Paper 337-2012

# Introduction to Predictive Modeling with Examples

David A. Dickey, N. Carolina State U., Raleigh, NC

## 1. ABSTRACT

Predictive modeling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or "dependent" variable and various predictor or "independent" variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable. Because these relationships are never perfect in practice, it is desirable to give some measure of uncertainty for the predictions, typically a prediction interval that has some assigned level of confidence like 95%. Another task in the process is model building. Typically there are available many potential predictor variables which one might think of in three groups: those unlikely to affect the response, those almost certain to affect the response and thus destined for inclusion in the predicting equation, and those in the middle which may or may not have an effect on the response. For this last group of variables, techniques to test whether to include those variables have been developed and research on this "model building" step continues today. This paper addresses some basic predictive modeling concepts and is meant for people new to the area. Predictive modeling is arguably the most exciting aspect in the emerging and already highly sought after field of data analytics. It is the way in which big data, a current buzz word in business applications, are used to guide decisions for smart business operations.

## 2. INTRODUCTION

This paper begins with an interesting example of simple linear regression in which the need for statistical inference, that is, the decision as to whether a potential predictor is or is not statistically significant, is demonstrated. The example consists of real data in which age at death is "predicted" by the length of the lifeline, a crease in the palm of the hand that superstition suggests as a predictor of life length. Most predictive models involve more than one predictor and this brings into play the possibility of multicollinearity which is simply an overlap or strong correlation between two of the predictors. In a bivariate example, the problems associated with this phenomenon are graphically illustrated and the effect on the statistical analysis is displayed. In data taken monthly, such as retail sales, hospital admissions, criminal activity, and environmental measurements, different monthly effects are often observed. In addition, there are sometimes level shifts in such data associated with events such as disease outbreaks, new regulatory legislation, strikes, or natural disasters. For modeling, indicator or "dummy" variables can be used to capture these effects. An example using traffic accident statistics in North Carolina will show how useful these indicator variables can be.

Up to this point all of the examples used will have involved target variables which, conditional on the values of the predictors, are assumed to be approximately normally distributed. This is not always a reasonable assumption. For example, the response may be binary, that is, a two level response. As an example, a historic data base of bank customers might include some who defaulted and many who didn't. Interest would lie in predicting the probability of a default. Methods for doing so lie in the realm of so-called generalized linear models. Again the idea will be to introduce and illustrate rather than to delve into the mathematical underpinnings of the methodology. Interesting historical data sets used here will include survival statistics from the sinking of the ship Titanic and data on the space shuttle missions leading up to the Challenger disaster, a data set also used in paper 263-2010 from the 2010 SGF.

## 3. SIMPLE LINEAR REGRESSION

The name simple linear regression is somewhat misleading. It is the model, not the method of fitting, that is simple. The model is of the form $Y=\alpha+\beta X+e$ where $\alpha$ and $\beta$ are the intercept and slope of a line relating Y to X and e is an error term accounting for the fact that in most practical situations, the (X,Y) points are not arrayed exactly in a straight line. The assumption is that the errors e have a normal distribution with mean 0 and some variance $\sigma^2$ that is to be estimated along with the $\alpha$ and $\beta$ coefficients. The data used here appear originally in Wilson and Mather (1974) and consist of Y=age at death and X=length of the so-called life line in the palm of the hand. The life line is a crease in the palm that superstition suggests may be a predictor of length of life. Shown on the right below is a plot of these data using the new graphics procedure PROC SGPLOT in SAS® software.

```
proc sgplot;
scatter Y=age X=line;
reg Y=age X=line;   run ;
```

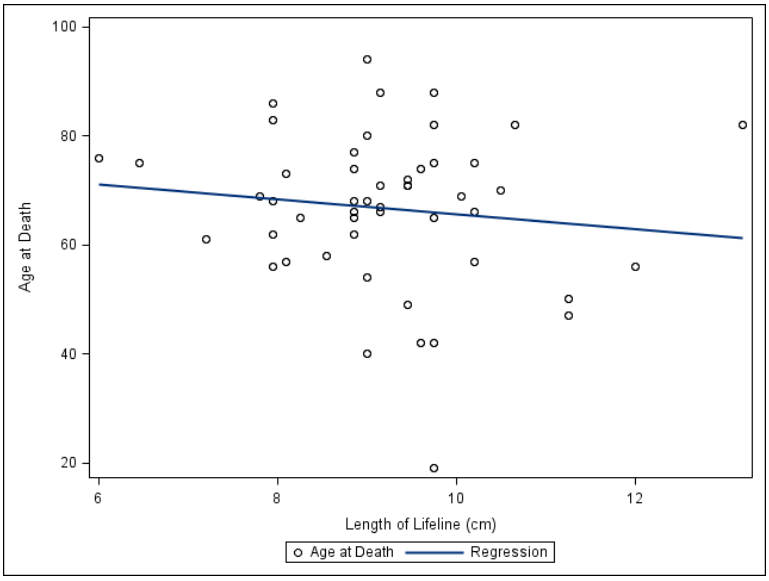


Figure 1: Palm reading example

The equation of the line shown is Predicted Age at Death = 79.24-1.367(Lifeline in cm.) . This implies that one looses 1.367 years of life for each centimeter of lifeline length. It is clear that there is a lot of variation around this line. Would another sample give a similar line? Perhaps another sample would give a horizontal line, indicating no relationship between the variables, or perhaps even an increasing rather than decreasing age associated with increasing lifeline length. Statisticians think about a true line that relates the two variables for all people and the behavior of sample estimates of that line. The sample estimates will vary around that true line. Assuming the deviations around the true population line are normal and the relationship is really linear, the sample slopes vary around the true slope with a standard deviation that can be estimated from a single sample. This estimated standard deviation is known as the standard error of the slope and it can be proven that t=(b-β)/se has a known distribution, the t distribution, where b is the estimated slope, β is the true population slope, and se is the standard error. Under the hypothesis that there is no relationship between the lifeline length and actual life the slope b would be 0 and it would be (b-0)/se that has a t distribution. Because t has a known distribution, we can compute the probability that, if β is really 0, t=(b-0)/se would exceed the t computed from our observed sample. This probability is referred to as the “p-value” for the t test and custom assigns the term “statistically significant” to t statistics with p-values less than 0.05. If the data are less likely than 5% to have occurred under our hypothesis (that there is no association) then we reject that hypothesis. Estimates of the slope and intercept along with standard errors, t ratios, and p-values are standard output in modern regression programs.

For the lifeline data, the SAS code

```
proc reg;
model age=line;run;
```

and (partial) output

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 79.23341 | 14.83229 | 5.34 | <.0001 |
| line | | 1 | -1.36697 | 1.59782 | -0.86 | 0.3965 |

give b=−1.367 as the estimated slope, standard error se=1.598, t ratio -1.367/1.598 = -0.86 and p-value 0.3965. Using our customary 0.05 to define unlikely, this slope is not at all unlikely to occur when the true slope is 0. Although a loss of 1 1/3 years of life per centimeter of lifeline may seem to be a large number, there is enough uncertainty in it that it is statistically insignificant. This sample does not provide sufficient evidence of a relationship between the length of the palm's lifeline and the actual duration of life.

How are the estimates of the intercept and slope determined? The principle is that of least squares, that is, for any potential intercept a and slope b, deviations or "residuals" (actual life – a – b(lifeline length)) are squared and added together across all the data points. The a and b that minimize this residual sum of squares are chosen. While these are actually computed from a formula, it is illustrative to show the sum of squares across a grid of (a,b) values. For this graph, the lifeline length is replaced by X = lifeline length-10 and the resulting regression line is Predicted Age at Death = 65.564 – 1.367X = 65.564-1.367(lifeline length-10), the same as before. There is a value M of this error sum of squares for which combinations (a,b) giving residual sum of squares exceeding M correspond to true intercept and slope combinations that have less than a 5% chance of delivering data like those observed. By truncating such sums of squares at M, the resulting graph shows a horizontal plane at height M with an elliptical hole representing a 95% confidence ellipse for the true intercept and slope pair. The lowest point on the plot is the minimum residual sum of squares and appears just above the (65.564, -1.367) least squares estimate of the (intercept, slope) pair. The area of this elliptical hole shows the joint uncertainty in the pair (a,b).

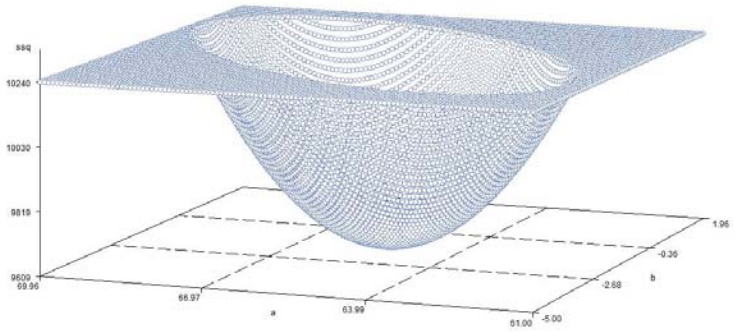


Figure 2: Error sum of squares versus intercept and slope estimates.

The main point here is that it is very important to consider the uncertainty in model parameter estimates. Even estimates that seem large may be indistinguishable from 0 in the presence of sufficiently large error terms.

Why did we change our predictor variable from lifeline length to (lifeline length -10)? Had we not done so, the ellipse above would be extremely narrow and long. While this makes the picture less appealing, it is not a *statistical* problem when one coefficient is the intercept and the other a slope for some variable X. However, this same sort of very long and narrow confidence ellipse can occur when slopes on two input variables are considered, and that can be an indication of a problem called multicollinearity. Muticollinearity results in estimates with very large standard errors and thus lots of uncertainty. This is exemplified by a long narrow confidence ellipse which indicates there are many values for one slope, some quite extreme, that are acceptable as long as the other slope falls within the narrow ellipse at that point.

## 4. Regression with multiple inputs

When there are 2 or more predictors, additional problems can arise, in particular the phenomenon known as multicollinearity. To illustrate in a dramatic fashion we use a rather extreme concocted example in which stores in a national chain choose to spend their advertising allocation on radio and television media in whatever proportions their managers choose. Here, in black, is a 3-D graph of the response (Y=sales) versus $X_1$ = radio advertising and $X_2$=TV advertising.

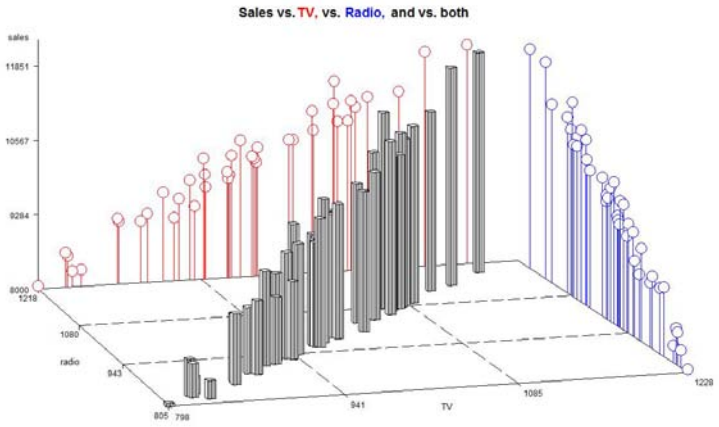


Figure 3. Sales versus advertising allocation (artificial data)

Along with the 3-D graph there are the two 2-D graphs (Sales versus TV in red on the back wall, and Sales versus Radio in blue on the right wall) projected onto the walls of the room in which the 3-D plot resides. We can see that the 3-D prisms have bases lying almost on the line where TV=Radio, that is, the store managers, receiving different total allocations for advertising, all tend to split them roughly equally between radio and TV advertising. In other words historically we have observed that, approximately, $X_1=X_2$ where these are the amounts spent on TV and radio. It follows that if $X_1=X_2$ exactly and sales = $\beta_0 + \beta_1X_1+\beta_2X_2$ we could substitute $X_1=X_2$ to get sales = $\beta_0 + 0\ X_1+(\beta_1+\beta_2)X_2$ or sales = $\beta_0 + (\beta_1+\beta_2)X_1+0X_2$ or sales = $\beta_0 + 5X_1+(\beta_1+\beta_2-5)X_2$ or any of an infinite number of equally good models of the form sales = $\beta_0 + CX_1+(\beta_1+\beta_2-C)X_2$ and we could pick any C we like, for example C= $\beta_1+\beta_2$ or C=5. It is the fact that $X_1$ is only *approximately* equal to $X_2$ that allows our least squares algorithm to distinguish at all between these possibilities but even so it is very *hard* to distinguish between them. A last point here is that this phenomenon would occur whenever $X_1$ is any multiple of $X_2$, for example if all store managers tended to spend 3 times as much on TV as radio.

The three graphs below help illustrate this idea. Clearly any of them would be a good predictor as long as the historical relationship between TV and radio advertising amounts were preserved. The graph on the left illustrates the least squares fit. The middle one has the radio coefficient (the slope along the left hand axis) set to 0 and the rightmost graph has the TV coefficient (the slope along the front axis, closest to the viewer) set to 0. Of course a coefficient 0 means that variable can be omitted. Clearly drastically different slopes in the two directions can still give good fits above locations on the floor where data have historically been. This phenomenon occurs because the footprints of the plotted prisms, the coordinates in the floor, are almost collinear.

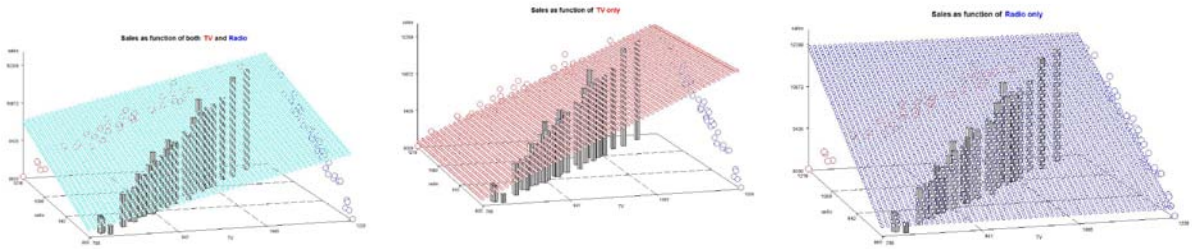

Figure 4. Effects of strong multicollinearity

It is obvious that none of these planes is horizontal – they all slope in some direction. The predictions for sales in the back right corner, where expenditures are high for both radio and TV advertising, should be higher than those with low expenditures in the front left parts of the graphs. We thus expect the overall model F test to be highly significant as it tests that the slopes of *both* TV and radio are 0, that is, it tests the hypothesis that a horizontal plane is OK. Moving to the individual t tests, we see that *individually* either variable can be omitted because the second and third prediction surfaces, which do just that, seem to provide a fine fit to the data. In fact the residuals from these three models (not shown) are very similar. Although there is clearly a positive combined effect of advertising, we have virtually no ability to separate the effects of the two forms of advertising. Notice how the two rightmost planes each intersect one wall in a line that forms the simple linear regression of sales on the associated variable, ignoring the other one. That is why the points were projected onto those walls in presenting the data in Figure 3.

As an analogy, imagine you are about to make a sandwich. You look in the cupboard and see a loaf of bread and some buns. If the bread is moldy, you'll throw it out and use a bun. If the buns are moldy, you'll use a couple of slices of bread. The "test" for bread and that for buns are both insignificant and these correspond to regression t tests. We can't interpret these jointly. With nothing else available (nothing else in our model) we cannot throw away both the bread and the buns and still make a sandwich. This corresponds to a significant F test in regression.

To perform the regression PROC REG can again be used. The procedure allows for multiple models with optional labels.

```
Title h=1.5 "Sales as function of " c=red " TV " c=black "and/or" c=blue " Radio";
PROC REG data=Stores;
  Both: model sales = TV radio/ss1 ss2;
  Both: model sales = radio TV/ss1 ss2;
  TV: model sales = TV;
  Radio: model sales = Radio;
run;
```

Note the ss1 and ss2 options as well as the difference in order of the model inputs in the first two models. The option ss1 requests the sequential or Type I sum of squares, that is, the reduction in error sum of squares resulting from adding each variable *in sequence*. Here is a partial output from the first model. Ignoring the intercept we see that this type I sum of squares is very large for TV, but with TV in the model, the additional reduction in error sum of squares (or increase in fit) is 40576, a much smaller number. The ss2 option asks what the contribution of a variable would be if it were fitted last. It is called the partial or Type II sum of squares. TV would only reduce the error sum of squares by 45254 if it were the last thing added to the model. As expected the F test is very significant while neither t test is significant. I cannot drop both TV and radio from my model but could drop either one with very little effect on my prediction accuracy. While there may be major differences in the two types of advertising, this particular data set sheds almost no light on the issue.

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 32660996 | 16330498 | 358.84 | <.0001 |
| Error | 37 | 1683844 | 45509 | | |
| Corrected Total | 39 | 34344840 | | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 531.11390 | 359.90429 | 1.48 | 0.1485 | 3964160640 | 99106 |
| TV | 1 | 5.00435 | 5.01845 | 1.00 | 0.3251 | 32620420 | 45254 |
| radio | 1 | 4.66752 | 4.94312 | 0.94 | 0.3512 | 40576 | 40576 |

The second model contains the same predictors as the first, just in a different order. The analysis of variance portion of the output is thus exactly the same as in the first model. Here is the parameter estimates table for that second model:

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 531.11390 | 359.90429 | 1.48 | 0.1485 | 3964160640 | 99106 |
| radio | 1 | 4.66752 | 4.94312 | 0.94 | 0.3512 | 32615742 | 40576 |
| TV | 1 | 5.00435 | 5.01845 | 1.00 | 0.3251 | 45254 | 45254 |

Note that the Type II sums of squares are the same as those from the first model, just in a different order. As always the Type I sum of square for the last variable in the model is the same as its Type II sum of squares. The Type I sum

of squares for radio is very large in the second output as it is coming into the model before TV. The error sum of squares is 1683844 for both of these models so if we were to leave TV out of the model, its Type II sum of squares indicates that the error sum of squares would increase to 1683844+45254=1729098 as is verified in the (partial) output from the model labeled "radio," shown below

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 32615742 | 32615742 | 716.79 | <.0001 |
| Error | 38 | 1729098 | 45503 | | |
| Corrected Total | 39 | 34344840 | | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 612.08604 | 350.59871 | 1.75 | 0.0889 |
| radio | 1 | 9.58381 | 0.35797 | 26.77 | <.0001 |

Even though the error sum of squares increased from the two variable model, the error mean square actually decreased. Note too that once we omit TV from the model, radio becomes highly significant according to its t test. Similarly, the model with only TV shows a very strong TV effect (output omitted). Either one of these can serve as a predictor but we do not need both.

Diagnosing multicollinearity, a strong linear relationship among predictor variables, can be difficult with many predictors. Illustrating with our 2 predictor model, Figure 5 is a plot of the TV (horizontal axis) and radio (vertical axis) expenditures for the 40 stores in our made up example. The strong correlation between TV and radio (0.997) indicates a high degree of collinearity as is displayed in the plot. On that plot a new pair of axes, called principal components, is shown. They are at 90 degrees to each other and the spread of the data along the first, positively sloping one is much longer than along the second.

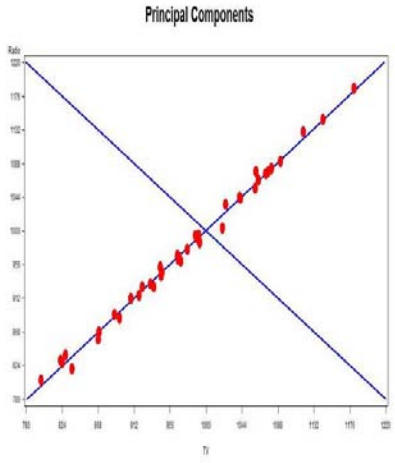


Figure 5. Principal component axes plotted on radio (vertical axis) and TV (horizontal) axes with data (dots) showing drastically greater spread on the first principal component than on the second.

The coordinates or "scores" on the principal component axes are independent. In other words if we rotated the graph so that the first principal component axis were horizontal then the point cloud would appear as a random scatter (with little spread in the component 2 direction). It is common to scale the original data so that the variance is 1 in each direction. In that case the distributions along the principal component axes have variances given by the eigenvalues of the correlation matrix of the original data. The sum of the diagonal elements (all 1s) of a correlation matrix is the number of variables, 2 in our case, and represents the total variation in both dimensions. The eigenvalues, give the amounts attributable to each principal component and these can be derived from PROC PRINCOMP as follows:

```
proc princomp data=stores;
   var TV radio;run;
```

Correlation Matrix

| | TV | radio |
|---|---|---|
| TV | 1.0000 | 0.9974 |
| radio | 0.9974 | 1.0000 |

Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 1.99737404 | 1.99474809 | 0.9987 | 0.9987 |
| 2 | 0.00262596 | | 0.0013 | 1.0000 |

Eigenvectors

| | Prin1 | Prin2 |
|---|---|---|
| TV | 0.707107 | 0.707107 |
| radio | 0.707107 | -.707107 |

The extreme correlation 0.9974 between TV and Radio is displayed and the total variation 2 is broken into 1.9964 along the first principal component axis and the remainder, 0.00263 along the second principal component axis. The eigenvectors indicate that 0.707T*+.707R*, or 0.707 times the sum of the normalized TV and radio amounts T* and R*, is the first principal component and a multiple of the difference is the second. The principal components are also called eigenvectors of the correlation matrix. The ratios of the largest eigenvalue to the others are indicators of the relative spread of the data in the associated principal component directions. In our case the relative spread 1.99737/0.002626 = 760.6 is quite large consistent with the small spread along the second principal component axis relative to the first. Its square root, 27.58, is a ratio of standard deviations along the two principal component axes and is referred to as a condition index. As a rule of thumb (see Rawlings et al 1998), indices larger than 30 are considered problematical in terms of the instability in the fitted regression surface so we are close enough to that problematical region to be concerned. This instability results in inflated variances of the parameter estimates and the way each parameter is affected by a large condition index is displayed when the COLLINOINT option in PROC REG's MODEL statement is used. For our data the displayed collinearity diagnostics show one bad condition index 27.5795 as expected, and it is seen that over 99% of the variation in each parameter estimate is caused by this multicollinearity problem. Here are a program and the relevant output:

```
PROC REG data=Stores;
  MODEL sales = TV radio/collinoint; run;
```

Collinearity Diagnostics (intercept adjusted)

| | | Condition | --Proportion of Variation- | |
|---|---|---|---|---|
| Number | Eigenvalue | Index | TV | radio |
| 1 | 1.99737 | 1.00000 | 0.00131 | 0.00131 |
| 2 | 0.00263 | 27.57948 | 0.99869 | 0.99869 |

Excess variation is a problem in estimation. The VIF option in PROC REG's MODEL statement shows a "Variance Inflation Factor," a multiplier comparing the variance we have to what would have happened had our predictors been orthogonal to (uncorrelated with) each other. Our VIF output looks like this

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
|---|---|---|---|---|---|---|
| Intercept | 1 | 531.11390 | 359.90429 | 1.48 | 0.1485 | 0 |
| TV | 1 | 5.00435 | 5.01845 | 1.00 | 0.3251 | 190.65722 |
| radio | 1 | 4.66752 | 4.94312 | 0.94 | 0.3512 | 190.65722 |

The variances of our parameter estimates are 190 times what they would have been if the inputs had been un correlated with each other. In this extreme case, replacing the two input variables by just one of them or by their average (which happens to be a multiple of the first principal component in this example) will give a coefficient approximately 9.671, the sum of the two coefficients above. For example our earlier regression on radio alone gave

slope 9.58. Regressing on all the principal components gives the same predictions as regressing on the original data and one proposed cure for multicollinearity is to omit principal components with high condition indices (relatively small eigenvalues). Another is to identify, perhaps with VIF or condition indices, sets of variables that are highly collinear and select one of them to represent the whole group.

## 5. Regression with interactions

Another nice, but concocted, teaching example is this one involving study time, IQ, and grades. Code and partial output are as follows:

```
Data tests; input IQ Study_Time Grade; IQ_S = IQ*Study_Time;
cards;
105      10     75
110      12     79
120      6      68
116      13     85
122      16     91
130      8      79
114      20     98
102      15     76
;
Proc reg data=tests;  model Grade = IQ;
Proc reg data=tests;  model Grade = IQ  Study_Time; run;
```

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 62.57113 | 48.24164 | 1.30 | 0.2423 |
| IQ | 1 | 0.16369 | 0.41877 | 0.39 | 0.7094 |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.73655 | 16.26280 | 0.05 | 0.9656 |
| IQ | 1 | 0.47308 | 0.12998 | 3.64 | 0.0149 |
| Study_Time | 1 | 2.10344 | 0.26418 | 7.96 | 0.0005 |

The first regression indicates no effect of IQ on the test results, which seems counterintuitive, but in the presence of study time, IQ becomes significant. One possible explanation for this is that the students with higher IQ studied less, counting on their native intelligence, while the others studied more. Adjusted for study time, that is, looking at what the second model says will happen when students study the same amount of time, there is a significant effect of IQ. This shows that a given variable can become *more* important in the presence of another variable whereas the radio and TV example had the opposite effect.

Another opportunity afforded by this little example is the chance to talk about interaction. As it stands, one extra unit (hour) of study is predicted to raise a person's grade by 2.103 points. This is true regardless of IQ according to the model and since that does not make much sense, we might propose an improved model in which the number of points gained from an extra hour of study depends on IQ. We can do that by simply including a cross product between study time and IQ which we will call IQ_S. The new term is insignificant but remember that insignificance does not imply the true parameter is 0, it only says there is not enough *evidence* in this small data set to convince a nonbeliever that the parameter is not 0. Having a common sense reason for including the interaction, we proceed to explore the resulting model. Note that, as before, no t test is significant and yet the F test tells us we cannot leave out all of our variables.

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 610.81033 | 203.60344 | 26.22 | 0.0043 |
| Error | 4 | 31.06467 | 7.76617 | | |
| Corrected Total | 7 | 641.87500 | | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 72.20608 | 54.07278 | 1.34 | 0.2527 |
| IQ | 1 | -0.13117 | 0.45530 | -0.29 | 0.7876 |

| | | | | | |
|---|---|---|---|---|---|
| Study_Time | 1 | -4.11107 | 4.52430 | -0.91 | 0.4149 |
| IQ_S | 1 | 0.05307 | 0.03858 | 1.38 | 0.2410 |

Interpretations of main effects that are involved in interactions is ill advised. The negative coefficients on Study time and on IQ suggest that being more intelligent is a detriment and you should study as little as possible. A look at the model shows the fallacy in this argument. Using S for study time, notice that our model is

$$\text{Grade} = (72.206 - 0.13117\text{IQ}) + (0.05307*\text{IQ} - 4.11107)*\text{S}$$

So that a person with IQ 100 would have predicted grade (72.206-13.117) + (5.307-4.111)S = 59.08908 + 1.19593S, thereby predicting a score of 59.09 with no studying plus a gain of 1.20 points per hour of study. For IQ 122 we have predicted grade 56.20334 + 2.36347S. Here is a graph of the predictions for the model with interaction the front edge of the surface is at IQ 100 and the back at IQ 122. Within this range of IQs, all study time (the front axis) slopes are positive.

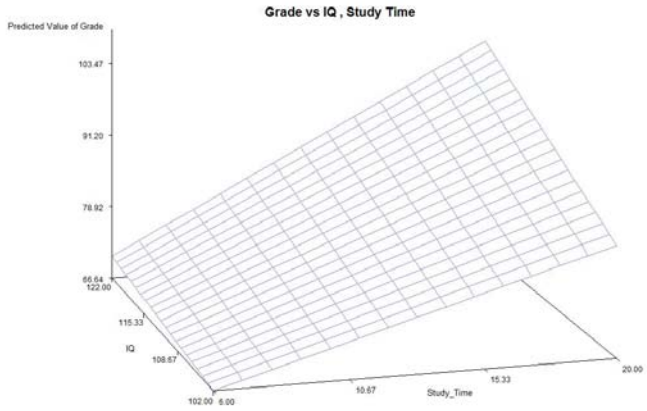


Figure 6. Predicted grade as a function of IQ (front to back) and study time (left to right) for interaction model.

## 6. Regression with indicator variables

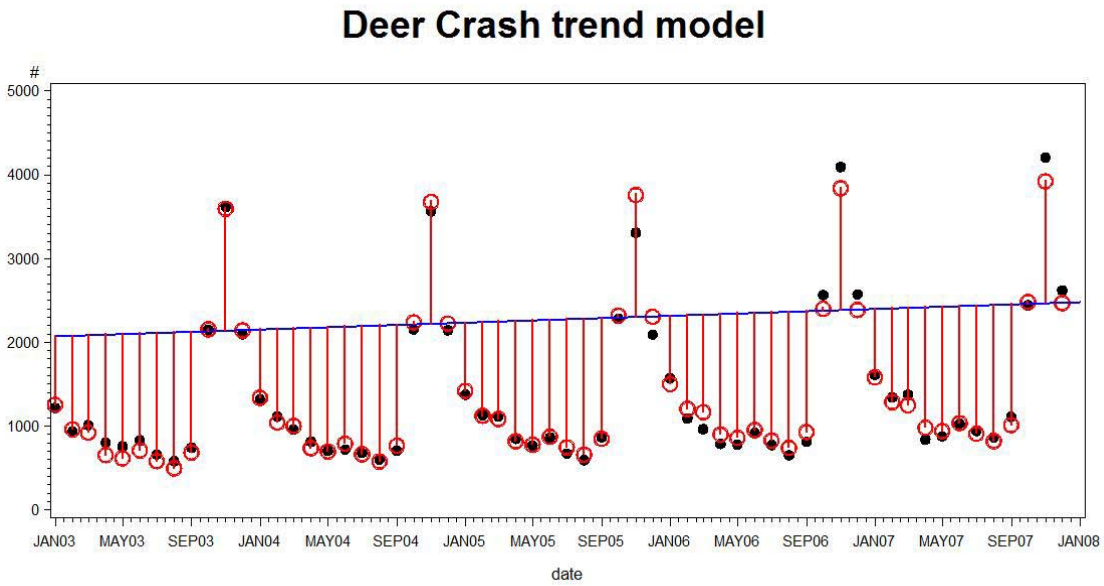


Figure 7. Monthly car accidents in North Carolina involving deer. Counts are dots, predictions are circles.

Figure7 is a plot of the number of automobile accidents in North Carolina in which the accident report contained the word “deer” such as “I swerved to avoid a deer” along with predictions (the circles) from a model to be discussed.

The data are taken over time and we see a slight increase, possibly due to increased driving, a growing deer population or both. The data are monthly and a strong repeating monthly pattern is evident. This corresponds to deer mating patterns, giving high values around November during the mating season. Effects such as this can be modeled using indicator or “dummy” variables. For example, imagine making up a variable NOV which is 1 in November and 0 everywhere else. A model of the form Accidents = 100 + 20*NOV would predict 100+0 accidents for all months other than November and 120 for November. Likewise a model of the form Accidents = 100 + 3*t+ 20*NOV would predict that accidents increase linearly in time t with the November values sitting 20 units above that line. The line can be thought of as a base line and indicator variables for the months can be included, allowing each to deviate from the line. Notice that if we use 12 indicator variables, one for each month, then the base line would have an arbitrary height, that is, we could increase the intercept by any number we like and then subtract that number from each monthly deviation, resulting in the same predictions. With all 12 indicator variables, the coefficients are not uniquely determined which is a problem. One solution is to eliminate one monthly indicator, for example the December one, so that the line predicts December accidents and the other 11 indicators have coefficients giving the deviations from that December line for the other months. In Figure 7 is a plot showing the data as dots. The predictions are circles, and the line runs through the December predictions. The high values each year are the November (mating season) values from which the December dates are easily identified. Note that the October deviations (from the December line) happen to be close to 0 for these particular data.

The analysis code and output give some insight into the elements of the graph. Using

```
Proc reg data=deer; model deer = X1-X11 date;
  Label X1="Jan-Dec" X2="Feb-Dec" X3="Mar-Dec"  X4="Apr-Dec"  X5="May-Dec"
   X6="Jun-Dec"  X7="Jul-Dec" X8="Aug-Dec" X9="Sep-Dec" X10="Oct-Dec"
   X11="Nov-Dec"; run;
```

we obtain these parameter estimates:

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -1439.94000 | 547.36656 | -2.63 | 0.0115 |
| X1 | Jan-Dec | 1 | -811.13686 | 82.83115 | -9.79 | <.0001 |
| X2 | Feb-Dec | 1 | -1113.66253 | 82.70543 | -13.47 | <.0001 |
| X3 | Mar-Dec | 1 | -1158.76265 | 82.60154 | -14.03 | <.0001 |
| X4 | Apr-Dec | 1 | -1432.28832 | 82.49890 | -17.36 | <.0001 |
| X5 | May-Dec | 1 | -1478.99057 | 82.41114 | -17.95 | <.0001 |
| X6 | Jun-Dec | 1 | -1392.11624 | 82.33246 | -16.91 | <.0001 |
| X7 | Jul-Dec | 1 | -1525.01849 | 82.26796 | -18.54 | <.0001 |
| X8 | Aug-Dec | 1 | -1618.94416 | 82.21337 | -19.69 | <.0001 |
| X9 | Sep-Dec | 1 | -1436.86982 | 82.17106 | -17.49 | <.0001 |
| X10 | Oct-Dec | 1 | 27.42792 | 82.14183 | 0.33 | 0.7399 |
| X11 | Nov-Dec | 1 | 1459.50226 | 82.12374 | 17.77 | <.0001 |
| date | | 1 | 0.22341 | 0.03245 | 6.88 | <.0001 |

Recall that dates in SAS are counted in days so the slope of the December line shows an increase of 0.22341 accidents *per day* and the intercept (date 0 is Jan 1, 1960) is -1439.94. Labels remind us that the effects for months are really differences between that month and December. The only large positive deviation is for November, the October deviation from the December line is close to 0 and the other months deviate from the December line by substantially large negative values. Since the arbitrary choice of December as a baseline is what gives the insignificant October deviation, we recommend not omitting the October indicator variable. For example, had we used April as the baseline, October would have had a significant deviation but all the predicted responses would be exactly the same. The R-square for this model exceeds 0.98 where R-square measures the proportion of variation explained by the model.

## 7. COUNT DATA

When the response variable is a category, such as survived or did not survive, and the predictors are also categories, contingency table approaches are appropriate. In many instances categories are binary, though the procedures described herein can handle an arbitrary number of response categories. Our example here is the survival data from the wreck of the Titanic ocean liner. We want to predict survival probabilities for each of 4 status

categories (first, second, and third class, and crew) and test to see if they are all the same. A cross tabulation of survival by status counts is called a contingency table and these are the results for the Titanic:

| Status→ Survived? | Crew | First class | Second class | Third class | Totals |
|---|---|---|---|---|---|
| Alive | 212 | 202 | 118 | 178 | 710 |
| Dead | 673 | 123 | 167 | 528 | 1491 |
| Totals | 885 | 325 | 285 | 706 | 2201 |

We see that the overall survival rate was 710/2201 and if that proportion survived in each category, then in a category with 706 passengers, we would expect 706(710/2201) = 227.74 rather than the 178 that we observed. The quantity (observed-expected)$^2$/expected, when computed for each cell of the table and summed, form a statistic that has a Chi-squared distribution under the hypothesis that people in the different status categories have theoretically the same probability of survival. The proportion surviving in each category, 212/885=24% for crew and 62%, 41%, 25% for first through third class are the predictions of the theoretical survival probabilities. The following code produces the predictions, the Chi-squared test that survival does not depend on status, and p-values.

```
data titanic;
input type$ fate$ count @@;
datalines;
crew alive 212 crew dead 673
first alive 202 first dead 123
second alive 118 second dead 167
third alive 178 third dead 528
;
proc freq; weight count;
tables fate*type / chisq nocol nopct; run;
```

```
                The FREQ Procedure
               Table of fate by type
fate      type

Frequency|
Row Pct  |crew    |first   |second  |third   |  Total
---------+--------+--------+--------+--------+
alive    |    212 |    202 |    118 |    178 |    710
         |  29.86 |  28.45 |  16.62 |  25.07 |
---------+--------+--------+--------+--------+
dead     |    673 |    123 |    167 |    528 |   1491
         |  45.14 |   8.25 |  11.20 |  35.41 |
---------+--------+--------+--------+--------+
Total         885      325      285      706     2201

Statistics for Table of fate by type

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     3    187.7932    <.0001
Likelihood Ratio Chi-Square    3    178.4143    <.0001
Mantel-Haenszel Chi-Square     1      0.0000    0.9982
Phi Coefficient                       0.2921
```

The first two Chi-Square tests agree that there is a highly statistically significant difference in survival rate among the 4 status categories and those tests have 3 degrees of freedom as expected from the formula df=(r-1)(c-1) where r and c are the number of rows (2) and columns (4) of the table. The Mantel-Haenszel Chi-Square has only 1 degree of freedom and does not find any differences in survival across the categories. Why? This statistic is based on the correlation between two columns of numbers, one having value 1 or 0 (alive or dead) for each passenger and the

other giving the rank (1 through 4) of their status. Ranks are assigned by the labels in alphabetical order (crew=1, first=2, second=3, third=4). in other words, the status category ranks are being treated as a continuous variable and since the crew and third class passengers both had about 24% survival there seems to be very little linear association. Note also the 1 degree of freedom as is typical of a continuous predictor variable. Changing the alphabetical order by giving crew a label z_crew (or using a format) the Mantel_Haenszel changes the output to read

```
Mantel-Haenszel Chi-Square     1    160.0011    <.0001
```

Sometimes the predictor variables are continuous to start with and in that case a logistic regression approach is often appropriate. Probabilities p and 1-p of the two outcome categories at given settings of the predictor variables are of interest here. The strategy here is to fit a linear model to L=ln(p/(1-p)) where L is called the logit. For example, with one predictor X, we might find the predicting equation L=0.7 + 0.5X. Depending on X, L could take on any value, positive or negative and of course probabilities must stay between 0 and 1. Solving L = ln(p/(1-p)) for p shows that $p=e^L/(1+e^L)$ and we note that for any L, $e^L>0$ so the predicted p using this formula will always be positive, even if L is negative. Furthermore, $e^L/(1+e^L)$ is a positive number divided by itself plus 1 so the ratio cannot exceed 1. Roughly the idea is to model L with a regression like procedure then convert to p.

A real data example is provided by the US space shuttle program. Prior to the Challenger disaster, 23 missions were flown and the O-rings were inspected upon recovery of the rocket components. Some O-rings showed flaws denoted as erosion and blowby. We will call these “failures.” We will assign a category 1 if either flaw was observed and 0 otherwise. There are 6 O-rings per mission and it is possible, using PROC GLIMMIX, to treat mission as a random effect but we omit that case from the discussion here, referring the reader to Dickey (2010) . Ultimately, the investigation of the Challenger disaster suggested that the cold temperature at launch affected the O-rings on that mission which led to the mixing and ultimate ignition of the fuel components thus exploding the shuttle and killing all on board. We here model the flaw status, 0 or 1, in terms of temperature using logistic regression. In other words we model the logit L of the probability of a flaw as a linear function of launch temperature. We will use both the launch temperature and the mission number as (continuous linear) inputs to the model. Here is the SAS code for doing so:

```
proc logistic data=O_ring; title3 "Logistic Regression";
model fail(event='1') = temp launch;
output out=out1 predicted = p; run;
```

The event = ‘1’ option tells PROC LOGISTIC to model the probability of a flaw (fail=1). The status of each O-ring, 0 or 1, is recorded in the data set as is the launch temperature (temp) and launch number (launch) for the mission. Here is some partial output.

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | 4.0577 | 3.0314 | 1.7917 | 0.1807 |
| temp | 1 | -0.1109 | 0.0445 | 6.2122 | 0.0127 |
| launch | 1 | 0.0571 | 0.0563 | 1.0311 | 0.3099 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| temp | 0.895 | 0.820 | 0.977 |
| launch | 1.059 | 0.948 | 1.182 |

We see that there is a significant negative effect of temp on L and hence on the probability of failure. This means that L and the probability of failure will decrease as the temperature increases and increase as the temperature gets colder. What are those “odds ratios” in the output? Note that if a horse has a p=2/3 chance of losing and a (1-p)=1/3 chance of winning, the odds against that horse are (2/3)/(1/3)= p/(1-p) or 2 to 1. This means that L=ln(p/(1-p)) is the logarithm of the odds for given values of the predictors. Now if one of those predictors increases by 1, then L will increase by the coefficient b of that variable. Furthermore, if L=ln(p/(1-p)) increases by b, then $e^L$ = p/(1-p), the odds, gets *multiplied* by $e^b$, that is, $e^{L+b} = e^L e^b$ and so $e^b$ is called the odds ratio. For values of b near 0, $e^b$ wil be quite near 1+b and this holds pretty closely for our two input variables. Multiplying by 0.895 decreases the odds (as temperature increases by 1) while multiplying by 1.059 increases the odds (as we progress to the next launch)

however this launch effect is insignificant - there is insignificant statistical evidence of increasing failure probabilities over time, adjusting for temperature. Leaving launch out of the model, the prediction of L becomes 5.085 - 0.1156(temp) which we will use to make a plot. The predicted values of a prespecified O-ring having a flaw along with the original data are plotted against temperature in Figure 8 where temperature is allowed to go down to 31 degrees F, the launch temperature for the tragic Challenger mission.

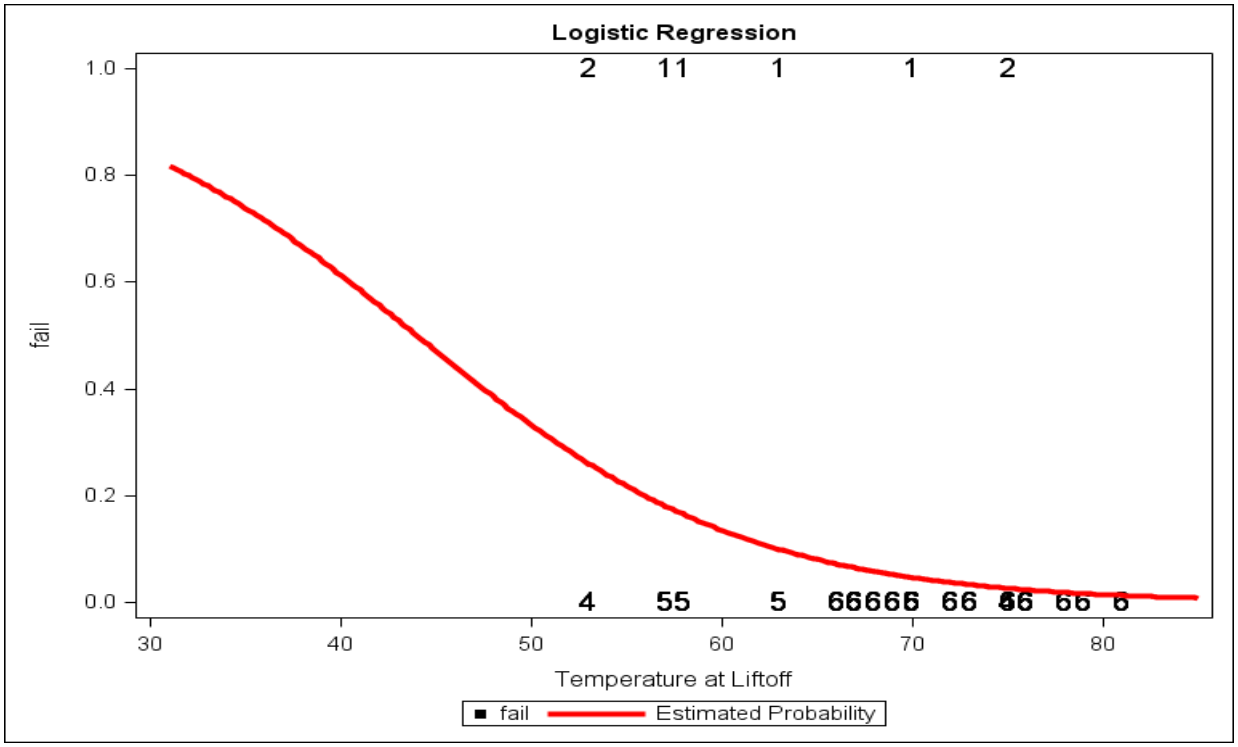


Figure 8. PROC SGPLOT for logistic regression on space shuttle O-ring data.

Note that we are assuming the model holds at temperatures far below those observed in the first 23 missions and further note that erosion and blowby are not necessarily fatal, but ignoring those important caveats, the model suggests that each of the six O-rings on the Challenger mission has a very high probability of a flaw when the temperature is 31 degrees. Note also that the graph shows the number of O-rings that each point represents as a symbol along with a series plot of the predicted values. This was done in the new SAS graphics procedure PROC SGPLOT with the following code (the variable "symbol" is the number of O-rings represented).

```
proc sgplot;
  scatter Y=fail X=temp/markerchar=symbol markercharattrs=(size=12 color=black);
  series Y=p X=temp/lineattrs=(color=red thickness=3);  run;
```

While a similar plot could have been produced in PROC GPLOT, more code would be involved. For example, the use of the Y*X=symbol syntax on the PLOT statement cannot be used when overlaying so the template facility might have to be employed. An axis statement and option would be needed to turn the "fail" label vertical.

## 8. CONCLUSION

Predictive modeling is introduced through a series of examples in which some of the power of SAS software is illustrated. With that power comes a responsibility on the part of users to continually educate themselves on the proper interpretation of the results, the appropriate tool for the data at hand, and hidden dangers such as multicollinearity, outliers, interactions, etc. Properly used, the analytic predictive tools shown herein can provide valuable insights into data of all kinds. Predictors and/or responses can be categorical or continuous and the proper tool depends on these data properties. Examples of all combinations have been given.

## 9. REFERENCES

Dickey, D. A.(2010) Ideas and Examples in Generalized Linear Mixed Models. *Proceedings of 2010 SAS Global Forum*, paper 236.

Rawlings, J. O., S. G. Pantula and D. A. Dickey (1998). Applied Regresson Analysis: A Research Tool. Springer-Verlag, N.Y.

Wilson, M. E. and L. E. Mather (1974) . *J. Amer. Med. Assn.* 1229(11):1421-1422.

## CONTACT INFORMATION

Name: Professor David A. Dickey
Enterprise: Department of Statistics
Address: Box 8203, North Carolina State University
City, State ZIP: Raleigh, NC 27695-8203
E-mail: dickey@stat.ncsu.edu
Web: http://www4.stat.ncsu.edu/~dickey/