Paper 364-2008

Introduction to Correlation and Regression Analysis

Ian Stockwell, CHPDM/UMBC, Baltimore, MD

ABSTRACT

SAS® has many tools that can be used for data analysis. From Freqs and Means to Tabulates and Univariates, SAS can present a synopsis of data values relatively easily. However, there is a difference between what the data are, and what the data mean. In order to take this next step, I would like to go beyond the basics and introduce correlation and hypothesis testing using regression models. A brief statistical background will be included, along with coding examples for correlation and linear regression.

INTRODUCTION

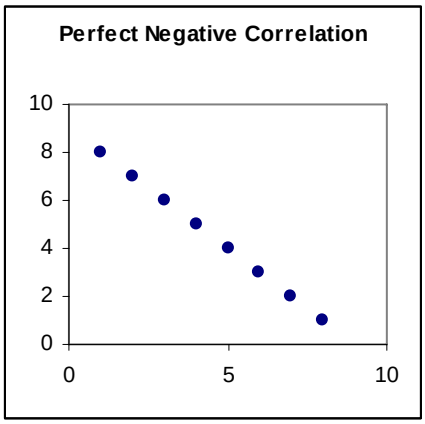
When analyzing vast amounts of data, simple statistics can reveal a great deal of information. However, it is often more important to examine relationships within the data, especially in medical and social sciences. Through correlation measures and hypothesis testing, these relationships can be studied in depth, limited only by the data available to the researcher. This paper and corresponding presentation will attempt to explain these powerful tools with a statistical background and programming examples.

CORRELATION: STATISTICAL BACKGROUND

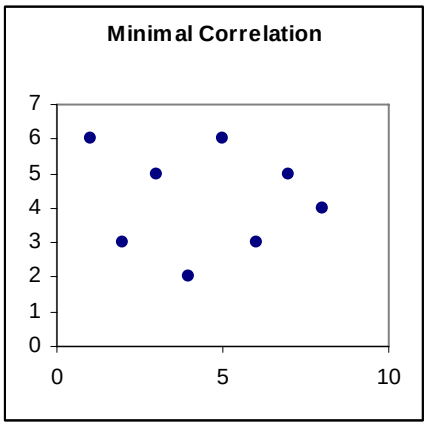
In layman's terms, correlation exists when two variables have a linear relationship beyond what is expected by chance alone. When examining data in SAS, correlation reveals itself by the relationship between two variables in a dataset. The most common measure of correlation is called the "Pearson Product-Moment Correlation Coefficient". It is important to note that while more than two variables can be analyzed when looking for correlation, the correlation measure only applies to two variables at a time, by definition:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

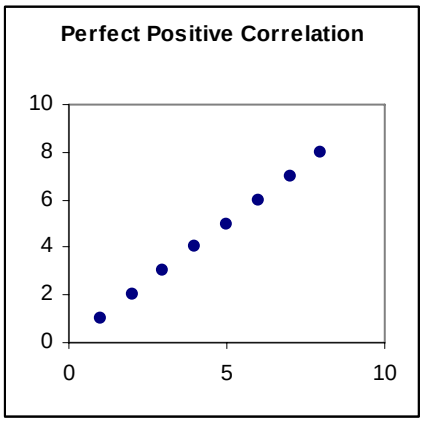
It is apparent when examining the definition of correlation that measures from only two variables are included, namely the covariance between the two variables {cov(x,y)} and the standard deviation of each ($\sigma_x\sigma_y$). The result of this calculation is the correlation between the two variables. This measure can range from -1 to 1, inclusive. A value of -1 represents a "perfect negative correlation", while a value of 1 represents a "perfect positive correlation". The closer a correlation measure is to these extremes, the "stronger" the correlation between the two variables. A value of zero means that no correlation is observed. It is important to note that a correlation measure of zero does not necessarily mean that there is no relationship between the two variables, just that there is no linear relationship present in the data that is being analyzed. It is also sometimes difficult to judge whether a correlation measure is "high" or "low". There are certain situations where a correlation measure of 0.3, for example, may be considered negligible. In other circumstances, such as in the social sciences, a 0.3 correlation measure may suggest that further examination is needed. As with all data analysis, the context of the data must be understood in order to evaluate any results.



This graph illustrates perfect negative correlation. The two variables of interest are on the x and y axis, respectively. When graphed this way, it is apparent that a (negative) linear relationship exists between the two variables, i.e. the variables "move together".



This graph illustrates minimal correlation. The two variables of interest are on the x and y axis, respectively. When graphed this way, it is difficult to establish any visual linear relationship between the two variables. In fact, this set of data points has a slight negative correlation (-0.11757927).



This graph illustrates perfect positive correlation. The two variables of interest are on the x and y axis, respectively. When graphed this way, it is apparent that a (positive) linear relationship exists between the two variables.

It is important to note that a strong (or even perfect) correlation does not imply causation, as other variables may be affecting the relationship between the two variables of interest.

CORRELATION: SYNTAX

In order to measure correlation in SAS, the proc corr procedure can be used. This procedure will provide correlation measures for multiple variables, in a cross-tabular format. The syntax for the procedure is as follows:

```
proc corr data=dataset;
  by byvars;
  freq freqvar;
  var varlist;
  weight weightvar;
run;
```

where:

- ***dataset*** is the name of the dataset to be analyzed, either temporary or permanent.
- ***byvars*** is a list of all variables to be used to create "by" groups for processing. This option is common among most procedures.
- ***freqvar*** is the numeric variable which contains the number of times an observation is to be "counted" for the analysis. Similar to weightvar.
- ***varlist*** is a list of all numeric variables to be included in the analysis. At least two numeric variables should be chosen. If no variables are listed, all numeric variables contained in the dataset will be used.
- ***weightvar*** is the numeric variable which contains the weight for each observation. Similar to freqvar.

Note that there are more options that can be used with this procedure for less common (but still useful) correlation measurements. See SAS documentation for further details.

CORRELATION: OUTPUT

The output produced by proc corr contains a great deal of useful information. The first information displayed is a list of the variables included in this analysis. This is especially useful when no variables were included in the *var* statement, so all numeric variables were included. Next is a list of "simple statistics" for each variable. This list contains the number of observations, mean, standard deviation, sum, minimum and maximum. After this section, each variable in the analysis and their label is listed. Finally, the correlation measures are presented. Unless a different correlation measure is requested, this section will be labeled "Pearson Correlation Coefficients". Results are provided in a cross-tabular format, with values of one on the diagonal (a variable will always have a perfect positive correlation with itself). Along with the correlation coefficients, p-values are listed, as are the number of observations (if different).

CORRELATION: EXAMPLE

In this example, the measure of interest is body mass index (BMI). The data being analyzed is a subsection of the 1999-2000 National Health Assessment and Nutrition Examination Survey (NHANES). These are public use datasets available through the National Center for Health Care Statistics. The NHANES is comprised of 87 separate data files, each covering a different aspect of the survey. The necessary variables were extracted from these data files, and merged to create a master file.

In order to test for correlation between BMI and age, for example, the following code would be used:

```
proc corr data=nhanes;
  var age bmi;
run;
```

And the following output would be produced:

```
                                The SAS System

                              The CORR Procedure

                          2  Variables:    age      bmi


                              Simple Statistics

Variable       N          Mean       Std Dev           Sum       Minimum       Maximum

age         3130      37.24473      15.52242        116576      16.00000      85.00000
bmi         2891      27.79065       6.38912         80343      14.42000      63.91000

                              Simple Statistics

                  Variable    Label

                  age         Age at Screening Adjudicated - Recode
                  bmi         Body Mass Index (kg/m**2)


                         Pearson Correlation Coefficients
                            Prob > |r| under H0: Rho=0
                              Number of Observations

                                                        age           bmi

          age                                       1.00000       0.17397
          Age at Screening Adjudicated - Recode                     <.0001
                                                      3130          2891

          bmi                                       0.17397       1.00000
          Body Mass Index (kg/m**2)                 <.0001
                                                      2891          2891
```

The reported correlation measures show a slight (but significant, according to the p-value) positive correlation between age and BMI, but caution must be taken when interpreting these results. Remember, correlation shows linear relationships between two variables, not causation. Correlation measures may be clouded by relationships that exist with other variables. For example, increasing age may lead to increased convenience food consumption and decreased exercise, which in turn would increase BMI. Therefore it is not necessarily an increase in age that causes an increase in BMI, but a change in factors that are related to age. This is known as a "spurious relationship".

REGRESSION ANALYSIS: STATISTICAL BACKGROUND

In order to combat these "spurious relationships", regression analysis allows multiple variables to be examined simultaneously. The most widely used method of regression analysis is Ordinary Least Squares (OLS) analysis. OLS works by creating a "best fit" trend line through all of the available data points. First, the variables to be included in the analysis must be chosen, and incorporated into the appropriate model (in this case, a linear model):

$$Y = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \ldots + \beta_n(x_n) + \varepsilon$$

where:

- **Y** is the dependent variable.
- $x_1 \ldots x_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1 \ldots \beta_n$ are the coefficients of interest.
- ε is the error.

In this model, **Y** represents the variable of interest in the analysis, while $x_1 \ldots x_n$ represent the variables to be tested. The coefficients on each of the independent variables ($\beta_1 \ldots \beta_n$) represent the effect that a change in the independent variable ($x_1 \ldots x_n$) on the dependent variable (**Y**). The sign on the coefficient reflects the direction of the relationship. Next, a testable hypothesis must be developed:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

where:

- H_0 is the hypothesis to be tested, or the "null hypothesis".
- H_1 is the alternative hypothesis.

This hypothesis tests whether the coefficient of a given dependent variable equals zero (remember: any variable multiplied by zero equals zero, so a coefficient of zero cancels out the related variable). Therefore, if the analysis finds that the null hypothesis can be rejected (i.e. that the coefficient of interest does not in fact equal zero), then that variable has a significant effect on the dependent variable (**Y**).

REGRESSION ANALYSIS: SYNTAX

In order to perform regression analysis in SAS, the proc reg procedure can be used. This procedure will provide regression analysis (OLS) measures for multiple variables, in a cross-tabular format. The syntax for the procedure is as follows:

```
proc reg data=dataset;
  by byvars;
  model depvar=indepvars;
  freq freqvar;
  weight weightvar;
run;
quit;
```

where:

- ***dataset*** is the name of the dataset to be analyzed, either temporary or permanent.
- ***byvars*** is a list of all variables to be used to create "by" groups for processing. This option is common among most procedures.
- ***depvar*** is the name of the dependent variable to be used in the analysis (**Y** above).
- ***indepvars*** is a list of all independent variables to be used in the analysis ($x_1 \ldots x_n$ above).
- ***freqvar*** is the numeric variable which contains the number of times an observation is to be "counted" for the analysis. Similar to *weightvar*.
- ***weightvar*** is the numeric variable which contains the weight for each observation. Similar to *freqvar*.

Note that there are more options that can be used with this procedure for less common (but still useful) correlation measurements. See SAS documentation for further details.

REGRESSION ANALYSIS: OUTPUT

The output produced by proc reg contains a great deal of useful information. The first information displayed is a count of the observations included in this analysis. The number of observations "used" in calculating the coefficients is included because missing values for either the dependent or independent variables are excluded from the analysis. Next is an "Analysis of Variance", including degrees of freedom, sum of squares, and the f value ("Pr > F" shows the probability that all coefficients of the independent variables are equal to zero). Also of interest in this section is the R-square and adjusted R-square values. These values represent the percentage of variation that is being captured by the regression model. Finally, the parameter estimates ($\beta_0 \ldots \beta_n$) are presented. Along with an estimate of the value, t values are included to test significance ("Pr > |t|" is the probability that a given coefficient is not statistically significant).

REGRESSION ANALYSIS: EXAMPLE

This example will build on the previous correlation analysis. In order to overcome a possible spurious relationship between BMI and age, variables for fast food consumption (food) and exercise (ex) will be added to the analysis. Since these variables are now being "controlled" for, the analysis can examine the effect of age independent of the corresponding effects of fast food consumption and lack of exercise.

$$BMI = \beta_0 + \beta_1(age) + \beta_2(food) + \beta_3(ex)$$
$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

where:

- **BMI** is the independent variable.
- β_0 is referred to as the intercept.
- **age,food,ex** are the dependent variables.
- H_0 is the null hypothesis that the coefficient on the variable **age** is zero.
- H_1 is the alternative hypothesis that the coefficient on the variable **age** is not zero.

In order to test the hypothesis that age has an affect on BMI independent of fast food consumption and exercise, the following code would be used:

```
proc reg data=nhanes;
  model bmi = age food ex;
run;
```

And the following output would be produced:

The REG Procedure
Model: MODEL1
Dependent Variable: bmi Body Mass Index (kg/m**2)

Number of Observations Read	3130
Number of Observations Used	2891
Number of Observations with Missing Values	239

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5960.27767	1986.75922	51.21	<.0001
Error	2887	112012	38.79874		
Corrected Total	2890	117972			

Root MSE	6.22886	R-Square	0.0505
Dependent Mean	27.79065	Adj R-Sq	0.0495
Coeff Var	22.41352		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	26.50333	0.35767	74.10	<.0001
age		1	0.05085	0.00791	6.43	<.0001
food		1	-0.02332	0.27055	-0.09	0.9313
ex		1	-2.14263	0.27334	-7.84	<.0001

The reported parameter estimate of the **age** variable is 0.05085, and is significant at the 1% level. The results, therefore, show that age has a positive effect on BMI when fast food consumption and exercise are controlled for. Keep in mind, however, that there could be other factors involved (both measurable and immeasurable) that should be accounted for if possible, as deemed necessary by the researcher. This analysis could also be completed by gender, for example, using the appropriate options.

There are also other forms of regression analysis that can be used, depending on the data available and the relationships between the variables. A few of these regressions and their applications are as follows:

- **Probit / Logit** – Used when the dependent variable is dichotomous (binary).
- **ANOVA**– Used when the independent are categorical (as opposed to continuous as in OLS).
- **Two Stage Least Squares** – Used to mitigate the effects of independent variable interaction.

Each of these types of regression (and others) are available to the researcher, and one must be very careful to chose the method of analysis that is most applicable to the situation.

CONCLUSION

Both correlation and regression analysis are excellent analytical tools when implemented correctly. While learning to use these methods, it is important to carefully specify the problem being studied, and the statistical methods being used. Once the programmer is comfortable, though, using these methods will become second nature. Remember to always have a SAS reference guide handy, and good luck!

CONTACT INFORMATION

Please feel free to contact me with any questions or comments:

Ian Stockwell
Research Analyst / SAS® Certified Professional
The Center for Health Program Development and Management / UMBC
1000 Hilltop Circle
Baltimore, MD 21250
Phone: 410.455.6728
Fax: 410.455.6850
E-mail: istockwell@chpdm.umbc.edu
Web: www.chpdm.org