

Statistics & Regression: Easier than SAS®

Vincent Maffei, Anthem Blue Cross and Blue Shield, North Haven, CT
Michael Davis, Bassett Consulting Services, Inc., North Haven, CT

ABSTRACT

In this paper the basics of estimation and hypothesis testing are covered without the associated probability theory and mathematical derivations of formulas. Basic concepts are explained logically and graphically. Without the mathematical wizardry, the simple underlying thread is exposed.

With this basic understanding, the analysts will know the methodology to use to generate estimates with valid error margins, and to set up tests of hypothesis (i.e. unsubstantiated claims about performance, effectiveness, cost savings,?) using either univariate statistics or regression analysis.

INTRODUCTION

PROC REG or PROC GLM allows SAS users to perform multivariate regression. SAS programmers are adroit enough to navigate their way through the code and successfully generate a load of statistical output. Unfortunately, many do not understand the statistical concepts of estimation, hypothesis testing, regression and its pitfalls well enough to properly interpret the statistical output and to be confident in their conclusions.

The intent of this paper is to cover the basics of estimation and hypothesis testing using regression, and to explain some of the more common pitfalls and how to avoid them. The authors assume that the audience has been exposed to inferential statistics at some point in their education, knows the concepts of arithmetic mean and standard deviation, has had some experience using z & t tables, and at least a vague notion of confidence intervals and hypothesis testing.

The discipline that we adhere to is that of classical statistics. Much of classical statistics is built upon the assumption of normality. This is fortunate since normal distributions are quite prevalent. Even if an underlying population is not normal, estimates (such as the sample mean) that are drawn from a non-normal population are usually normally distributed. Even though we usually only take one sample (and calculate one sample mean) we must recognize that our sample is but one of perhaps millions of different

possible samples that could have been taken. Different timing, different turn of events, different pollster, etc., yield different samples, each with its own sample mean. The center of that distribution (i.e. the mean of the distribution of all possible sample means) is the mean of the parent population (μ), as shown in Figure 1. The standard deviation of that theoretical distribution of all possible sample means is the standard deviation of the parent population divided by the square root of the sample size (σ/\sqrt{n}).

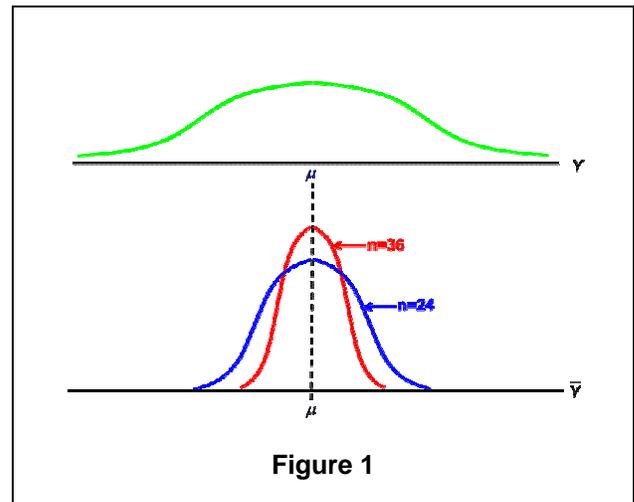


Figure 1

This concept, that a sample estimate is drawn from a theoretical distribution of millions of different possible estimates extends to all types of estimates, proportions, paired differences, and even regression coefficients.

BASIC HYPOTHESIS TESTING

The steps in hypothesis testing are fairly simple. Based on some claim or proposition, we formulate a Null & Alternative hypotheses (H_0 & H_a). Next, we take a sample and calculate the sample statistic and its variance. Finally, we check to see if the difference between the sample statistic and the hypothesized 'true' value is small enough to be attributed to random variation, or is so large that we chose to believe the sample statistic came from a distribution that is centered over a value very much different than the hypothesized one. Whatever

situation prevails will allow us to draw a conclusion about the initial proposition.

Paradoxically, classical statistics usually starts by formulating a null hypothesis that is contrary to the original proposition. The alternative hypothesis is customarily expressed in a manner that is consistent with the initial proposition. For example, if the authors were to claim that SAS consultants with at least five years of experience make more than \$150 an hour on average, you would set $\mu = \$150$ in your null hypothesis (H_0) and $\mu > \$150$ as your alternative (H_a). If you wanted to be 95% confident in the correctness of your conclusion you would select a 'critical' z value such that only 5% of all possible sample means would generate z values 'above' that critical z value. From the z table that value is 1.645 (see Figure 2).

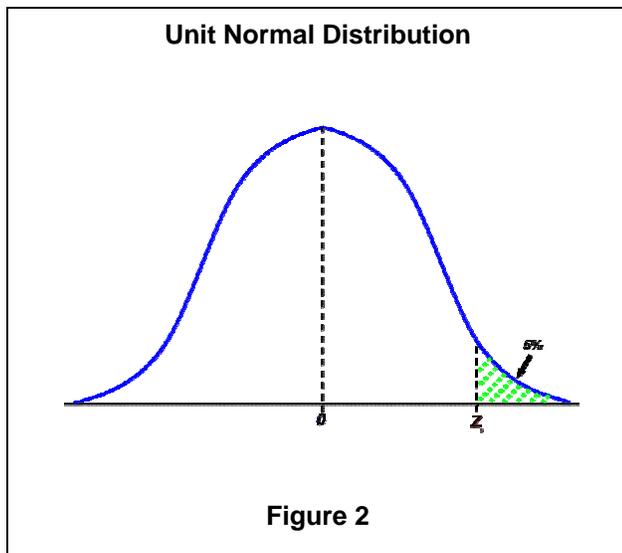


Figure 2

To test this claim you take a sample of 36 and find the sample mean to be \$154. Assume for the time being that you know (from another study) that the population standard deviation is \$18. All you need do now is calculate the z value that corresponds to our sample mean and see if it is above or below the 'critical' z value.

The formula for calculating a z value from the sample statistics is always the 'observation' minus the mean of the distribution, divided by the standard deviation. (The z value calculated from the sample statistics is not to be confused with the critical z_c value that you look up in the tables.) In this example you observe a sample mean from the theoretical distribution of all possible sample means. The mean of that distribution is the hypothesized population mean, (until proven false). The standard deviation is σ/\sqrt{n} . The z value corresponding to our sample

mean is $z = (154 - 150)/(18/\sqrt{36}) = 1.33$. Since this is less than the critical z of 1.645 you accept the null hypothesis that $\mu = \$150$ and reject the alternative that $\mu > \$150$. Your rationale is that the difference between the sample mean and the hypothesized population mean is small enough to be produced by random variation. A population with a mean of \$150 can easily generate samples (of size 36) with sample means around \$154. Therefore, a mean of \$154 is not sufficient evidence to accept the authors' claim. (See Figure 3.)

$H_0: \mu = \$150$
 $H_A: \mu > \$150$

$Z = \text{observation} - H_0 \text{ value} / \text{std deviation}$

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$\alpha = 5\% \text{ and } n = 36 \implies Z_c = 1.645$

$1.33 = \frac{(154 - 150)}{18/\sqrt{36}}$

If $Z \leq 1.645$ then accept H_0

Figure 3

If the sample mean had been \$158 instead, the z value would have been $(158 - 150)/(18/\sqrt{36}) = 2.67$. Since this is more than the critical of 1.645 you would reject the null hypothesis and accept the alternative. Your rationale this time is that the difference between the sample mean and the hypothesized population mean is too large to be explained by random variation. A population with a mean of \$150 is not likely to generate samples (of size 36) with sample means as high as \$158. (The probability is less than the predetermined 5%.) It is more likely, that this sample mean came from a population who's mean is significantly greater than \$150. In this situation, a mean of \$158 would be considered sufficient evidence to accept the authors' claim.

In almost all real situations you do not know the population mean (μ) or standard deviation (σ). You must substitute the sample standard deviation (s) for the unknown population standard deviation. This introduces error into our calculation of the z value, since s is only an estimate of σ . The error is the difference between σ and its sample estimate. This

additional error necessitates that you switch from the **z** table to the **t** table. The **t** distribution is similar in concept to the **z**, except that it has more area in the tails to compensate for the error introduced by substituting **s** for σ . (See Figure 4).

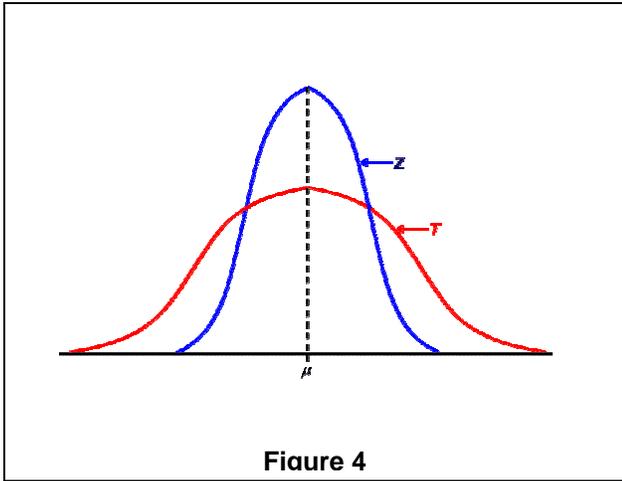


Figure 4

REGRESSION ANALYSIS

By employing regression we are postulating that the value of one of the variables (referred to as the dependant variable and symbolized by Y), is explained by the values of the other variables (referred to as the explanatory variables and symbolized by X 's). In the simplest model $Y = \alpha + \beta X + \varepsilon$. α is the value of the vertical intercept, β is the slope coefficient (the change in Y for a unit change in X), and ε is a normally distributed random disturbance term. If we could freeze X at the value X_1 , and then sample repetitively, we would discover that the resulting values of Y would be normally distributed around the value $\alpha + \beta X_1$. In concept, there is a whole set of normal distributions centered over the regression line (as in Figure 5). Given that the underlying distribution is normal, the concepts of hypothesis testing discussed earlier still apply.

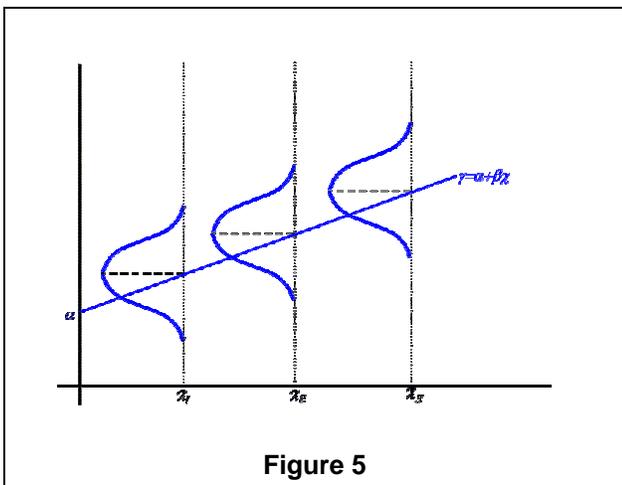


Figure 5

Typical uses of regression analysis include hypothesis testing, parameter estimation, and prediction. An example of hypothesis testing would be to test the claim that each \$1 spent on preventive care this year, trims \$10 from total health care costs next year. There are many factors driving health care costs, and changes in any one of them will cause changes in next year's health care costs. We need to account for all these factors in order to separate out the impact of a change in preventive care spending. Regression analysis is ideally suited for this task as we can include all the major factors that affect health care costs in the regression equation.

An example of estimation would be to re-estimate the Marginal Propensity to Consume (the fraction of each dollar of income that households spend, on average). Old economics textbooks placed the MPC at approximately 90%. Following the bursting of the stock market bubble and the events of 9/11 there is ample evidence to indicate that households are spending far less than they used to. In order to formulate sound economic policy, it would be necessary to re-estimate the MPC to determine its new value.

Having re-estimated the MPC, or perhaps the consumer demand curve in a particular market, a retail firm may want to predict consumer demand for their product so that they can plan production accordingly. The predicted value for our dependant variable is $\hat{y} = a + bx$, where **a** is our estimate of α , and **b** is our estimate of β . (Since ε is a normally distributed random disturbance our best estimate of it will be 0, which we do not need to note in the equation for predicted values.) Predictions are, of course, subject to error. Error results from the fact that our estimates of the intercept and slope coefficient, (**a** & **b**) will differ somewhat from the true parameters α & β , and because that random disturbance term only equals zero on average. A detailed discursion of prediction error is beyond the scope of this paper. Suffice it to say that prediction error is something that we would want to minimize.

VALID TESTS & RELIABLE ESTIMATES

Do not trust parameter estimates when your explanatory variables cover only a small range of values. If these variables are tightly packed then there are many different lines that could provide a 'good' fit for the data. Granted, the regression technique will pick only one line, and that one will have the 'best' fit, however, there will be many other possible lines through the data that are nearly as good. Updating your data with just one observation could cause the 'best fit' to jump to one of those other lines, perhaps radically change the intercept &

slope of your estimated line (see Figure 6).

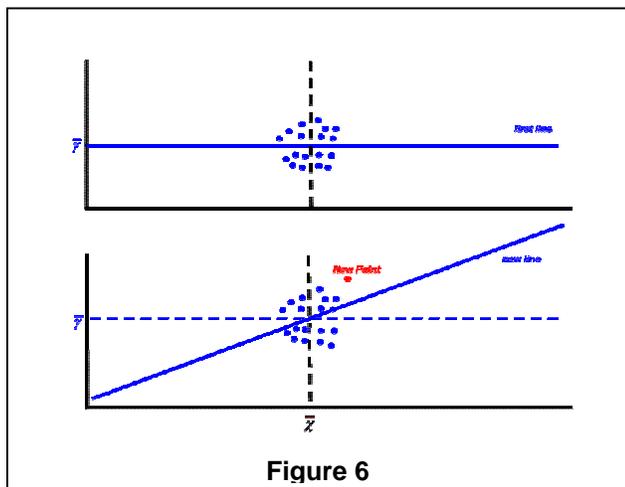


Figure 6

If your explanatory variables span a large range of values then all possible lines that could be said to provide a reasonably ‘good’ fit will be highly similar. Updates to your data will not significantly alter the intercept and slope of your estimated line, (as shown in Figure 7).

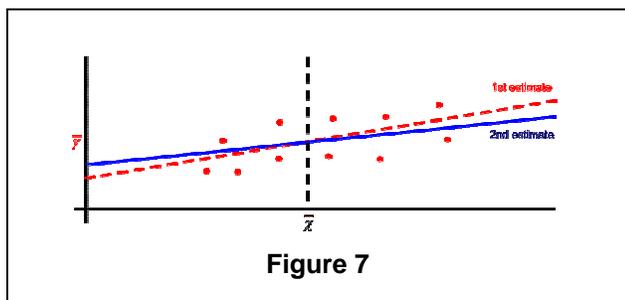


Figure 7

Do not trust your predictions for values of the explanatory variables well beyond the range you used to estimate the regression line. Beyond that range the estimated relationship between X and Y may not hold. For example, if you estimated that \$100 per patient spent on preventive care will save, on average, \$500 per patient next year, do not expect that \$1000 spent on preventive care would save \$5000 per patient. (Especially since \$5000 currently exceeds the average total cost per patient for all health care costs.) See Figure 8.

Hypothesis testing using regression employs the same basic concepts as previously discussed in the univariate case. The regression output generated by the software provides all the information necessary to do the tests. The steps are the same. Formulate H_0 & H_a , calculate the t statistic from the sample data, then compare the resulting t value to the critical t value. The formula for the t statistic is the same as before, $t = \text{observed value minus the}$

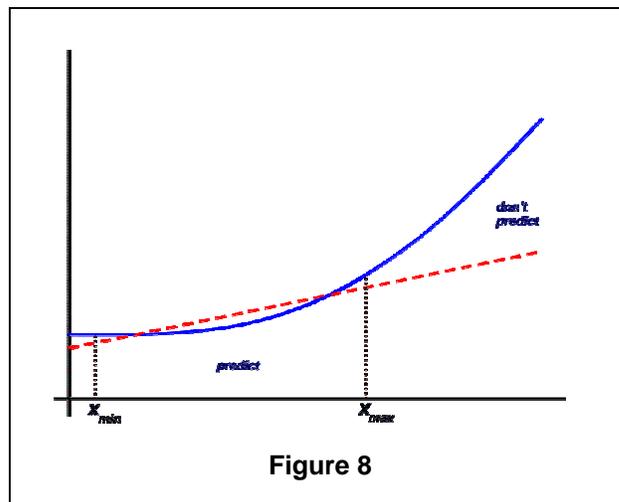


Figure 8

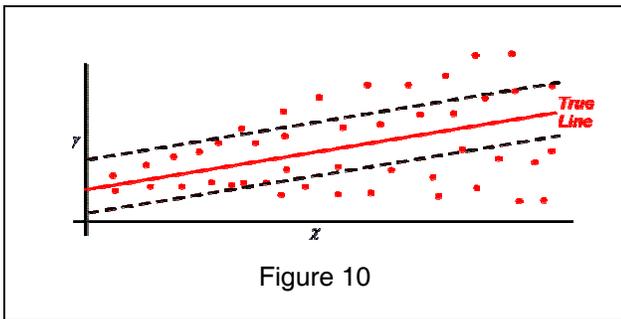
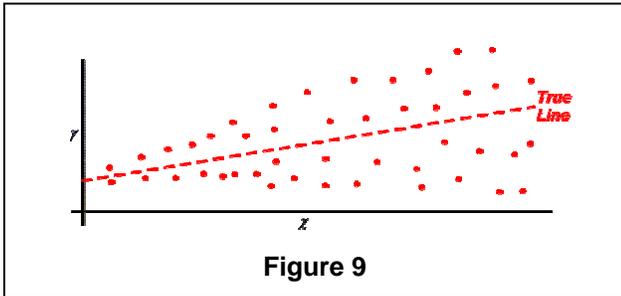
hypothesized value, divided by the standard deviation. If we are testing a hypothesis about the impact of one of the explanatory variables on the dependent variable, (i.e. about the value of one of the β 's), we observe a sample estimate (b). The standard deviation for the distribution of all possible sample estimates is the ‘Standard Error’ which is printed next to the ‘Parameter Estimate’ in the SAS[®] regression output.

The regression output also generates the t statistic for the standard null hypothesis that $\beta = 0$. However, the standard test is not always appropriate. For example, you may want to test the hypothesis that the MPC = 90% vs. the alternative that the MPC < 90%. In this situation the t statistic is calculated as $t = (b - .9) / \text{std error}$. The t statistics in the regression output are for the special case in hypothesis testing where H_0 specifies that the hypothesized value is zero.

AVOIDING COMMON PROBLEMS

The simplest regression technique, Ordinary Least Squares (OLS), makes assumptions, which may not be valid. First, it assumes that the random disturbance term has the same amount of variation, (measured by the ‘Standard Error of the Regression’), across the entire range of explanatory variables. Unfortunately, heteroscedasticity, (the label applied to changing variation in the disturbance term), is a common problem in cross-sectional data where variation usually increases as the explanatory variable (or the dependant variable) increases. Second, OLS assumes that the disturbance terms are independent of one another as you go from one observation to the next. Unfortunately, serial correlation (the label applied to this type of dependence) is a common problem in time series data.

Heteroscedasticity does not cause unbiased estimates of your regression parameters. What it does is undermine the reliability of your predictions of the dependent variable. Suppose the variance of the random component increases as the explanatory variable increases, (as in Figure 9). If you did not compensate for heteroscedasticity, then the confidence interval for your predictions will be too wide for the low values of X, and too narrow for the large values of X. (See Figure 10.)



To check for heteroscedasticity code **PROC REG** {options}; **MODEL** {options} / **SPEC**;. If the diagnostics indicate the presence of heteroscedasticity you need to determine what your residuals (i.e. error terms) are best correlated with, with Y, X, or even better, with X^2 . Once you have determined what the variance of the error term is related to, run a weighted regression with the inverse of the relationship as the weight. For example,

$$\text{if } \sigma_{\epsilon}^2 = X^2 \sigma_u^2$$

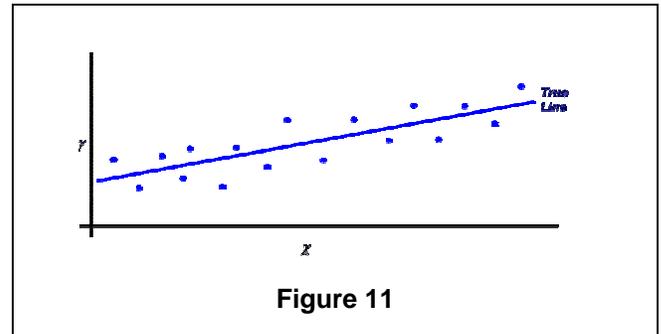
where u is an independently, identically distributed error

Then code:

```
PROC REG {options};
MODEL {options};
WEIGHT WT;      where WT=1/X2
```

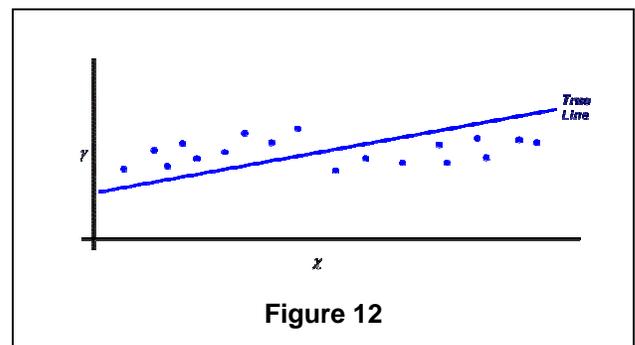
Serial Correlation results when the error from one time period determines a portion of the error in the next period. Mathematically speaking, $\epsilon_t = \lambda \epsilon_{t-1} + u_t$,

where u is a random and independent disturbance term, and λ is the portion of the previous error that influences the current error. Serial correlation comes in two flavors negative and positive. With negative serial correlation, ($-1 < \lambda < 0$), a positive error (an observation above the line) is followed by a negative error (an observation below the line) as in Figure 11.



Negative serial correlation is a common problem in health care. If doctor visits and elective surgery are postponed one month because of bad weather or holidays, the next month will experience a surge in utilization of services as patients scramble to make up for 'lost time'. Negative serial correlation is not as bad a problem as positive serial correlation. It does not cause biased estimation. Neither does it lead to poor predictions if ignored. (However, do not ignore the 'problem' of negative serial correlation. Knowledge about the degree to which the last period's error term will spill over into the next period, will help improve your predictions of future values of Y.)

With positive serial correlation ($0 < \lambda < 1$) the error term tends to stay on one side of the regression line until an unusually large random disturbance (u_t) knocks it to the opposite side, where it will reside for several periods, until an unusually large (See Figure 12.)



Unlike the situation with negative serial correlation, positive serial correlation can lead to really bad parameter estimates. If you have an unbalanced

number of positive, serially correlated error terms around the regression line, it can 'twist' the line OLS estimates away from the true line, as in Figure 13. Mathematically OLS still produces unbiased estimates in the presence of positive serial correlation, since it is equally likely that your estimated line will be twisted upward as downward. However, the end result is that your estimated line will still be well off the mark.

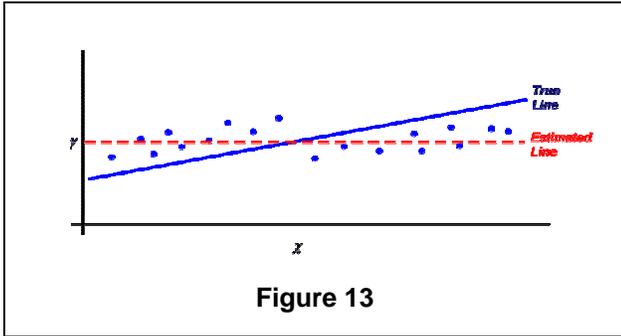


Figure 13

To test for serial correlation code `/DW` as an option in the **MODEL** statement. A Durbin-Watson statistic close to 2.0 indicates that serial correlation is not present. Values for the DW statistic less than 1.6 or greater than 2.4 indicate the presence of serial correlation. To correct for serial correlation code: **PROC AUTOREG; MODEL {options} /NLAG = n;** where n is the order of the correlation lag. In the example used above where $\epsilon_t = \lambda\epsilon_{t-1} + u_t$, we have first order serial correlation and n would equal 1. If an error term has repercussions beyond the next period, e.g. $\epsilon_t = \theta\epsilon_{t-2} + \lambda\epsilon_{t-1} + u_t$, we would have 2nd order serial correlation and n would equal 2.

Multicollinearity occurs when one of the explanatory variables is a linear combination of one or more of the other explanatory variables, (e.g. $X_3 = 0.4X_1 + 0.6X_2$). Perfect multicollinearity, where the value of one of the explanatory variables is completely determined by the other explanatory variables, is not a big problem. You will be informed of its presence by 'friendly' error messages right in your output. The SAS error messages will tip you off as to which variables are co-linear. Get rid of the one that is a linear combination of the others. If there is some question about which ones are collinear and to what extent, you can make that determination by regressing each of the explanatory variables, one at a time, on the remaining explanatory variables. (Go with the regression that has the highest R^2 .)

A high degree of, (but less than perfect), multicollinearity is a problem. It is harder to detect because there will not be any friendly warning messages. Clues to its presence are 'Standard

Errors' which are unusually large, parameter estimates which are inexplicably insignificant, and parameter estimates which change dramatically when you make minor changes in the model or slice the data different ways. Once you suspect multicollinearity, you can confirm it by coding **/COLLIN** in the **MODEL** statement. If multicollinearity is present the best approach is to eliminate one of the collinear variables from the model, preferably the one with the highest R^2 when regressed on the other explanatory variables. As long as the collinear variable continues to move in unison with the other explanatory variables, removing it from the model will have minimal affect on R^2 and on your predicted values.

CONCLUSIONS

This concludes our brief trip through basic regression and common problems. The authors hope you found the discussion useful. If the reader would like further information, we recommend '*Econometric Models & Economic Forecasts*', by Pindyck & Rubinfeld; McGraw Hill, 4th edition.

REFERENCES

Pindyck, Robert S. and Rubinfeld, Daniel L., (1997), *Econometric Models & Economic Forecasts*, Fourth Edition, New York: McGraw-Hill, Inc.

ACKNOWLEDGMENTS

SAS is a Registered Trademark of the SAS Institute, Inc. of Cary, North Carolina.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Vincent Maffei
 Anthem Blue Cross and Blue Shield
 370 Bassett Road
 North Haven CT 06473
 Phone: 203-985-7188
 Email: vincent.maffei@anthem.com
 Web: <http://www.anthembcbsct.com>

Michael Davis
 Bassett Consulting Services, Inc.
 10 Pleasant Drive
 North Haven CT 06473
 Phone: 203-562-0640
 Fax: 203-498-1414
 Email: michael@bassettconsulting.com
 Web: <http://www.bassettconsulting.com>