Paper 245-26

A Non-mathematical Introduction to Regression Concepts Using PROC REG

Mel Widawski, UCLA, Los Angeles, California

ABSTRACT

This paper will help you build a competence and confidence in linear regression as a technique through simple examples. We will start with a simple 7-observation data set and see how PROC REG ferrets out the relationship between the variables. You will develop a feel for Linear Regression as a technique through subsequent manipulations to this small data set.

INTRODUCTION

Linear regression is a useful technique for creating models to explain relationships between variables. For use in multiple regression variables must be numeric, and dependent variables need to have meaningful numeric values. That means if one person has a score of 4 and another a score of 3, then the person with the score of 4 has more of what is being measured than the person with a score of 3. This is what people mean when they say data is ordinal.

Actually, there is another requirement that needs to be applied. This especially is true of the dependent variable, which should be an interval variable. This means that a score of 5 is the same amount greater than a score of 4 on what is being measured as a score of 4 is greater than a score of 3.

You may have heard of a ratio scale. This means that a score of 4 can be assumed to have twice as much of what is being measured as a score of 2. This is also to be desired when doing regression. Independent variables have a little more latitude, but what can be said about the results depends on whether the predictor is an ordinal, interval, or ratio variable.

I will assume that you know how to read instream data with a DATA STEP. I have taken the liberty of condensing the output for ease of display and underlining important parts of the output.

We will start with a small data set of 7 values. It would be helpful if you actually were doing this yourself and treating it like an experiment to determine what regression does. We will then run PROC REG in SAS to see if the one to one relationship in the data can be detected.

We will then proceed by modifying this small data set to show the effect of two predictor variables, explore the concept of collinearity, orthogonallity, and the importance of error in the use of regression for model building.

Finally, we will see the results of regression when all of the assumptions are met.

REGRESSION: A SIMPLE EXAMPLE

In this example we will input a 7 line data set and then run **PROC REG** to perform linear regression. The **DATA STEP** below will input the data.

```
DATA first;
   INPUT DV IV;
CARDS;
1 1
2 2
3 3
4 4
5 5
6 6
7 7
;
RUN;
```

Notice that both the predictor **IV** and the criterion variable **DV** contain the same values. This is a one-to-one relationship. The dependent variable is predicted exactly by the independent variable. If you know that the value of **IV** is 1 then the value of **DV** will be 1, a value of 7 predicts a value of 7 for the dependent variable.

Now, let us see if **PROC REG** can detect this relationship. The program to do this follows.

```
PROC REG DATA=first;
        MODEL DV=IV;
        PLOT  DV*IV;
RUN;
```

We identify the data set being used with **DATA=first**. The **MODEL** statement tells the procedure to try to predict **DV** ffrom **IV**. The **PLOT** statement will be explained below.

The ensuing output has two sections; one labeled **Analysis of Variance**, and the next labeled **Parameter Estimates**.

```
Model: MODEL1
Dependent Variable: DV

Analysis of Variance

              Sum of      Mean
Source   DF   Squares     Square   F Val   Prob>F

Model     1   28.00000   28.00000       .       .
Error     5          0           0
C Total   6   28.00000

 Root MSE 0.00000   R-square 1.0000
 Dep Mean 4.00000   Adj R-sq 1.0000
 C.V.     0.00000
```

The first section gives over all model information. In this section, the **Error** for the model is zero, and neither an **F value** nor a **Prob>F** are calculated, both are shown as missing values. Since there is no error in prediction, it is impossible to test the significance of the prediction.

Perfect prediction means you can't publish because you don't have the necessary **p** value. Statistics are used to try to determine the probability of finding an effect as strong as the one you found in an imperfect world. Inferential parametric statistics require error.

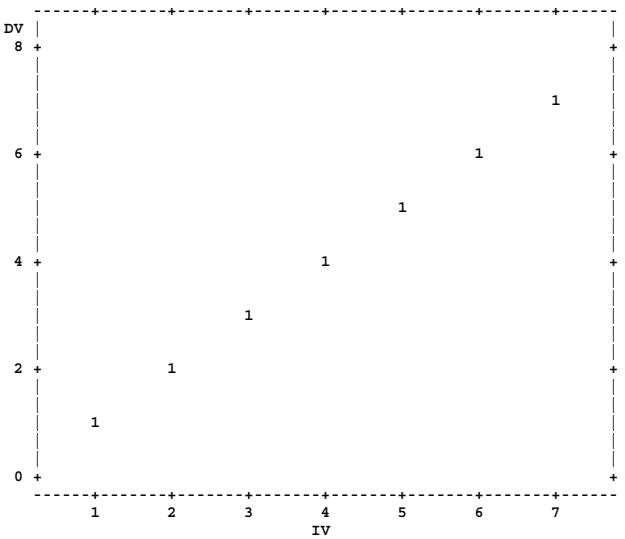
However, if you notice the **R-square** in the output is 1.00, that means that the regression equation explains 100% of the variability in **DV** through the use of **IV**.

Parameter Estimates

Variable	DF	Param Estim	Stand Error	T for H0: Param=0	Prob>|T|
INTERCEP	1	0	0.0000	.	.
IV	1	1.000	0.0000	.	.

The **Parameter Estimates** presented above show the relationship between the two variables. Note that the estimate for the intercept **INTERCEP** is 0, and the estimate of the parameter for **IV** is 1. This is also called the coefficient of **IV** in the regression equation.

This means that the regression equation is **DV=0+1*IV**, or the criterion equals the **INTERCEP** plus the coefficient for the predictor times the predictor itself. It can be seen that for every line in the data set above the corresponding value for **DV** will be obtained if you use any of the 7 values of **IV**. Since there is no error in measurement, no statistical tests of the reliability of these estimates is possible. Thus, the values for **T** and **P** are missing. The relationship can also be shown by looking at the plot of **DV*IV**.



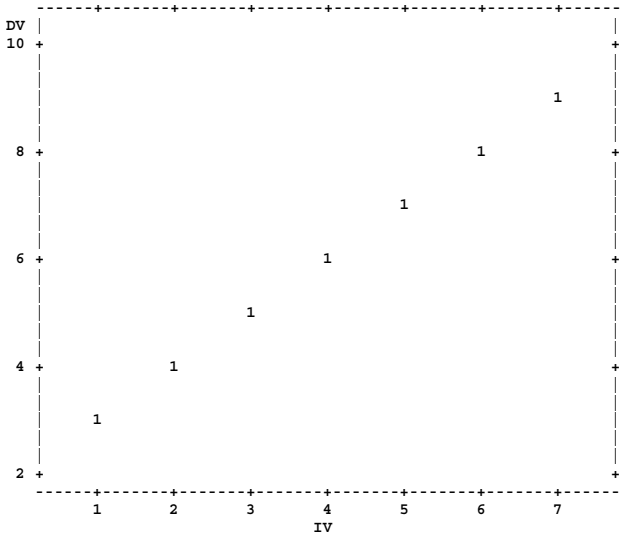
Notice that as **IV** increase by 1 so does **DV**. When **IV** goes from 3 to 4, then **DV** goes from 3 to 4. This is another way of saying what the coefficient, or parameter estimate does. Each 1 point increase in the predictor results in an increase in the criterion variable equal to the size of the coefficient.

What If We Add A Constant?

Consider the following data set. This is the same as the first, except each of the values for **DV** are increased by two. The values of **DV** range between 3 and 9. Is this still perfect prediction?

```
DATA second;
   INPUT DV IV;
CARDS;
3 1
4 2
5 3
6 4
7 5
8 6
9 7
;
RUN;

PROC REG DATA= second;
        MODEL DV=IV;
 PLOT  DV*IV;
RUN;
```

First, let's look at the plot of **DV** and **IV**. It looks the same as the plot opposite except that if you extend the line then this plot would have a value of 2 for **DV** when **IV** would have a value of zero.



The intercept is the value you would obtain for the criterion when the value of the predictor is set to zero.

Lets look at what can be detected in the output for this **PROC REG**. We will only look at the parameter estimates.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error
INTERCEP	1	2.000000	0.00000000
IV	1	1.000000	0.00000000

Notice that the coefficient for **IV** is still 1, but the value for the intercept is now 2. Thus, the relationship that was detected was **DV=2+1*IV**. Since the prediction is perfect, we still have no statistical estimate of how likely this finding is by chance.

While we are on the subject, why don't we create a new dependent variable using the formula we detected above, and see what happens.

```
DATA third;
   SET second;
   DV2=IV+2;
RUN;

PROC PRINT DATA=third noobs;
   VAR DV2 DV IV;
RUN;
```

DV2	DV	IV
3	3	1
4	4	2
5	5	3
6	6	4
7	7	5
8	8	6
9	9	7

Notice that **DV2**, which was created with the equation, is exactly the same as **DV**. The variable **DV2** is the predicted value of the regression equation. This can also be obtained by request from **PROC REG**. So far **PROC REG** has done a pretty good job of detecting the relationship even when the dependent variable is shifted by a constant.

REGRESSION: COLLINEARITY

In this example we will add another variable **IV2** to our original data set. See if you can detect a strange relationship between the variables **IV1** and **IV2**. The new dependent variable **DV** is a simple sum of the independent variables.

DV	IV1	IV2
2	1	1
4	2	2
6	3	3
8	4	4
10	5	5
12	6	6
14	7	7
16	8	8

You might notice that variables **IV1** and **IV2** are identical to each other. This is what is referred to as collinearity. This can cause problems for regression programs.

We can simply name both variable on the model statement and see if the relationship can be detected. This would be a simple sum of the two variables.

```
PROC REG DATA = colin;
  MODEL DV = IV1 IV2 ;
RUN;
```

Notice the following results.

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	168.00000	168.00000	Infty	<.0001
Error	6	0	0		
C Total	7	168.00000			

In addition the following notice appears in the output.

```
NOTE: Model is not full rank. Least-squares solutions
      for the parameters are not unique. Some statistics
      will be misleading. A reported DF of 0 or B means
      that the estimate is biased.
NOTE: The following parameters have been set to 0,
      since the variables are a linear combination of
      other variables as shown.
```

You are warned that the Model is not of full rank, and that results may be misleading. Also you are told that some of the variables are linear combinations of other variables. We know this to be true as variables **IV1** and **IV2** are identical.

Now let us look at the parameter estimates for the model.

Parameter Estimates

Var	DF	Param Estim	Stand Error	t Val	P>|t|
Interc	1	0	0	.	.
IV1	B	2	0	Infty	<.0001
IV2	0	0	.	.	.

Notice that only one parameter is estimated **IV1** has a parameter of **2**. This means that the relationship **DV=2*IV1** has been detected. This actually is a true relationship since variables **IV1** and **IV2** are identical.

Complete collinearity results when any of the independent variables can be predicted completely by any of the other independent variables. If only two predictors are involved then this would be indicated by a perfect correlation. A similar situation exists when one of the independent variables is related to more than one other independent variable this is referred to as multi-collinearity.

The relationship does not have to be perfect, but extremely high for this situation to exist.

REGRESSION: TWO PREDICTORS

In this example we will add another variable **IV2** to our original data set. I will make sure that this new variable is completely independent of the original variable which has been renamed **IV1**. We will create a new **DV** by summing these two variables together, and adding a constant of 2 for good measure. The resulting equation will look like **DV=2+IV1+IV2**. Notice we also added one more line of data.

```
DATA fourth;
   INPUT IV1 IV2;
   DV=IV1+IV2+2;
CARDS;
1  3
2 -3
3 -2
4  2
5  2
6 -2
7 -3
8  3
;
RUN;
```

The resulting data set looks like the following.

```
DV IV1 IV2
 6  1    3
 1  2   -3
 3  3   -2
 8  4    2
 9  5    2
 6  6   -2
 6  7   -3
13  8    3
```

Both predictors can be named in the model statement of the **PROC REG**.

```
PROC REG DATA=fourth;
        MODEL DV=IV1 IV2;
RUN;
```

In the following results we can see that both variables were detected as predicting **DV**.

```
Parameter Estimates

                  Parameter      Standard
Variable  DF       Estimate         Error

INTERCEP   1       2.000000    0.00000000
IV1        1       1.000000    0.00000000
IV2        1       1.000000    0.00000000
```

Notice that the equation discovered is **DV=2+1*IV1+1*IV2**. This is exactly the same as the one used to create **DV** in the first place.

The Concept of Error

What would happen if we were to omit this new variable from the model as in the example below.

```
PROC REG DATA=fourth;
        MODEL DV=IV1;
RUN;
```

Now when we look at the output from the procedure, we see that we have additional information that we never had before.

```
Model: MODEL1
Dependent Variable: DV

Analysis of Variance
              Sum of      Mean
Source   DF  Squares    Square F Value  Prob>F

Model     1 42.00000 42.00000   4.846  0.0700
Error     6 52.00000  8.66667
C Total   7 94.00000

 Root MSE  2.94392  R-square 0.4468
 Dep Mean  6.50000  Adj R-sq 0.3546
 C.V.     45.29108
```

```
Parameter Estimates

             Param  Standard T for H0:
Variable DF  Estim    Error  Param=0  Prob>|T|

INTERCEP  1  2.000 2.2938842  0.872   0.4168
IV1       1  1.000 0.4542567  2.201   0.0700
```

In the output above we now have a statistical tests for the first time. This is because without our second variable we no longer explain **DV** completely. There is some variability in the dependent variable. Notice that the **RSQUARE** has been reduced to **.4468**, which means that **IV1** explains about **44.7%** of the variability of **DV**. The p value for the model is only **.07** so we cannot say we have a statistically reliable result. Notice that since there is only one predictor, the p value for the test of the overall model and the parameter estimate for the predictor are the same. This is always true with a single predictor.

It is important to realize that any variability that could be explained by known variables becomes error when those variables are not included as predictors. That is why better models will include the relevant variables.

Even though the statistical test is not necessarily adequate, notice that the coefficient for **IV1** is still correctly estimated as 1.

What Does Orthogonal Mean?

I mentioned the new variable **IV2** was orthogonal to the first variable **IV1**. This means simply that they do not relate to each other in a linear fashion. This can be show with a simple correlation between the two.

```
PROC CORR DATA=fourth;
   VAR  IV1 IV2;
RUN;
```

This yields a correlation of zero as can be seen below. The two variables are completely uncorrelated. Prediction is usually best when the predictors are not correlated with each other, but each does related to the dependent variable.

```
Pearson Correlation Coefficients / Prob > |R| under Ho:
Rho=0 / N = 7

          IV1       IV2

IV1   1.00000   0.00000
       0.0       1.0000
```

Predicted Values and Residuals

If we take the equation for predicting **DV**, **pred=2+1*IV**, and apply it in the program below, we should come up with a predicted value for the **DV** from the equation with one predictor. Now if we subtract **pred** from **DV** we will see how far we missed in our prediction. This difference for each case is the residual. The variance of the residual is your sum of squares for the error term.

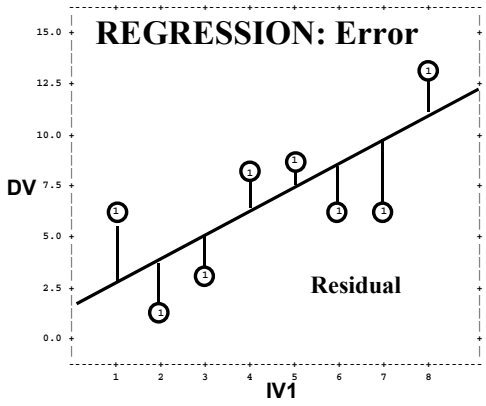
```
DATA fourthb;
     SET fourth;
   pred=1*IV1+2 ;
   resid=DV-pred;
RUN;
```

Notice the residual below. I created the data so that **IV2** was completely unrelated to **IV1,** and so that **IV2** has a mean of zero. Thus **resid** is a direct copy of **IV1**. If the mean of **IV1** were not zero, then **resid** and **IV1** would still be perfectly related, but the values would be slightly different.

DV	PRED	IV1	IV2	RESID
6	3	1	3	3
1	4	2	-3	-3
3	5	3	-2	-2
8	6	4	2	2
9	7	5	2	2
6	8	6	-2	-2
6	9	7	-3	-3
13	10	8	3	3

If an important variable is left out of the model, then it will be related to the residuals from the regression equation.

In the following graph you will see a plot of this data. Notice that the residuals are the distance from each actual point to the regression line. It is also important to note that these distances are measured perpendicular (at right angles) to the X-axis (the line at the bottom). The variance of these residuals is the error variance for the regression.



Both the predicted value and residuals may be produced by **PROC REG** with the following request.

```
PROC REG DATA=fourth;
        MODEL DV=IV1;
     OUTPUT OUT=new4th P=pred R=resid;
RUN;
```

The data set **new4th** will contain all of the variables plus **resid** and **pred** which are created in the **OUTPUT** statement.

What If the Coefficients Are Not Equal?

Do you think that **PROC REG** will find the correct coefficients when they are not equal. Let us use the same data, but multiply **IV1** by 2 instead of one.

```
DATA fifth;
   SET fourth;
   DV=IV1+2*IV2+2;
RUN;
```

Now when we run the same two-variable regression presented above, we see that the correct relationship is still detected.

Parameter Estimates

Variable	DF	Parameter Estimate
INTERCEP	1	2.000000
IV1	1	1.000000
IV2	1	2.000000

The equation obtained would be **DV=2+1*IV1+2*IV2** when you look in the Parameter Estimate column of the output.

REGRESSION: TWO PREDICTORS PLUS ERROR

Let's add an error term that is unrelated to the original variables. All we have to do is copy the lines of the previous data set after it, and add a new column that has a value of minus one for the first set, and one for the second set. The new data set follows.

```
DATA seventh ;
   INPUT IV1 IV2 ERR @@;
   DV=IV1+2*IV2+2 +ERR;
CARDS;
1  3 -1      1  3  1
2 -3 -1      2 -3  1
3 -2 -1      3 -2  1
4  2 -1      4  2  1
5  2 -1      5  2  1
6 -2 -1      6 -2  1
7 -3 -1      7 -3  1
8  3 -1      8  3  1
;
RUN;
```

This data set there is a little uncertainty of 1 either side of the prediction line. The following procedure demonstrates this. The new variable **ERR** is left out of the model.

```
PROC REG DATA=seventh;
        MODEL DV=IV1 IV2/stb;
        PLOT DV*P.;
RUN;
```

The output now includes statistical tests. We will also be able to see a new statistic, the standardized regression coefficient **stb**. This gives us an estimate of the relative amount of the variability of the criterion explained by each predictor.

The un-standardized regression coefficients are sensitive to how the predictor is scaled, while the standardized coefficient is not.

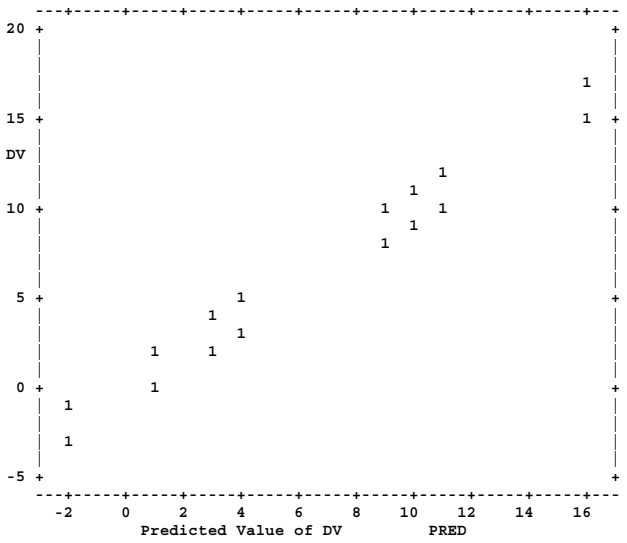
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	500.000	250.000	203.125	0.0001
Error	13	16.000	1.230		
C Total	15	516.000			

Root MSE	1.10940	R-square	0.9690
Dep Mean	6.50000	Adj R-sq	0.9642
C.V.	17.06770		

```
Parameter Estimates

                Param   Stand   T for H0:             Stand
Variable  DF  Estim    Error   Param=0   Prob>|T|    Estim

INTERCEP   1  2.000   0.6112     3.272     0.0061     0.000
IV1        1  1.000   0.1210     8.261     0.0001     0.403
IV2        1  2.000   0.1087    18.385     0.0001     0.897
```

We can see how the error is distributed by looking at a plot of the dependent variable by the predicted variable, as requested in the procedure above.



There is more than one possible observed value for the each predicted value. This constitutes the error in estimation or the residual.

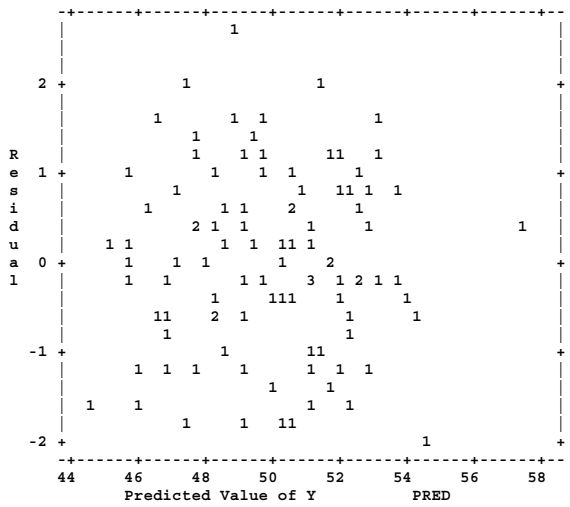
REGRESSION: A MORE REALISTIC EXAMPLE

We can make use of some additional functions in SAS to create a more realistic example, which would be an ideal example with truly normal variates and error. The following **DATA STEP** uses the **RANNOR** call to produce these variables.

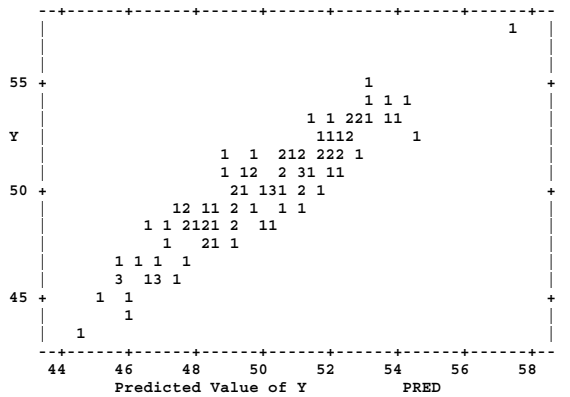
```
DATA random;
  RETAIN
     SEED1      7936414
     SEED2      1434021
     SEED3      8252093;
  Z1=0;Z2=0;Z3=0;
  DO ID=1 TO 100;
     CALL  RANNOR(SEED1,Z1);
     CALL  RANNOR(SEED2,Z2);
     CALL  RANNOR(SEED3,Z3);
         Y=Z1+2*Z2+Z3+50;
         output;
  end;
RUN;

PROC REG DATA=random;
        MODEL Y=Z1 Z2 /stb;
        PLOT R.*P. Y*P.;
RUN;
```

The following output shows a that **Z1** and **Z2** significantly predict **Y**. The equation of **Y=50.15+1.08*Z1+2*Z2** is very near the formula that created the data for **Y**.

```
Analysis of Variance

                Sum of      Mean
Source     DF   Squares     Square      F Value    Prob>F

Model       2  608.3080  304.15402  295.662    0.0001
Error      97   99.7859    1.02872
C Total    99  708.0940

 Root MSE  1.01426  R-square 0.8591
 Dep Mean 49.93456  Adj R-sq 0.8562
 C.V.      2.03118

Parameter Estimates

                Param    Stand    T for H0:             Stand
Variable  DF   Estim    Error    Param=0    Prob>|T|   Estim

INTERCEP   1   50.15   0.1025    489.198     0.0001    0.000
Z1         1    1.08   0.0952     11.351     0.0001    0.434
Z2         1    2.02   0.0993     20.351     0.0001    0.779
```



We have also asked for a residual plot by specifying **R.*P.** on the plot statement. This plot shows that the residual error reasonably meets the assumptions of regression. This plot should have an approximately oval shape. The plot of the observed values by the predicted values follows.



This plot shows reasonably constant variability along the regression line.

REGRESSION: Scaling

Now we will investigate what happens when one of the independent variables is scaled by multiplying it by a constant. And then the scaled version is used in the analysis.

```
DATA random2;
  SET random;
  z1a=z1*10;
RUN;

PROC REG DATA=random2;
        MODEL Y=Z1A Z2 /stb;
        *PLOT R.*P. Y*P.;
RUN;
```

The following output shows that the parameter for **Z1A** is the same as the previous parameter for **Z1** divided by 10. And the standard error of that parameter is the standard error for **Z1** divided by 10 as well. All of the other statistics are the same. And the T and significance for **Z1A** is the same as the previous for **Z1**.

```
Analysis of Variance

              Sum of      Mean
Source    DF  Squares    Square      F Value    Prob>F

Model      2  608.3080  304.15402  295.662    0.0001
Error     97   99.7859    1.02872
C Total   99  708.0940

 Root MSE  1.01426  R-square 0.8591
 Dep Mean 49.93456  Adj R-sq 0.8562
 C.V.      2.03118

Parameter Estimates

              Param    Stand    T for H0:            Stand
Variable DF   Estim    Error    Param=0    Prob>|T|  Estim

INTERCEP  1   50.15   0.1025    489.198     0.0001    0.000
Z1        1    .108   0.0095     11.351     0.0001    0.434
Z2        1   2.02    0.0993     20.351     0.0001    0.779
```

Scaling an independent variable only changes parameter estimates for that variable and has no effect on significance, R-square, or other parameters. It is sometimes useful when a variable with a large range is to Be used as a predictor in regression.

CONCLUSION

I hope you have developed some confidence in regression as a tool for detecting the relationships between variables. One way to learn more about the technique is to do more of what we have done here. Create variables and modify them so that you can see what happens with the technique under various circumstances.

I will leave you with a finally problem. Create a data set with both columns identical, and create a new variable by adding the columns together. Then try a two-variable model. This will demonstrate sever collinearity.

REFERENCES

SAS Institute Inc. (1989), *SAS/STAT® User's, Version 6, Fourth Edition,* Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

Thanks to Barbara Widawski without whose editing this manuscript would be illegible.

CONTACT INFORMATION

Mel Widawski
UCLA
Los Angeles, CA

MEL@UCLA.EDU