2003

# A Comparison of Multiple Regression Models to Help Predict Road Race Performance for Two Runners

Lynn Brown
*University of North Florida*

# A Comparison of Multiple Regression Models to Help Predict Road Race Performance for Two Runners

Lynn Brown

Faculty Sponsor: Dr. Craig G. Harms, Assistant Professor of Management, Marketing and Logistics

**Abstract**

**As a student I start the new semester armed with my course schedule, course descriptions straight from the undergraduate catalogue, and the all-important feedback from students who have taken the course before me. Introduction to Management Science (QMB 4600) is "A study of selected mathematical and statistical models used to aid managerial decision making." It sounds fairly difficult and potentially boring, but a necessary evil, none-the-less. Imagine my surprise when the subject is in fact interesting, engaging, and yes, fun. This paper presents the development and analysis of one of the cases in this course. It involves the creation of three multiple regression models from two different sources (runners) to help predict the pace in minutes per mile a runner is expected to run a five kilometer race. Students are involved in the process from the ground up, from model development to model validation, thus creating an atmosphere for meaningful learning.**

## Introduction

It is not easy to get excited about the statistical analysis of 'widgets'. Today's university students are products of the digital age. We are accustomed to instant messaging and real-time data streams. Most students are comfortable with computers yet have a difficult time adapting to powerful statistical programs as applied to large empirical research topics. For a group of students to stay engaged in a multiple regression project, the database, variables, and hypotheses must have real meaning. What is unique to the project being presented in this paper is that the database is the professor. Brainstorming the independent variables that will be investigated is a 'one on one' process involving the students and the "corporation." There is realism in the modeling process and the students are able to experience each step in the research process with the guidance and direction of the professor.

## Research Goal

The goal of this paper is to develop a multiple regression model to help predict the time (pace in minutes per mile) a runner may run a five-kilometer race this coming weekend. This paper will also include a comparison of two runner's prediction models. The regression models will be used to predict the race pace and then will be compared to the runner's actual performance in several races. [Y-hat vs. Y(i)]

## Vertically Integrated Case Approach

Table One presents a list of tasks for the faculty member and the students to perform during the five weeks.

**Table One**

<u>Steps to Perform in This Vertically Integrated Case</u>

1. General overview of a causal model (class #1) statement of dependent variable and the various independent variables.
2. Hypothesize the quantitative model and the relationship of each independent variable to the dependent variable (class #1).
3. Determine the goals of the research — Three goals are developed:

a) to collect empirical data from two runners for a group of independent variables;

b) to investigate and develop several good statistical models that can be used to predict running pace for a five-kilometer race;

c) to decide on the "best" model to use to predict upcoming times; and

d) to compare the predictive model for the two runners (to be discussed during class #1).

4. Collect the data and build the databases:

a) this author's database is expanded from 398 [ 3 ] observations to 417. The 417 observations are divided into two databases—209 and 208 observations respectively (A209 and B208);

b) the dependent and independent variables are collected for the second runner's database (FRAZ116).

5. Distribute the database to students. Each student receives a disk with A209, B208, and the FRAZ116 database (class #2, during second week).

6. Analyze data: model validation & statistical tests:

a) perform linear regression with each independent variable (after class #2, third week);

b) look for non-linear relationships and perform transformations (after class #2, third week);

c) perform a stepwise regression (after class #3, fourth week);

d) develop several multiple regression models (after class #3, fourth week);

e) perform statistical tests (after class #3, fourth week);

f) make necessary modifications and repeat (after class #3, fourth week);

7. Compare runner's models:

a) same variables?

b) same B-coefficients?

c) same accuracy? (class#4)

8. Draw conclusions – review hypotheses (class #4):

a) test model and predict time for five-kilometer race on weekend;

b) compare model with actual time.

**Developing a Causal Model**

Developing a causal model can be accomplished in a number of ways. A professor can lecture the class about the dependent and independent variables, the equations for predicting the outcome, and statistical measures of the validity and value of the model. Or the professor can create a forum for interaction by asking the right questions, thus leading the students down a path to meaningful learning. Using this method, any student is welcome and encouraged to offer a suggestion when asked, "What independent variables will affect how fast a runner will run a five kilometer race this weekend?" It is a safe question that elicits many "common sense" answers from the class, such as "age." The class and the "corporation" enjoy a two-way interchange as opposed to the far less effective one-way lecture approach.

Now that everyone is awake and participating we learn that an independent variable must pass three tests:

The proposed independent variable must be logical. We must be able to sit back and nod, "Yes, that makes sense to me."

The proposed variable must be quantifiable. I must be able to develop a number to represent the variable value; and

the variable must be obtainable. Beyond overcoming the proprietary problems in many corporate databases, I must be able to get my hands around the variable. In this case there is no "proprietary."

"Age" is a wonderful independent variable. It is an easy concept that everyone can understand and relate to. We must then determine the hypothesized relationship between the dependent and independent variable. The relationship between "pace" and age is age is direct and positive, i.e., the older the runner, the higher the pace per mile (the slower the runner will run). The relationship is positive, however the information is not good news for the older runner.

"Weight" and "training" are suggested as independent variables. The more a runner weighs the slower the pace, which is a positive, direct relationship. Training is an independent variable that can be quantified by the number of training miles in the week or two weeks prior to competition.

"Weather" has an obvious contribution to the model, but the question posed to the group is how can weather be quantified? Temperature and humidity are examples of weather conditions that would be excellent variables, however this information is not part of this particular database and not obtainable.

One of the students suggests that the variable "hilly course" should be added to the model. It is not a variable that can be quantified easily, but we can use a dummy variable to represent it. "Hilly course" is a "qualitative" variable that can be used to reflect "adverse conditions" in general. This could encompass and extremely hot or humid day as well as difficult (hilly or grass) terrain. It is an important variable to include because it is useful for reducing some noise in the model.

Ideas are exchanged, relationships are hypothesized, and the model is developed with total involvement of the students. Table Two presents the list of independent variables for the initial database. The task is meaningful and it all makes sense. For students, what makes sense tends to sink in.

**Table Two**

List of Independent Variables

| # | Variable<br>Expected Coefficient Sign |
|---|---|
| 1) | Age<br>positive (a bad thing) |
| 2) | Training/miles previous 7 days negative (a good thing) |
| 3) | Training/miles previous 14 days negative (a good thing) |
| 4) | Weight<br>positive (a bad thing) |
| 5) | Adverse Racing Conditions<br>positive dummy (1/0), 1=adverse |
| 0) | Pace in Minutes per Mile = Dependent Variable |

The ability to compare the two runner's predictive models presupposes that the variables in the original model (and paper [3]) are also collected by the other runner. In this particular case the same variables are collected and in addition, temperature at race time was recorded. Thus one of the desired variables chosen during the brainstorming session is available for the second database. Regrettably, the first runner did not record that data. Therefore, the above five independent variables are used to develop both models.

**Data Collection**

The two runners collected the values for the dependent and independent variables from their respective running logs. The larger database (417 observations) was divided into two parts to allow for more variety and to guarantee that students would have a richer comparison of models. The FRAZ database was smaller, only 116 observations with six years of data missing. However, this did not adversely affect the model.

The availability of three databases enriches the project by allowing for many more statistical questions to be posed, such as "are the databases the same?" For each model, a test of means, a comparison of the B-coefficients, and a t-test can be done on each of the variables. In addition, one more requirement was added to the assignment. In the past the corporate databases used in this course have contained between eight and 19 independent variables. The solution process is an intelligent search procedure trying to find a subset of good variables. With only five variables, each can be explored more carefully. Therefore students are asked to perform a complete step-wise regression and also to look at the possibility that the independent variables are related to the dependent variable in a non-linear way. This requires observing a scatter plot and performing a transformation on each of the variables that "look" non-linear. The only transformation allowed was a squaring of the "X" variable—just to make the job manageable.

## WE HAVE A MODEL, NOW WHAT?

The solution process is well underway but students will not get far without proper guidance. First, we are cautioned to test for muticollinearity. Obviously, there will be a large correlation between miles/7 days and miles/14 days (.953, .944, and .841 in the three databases). These two variables may not be used in the same model. In addition, many of the other pairs of variables may have some multicollinearity. Therefore it is necessary to pay close attention to the t-tests, the values of the B-coefficients, and the standard deviation of the model. If the t-values are very poor, if the B-coefficients make no sense, or the standard deviation for a particular model is relatively poor (large) compared to other models, some of the independent variables may be too correlated and consequently should not be used together. Researchers cannot be too "hard" on independent variable correlation, however, because viable models may be reduced or eliminated if the concept of "acceptable multicollinearity" is not relaxed. In previous cases, the correlation between independent variables of greater than 0.7 or less than -0.7 is considered unacceptable multicollinearity. In this case that measure may have to be relaxed.

Our professor also cautioned us about the use of transformations. It is quite possible that ALL continuous variables may have some quadratic relationship to the dependent variable, making the final model development very tedious. Our guideline suggested that a quadratic term might be used in final model development if it significantly improves the statistical measures. In practical terms, if the transformed variable in a single variable model as shown in equation (1) increases the value of the adjusted R-square by 0.10 versus the simple linear regression model, then it can be included in the final model search process. Otherwise, although marginally helpful, the transformed variable is not researched.

Y [hat] = B(0) + B(1)X(n) + B(1)X(n)2 (1)

## STARTING THE STEP-WISE PROCESS

All students are required to begin the solution process in the same manner—using this author's databases "A" and "B." Table Three presents the simple linear regressions for each variable for both databases.

**Table Three**

Database A

| | Variable Values | | | | Model Values & Measures | | | |
|---|---|---|---|---|---|---|---|---|
| Var # | Mean | Std.Dev | r | F | B | t-stat | R2 | σ |
| 0 | 17.251 | 1.803 | | | | | | |
| 1 | 430.086 | 94.092 | .838 | 487. | .0120 | 22.076 | .702 | .735 |
| 2 | 88.299 | 26.790 | -.700 | 198. | -.0351 | -14.102 | .490 | .961 |
| 3 | 174.213 | 51.134 | .700 | 198. | -.0184 | -14.105 | .490 | .961 |
| 4 | 149.924 | 6.808 | .718 | 220. | .1417 | 14.8517 | .516 | .937 |
| 5 | .187 | .391 | .119 | 2.954 | 4079 | 1.7187 | .014 | 1.337 |

Database B

| | Variable Values | | | | Model Values & Measures | | | |
|---|---|---|---|---|---|---|---|---|
| Var # | Mean | Std.Dev | r | F | B | t-stat | R2 | σ |
| 0 | 17.250 | 1.312 | | | | | | |
| 1 | 429.923 | 93.486 | .818 | 417. | .0115 | 20.423 | .669 | .756 |
| 2 | 86.497 | 26.030 | -.673 | 170. | -.0339 | -13.0654 | .453 | .972 |
| 3 | 170.675 | 51.238 | -.669 | 167. | -.0171 | -12.9299 | .448 | .977 |
| 4 | 149.937 | 6.987 | .794 | 352. | .1491 | 18.7656 | .631 | .799 |
| 5 | .221 | .416 | .001 | .000 | .0035 | .0159 | .000 | 1.315 |

Residual plots are examined for every variable in both databases. It is obvious that that residual of the age variable shows a strongly quadratic plot as shown in Exhibit One. The other residual plots reveal some non-linear relationships. Transformations are performed on each variable of both databases. Table Four presents the statistics for each transformed model.

## Exhibit One

Residual Plot of Age Versus Pace for Database A

```
  2.82:-------------------------------------------------------o----
  2.62:-----------------------------------------------------------o
  2.41:------------------------------------------------------------
  2.21:------------------------------------------------------------
  2.00:o-------------------------------------------------------oo--
  1.80:------------------------------------------------------------
  1.59:----------------o-------------------------------------------
  1.39:-oo---------------------------------------------------------
  1.18:-o--o---------------------------------------------------o---
   .98:----------------------------------------------------o------o
   .77:--oo---o-oo-oo--o--------------------------oo-------------o-
   .57:-------o--oo-o--o-------o--o-------------------------------o
   .36:--------o--o-oo---------------------------o--o----------ooo-
   .15:--------o------ooooo-------o-------o-o---o-oo----o-------oo-
  -.05:---------ooo--oo--oo-------oo------oooo--o-o--o-o---------o-
  -.26:-------------oo-o-o-ooo----o----o----o-o--ooo-oo---------o--
  -.46:--------------o-----o-o-o---oo----o-ooo--ooo--o-o-----------
  -.67:--------------oo-oo--oo-ooo-oooo-------ooo-o----o--------oo-
  -.87:--------------oo----ooooo-o--ooo-o----o-o--o----------------
 -1.07:--------------------o-o-------oooo---ooo--------------------
      ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^

     242  268  299  330  362  393  424  456  487  518  550  581  612
                                                    age in months
```

### Table Four

Model Statistics for Transformed Models

| Database | Variable | F | X t-stat | X sqrd | t-stat | adjR2 | σ | R2/linear R2+Δ |
|---|---|---|---|---|---|---|---|---|
| A | age | 566 | -10.2539 | 13.8928 | .844 | .529 | .702 | .142* |
| B | age | 457 | -9.62828 | 12.8657 | .814 | .564 | .669 | .145* |
| A | miles/7 | 132 | -8.82675 | 5.88736 | .557 | .892 | .490 | .067 |
| B | miles/7 | 134 | -9.98392 | 7.33043 | .560 | .868 | .453 | .107* |
| A | miles/14 | 141 | -9.34132 | 6.56725 | .572 | .876 | .490 | .082 |
| B | miles/14 | 125 | -9.48953 | 6.85819 | .544 | .883 | .448 | .096 |
| A | weight | 109 | .18496 | .18939 | .509 | .939 | .516 | .007 |
| B | weight | 179 | -1.13553 | 1.67059 | .631 | .795 | .631 | .000 |

*passes arbitraty test of "good enough to use"

During the third class the linear models and the single variable transformed models are discussed. To this point every student has performed the same steps with either database A or database B. Some obvious conclusions are:

1) The two databases are very similar.
2) In addition to understanding the extremely high multicollinearity between miles/7 and miles/14, the statistical measures using these two variables are also extremely similar. Thus, either of these two variables could be used in the final model, but not both. Miles/14 will be used in database A and miles/7 and miles/7 squared will be used in database B.

3) The "Adverse Conditions" variable is quite meaningless according to the linear statistical model; however, we are advised that in the search for a good model, we "don't trash anything." Although by itself, a variable may be seemingly worthless, in combination with other variables, it may help the overall model and become statistically significant.

**Now It Is Our Turn**

After the third class, we are turned loose to investigate a variety of models. Two final models are required for the first runner using databases A209 and B208.The second runner's model is to be build using the same procedure, but with no faculty help or guidance. With the knowledge that both training variables and the weight variable are significant in the simple regression model, it is hoped that they will also be significant in the multiple regression model. Even though the adverse condition variable was almost worthless in the linear model, it seemed logical and would be added as a last variable to the multiple regression model.

**Final Presentation Day and Model Validation**

Class #4 is a presentation day. We have worked independently and it is amazing to see how many acceptable models are developed. Three models are presented in this paper, one from each database. Table Five presents a statistically strong model using the "A" database. Table Six is the model from Database B, which contains two transformations, age square and mile/7 square.

**Table Five**

ANOVA from Database A

| Source | Sum of Squares | df | Mean Square | F |
|---|---|---|---|---|
| Model | 332.565 | 5 | 66.513 | 317.525 |
| Error | 42.523 | 203 | .209 | |
| Total | 375.089 | 208 | | |

| Variable | Estimated Coefficient | t-statistic |
|---|---|---|
| Intercept | 18.31056 | |
| age/month | .035877 | -12.11144 |
| age squared | 000050 | 15.25510 |
| weight | .034290 | 4.86051 |
| miles/14 | -.003552 | -3.59995 |
| adv cond | .364263 | 4.45360 |

adjusted R-square: .8833 standard deviation: .4577 minutes

The residual plot for the model in Exhibit Two shows no quadratic relationship remaining after the two transformations. This plot represents a random scatter and therefore the model has been explained very well. Using the hands-on approach along with our heads every step of the way make the appearance of the scatter plot a real triumph and the atmosphere in the classroom is one of real excitement and achievement.

**Table Six**

ANOVA from Database B

| Source | Sum of Squares | df | Mean Square | F |
|---|---|---|---|---|
| Model | 320.225 | 6 | 53.371 | 297.787 |
| Error | 36.024 | 201 | .179 | |
| Total | 356.249 | 207 | | |

| Variable | Estimated Coefficient | t-statistic |
|---|---|---|
| Intercept | | 15.33234 |
| age/month | -.031482 | -10.45741 |
| age squared | .000044 | 12.99123 |
| miles/7 | -.029149 | -3.96354 |
| miles/7 square | .000126 | 3.35568 |
| weight | .0560068 | .10031 |
| adv cond | .297184 | 4.16197 |

adjusted R-square: .8953 standard deviation: .4233 minutes

**The second runner's model**

This project took on a whole new dimension with the addition of the database recorded and painstakingly compiled by Mr. Frank Frazier. The same five independent variables are in the database, but only 116 observations are available. Unfortunately, there are six years of data missing. We will follow the same procedures by performing a simple regression on each of the five variables. The analysis is much easier the second time around having gained a lot of experience while building the first model.

A comparison of the two runner's linear models is now appropriate:

1) Weight is virtually the same.

2) Training miles for Fraz is quite a bit smaller.

## Exhibit Two

Residual Plot of Age Versus Pace for Database B

```
 1.33:-------------------------------------------------------o-o--
 1.20:---------------------------------------------------------o-o
 1.07:---------------------------o--------------------------------
  .94:----------------o-------------------------------------------
  .81:----------------o-o-----------------------------------------
  .68:----------------------------------oo------------------------
  .55:----------------o----------o----------o----o----------------
  .42:------------o-------o----o--------o-------o-----------------
  .29:-o-----------o----ooo-o-o---o-o------o----o----------o----oo
  .16:----------o-o-oo-oo--oo--ooooo----o--o-o---o----o----------o
  .04:-------o---o-oooooo--o-oo--ooo--o----oo-o-oo--o---------o-o-
 -.09:---oo--o-o----o-o---ooo-o---o--o----oo-oo--o----------------
 -.22:-o--------o---oo-oo--oo-o-o--o-ooo---ooo-o-oo------------ooo
 -.35:---o----ooo---ooo---o-o-------oo---oo----o-o--------------oo
 -.48:-o------------oo--------o-----o-oo-oo-oo-o------ooo------oo-
 -.61:--------o-----------o----o----oo----------o-o-oo------------
 -.74:o-----------------------------o----------o---------------o--
 -.87:-----------------------------------------------o------------
-1.00:------------------------------------------------o-----------
-1.12:--------o----------------------o----------------------------
      ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^
     243  268  300  331  362  393  424  455  487  518  549  580  611
                                                       age in months
```

## Table Seven

Database Fraz

| | Variable Values | | | | Model Values & Measures | | | |
|---|---|---|---|---|---|---|---|---|
| Var # | Mean | Std.Dev | r | F | B | t-stat | R2 | σ |
| 0 | 19.046 | 1.028 | | | | | | |
| 1 | 594.483 | 71.648 | .787 | 185. | .011295 | 13.63659 | .620 | .634 |
| 2 | 33.888 | 15.418 | -.433 | 26. | -.028894 | -5.13520 | .188 | .930 |
| 3 | 68.147 | 25.320 | -.464 | 31. | -.018817 | -5.58575 | .215 | .915 |
| 4 | 51.664 | 3.565 | .530 | 44. | .152882 | 6.67904 | .281 | .875 |
| 5 | .060 | .239 | .107 | 1.32 | .460066 | 1.14960 | .0115 | 1.026 |

3) Age is similar, Fraz slightly older.

4) Although the t-statistics, R-square, and F-statistic are weaker for the Fraz database, any reasonable test passes with flying colors in both models for variables one through four.

5) The standard errors for the respective variables are very, very similar. This is the most important variable as far as actually applying the model because this represents the average error in minutes for the race. Obviously, both runners want this number to be small.

6) The authors are very pleased that these variables seem to be very important to both runners. A possible conclusion might be drawn that many runners would be able to predict good quantitative models if they keep these particular variables in their running logs.

Residual plots are examined for each variable in the Fraz database. While age shows a definite quadratic relationship as shown in Exhibit Three, the training and weight variables do not show the slightest non-linear relationship. They exhibit random scatters. Therefore the model that is used for the author's "A" database is the model that will be developed for the Fraz database. A transformation, using age squared, is calculated and the model determined that contains age and age squared. These models are compared in the ANOVA table in Table Eight.

Although the F-stat, t-stats, and adjusted R-square are weaker in the Fraz model, the key measure of performance, the standard error, is very similar. This is good news. Next, the final multiple regression model is developed. The model with the variable, "miles/14" is slightly stronger and is presented in Table Nine.

**Table Eight**

Model Statistics for Age and Age Squared

| Age and Aged Squared | F | A t-stat | A2 t-stat | adj R2 | σ |
|---|---|---|---|---|---|
| Fraz116 | 131 | 4.99877 | 5.49737 | .692 | .5678 |
| Database A | 566 | -10.2539 | 13.8928 | .844 | .529 |
| Database B | 457 | -9.6283 | 12.8657 | .814 | .564 |

## Exhibit Three

Residual Plot of Age Versus Pace for Fraz116

```
  1.69:o-----------------------------------------------------------
  1.53:------------------------------------------------------------
  1.36:----o-------------o-----------------------------------------
  1.20:---------------------------------------------------------o--
  1.04:-----------------------------------------------------oo-----
   .87:------o--------o-oo----------------------------------------o
   .71:------o-----------o-------------------------------oo----o---
   .54:----------------o-------------------------------o-----o-o---
   .38:-----o---o---------o---------------------------o------o-o-o-
   .21:---------------o--oo---------------------------------o-ooo-o
   .05:-o------o-------ooo------------------------------------o-o--
  -.12:-oo-----o-----o--o---oo-----------------o----------ooo-o----
  -.28:---------o-o--oo-------------------------------oo-o-o--o----
  -.45:--------o-----o-o------------------------oo-----o--oo-------
  -.61:--------o--oo-oo---o----------------------------o---o-------
  -.78:----------ooo--o-------------------------oooo---------------
  -.94:-----------------------------------------o-o-oo------o------
 -1.11:------------------------------------------------------------
 -1.27:------------------------------------------o-oo--------------
 -1.43:--------------------------------------------oo--------------
     ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^
    472  487  505  523  541  559  577  595  614  632  650  668  686
                                                 age in months
```

## Table Nine

ANOVA from Fraz116

| Source | Sum of Squares | df | Mean Square | F |
|---|---|---|---|---|
| Model | 94.691 | 5 | 18.938 | 77.746 |
| Error | 26.795 | 110 | .244 | |
| Total | 121.486 | 115 | | |

| Variable | Estimated Coefficient | t-statistic |
|---|---|---|
| Intercept | 31.59803 | |
| age/month | -.084111 | -3.98992 |
| age squared | .000080 | 4.45252 |
| weight | .059928 | 3.92820 |
| miles/14 | -.004884 | -2.28159 |
| adv cond | .585396 | 3.00486 |

adjusted R-square: .7673 standard deviation: .4935 minutes

A final residual plot reveals a random scâtter. Regrettably six years of data are missing, but the plot seems reasonable—we could guess at the residual in a general sense for the missing data. The residual plot is presented in Exhibit Four.

### Using the models to predict race time

The last step in this research is to use the models to predict a runner's pace for a race this coming weekend. A total of twelve models are "put to the test" to see how well they predict a runner's pace. It would be advantageous if a large number of races are available to check the predictability of the models. However, only four observations are available—two for each runner. On the morning of the race, each runner inputs the values of the various independent variables into a spreadsheet that calculates the expected pace pace (Y-hat) for a large group of acceptable models. Table Ten and Table Eleven show the twelve models, the predicted time and the actual race time (Y-i).

## Exhibit Four

Final Residual Plor for Database Fraz

```
 1.15:o----------------oo-------------------------------o---------
 1.02:---------------------------------------------------------o--
  .88:------------------------------------------------------------
  .75:---------------------------------------------------o--------
  .62:------o---------o-o-----------------------------------------
  .48:------o--------o---o---------------------------o----o---o--o
  .35:--------o----------o----------------------------o--o--o-----
  .22:--------------o-o-oo--o-----------------oo-----oo-------o---
  .08:---------o----o--o---------------------------------oo--oo---
 -.05:----oo-----------oo--o--------------------o-o-------o-o-o---
 -.18:--o-----------oo-------------------------o-o----o-o-oo-oooo-
 -.31:--------oo-o---o-------------------------o---o--o--o-ooo----
 -.45:-o-------------oo--o----------------------o---o-----oo-o-o-o
 -.58:-----------o--------------------------------oo---------o----
 -.71:---------o-oo------------------------------o-o--------------
 -.85:--------o-o-o-----------------------------------------------
 -.98:--------------------------------------------o---------------
-1.11:------------------------------------------------------------
-1.25:---------------------------------------------o--------------
-1.37:-----------------------------------------------------o------
     ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^
    472  487  505  523  541  559  577  595  614  632  650  668  686
```

**Table Ten**

Comparison of Predicted Pace and Actual Pace for Author for 12 Models

| Model# | Model Database: Indpt. Vars. | Race #1 Y-hat | Race #1 Y-i | Race #2 Y-hat | Race #2 Y-i |
|---|---|---|---|---|---|
| 1. | A:1,6,3,4,5* | 20.09 | 21.06 | 20.09 | 19.37 |
| 2. | B:1,6,2,7,4,5 | 19.96 | 19.96 | | |
| 3. | A:1,3,4,5 | 19.00 | 19.00 | | |
| 4. | B:1,2,4,5 | 18.68 | 18.68 | | |
| 5. | A:1,6,3,8,4,5 | 20.17 | 20.17 | | |
| 6. | B:1,6,2,4,5 | 19.77 | 19.77 | | |
| 7. | A:1,2,7,4,5 | 19.17 | 19.17 | | |
| 8. | B:1,3,8,4,5 | 19.02 | 19.16 | | |
| 9. | A:1,6,3,8,4,9,5 | 20.33 | 19.02 | | |
| 10. | B:1,6,2,7,4,9,5 | 19.72 | 20.33 | | |
| 11. | A:1,2,7,4,9,5 | 19.10 | 19.72 | | |
| 12. | B:1,3,8,4,9,5 | 18.94 | 18.94 | | |
| | Average: | 19.50 | 19.50 | | |

*legend:

| independent variable number | variable name |
|---|---|
| 1 | age |
| 2 | miles/7 |
| 3 | miles/14 |
| 4 | weight |
| 5 | adverse conditions |
| 6 | age square |
| 7 | miles/7 square |
| 8 | miles/14 square |
| 9 | weight square |

How well did the models perform? That is a very difficult call. They did not predict the race times to the second. The fact is that Frank Frazier performed better in both races than any of the models predicted. In addition, Frank is faster than Y-hat-1$\sigma$ for all of the models in the first race. Frank's comment about the first race, without knowing what times the models predicted, is that he had a "great race." For the author the situation is both good and bad. In the first race all of the models predicted a

**Table Eleven**

Comparison of Predicted Pace and Actual Pace for Fraz116 for 12 Models

| Model Database: Model# | Indpt. Vars. | Race #1 Y-hat | Race #1 Y-i | Race #2 Y-hat | Race #2 Y-i |
|---|---|---|---|---|---|
| 1. | 1,6,3,4,5 | 21.44 | 19.80 | 21.35 | 20.02 |
| 2. | 1,6,2,7,4,5 | 21.56 | | 21.48 | |
| 3. | 1,3,4,5 | 20.50 | | 20.34 | |
| 4. | 1,2,4,5 | 20.51 | | 20.37 | |
| 5. | 1,6,3,8,4,5 | 21.14 | | 21.05 | |
| 6. | 1,6,2,4,5 | 21.52 | | 21.44 | |
| 7. | 1,2,7,4,5 | 20.52 | | 20.38 | |
| 8. | 1,3,8,4,5 | 20.47 | | 20.32 | |
| 9. | 1,6,3,8,4,9,5 | 21.46 | | 21.41 | |
| 10. | 1,6,2,7,4,9,5 | 21.31 | | 21.27 | |
| 11. | 1,2,7,4,9,5 | 20.59 | | 20.49 | |
| 12. | 1,3,8,4,9,5 | 20.57 | | 20.45 | |
| | Average: | 20.97 | | 20.86 | |

much, much faster time than is actually attained. In the second race the actual outcome is slightly better than predicted. Maybe, the models should be used as a "goal" to attempt to attain, and prepare to compete at that predicted level, rather than expecting to hit the Y-hat "right on."

## Conclusion

The time had finally come to wrap it all up. As a class we had come a long way. The goal of learning was reinforced by the process of discovery. When everyone is involved in building the model, no one is left behind. The interaction of the students, the professor, and the database itself made the project interesting and engaging. We are not spoon fed a sterilized, meaningless, laboratory tested batch of data from a textbook. Somehow, we all are invested in the project and the enthusiasm for the subject infected the entire class. We performed our analysis of the data in both a structured (the step-wise process) and non-structured way (the final decision). If the students could defend their models, the professor was pleased. Students developed 45 different models in detail, looking at t-statistics, adjusted R-squares, value of B-coefficients, standard deviations, and residual plots. After intense scrutiny and group discussion, we did not come to a consensus regarding the best model. The beauty of this approach is that we are not told (only to forget) what is the best model and why.

We took real and meaningful data, determined the independent variables, built and analyzed our models, and made informed decisions we could support with statistical measurements. There is little fear of being right or wrong, although oral presentations are always stressful! We are allowed to try our own ideas, and in the process we learned how to think. Finally, even though it was challenging material, we managed to have fun with it!

Here are just a few reasons why this case is such an effective way to learn a quantitative lesson in the classroom:

1) The data is real and timely.
2) The situation is realistic and not just a "classroom exercise."
3) Students are encouraged and expected to interact throughout the process.

4) The computer is used extensively.
5) Sophisticated models are developed using the computer.
6) Many steps are needed to reach a conclusion.
7) There is not one and only one correct answer.
8) The students enjoy the realistic and far-reaching discussions.
9) It is fun to compare models during the oral presentations.
10) When it makes sense, it sinks in!

**References**


Andrews, Robert L., "A Data Collection and Analysis Project for a Statistics or Quality Management Class." *Proceedings of the Southeast Decision Science q Institute,* 1995, pg. 127.

Coleman, B. Jay, *The Analysis of Statistical Relationships,* 1993.

"A Multiple Regression Model to Predict Road Race Performance." *Proceedings of Southeast INFORMS,* October, 2001.

Neter, John & Wasserman, William, *Applied Linear Statistical Models*, Richard D. Irwin, 1974.