at least 600 cfm. Also, most race vehicles weigh between 2700 and 3800 lbs. The characteristics of engines and cars within these ranges of assumed parameters are apparently plugged into the linear regression calculations as weight variances. Remember that the core of all the predictive software is least squares linear regressions and apparently for the value of many of the variables is weighted least squares, that is the relative value of the effect of the variable on the calculations is given a “weighting” coefficient to help make the calculations more accurate. These weighting coefficients are added to reduce the number of calculations on the front end to determine the weighting factor. Our textbook and our work in class pointed out that these weighting factors need to be determined by experienced people when dealing with business predictions. In this case the weighting factors have apparently been determined by the industry experience of what is applicable in a given range of vehicle definitions. So, as the racer deviates outside the range of the assumed “standard” race car, the value and accuracy of the predictions declines, that is, the statistical “\( R \)-value” becomes less and less. A racer with a car with a little engine that breathes through a small carburetor has less and less possibility of being able to predict his next run using the common products. The problem is often described by the racers as the fact that changing atmospheric conditions and wind parameters have a greater effect on such cars than cars with big engines and large carburetors, and they are correct. However, the performance of the software is not the fault of the car; it is that the inaccuracy of the weighting factors in the weighted least squares linear regression formulas! Possibly some of the more expensive software products may overcome this issue, but a simple solution would be for someone to offer a weather station product with specific parameters for differing varieties of cars.
Common Ground with Business Strategy

As a second side note, the adaptation of the field of linear regression to business is obviously a valuable tool. In fact, linear regression is one of the most widely used statistical tools used across all disciplines. It is especially important in the fields of biological, behavioral and social sciences and is one of the more useful tools in those disciplines. Combining the applications from the behavioral and social sciences with business marketing research, large businesses with a chain of stores are using regression analysis to help forecast the success of new store locations. A case study of fast food giants McDonald’s and Burger King illustrates how they are combining regression analysis with data collected from geospatial information systems (GIS) to help predict locations for new stores. GIS adds a new set of variables to be plugged into linear regression calculations that look very similar to those used in race car prediction software. Their experience helps them establish weighting values in determining the potential success rate of building in these new markets. The concept of adding expanded data variables like GIS information for accurate forecasting seems like new business technology, but the approach is basically the same as predicting how fast my race car will be today.
APPENDIX

Output for an engine definition from Engine Pro, a software program marketed by Racer Support LLC. See www.quarterjr.com

Output for a vehicle from Quarter Pro, a software program marketed by Racer Support LLC. See www.quarterjr.com
REFERENCES


Freeman, David (2009). Statistical Models: Theory and Practice. Cambridge University press. A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has several explanatory variables on the right hand side, each with its own slope coefficient.


Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 1–2, ISBN 9789812834119. Regression analysis ... is probably one of the oldest topics in mathematical statistics dating back to about two hundred years ago. The earliest form of the linear regression was the least squares method, which was published by Legendre in 1805, and by Gauss in 1809 ... Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun.