

Regression and stepwise regression

An alternative medical example of a multiple regression problem

One hundred and ten patients diagnosed as having the eating disorder, bulimia, are recruited to a study designed to evaluate potential predictors of SUCCESS of a ten week course of treatment using cognitive behavioural therapy (CBT). The DV is a success score based on responses to patient and GP post-treatment questionnaires. The potential predictors include scores on three variables that may be regarded as providing good approximations to interval scales: (1) score on a self esteem inventory (SELFEST), (2) SEVERITY of the eating disorder based on family and patient interviews and questionnaires and (3) score on a DEPRESSION inventory. This is an example of an attempt to use regression for prediction. Of course it may also contribute to an explanation, and even in a small way to control. If it turns out that you can predict how well patients do using measures of self esteem and depression (as well as severity), it may be that the illness could be treated by efforts to raise self esteem, or else by treating the patients for depression.

Multiple regression: some (fabricated) data

The first few rows of data are shown in Table 4.1 (the full dataset can be found as med.regression.stepwise.sav on the book website). Also in Table 4.1, we have two categorical variables, GP and FAMILY, which record how the patient was referred and will be used later, but for the moment we will ignore these and consider only the first four columns. They should be entered into the SPSS datasheet in the usual way, one row for each case, so we have six columns, including the two variables we use later.

Table 4.1

The first few rows of data on treatment success, self esteem, severity and referral (the full dataset can be found as med.regression.stepwise.sav on the website)

| success | selfest | severity | depression | GP | family |
|----------------|----------------|-----------------|-------------------|-----------|---------------|
| 68 | 117 | 104 | 27 | 0 | 0 |
| 36 | 93 | 90 | 43 | 0 | 0 |
| 25 | 101 | 96 | 48 | 1 | 0 |
| 36 | 116 | 108 | 59 | 0 | 0 |
| 35 | 103 | 92 | 45 | 1 | 0 |
| 35 | 101 | 95 | 38 | 0 | 1 |

Multiple regression: inspecting the correlation matrix

The usual starting point for a regression analysis is inspection of the correlation matrix for all of the variables. This gives you an initial idea of the relationships among the variables, but not everything of interest is necessarily apparent from the pattern of correlations. Sometimes plotting the DV against the IVs one at a time can suggest when a relationship may exist but be non-linear and may give you a better indication of whether a regression analysis will be helpful. However, we need to be careful about drawing conclusions based on plots involving one IV at a time because, unfortunately, with several independent variables, looking at them one at a time gives little indication of how effective they may be as predictors together. Look at Figure 4.2, which shows SUCCESS plotted against SEVERITY. This graph does not look promising: it seems unlikely that SEVERITY will be a useful predictor of SUCCESS. Nevertheless, it turns out that, in the presence of SELFEST, it does make a significant contribution to predicting SUCCESS.

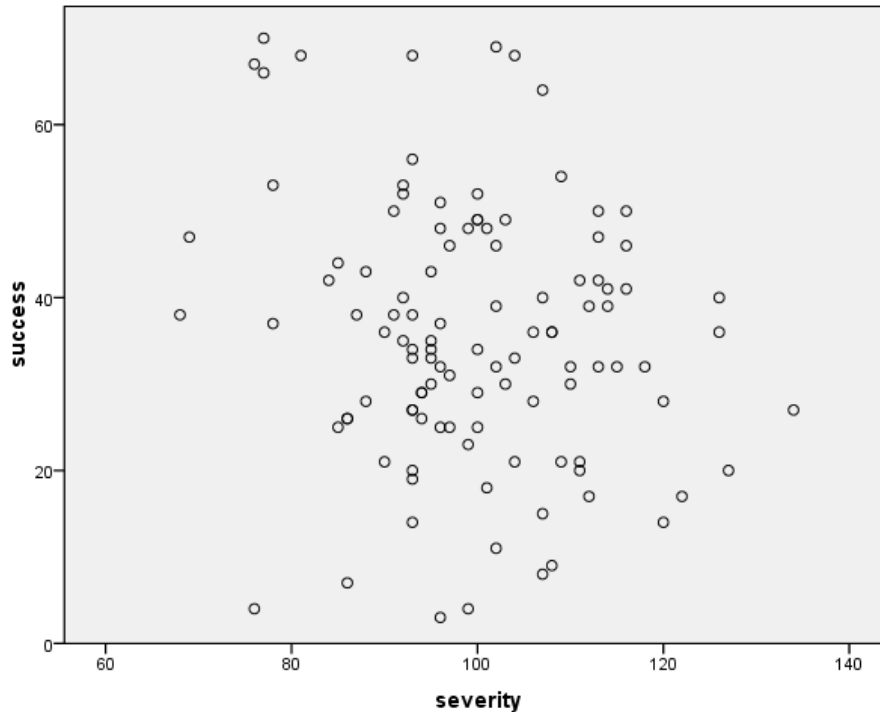


Figure 4.2. Dependent variable plotted against one of the IVs

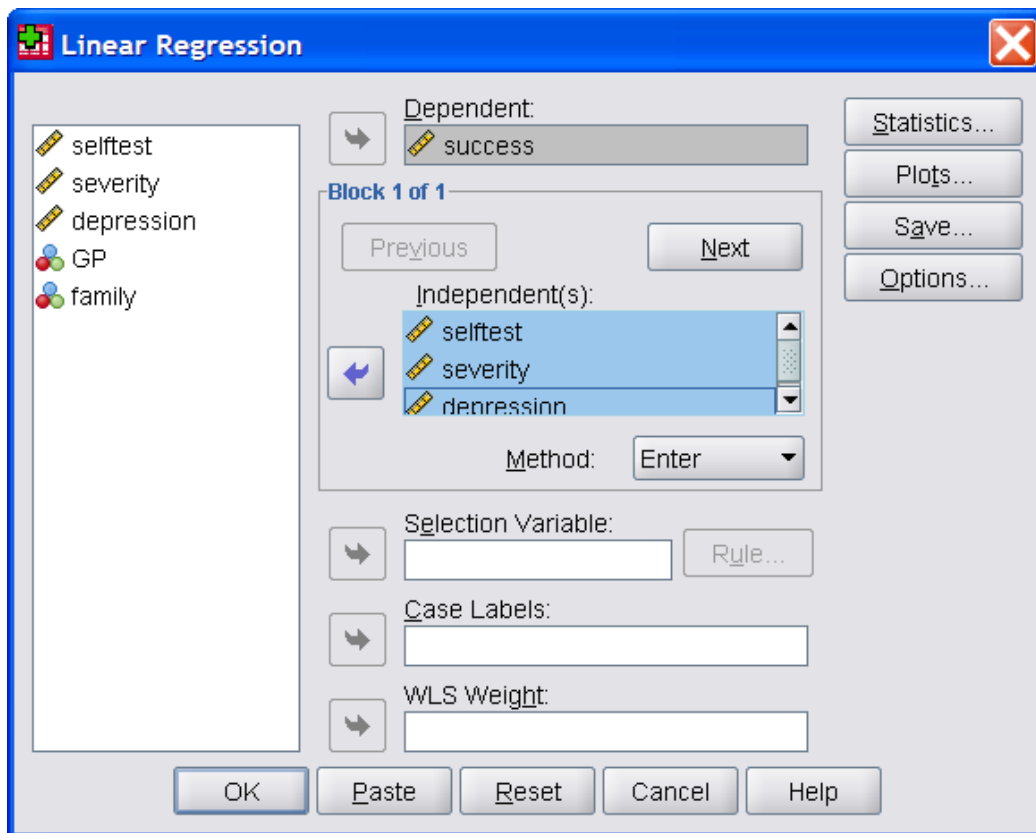
Multiple regression: statistical analysis

It's not hard to do a regression analysis, in fact if you only have one IV you can easily do it on a calculator, but deciding whether your regression model is a good enough fit, and whether it is of any practical use, is more difficult. If it does look useful, then users must remember that it should not be applied outside the range of values of the IVs that was covered in the original dataset. Also it can be tempting to fall into the trap of assuming that successful prediction implies causation. Every scientist knows that association does not imply causation, but somehow a neat regression analysis that results in successful prediction sometimes seems to lure people into believing they have identified causes and not just predictors. SPSS provides many ways to check on the fit, and we will look at several of them, but only your good scientific sense will keep you from the other errors.

Standard multiple regression: requesting the analysis in SPSS

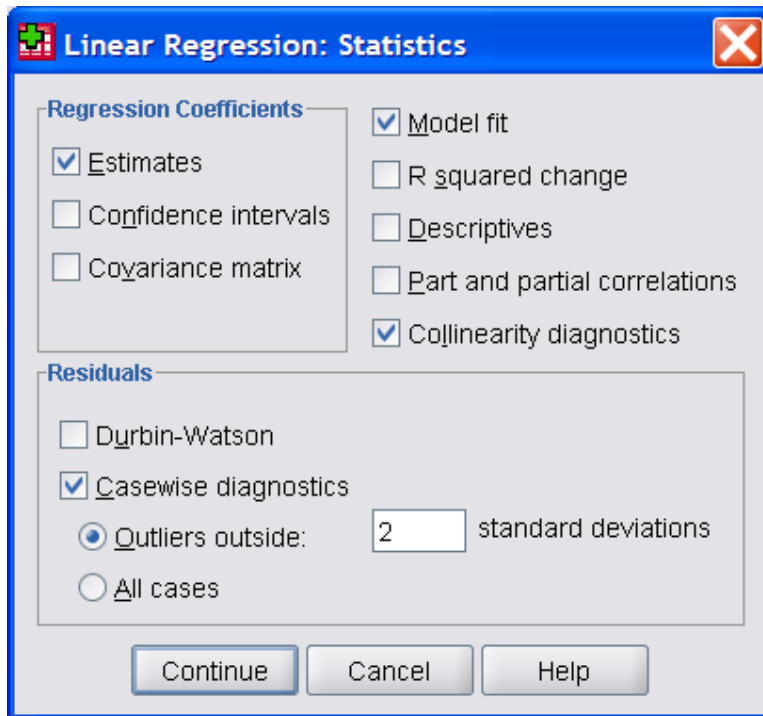
From the menu bar choose **Analyze**, then **Regression**, then **Linear** to get SPSS

Dialog Box 4.1. Select the DV (SUCCESS in our case) from the list on the left and use the arrow to put it in the **Dependent** box. Then put the IVs (SELFEST, SEVERITY and DEPRESSION) into the **Independent(s)** box in the same way. You may want to try out regressions on each IV one at a time before you get into multiple regression, but here we will proceed straight to using all our IVs. Make sure the **Method** box reads **Enter** (we look at a stepwise alternative later). The dialog box now looks like SPSS Dialog Box 4.1, and you could accept all the defaults and click **OK**. However, we will make use of the buttons at the side to get some extra information.



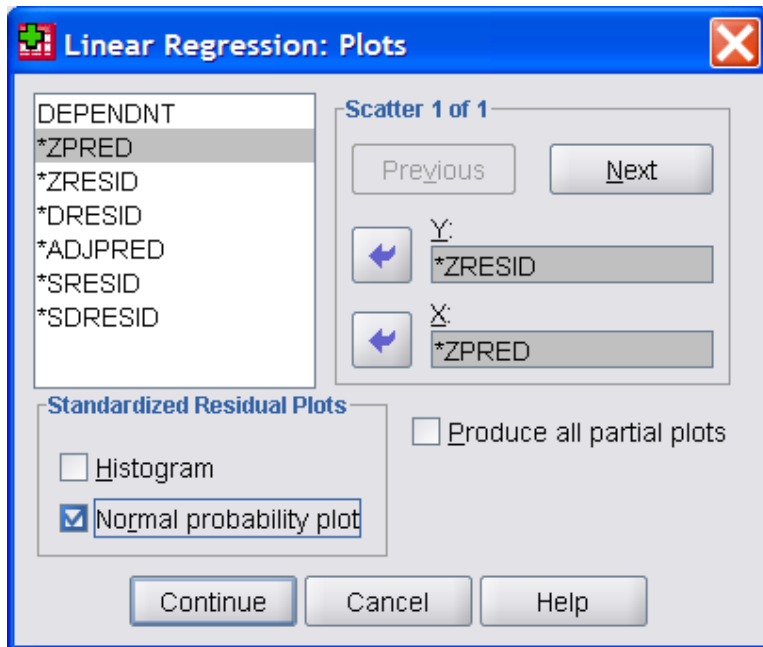
SPSS Dialog Box 4.1. First dialog box for regression

First click the **Statistics** button and get SPSS Dialog Box 4.2. The default settings will show the estimated regression coefficients (**Estimates**) and the **Model fit** in the Output window. **Confidence intervals** for the regression coefficients may also be useful (wide ones indicate that the estimates are rather imprecise). The **Durbin-Watson** test for serially correlated residuals is useful if data have been collected sequentially over time, so that they may not be independent observations. **Casewise diagnostics** (selected here) enables you to identify any cases that are badly fit by the model. Outliers outside two standard deviations should occur by chance in only about 5% of our cases, and larger deviations should be even more rare. Here we have requested a list of cases where the residuals exceed two standard deviations. The **R squared change** is one way of judging the effectiveness of each IV in the regression. It is, however, unnecessary to select it in this case because the information will be in the default output since we are entering all our IVs together. In later analyses where we enter IVs in stages you need to select this statistic. The **Descriptives** and the **Part and partial correlations** give information about the IVs, and the **Collinearity diagnostics**, selected here, indicate if there are linear relationships among the IVs that would make the estimates of the regression coefficients unreliable. Click **Continue** to return to SPSS Dialog Box 4.1.



SPSS Dialog Box 4.2. Dialog box for additional information about model fit

Now click the **Plots** button to get SPSS Dialog Box 4.3. A plot of the residuals against the fitted values is useful for checking the assumption that the variance of the DV is constant over the range of values of the IVs. This can also alert us to other failures of the model, as we will see when we look at the astronomy example later. The **ZRESID** and **ZPRED** are just the residuals and predicted values, each standardized by subtracting the mean and dividing by the standard deviation, so select them for plotting as Y and X respectively. The **Normal Probability Plot** will give us a check on the assumption of normality that we need for hypothesis tests. SPSS Dialog Box 4.3 shows these selections.



SPSS Dialog Box 4.3. Selecting plots from a regression analysis

Click **Continue** to return to SPSS Dialog Box 4.1. The **Save** button can be left for later and the **Options** button applies when we do a stepwise regression, again later. Click **OK** to get the analysis.

Standard multiple regression: understanding the main analysis output

First in the output is a table (not shown) that lists which variables are in the regression and shows whether a stepwise method was used. We just entered all our independent variables so this table is of little interest. We consider later the version obtained when we do a stepwise regression.

Next come the Model Summary and ANOVA tables shown in SPSS Output 4.1.

Under the model summary table are listed the independent and dependent variables, and in the table we see the value of R^2 , which tells us that in our example 0.302, or 30%, of the variance of SUCCESS was explained by the regression on SELFEST,

SEVERITY and DEPRESSION. R^2 always increases with the inclusion of additional predictor variables. Adjusted R^2 , which is also given, takes account of the number of predictor variables and also the number of cases. This statistic is intended to allow comparison between different regressions or even regressions fitted to different sets of data, and may be useful as an initial rather rough indicator for this purpose. Normally it is the unadjusted R^2 that is reported. The ANOVA table shows that for the overall regression, $F = 15.29$ with 3 and 106 degrees of freedom, with a probability (in the Sig column) well below 0.05. So the regression is significant.

Model Summary^b

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .550 ^a | .302 | .282 | 12.776 |

a. Predictors: (Constant), depression, severity, selftest

b. Dependent Variable: success

ANOVA^b

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|-----|-------------|--------|-------------------|
| 1 | Regression | 7486.951 | 3 | 2495.650 | 15.289 | .000 ^a |
| | Residual | 17302.540 | 106 | 163.232 | | |
| | Total | 24789.491 | 109 | | | |

a. Predictors: (Constant), depression, severity, selftest

b. Dependent Variable: success

SPSS Output 4.1. Model summary and ANOVA tables

Next is a table of coefficients, shown in SPSS Output 4.2. This (in the column labelled 'B') gives the estimated values of the regression coefficients and then their standard errors. From this we see that we can calculate the predicted value of SUCCESS for any case as:

$$\text{Predicted SUCCESS} = 15.570 + 0.595 * \text{SELFEST} - 0.342 * \text{SEVERITY} - 0.142 * \text{DEPRESSION}.$$

Higher scores on SELFEST predict higher scores on SUCCESS (the coefficient is positive), but higher scores on the other two variables predict lower scores on SUCCESS.

In the *t* and Sig columns we see the results of testing, for each coefficient, the null hypothesis that it is zero, assuming that the other variables are included in the regression. The probability of this value of the *t* statistic if the null hypothesis is true is also given. We can see that for our example, SELFEST and SEVERITY are significantly different from zero but DEPRESSION is not.

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|-------|------------|-----------------------------|------------|---------------------------|--------|------|-------------------------|-------|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 15.570 | 13.397 | | 1.162 | .248 | | |
| | selftest | .595 | .097 | .512 | 6.151 | .000 | .952 | 1.051 |
| | severity | -.342 | .100 | -.285 | -3.430 | .001 | .956 | 1.047 |
| | depression | -.142 | .112 | -.103 | -1.267 | .208 | .994 | 1.007 |

a. Dependent Variable: success

SPSS Output 4.2. Regression coefficients and collinearity check

Standard multiple regression: understanding the diagnostic output

Because we requested collinearity diagnostics in SPSS Dialog Box 4.2, we have two extra columns on the right. The first column, Tolerance, gives a value between zero and one, which is the proportion of a variable's variance not accounted for by the other independent variables in the regression. Values close to zero show that there are linear relationships among the independent variables that will cause computational problems. In that case, one or more of the independent variables should be removed. All our values are close to one, so the independent variables do not depend linearly on each other. The VIF (or variance inflation factor) is just the reciprocal of the tolerance. There is another table (not reproduced here) with further details on collinearity but the tolerance values in SPSS Output 4.2 will usually suffice.

Since we requested casewise diagnostics in SPSS Dialog Box 4.2, we get a table listing cases where the residual is more than two standard deviations (shown in SPSS Output 4.3). There are three of these, with the largest (in absolute value) just under 3,

which is no cause for concern with 110 cases. We would expect about 5% as large as 2, just by chance. SPSS can calculate a variety of transformations of the (raw or unstandardised) residuals. The **standardized residual** is the residual divided by its standard error. Hence observations with residuals with more than two standard deviations are those with a standardized residual greater than 2. The **studentized residual** is similar to the standardized residual but it is more sensitive to departures from the model. This is because its scaling factor takes into account how far the observation is from the mean. The deleted and studentized residuals for an observation are found as above but from the model that excludes that observation.

The next table is also shown in SPSS Output 4.3, and gives descriptive statistics for the predicted values and residuals, and the versions of them standardized to Z values.

Casewise Diagnostics^a

| Case Nu... | Std. Residual | success | Predicted Value | Residual |
|------------|---------------|---------|-----------------|----------|
| 67 | -2.903 | 4 | 41.09 | -37.094 |
| 84 | 2.056 | 64 | 37.74 | 26.262 |
| 97 | 2.095 | 69 | 42.23 | 26.771 |

a. Dependent Variable: success

Residuals Statistics^a

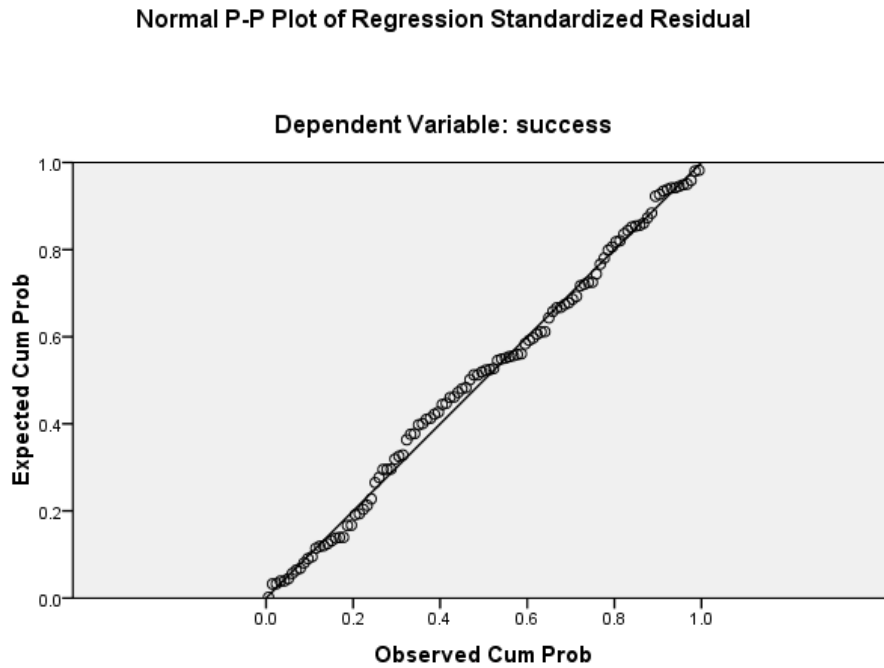
| | Minimum | Maximum | Mean | Std. Deviation | N |
|----------------------|---------|---------|-------|----------------|-----|
| Predicted Value | 11.61 | 60.33 | 35.51 | 8.288 | 110 |
| Residual | -37.094 | 26.771 | .000 | 12.599 | 110 |
| Std. Predicted Value | -2.883 | 2.994 | .000 | 1.000 | 110 |
| Std. Residual | -2.903 | 2.095 | .000 | .986 | 110 |

a. Dependent Variable: success

SPSS Output 4.3. Information about residuals and predicted values

The Normal probability plot we requested in SPSS Dialog Box 4.3 is shown in SPSS Output 4.4. If the assumption of Normality is correct, then the residuals form a random sample from a standard Normal distribution, and the plot shown will be a straight line from the origin (0,0) to top right (1,1). Our residuals are quite a good

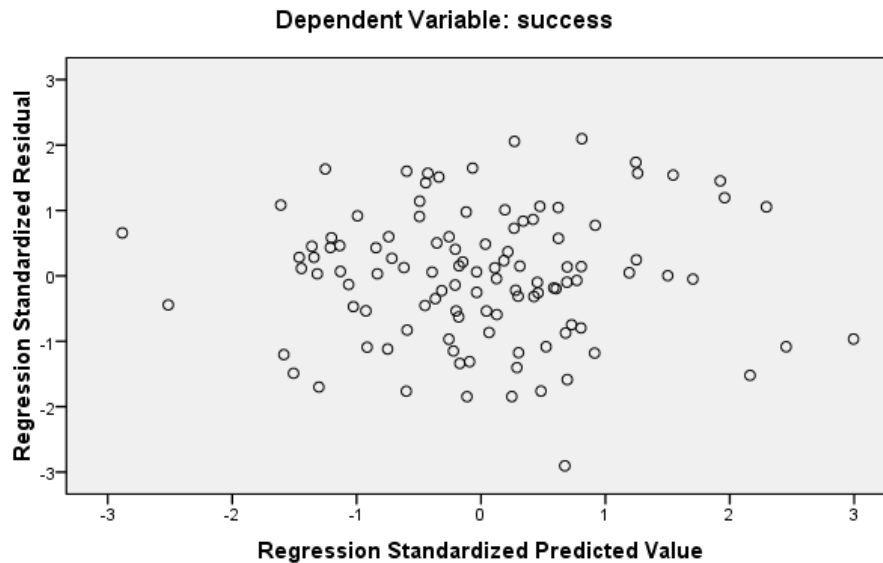
approximation to that, which means we can have confidence in the ANOVA and t tests on the regression coefficients.



SPSS Output 4.4. Normal probability plot of residuals

The residual plot that was also requested in SPSS Dialog Box 4.3 appears last and is shown in SPSS Output 4.5. If the assumption of constant variance is true, and if the linear regression model is a reasonable fit, the residuals should not show any dependence on any of the independent variables, and so not on the predicted values either. In fact if our assumptions about the data are correct the residuals form a random sample from a Normal distribution and so this plot should show a shapeless cloud of points: any discernible pattern is bad news for our model. But ours looks good. We will show you one later that displays the inadequacy of a regression model.

Scatterplot



SPSS Output 4.5. Residual plot: standardized residuals against standardized predicted values

Using a stepwise method

We have already seen that the t -test on the coefficient of DEPRESSION suggested that this variable contributes nothing to predicting SUCCESS. We could try the regression again, omitting this variable, and see whether the proportion of the variance of SUCCESS accounted for by the regression is significantly reduced. But SPSS offers a way to deal with this as part of the analysis. This may be useful if you have a longer list of independent variables, some of which may be useful predictors, but some of which are almost certainly useless. What you want is the best subset to do the job of prediction as well as possible with as few variables as possible. Trying out all possible subsets yourself would not only be tedious and time consuming, but at the end you would have difficulty sorting out the best from your long list. Even with only three possible independent variables, there are seven possible subsets containing at least one of them. One solution is to use a stepwise method, and SPSS offers three. We

would, however, like to include a note of caution here. The stepwise methods we are about to describe select a subset of variables entirely by statistical criteria. One variable may be included and another excluded when the difference between their ability to predict the DV is very small and perhaps an effect of random variation in the sample data. It is always best to select variables for possible inclusion in a regression on the basis of theoretical or practical considerations. Of course, when you try the regression you may have to reject some of these: your theory may not be as good as you hope and the relationships it proposes may be weak or even non-existent. But choosing a subset of variables for which you have no coherent theoretical underpinning is unlikely to be a very fruitful strategy. Nevertheless, a stepwise method may be a useful way to identify some promising candidate variables for further consideration, so we explain how to do it in SPSS. There are three methods.

Forward selection

Forward selection starts with the independent variable that is the best predictor of the dependent variable and checks that the coefficient is significantly different from zero, at the 5% level. Then it adds the one that improves the prediction the most, subject to a criterion, usually that the coefficient is significantly different from zero at the 5% level. The process continues until no more variables pass the criterion. SPSS allows adjustment of the criterion so you could add variables that are significant at, say, 10%, or else only those that are significant at 1%.

Backward elimination

Backward elimination starts with all independent variables in the regression, then removes the one with the smallest t statistic, provided that its p value is at least 0.10.

The process continues until no more variables are removed. Again, the criterion can be adjusted.

Stepwise regression

Stepwise regression combines forward selection and backward elimination. At each step, the best remaining variable is added, provided it passes the significant at 5% criterion, then all variables currently in the regression are checked to see if any can be removed, using the greater than 10% significance criterion. The process continues until no more variables are added or removed. This is the one we shall use. It is not guaranteed to find the best subset of independents but it will find a subset close to the best.

Stepwise regression: requesting the analysis in SPSS

In SPSS Dialog Box 4.1, in the **Method** box, replace **Enter** with **Stepwise**. You will also find in the list of methods one called **Remove**. This allows you to remove some variables from the equation as a block. You can also control the entry of successive IVs or blocks of them, rather than allowing statistical criteria to determine what gets entered and in what order. This is achieved with the **Enter** method selected and using **Next**, between the **Dependent** and **Independent(s)** boxes. We will consider this sequential method later.

The statistical criteria for adding and removing variables can be modified in a dialog box that is opened by clicking on the **Options** button, but we can accept the defaults (Enter if significant at 5% and remove if not significant at 10%, as described above). You can also force the regression equation to pass through the origin by not including the constant, but this is rarely required, and the default is to include the constant.

There are also different ways to deal with missing values but again we can accept the default.

Stepwise regression: understanding the output

When we use a stepwise method, the first table in the output, shown in SPSS Output 4.6, is more interesting. In our example it shows that SELFEST was the best predictor of SUCCESS and was entered first. Then SEVERITY was entered, and neither was removed. No more variables were entered or removed and so the final model has just SELFEST and SEVERITY to predict SUCCESS. The proportion of the variance of SUCCESS explained by this model is 0.291 (R^2 for Model 2 - the final model - in the second table in SPSS Output 4.6), which is almost the same as the 0.302 achieved with DEPRESSION included as well.

Variables Entered/Removed^a

| Mo... | Variables Entered | Variables Removed | Method |
|-------|-------------------|-------------------|---|
| 1 | selfest | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |
| 2 | severity | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |

a. Dependent Variable: success

Model Summary^c

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .460 ^a | .212 | .204 | 13.453 |
| 2 | .540 ^b | .291 | .278 | 12.812 |

a. Predictors: (Constant), selfest

b. Predictors: (Constant), selfest, severity

c. Dependent Variable: success

SPSS Output 4.6. Model summary for stepwise regression

Because DEPRESSION contributed so little to the regression, the diagnostics are similar to those we have already looked at. Only the table of coefficients, shown in SPSS Output 4.7, has any more to tell us. This gives the coefficients for the new regression model (Model 2) with only SELFEST and SEVERITY as IVs.

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|-------|------------|-----------------------------|------------|---------------------------|--------|------|-------------------------|-------|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | -18.343 | 10.087 | | -1.819 | .072 | 1.000 | 1.000 |
| | selftest | .535 | .099 | .460 | 5.383 | .000 | | |
| 2 | (Constant) | 9.225 | 12.460 | | .740 | .461 | .957 | 1.045 |
| | selftest | .605 | .097 | .520 | 6.249 | .000 | | |
| | severity | -.347 | .100 | -.289 | -3.474 | .001 | | |

a. Dependent Variable: success

SPSS Output 4.7. Coefficients for new regression

Categorical variables

So far we have only used variables on at least interval scales, but you can incorporate categorical variables as independents.

Creating dummy variables

To use a categorical variable in a regression analysis you need to create *dummy variables*. As well as the self esteem, severity and depression data on our patients, we have a record of part of their referral, namely whether they were referred by the GP or the family, or referred themselves. This information gives us a categorical variable with three unordered levels. We can only use this in a regression by forming two dummy variables (one less than the number of categories). The first dummy variable, which we have called GP, is coded 1 for patients referred by their GP and zero otherwise. The second, which we have called FAMILY, is coded 1 for those referred by the family, and zero otherwise. Those referred by neither the GP nor the family, will be coded zero on both these dummy variables. No case can be coded 1 on both dummy variables, since they represent mutually exclusive categories on the original categorical variable. Table 4.2 shows the codings for the two dummy variables and illustrates the application of the codes to three cases from the dataset (bolded in Table 4.1); Case 1 referred neither by the GP nor the family (so a self referral), Case 3 referred by the GP and Case 6 referred by the family. We could have used any two of

the three levels of the original categorical variable to define the two dummy variables, then those cases in the unused level would have been coded zero on both dummies.

For a binary variable, you only need one dummy. If you have several categorical variables each with three or more levels, you can see that you can easily end up with an unwieldy set of dummies.

Table 4.2
Codings for two dummy variables to represent three categorical variables

| dummy variable | GP referral | family referral | neither | case 1 | case 3 | case 6 |
|-----------------------|--------------------|------------------------|----------------|---------------|---------------|---------------|
| 1. GP | 1 | 0 | 0 | 0 | 1 | 0 |
| 2. FAMILY | 0 | 1 | 0 | 0 | 0 | 1 |

Stepwise regression with a categorical variable: requesting the analysis in SPSS

Now we will try our regression again, this time including the dummy variables that code for referral. They can be added to the **Independent(s)** box using the arrow in exactly the same way as the other independents (see SPSS Dialog Box 4.1). Before starting the stepwise regression it is good to perform a multiple regression with all candidate predictor variables, and check that the collinearity statistics give no cause for concern, since collinearity may affect the stages in the process. If we do that for our data we find that the lowest Tolerance is 0.76. That tells us that, for each independent variable, the proportion of variance not accounted for by other independent variables in the equation is high enough for us to continue with our stepwise regression. Make sure this is selected in the **Method** box.

Stepwise regression with a categorical variable: understanding the output

The model summary is shown in SPSS Output 4.8. You can see that SELFEST was entered, then GP, then SEVERITY. None was removed and no more variables were

entered. The resulting model has an R^2 of 0.394, so just under 40% of the variance of SUCCESS is explained by SELFEST, GP and SEVERITY, a little better than the 30% explained by SELFEST and SEVERITY (see SPSS Output 4.6).

Variables Entered/Removed^a

| Mo... | Variables Entered | Variables Removed | Method |
|-------|-------------------|-------------------|---|
| 1 | selftest | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |
| 2 | GP | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |
| 3 | severity | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |

a. Dependent Variable: success

Model Summary^d

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .460 ^a | .212 | .204 | 13.453 |
| 2 | .595 ^b | .354 | .342 | 12.236 |
| 3 | .627 ^c | .394 | .377 | 11.907 |

a. Predictors: (Constant), selftest

b. Predictors: (Constant), selftest, GP

c. Predictors: (Constant), selftest, GP, severity

d. Dependent Variable: success

SPSS Output 4.8. Stepwise regression summary with education variables included

The ANOVA shown in SPSS Output 4.9 tells us that the final regression model (Model 3) was significant at the at the 0.1% level. The regression coefficients are also shown in SPSS Output 4.9. You can also see there that the collinearity check gives no cause for concern, the tolerance for all variables included is close to one. From the B column we can see that an increase of 1 in the SELFEST score increases the estimated SUCCESS score by 0.52. Being coded 1 for GP (which means a GP referral) *reduces* the estimated SUCCESS score by 10.19 (the coefficient is negative). This means that on this measure of success, patients referred by the GP do worse. An increase of 1 in the SEVERITY score also decreases the estimated SUCCESS score (by 0.25). Judging by these results, the patients likely to do best are those with high SELFEST scores, not referred by the GP and low SEVERITY scores.

ANOVA^d

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|-----|-------------|--------|-------------------|
| 1 | Regression | 5243.584 | 1 | 5243.584 | 28.973 | .000 ^a |
| | Residual | 19545.907 | 108 | 180.981 | | |
| | Total | 24789.491 | 109 | | | |
| 2 | Regression | 8768.588 | 2 | 4384.294 | 29.282 | .000 ^b |
| | Residual | 16020.903 | 107 | 149.728 | | |
| | Total | 24789.491 | 109 | | | |
| 3 | Regression | 9760.179 | 3 | 3253.393 | 22.946 | .000 ^c |
| | Residual | 15029.312 | 106 | 141.786 | | |
| | Total | 24789.491 | 109 | | | |

a. Predictors: (Constant), selffest

b. Predictors: (Constant), selffest, GP

c. Predictors: (Constant), selffest, GP, severity

d. Dependent Variable: success

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|-------|------------|-----------------------------|------------|---------------------------|--------|------|-------------------------|-------|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | -18.343 | 10.087 | | -1.819 | .072 | | |
| | selffest | .535 | .099 | .460 | 5.383 | .000 | 1.000 | 1.000 |
| 2 | (Constant) | -5.986 | 9.521 | | -.629 | .531 | | |
| | selffest | .460 | .092 | .395 | 5.012 | .000 | .971 | 1.029 |
| | GP | -11.682 | 2.408 | -.383 | -4.852 | .000 | .971 | 1.029 |
| 3 | (Constant) | 12.495 | 11.605 | | 1.077 | .284 | | |
| | selffest | .520 | .092 | .447 | 5.644 | .000 | .912 | 1.097 |
| | GP | -10.190 | 2.410 | -.334 | -4.228 | .000 | .918 | 1.089 |
| | severity | -.252 | .095 | -.210 | -2.645 | .009 | .905 | 1.105 |

a. Dependent Variable: success

Casewise Diagnostics^a

| Case Nu... | Std. Residual | success | Predicted Value | Residual |
|------------|---------------|---------|-----------------|----------|
| 18 | 2.101 | 52 | 26.98 | 25.018 |
| 67 | -2.563 | 4 | 34.52 | -30.516 |
| 97 | 2.060 | 69 | 44.47 | 24.532 |

a. Dependent Variable: success

SPSS Output 4.9. ANOVA and Regression coefficients with education variables included, and cases with large residuals

The last table in SPSS Output 4.9 shows that the new model has three cases with standardized residuals greater than 2. However, none is very large and they still comprise less than 3% of 110 cases. The residual plot and normal probability plot (not shown here) are very similar to those we obtained with the previous model, and we conclude that the model with SUCCESS predicted by SELFEST, GP and SEVERITY is somewhat better than the previous one with SELFEST and SEVERITY.

Estimating the success of predicting new cases

The example we have been considering is typical of the way an attempt to predict scores on a DV using regression often turns out. Our regression is significant but it only explains about 40% of the variance of the variable we want to predict. As in our case here, this may give a useful general indication for action: the best prognosis for success of CBT arises when the patient scores relatively high on self esteem, when the referral is not initiated by a GP and when the self-reported severity of the illness is relatively low. However, if we were hoping to get good predictions of actual SUCCESS scores for individual cases we shall probably be disappointed. For this we probably need a regression with about two thirds of the variance accounted for, even more if we hope for fairly precise results. Recall that for a standard Normal distribution, 95% of values lie within 1.96 of the mean (zero). So for values with an approximately Normal distribution, 95% of values lie within about 2 standard deviations of the mean. Now look back to SPSS Output 4.3 and see that the standard deviation of the residuals is 12.599, so about 95% of predicted values will be within about 2×12.599 , or about 25, of the actual value. The full range of observed success values is only 67 (from 3 to 70), so although using the regression as a predictor is certainly better than just guessing (the fact that it is significant tells us that) it may not be of much practical use. You could not, for instance, confidently use it as a way to distribute treatment resources among particular cases.

In addition to the common problem discussed above, prediction on new cases will be a little less successful than for those used to calculate the regression coefficients.

Every dataset will have some random variation, and the calculated regression coefficients are optimal for this dataset, including its random components. New data

will have different random components and the regression coefficients will be slightly less than optimal. There are two ways to estimate how well your regression will predict for new cases, both used in practice.

Checking prediction success: waiting for actual DV values

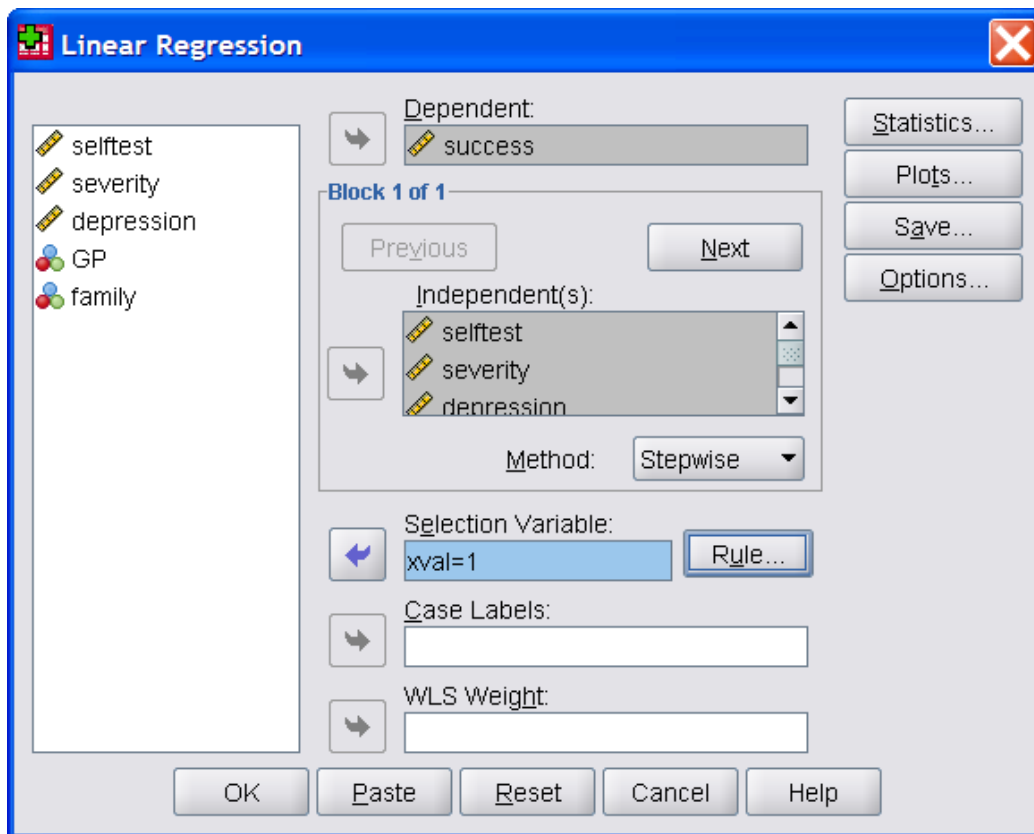
One way is to predict for new cases using the regression equation, and then to wait until you can get the actual DV values for these new cases, and calculate the standard deviation of the residuals (the lower the standard deviation, the better - the closer cases will be to the multiple regression prediction). This approach is only a problem if your prediction causes you to make some decision, perhaps about treatment, that makes it impossible to get a realistic actual value later because the decision you made will have influenced the result (e.g., those predicted to have a low SUCCESS score are given extra counselling in the hope of boosting their self esteem). This is sometimes referred to as *criterion contamination*.

Checking prediction success: using training and validation sets

Another approach is to randomly assign your cases to two datasets. The first, called the *training set*, is used to calculate the regression. The second is called the *validation set*. The predicted score is calculated for all the cases in the validation set, but as we already have their actual scores, we can find the residuals and their standard deviation. It's quite tedious to do this with a calculator but SPSS will do it easily if we use the **Selection Variable** box in the first regression dialog box.

First, we need an extra variable to indicate which cases are in the training set and which in the validation set. We could call this variable XVAL, for cross-validation, and code 1 for the training set and 2 for the validation set. It is best to assign cases

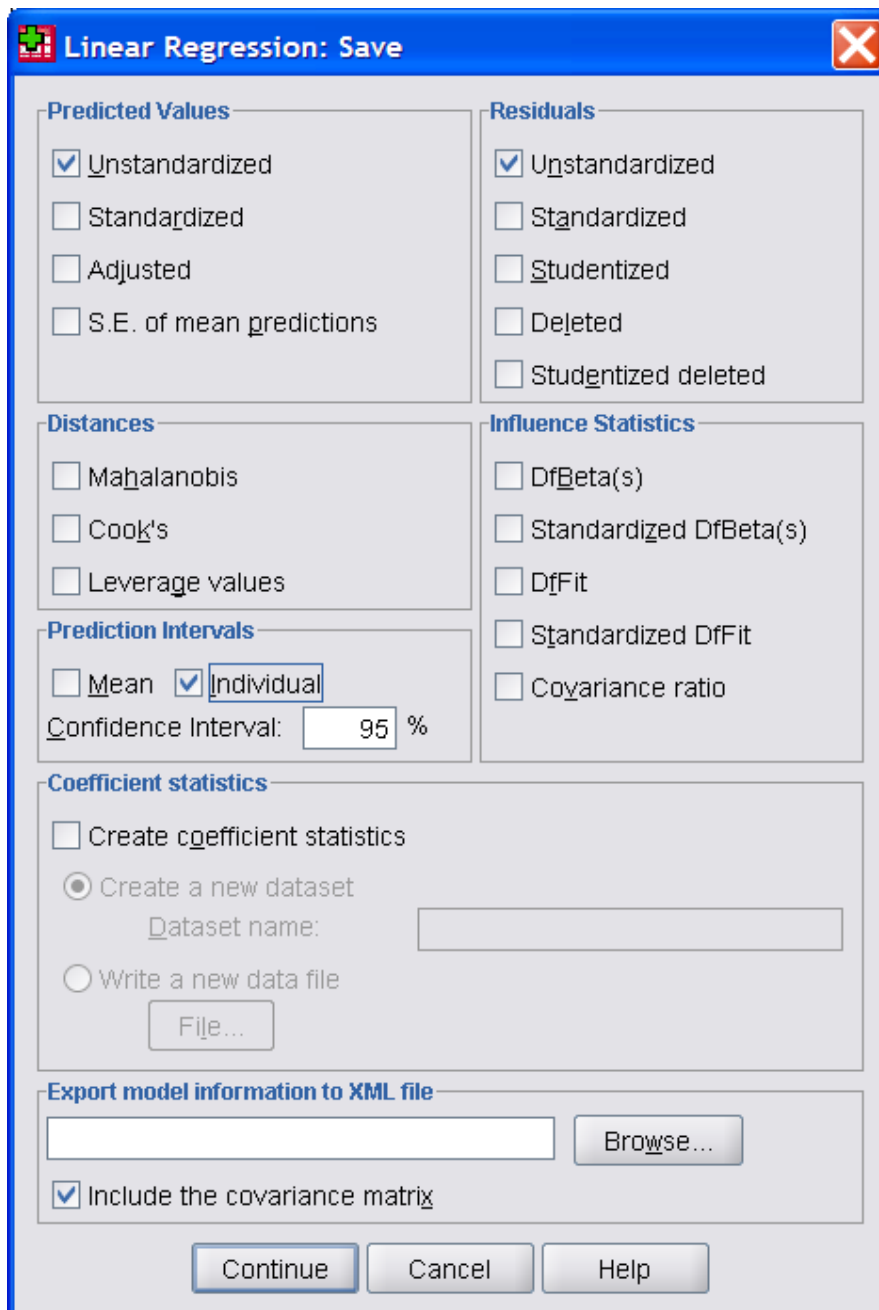
randomly to the two sets, but to make it easy for you to reproduce our results, we have assigned alternate cases, starting with the first, to the training set, so our extra variable is a column of alternating 1s and 2s. The dataset with this added column can be found on the website as med.regression.crossvalidation.sav. In SPSS Dialog Box 4.1, select the new variable XVAL and use the arrow to put it in the **Selection Variable** box. Then click the **Rule** button and enter **equal to 1** in the box. Now the regression will be calculated using only those cases for which $XVAL = 1$, our training set. The dialog box now looks like SPSS Dialog Box 4.4. Leave all the IVs in the **Independent(s)** box and leave the **Method** as **Stepwise**.



SPSS Dialog Box 4.4. Using the variable XVAL to select validation set cases for the regression

Even though the regression will be calculated using only those cases selected by the **Selection Variable**, we can use the **Save** button to keep the values predicted by the

regression for all of the data in the datasheet, the training set ($XVAL = 1$) and the validation set ($XVAL = 2$). Click the button and get the dialog box shown in SPSS Dialog Box 4.5. Save the **Unstandardized Predicted Values** and their **Residuals**, so that we have these in the datasheet as two new columns. You may be interested in the 95% confidence intervals for the predicted values and, if so, tick **Individual** in the **Prediction Intervals** box as shown. Click **Continue** and **OK**.



The image shows the 'Linear Regression: Save' dialog box in SPSS. The dialog box is titled 'Linear Regression: Save' and has a close button (X) in the top right corner. It is divided into several sections:

- Predicted Values:** Contains four checkboxes: Unstandardized, Standardized, Adjusted, and S.E. of mean predictions.
- Residuals:** Contains five checkboxes: Unstandardized, Standardized, Studentized, Deleted, and Studentized deleted.
- Distances:** Contains three checkboxes: Mahalanobis, Cook's, and Leverage values.
- Prediction Intervals:** Contains two checkboxes: Mean and Individual. Below these is a text field for 'Confidence Interval:' with the value '95' and a '%' symbol.
- Influence Statistics:** Contains five checkboxes: DfBeta(s), Standardized DfBeta(s), DfFit, Standardized DfFit, and Covariance ratio.
- Coefficient statistics:** Contains three radio buttons: Create coefficient statistics, Create a new dataset, and Write a new data file. Below the radio buttons is a text field for 'Dataset name:' and a 'File...' button.
- Export model information to XML file:** Contains a text field for the file name and a 'Browse...' button. Below this is a checkbox Include the covariance matrix.

At the bottom of the dialog box are three buttons: 'Continue', 'Cancel', and 'Help'.

SPSS Dialog Box 4.5. Saving predicted values

Training and validation sets: completing the analysis and understanding the output

In addition to the output already described which now refers only to the 55 cases in the training set, we have extra columns in the datasheet giving predicted values, residuals and confidence intervals for all of the cases in both datasets. You may want to sort the list on the XVAL variable and look to see how many of the validation cases have a SUCCESS score inside the 95% confidence interval for the predicted value (select **Data** from the top menu bar, then **Sort Cases**). (If you assign alternate cases to the training and validation sets as we did, you should find all but two of the validation cases inside the 95% confidence interval. The training set also has two cases falling outside the confidence interval). A simple way to compare results on the training and validation sets is to look at the standard deviations of the residuals. Choose **Analyze**, then **Compare means**, then **Means** from the menu bar. Put UNSTANDARDIZED RESIDUAL in the **Dependents List** and XVAL in the **Independents List**. The result is the table shown in SPSS Output 4.10. You can see that the standard deviation of the residuals for the validation set (13.04) is about 10% larger than that for the training set (11.87), and this suggests that prediction for new cases will be correspondingly less precise. After splitting your dataset like this to estimate precision in future prediction, it is best to use as the predictor the regression calculated on the full dataset.

Report

| Unstandardized Residual | | | |
|-------------------------|------------|-----|----------------|
| xval | Mean | N | Std. Deviation |
| 1 | .0000000 | 55 | 11.86713931 |
| 2 | -1.6144827 | 55 | 13.04430323 |
| Total | -.8072413 | 110 | 12.43875039 |

SPSS Output 4.10. Comparison of standard deviations of residuals for training (1) and validation (2) sets

Predicting for new cases

The Selection Variable box can also be used to calculate predicted values for any new cases. Just add the new cases to the bottom of the dataset already in the data window.

Of course their SUCCESS scores will be missing. You need a variable to select the original cases for the regression. You could make a new variable called CHOOSE, and code all the original cases as 1, and the new cases as 2. Then put CHOOSE = 1 in the **Selection Variable** box. Don't forget to use the **Save** button to save the predicted values for all cases, including the new ones, to the data window.