From: Norman, G. R., & Streiner, D. L. (1997).
PDQ Statistics. St Louis, MO: Mosby.

6

Relationship Between Interval and Ratio Variables: Linear Regression and Related Methods

Regression analysis deals with the situation in which there is one measured dependent variable and one or more measured independent variable(s). The *Pearson correlation* and the *multiple correlation coefficient* describe the strength of the relationship between the variables.

Despite the fact that you have been introduced to some statistical sledgehammers in the last few chapters, you might have noticed that conditions under which they could be applied were somewhat restrictive. One variable was always a nominal variable (e.g., reader-nonreader, clam juice-no clam juice), and the other variable was always interval or ratio. Although that fairly well describes the situation in many studies, there are two other combinations that frequently arise. The first is when both variables are nominal or ordinal (e.g., dead-alive, cured-not cured), in which case we must use **non-parametric statistics**. This situation will be dealt with in Chapter 11.

The second class of studies are those in which both independent variables (IVs) and dependent variables (DVs) are interval or ratio. This situation frequently arises when the researcher cannot institute an experiment in which some people get it and some don't. Instead, the experimenter must deal with natural variation in the real world, in which people may, of their own volition, acquire varying degrees of something, and then have more or less of the DV.

For example, suppose you want to examine the relationship between obesity and blood sugar. In the best of all possible worlds, you would take a sample of newborn infants and randomize them into two groups. Group A members would be raised on puréed pizza, milkshakes, and potato chips for the next 40 years, and Group B members would have small quantities of rabbit food during the same period. But the ethics committee wouldn't like it, and the granting agency wouldn't fund it. So a more likely approach would be to venture timorously out into the real world, grab a bunch of complacent and compliant folks, measure their skinfold and blood sugar, and plot a graph depicting the relationship between them. If the relationship were really strong, these points would lie on a straight line. Unfortunately, these relationships don't occur often in the real world because there are usually many variables, both known and unknown, that might affect blood sugar. So there is bound to be a great deal of scatter about the average line, and the first challenge may be determining where to draw the line.

If you recall your geometry, you might remember that a straight line equation is described as follows:

$$\text{Blood sugar} = a + b \times \text{skinfold}$$

The *y* intercept of the line is "a," and the slope is "b." The issue then is, "What is the best combination of "a" and "b" that yields the best fit?

The way statisticians approach this is to define "best" in a particular way. They determine the vertical distances between the original data (•) in Figure 6-1 and the corresponding point on the line (0), square these distances, and sum over all the data points. They then select a value of "a" and "b" that results in the least value for this sum, called a **least squares criterion.** This Sum of Squares, which is an expression of the deviation of individual data from the best-fitted line, is exactly analogous to the Sum of Squares (within) in ANOVA, and is called the **Sum of Squares (Residual).** A second Sum of Squares can then be calculated by taking the differences between the points on the line and the horizontal line through the two means, squaring, and adding. This one is analogous to the Sum of Squares (between) in ANOVA and is called the **Sum of Squares due to regression.**

TESTING SIGNIFICANCE AND STRENGTH OF RELATIONSHIP IN SIMPLE REGRESSION

Although you can do a test of significance on the Pearson correlation (see p. 58) to determine if there is a relationship between the IVs and DVs,

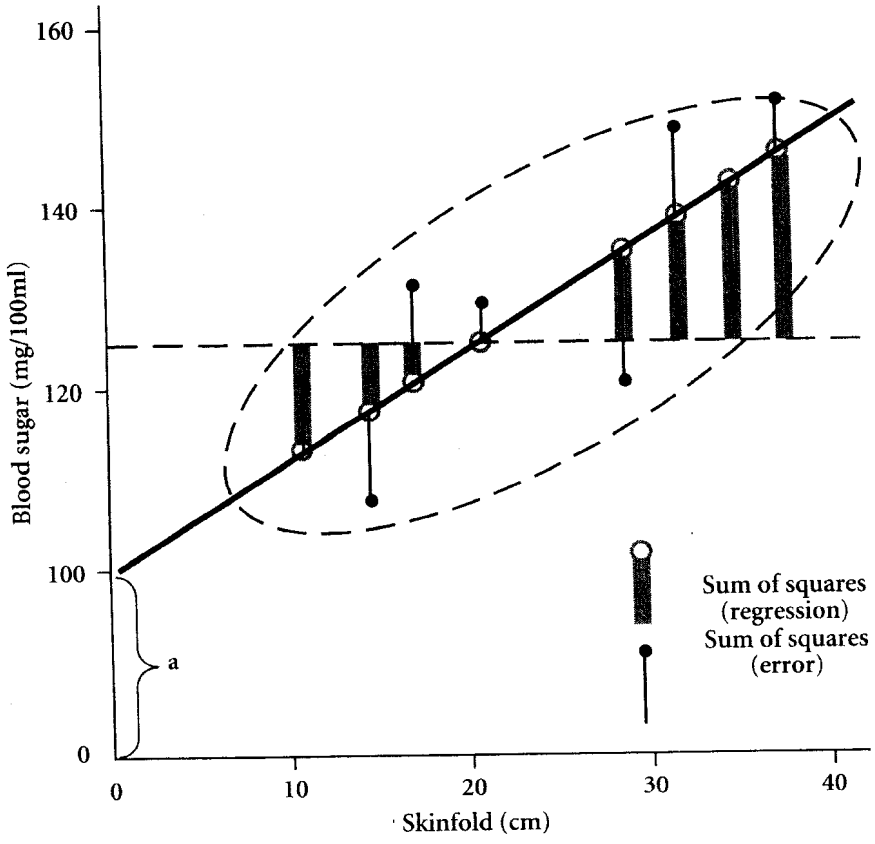

Figure 6-1 Relationship between blood sugar and skinfold.

Table 6-1
Output from Regression Analysis

Variable	Coefficient	SE	*t*	Significance
Constant	98.00	12.3	7.97	0.0001
Skinfold	1.14	0.046	24.70	0.0001

commonly this significance testing comes about in a different way. Frequently, the computer printout will list a table of numbers with headings like Table 6-1.

The coefficients are the intercept (Constant), which equals 98.0, and can be identified on the graph. Similarly, the second line of the table is the slope of the regression line. The computer also calculates the standard error (SE)

Table 6-2
ANOVA Table from Regression Analysis

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	*p*
Regression	2401.5	1	2401.5	610.1	<0.0001
Residual	270.0	18	15.0		

of these estimates, using complicated formulas. The *t* test is the coefficient divided by its SE, with $n - 2$ degrees of freedom (where n is the sample size), and the significance level follows.

Usually the computer also prints out an "ANOVA table." But, sez you, "I thought we were doing regression, not ANOVA." We have already drawn the parallel between regression and ANOVA in the previous section. These Sums of Squares end up in the ANOVA table as shown in Table 6-2. If you take the square root of the F ratio, it equals exactly the *t* value calculated earlier (as it should because it is testing the same relationship).

Finally, the strength of relationship could then be expressed as the ratio of SS (regression) to [SS(regression) + SS(residual)], expressing the proportion of variance accounted for by the IV. In fact, usually the square root is used and is called a **Pearson correlation coefficient**.

$$\text{Correlation} = \sqrt{\frac{\text{Sum of Squares (regression)}}{\text{Sum of Squares (regression)} + \text{Sum of Squares (residual)}}}$$

So, in the present example, the correlation is:

$$R = \sqrt{\frac{SS\,(reg)}{SS\,(reg) + SS\,(res)}} = \sqrt{\frac{2401.5}{2401.5 + 270.0}} = 0.95$$

We also could have tested significance of the relationship directly by looking up significance levels for different values of the correlation coefficient and different sample sizes. This is, of course, unnecessary at this point.

We can interpret all this graphically by referring back to Figure 6-1. In general, the individual data points constitute an ellipse around the fitted line. The correlation coefficient is related to the length and width of the ellipse. A higher correlation is associated with a thinner ellipse and better agreement between actual and predicted values.

TWO OR MORE INDEPENDENT VARIABLES: MULTIPLE REGRESSION

In the previous example we dealt with the relationship between blood sugar and skinfold. This is the simplest form of regression analysis in that there is

one IV, one DV, and a presumed straight line relationship between the two. A bit of reflection suggests that blood sugar is likely to be influenced by other variables, diet and heredity, for example. If these could be used, it seems likely that our ability to predict blood sugar levels would improve. There is a fair amount of mileage to be gained by using several IVs in predicting a DV. The technique is called **multiple regression** and is an extension of the previous approach.

Suppose, for example, you are chair of the admissions committee for the residency training program in pediatric gerontology at Mount Vesuvius Hospital. Every year you have interviewed all the applicants to the program, but you wonder if you might save some money and predict performance better using previous academic records. Performance in the program is based on a rating by supervisors. The academic record of applicants contains (1) grade point average in medical school (MDGPA), (2) National Board license examination results (NBE), and (3) undergraduate grade point average (UGPA). The regression equation using these IVs might look like the following.

$$\text{Performance} = a + (b \times \text{MDGPA}) + (c \times \text{NBE}) + (d \times \text{UGPA})$$

The statistical analysis is conducted by estimating values of the parameters $a \rightarrow d$ in such a way as to minimize the squared differences between the real data and the estimated points. Essentially, what the computer is doing is fitting a straight line in four-dimensional space (you will forgive us if we don't include a figure). And once again, the overall goodness of fit is determined by calculating a correlation coefficient from the ratio of the variance fitted by the line to the total variance. This correlation coefficient is called the multiple correlation (written R). The square of the **multiple correlation** can be interpreted directly as the proportion of the variance in the DV, ratings, accounted for by the IVs.

Of course, that isn't all the information obtained from the analysis. The computer also estimates the coefficients *a* to *d* and does significance testing on each one. These coefficients indicate the degree of relationship between performance and each IV after the effects of all other variables have been accounted for.

Suppose, for example, that the resultant regression equation looked like the following:

$$\text{Performance} = 0.5 + (0.9 \times \text{MDGPA}) + (0.04 \times \text{NBE}) + (0.1 \times \text{UGPA})$$

The estimated coefficients (0.9, 0.04, 0.1) are called **unstandardized regression coefficients**. Funny name because they look standard enough. It would appear that MDGPA predicts quite a bit and NBE very little. But let's take a closer look. Suppose MDGPAs have a mean of 3.5 and SD of 0.25. Then a change of one SD in MDGPA results in a change of $0.9 \times 0.25 = 0.225$ in

performance. By contrast, if NBE scores have a mean of 75% and SD of 20%, a change of one SD yields a change in performance of $20 \times 0.04 = 0.8$ units in performance. So the size of the coefficient doesn't reveal directly how predictive the variable is. To make life easier, we often transform these to **standardized regression coefficients** or **beta weights** by converting each variable to have a mean of 0 and SD of 1. The resulting weights can then be compared directly. In the present example, that would result in weights of 0.53 for NBE and 0.15 for MDGPA; thus NBE is approximately three times as strong a predictor as MDGPA.

You might have your interest piqued by these results to explore the situation a bit further. For example, if you can do nearly as well without the UGPA, you might be prepared to forego this requirement. The question you now wish to ask is, "How much do I gain in prediction by adding in UGPA?"

One approach to this question would be to fit another regression line using only MDGPA and NBE and determining the multiple correlation. The difference between the squared multiple correlations for the two equations, with UGPA in and out of the prediction, tells you how much additional variance you have accounted for by including the variable. This is the basic process in **step-wise regression**, a method where predictor variables are introduced one at a time into the regression equation, and the change in the multiple correlation determined. There are two ways of approaching stepwise regression: Either you can introduce the variables into the equation in some logical order specified by the experimenter, as in the current example, or you can let the computer decide the sequence.

The computer method is probably more popular. There are a number of esoteric criteria used to determine the order in which variables will be introduced. Basically, the computer enters them in an order of decreasing ability to account for additional variance, a sort of statistical law of diminishing returns, so that at some point you can determine at what point little is to be gained by adding more predictor variables. The cutoff can be based on the statistical consideration that the contribution of an additional variable is not statistically significant. Alternatively, it can rest on more pragmatic grounds, namely that the additional explained variance isn't worth the effort.

Although stepwise regression in which the computer does all the work is likely more popular because it avoids any deep thought on the part of the researcher, it has fallen on hard times at the hands of real statisticians. The reason of course is a statistical one.

To understand this, put yourself in the shoes of a computer—something that is lightning fast but kind of thick (like some of the guys on athletic scholarships). You are doing stepwise regression on someone's pet data base, consisting of 193 variables on 211 subjects. You have searched the data for the most significant relationship and put it into the equation. Now it's time

to find the next one, which you do by doing a whole bunch of new regression analyses with two variables in the equation, recalculating all the significance levels as if each of the remaining 191 variables was lucky enough to be chosen. That's 191 F tests. You put the lucky variable in the equation next, then repeat the procedure to find the next variable—another 190 significance tests.

The big trouble is that some of those significant F tests got that way by chance alone, based on random variation in the data set. The chances of replicating the findings with a new data base are the same as a snowball's survival time in Arizona in July. The solution is to go back to the old way, entering the variables according to a sequence laid out in advance by the experimenter. It's more intellectually satisfying, too. You can enter variables systematically according to a theory (e.g., family structure "causes" family dysfunction "causing" loss of self esteem, which, with life stress, "causes" depression) or on more pragmatic grounds (in the above example, GPA from medical schools is more standardized and relevant than UGPA). To ensure that everybody knows you're still doing science, you can glorify the whole process with a big term. It's called **hierarchical regression**.

Carrying the analysis of our example one step further, the data from the regression analysis might be presented in the style of Table 6-3.

The data show each step of the analysis on each successive line. The first step is just a simple regression, with the multiple R^2 equal to the square of the simple correlation $(0.50)^2 = 0.25$. The computer then calculated an F ratio, which proved to be statistically significant. At the second step, MDGPA was added, explaining an additional 8% of the variance, and again this was statistically significant. Finally, introducing UGPA explained only 2% more of the variance and was not significant.

One bit of subtlety: sometimes a variable that has a high simple correlation with the DV won't do anything in the multiple regression equation. Returning to our blood sugar example, an alternative measure of obesity might be kilograms above ideal weight, and it might correlate with blood sugar nearly as well as skinfold. But it's likely that the two are highly correlated, so if skinfold goes into the regression equation first, it is probable that kilograms won't explain much additional variance and may not be significant.

Table 6-3

Results of Stepwise Regression Predicting Resident Performance

Step	Variable	Multiple R2	Change in R2	F Ratio	Significance
1	NBE	0.25		13.78	<0.0001
2	MDGPA	0.33	0.08	3.02	<0.05
3	UGPA	0.35	0.02	1.78	N.S.

The message here is that a variable may not be a useful predictor of the DV, for two reasons. One, it has a low correlation with the DV. Two, it has a reasonable correlation with the DV, but is highly correlated with another IV that has higher correlation with the DV and enters the equation first.

So that's what multiple regression looks like. Beware the study founded on a large data base but with no prior hypotheses to test with an adequate experimental design. The investigators can hardly resist the temptation to bring out the statistical heavy artillery such as multiple regression to reach complex, glorious, and often unjustified conclusions about the relationships between variables.

C.R.A.P. DETECTORS

Example 6-1

A large data base gathered during a 10-year period in a certain Midwestern town ($n = 12{,}498$) was analyzed to determine predictors of high serum cholesterol. Twenty-eight different dietary factors were examined, and it was found that serum calcium levels correlated -0.07 ($p < 0.05$) with serum cholesterol levels. They concluded that low calcium causes high cholesterol.

Question. Will you drink more milk?

Answer. Not from these findings. A correlation of -0.07 is statistically significant because of the large sample, but it accounts for only $0.07^2 = 0.49\%$ of the variance. Anyway, we would expect that one of the 28 variables would be significant by chance alone. Finally, the association may be caused by something else.

C.R.A.P. Detector VI-1

Beware the large sample revisited. With large samples, statistical significance loses all relationship to clinical significance.

C.R.A.P. Detector VI-2

Watch out for fishing expeditions. The meaning of $p < 0.05$ applies equally well to correlation coefficients.

C.R.A.P. Detector VI-3

Correlation does not imply causation. Height and weight are highly correlated, but height doesn't *cause* weight. Researchers can get carried away

when they see large correlations and start to interpret them as evidence of a causal relationship.

Example 6-2

A teacher of dyslexic children interviewed the parents of 12 of his students and found that birth order was significantly correlated with reading test scores ($R = 0.65$, $p < 0.05$). He concluded that lack of parent stimulation in infancy, which is more likely in large families, is a cause of reading problems.

Question. Do you agree?

Answer. There are a few difficulties here. With only 12 kids, one or two from very large families could bump up a correlation, significance notwithstanding. Also the fact that birth order was correlated with dyslexia does not imply a causative relationship.

C.R.A.P. Detector VI-4

Beware the small sample revisited. It is easy to obtain correlations that are impressively large but cannot be replicated. A good rule of thumb is that the *sample size, or number of data points, should be at least five times the number of IVs.*

C.R.A.P. Detector VI-5

Regression equations may fit a set of data quite well. But extrapolating beyond the initial data set to values higher or lower is tenuous.

C.R.A.P. Detector VI-6

The multiple correlation is only an indicator of how well the initial data were fit by the regression model. If the model is used for different data, the fit won't be as good because of statistical fluctuations in the initial data.