Spring Term 2003

Research Methods II

Regression: Summary

Regression involves predicting

-a continuous variable, called the criterion variable or dependent variable

from

-one or more other (continuous or discrete) variables, called the predictor variables or independent variables.

Last year you learnt about simple regression, which is where there is just one independent variable (and one dependent variable of course). For example, you might want to predict the amount of happiness associated with different incomes. You would measure happiness on a continuous interval scale and measure income of the people in your sample. Then plot happiness against income and see if the relationship looks linear. If so, you could use regression to find the best fitting straight line: $H = a + b*I$. That is, the amount of happiness (H) is equal to a constant (a, called the intercept) plus the slope (b) times the amount of income (I). The intercept, a, tells you how happy you would be with no income, and the slope, b, tells you how much extra happiness each extra pound buys you.

Now imagine you want to predict the amount of people's happiness given their income and a measure of the degree of control they have in their work life (call this C). The relationship can be inspected graphically in the following way. Imagine a flat horizontal surface, like that on a table in front of you. The long side of the table will be our income (I) axis, and the other side going at 90 degrees will be our control axis (C). Any point on the table then represents a particular combination of income and control. The degree of happiness of the person with that combination of I and C can then be plotted in the axis going up to the ceiling. As we plot the data for all subjects we will be filling a three dimensional space above the table surface (and below it, if people in our sample get that unhappy!).

If when you look at the scatter of points in 3D space it looks like it is roughly described by a rigid plane then you could use multiple regression (i.e. regression with more than one independent variable) to find the best fitting plane: $H = a + b_1*I + b_2*C$. The intercept, a, is how happy you predict people to be when they have no income ($I=0$) and they have the level of control in their work corresponding to $C=0$ (this scale would just have an arbitrary zero point). b_1 is the slope of the plane along the income dimension, and b_2 is the slope along the control dimension.

What do these slopes mean? b_1 is how much extra happiness each pound buys you *when C is kept constant.* Satisfy yourself this is true from the equation - if variable I increased by 1 unit and C by no units, H would increase by an amount b_1. Imagine that when you conducted a simple regression $H = a + b*I$ you find a significant slope. That is, happiness and money were related. It may be that money makes you happy, or being happy enables you to earn lots of money, or it may be that money and happiness don't directly affect each other but both depend on a third variable. For example, people who have more control in their work may tend to both earn more money and to be more happy. Imagine it is really the control that causes the happiness, and income is only positively related to happiness because income is related to amount of control at work, and amount of control is related to happiness. That is, if you kept control constant there would be no relationship between happiness and money. That's just the sort of information that b_1 tells you: The regression slopes tell you the unique effect of each variable above and beyond the effect of the other variables in the equation. We can say we are testing the relationship between happiness and income *controlling for* or *taking account of* or *partialing out* the amount of control a person has at work. Notice that the multiple regression slope b_1 relating income to happiness can be different from its simple regression value b, because the IVs are themselves correlated. If the IVs were not correlated, the size of the slopes b_1 and b would not differ systematically.

When you find the best fitting equation of the plane, each person has both their actually measured happiness, H, and the amount of happiness the equation predicts they should have given their I and C: this predicted value or *fit* is often represented: $\hat{H}$. The differences between the observed H values and the fits are called the residuals. The residual is the amount of error in predicting a person's happiness. The variance of the fits is the amount of variance in the people's happiness that we could explain by the independent variables, and the variance of the residuals is the amount of variance in people's happiness that we couldn't explain by the independent variables. The proportion of the variance of H explained by the independent variables is variance of fits/variance of H. This give us a good idea of how well we can really predict happiness, and whether we should look for further independent variables to get a better prediction, if we need a better prediction. Another way of seeing how well we can predict H from our independent variables is to correlate the fits with the observed values, to get what is called *the multiple correlation coefficient* r. In fact, r^2 IS the proportion of variance explained, so these two methods are equivalent.

You can look at the residuals as the noise through which you trying to determine whether the plane really has non-zero slopes - it is identically the problem in ANOVA of looking through the noise of the within group variation to see if the different groups really have different population means. The more genuinely predictive variables you add to the regression equation, the smaller the residuals will be, the smaller the noise will be. Thus, even if your IVs are completely unrelated to each other, it can still be useful to use multiple regression to determine the significance of each IV because you will be assessing their slope values with less noise obscuring their true value.

Let us say you were to *standardize* all the variables before entering them into the equation (that is, for each variable express each person as a difference from the mean in standard deviation units). We will call the standardized variables h, I, and c. The equation becomes: $h = ß_1*i + ß_2*c$. (Note the intercept is always zero.) The ß (pronounced "beta") are the *standardized* regression coefficients. For example, $ß_1$ tells you how much happiness increases (as a fraction of its standard deviation) for one standard deviation increase in income. The standardized regression coefficients can give you a rough idea of the relative size of effect of different variables measured on a common scale. For example, based on the *raw* regression coefficients (i.e. the non-standardized ones) one might say that each pound increase in income increases happiness by 10 units (if that were the value of b_1) and each unit increase in control increases happiness by 3 units (b_2). But why juxtapose one pound of income with one unit of control? Why not one pence of income? In standardized units, one says that one standard deviation increase in income increases happiness by 0.2 standard deviations ($ß_1$) and one standard deviation increase in control increases happiness by 0.4 standard deviations ($ß_2$), both on a common scale.