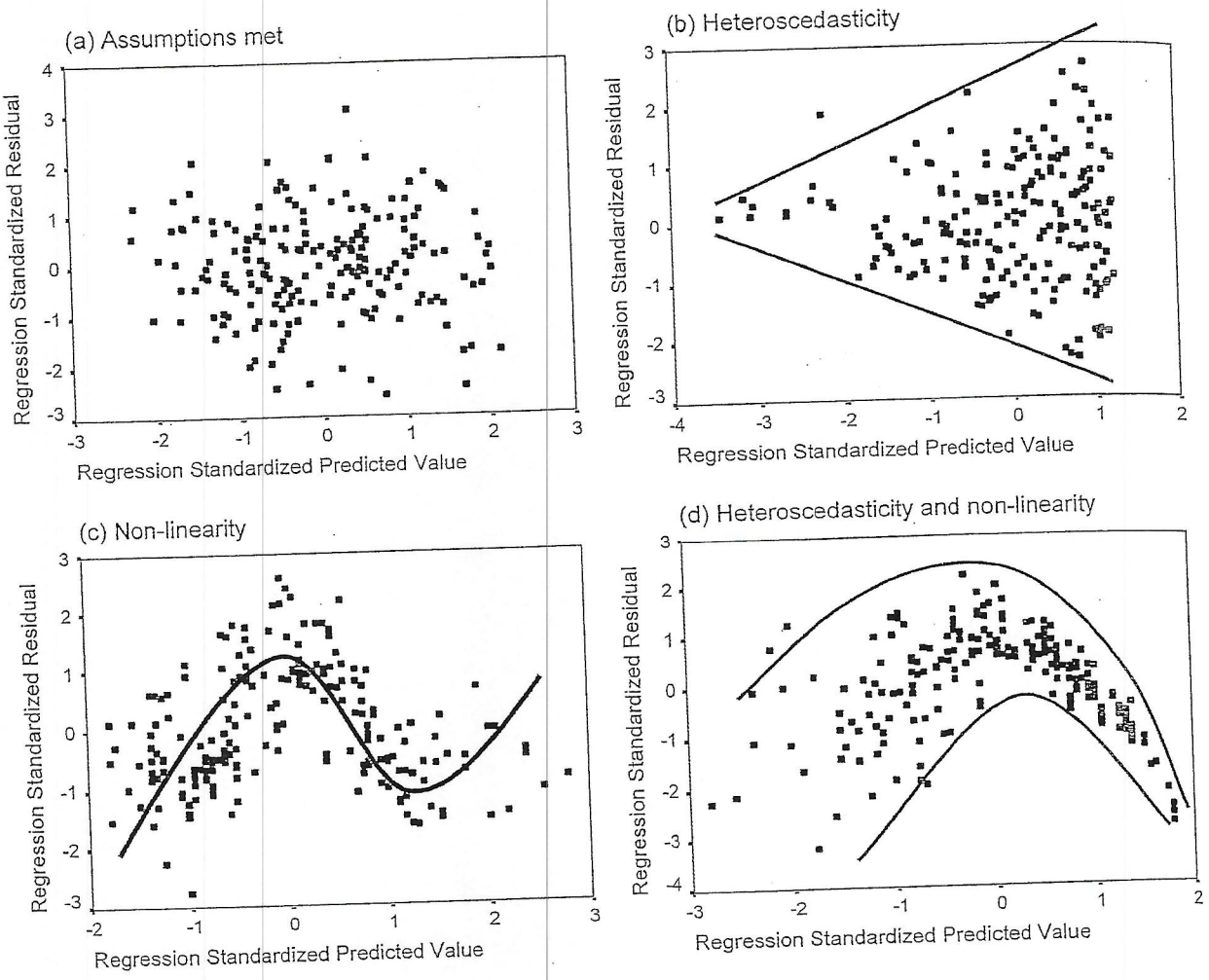


FIGURE 7.19 Plots of *ZRESID against *ZPRED

To test the normality of residuals, we must look at the histogram and normal probability plot selected in Figure 7.15. Figure 7.20 shows the histogram and normal probability plot of the data for the current example (left-hand side). The histogram should look like a normal distribution (a bell-shaped curve). SPSS draws a curve on the histogram to show the shape of the distribution. For the record company data, the distribution is roughly normal (although there is a slight deficiency of residuals exactly on zero). Compare this histogram to the extremely non-normal histogram next to it and it should be clear that the non-normal distribution is extremely skewed (unsymmetrical). So, you should look for a curve that has the same shape as the one for the record sales data: any deviation from this curve is a sign of non-normality and the greater the deviation, the more non-normally distributed the residuals. The normal probability plot also shows up deviations from normality (see Chapter 5). The straight line in this plot represents a normal distribution, and the points represent the observed residuals. Therefore, in a perfectly normally distributed data set, all points will lie on the line. This is pretty much what we see for the record sales data. However, next to the normal probability plot of the record sales data is an example of an extreme deviation from normality. In this plot, the dots

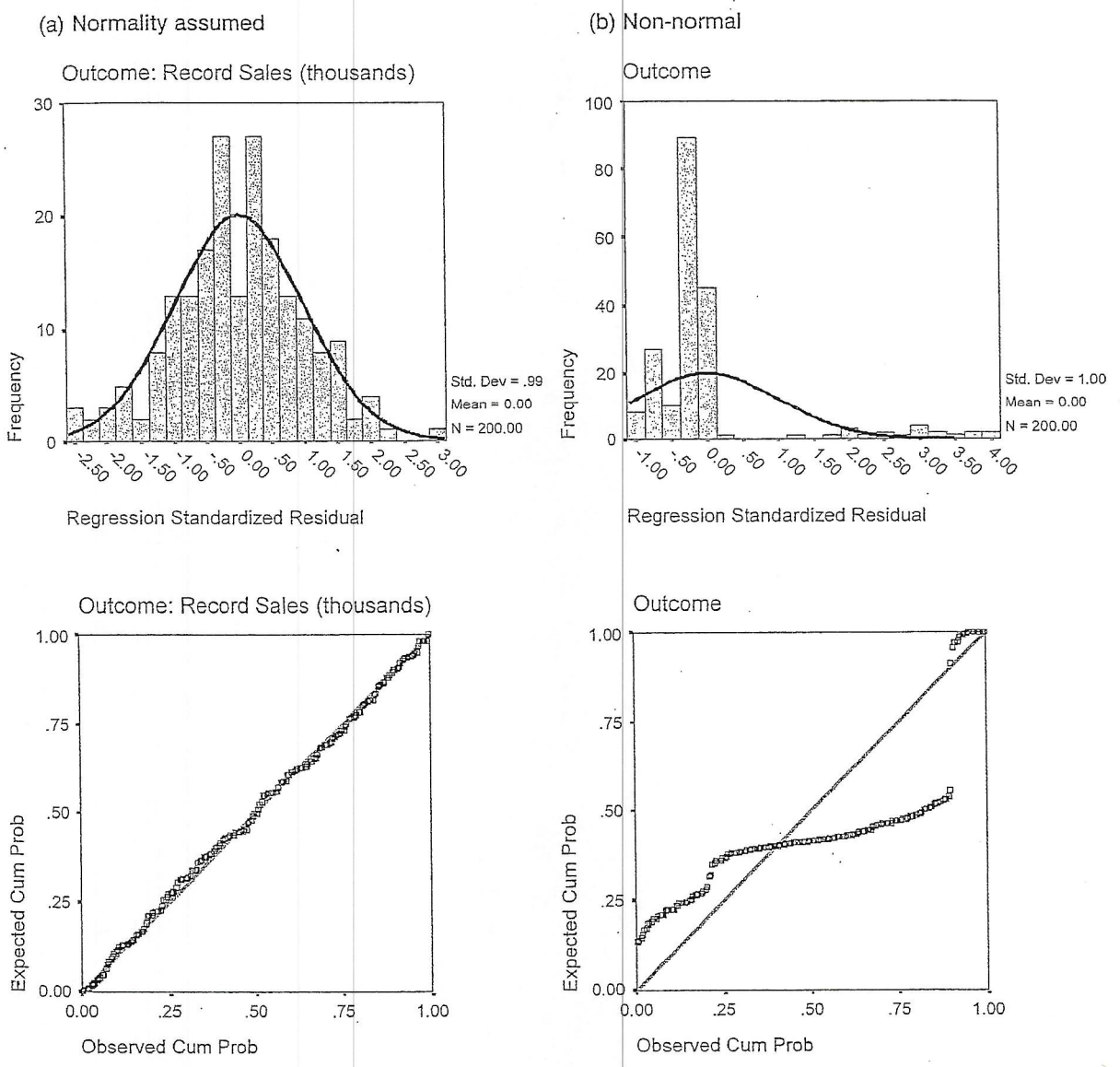
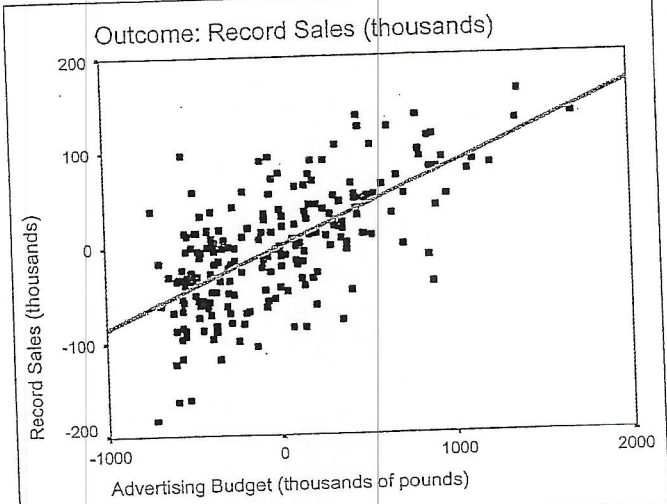


FIGURE 7.20 Histograms and normal P–P plots of normally distributed residuals (left-hand side) and non-normally distributed residuals (right-hand side)

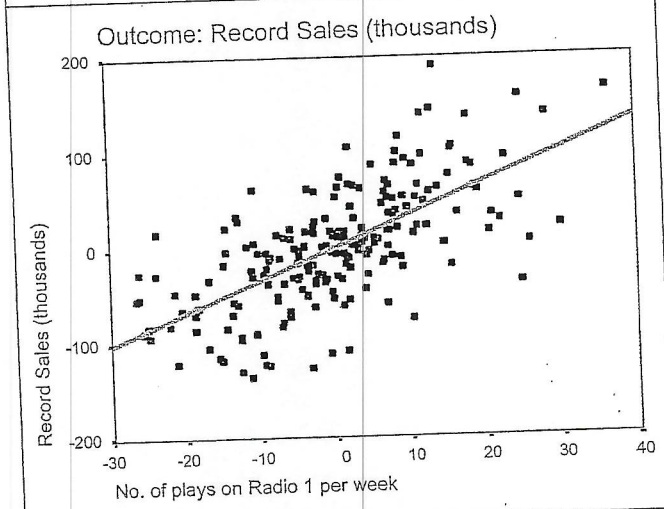
are very distant from the line, which indicates a large deviation from normality. For both plots, the non-normal data are extreme cases and you should be aware that the deviations from normality are likely to be subtler. Of course, you can use what you learnt in Chapter 5 to do a K–S test on the standardized residuals to see whether they deviate significantly from normality.

A final set of plots specified in Figure 7.15 was the partial plots. These plots are scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. I mentioned earlier that obvious outliers on a partial plot represent cases that might have undue influence on a

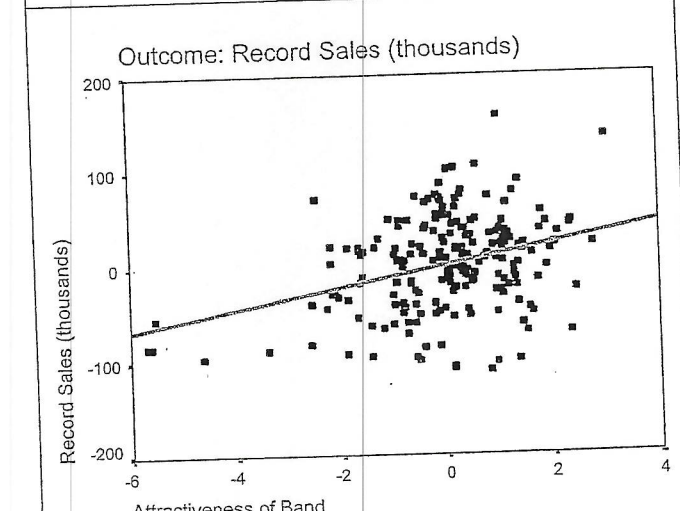
predictor's regression coefficient and that non-linear relationships and heteroscedasticity can be detected using these plots as well:



For advertising budget the partial plot shows the strong positive relationship to record sales. The gradient of the line is *b* for advertising in the model (this line does not appear by default). There are no obvious outliers on this plot, and the cloud of dots is evenly spaced out around the line, indicating homoscedasticity.



For airplay the partial plot shows a strong positive relationship to record sales. The gradient and pattern of the data are similar to advertising (which would be expected given the similarity of the standardized betas of these predictors). There are no obvious outliers on this plot, and the cloud of dots is evenly spaced around the line, indicating homoscedasticity.



For attractiveness the partial plot shows a positive relationship to record sales. The relationship looks less linear than the other predictors do and the dots seem to funnel out, indicating greater variance at high levels of attractiveness. There are no obvious outliers on this plot, but the funnel-shaped cloud of dots might indicate a violation of the assumption of homoscedasticity. We would be well advised to collect some more data for unattractive bands to verify the current model.

OLIVER TWISTED

Please, Sir, can I have some more ... robust regression?

'R you going to show us how to do robust regression?,' mumbles Oliver as he dribbles down his shirt 'I want to do one.' I've prepared a flash movie on the companion website that shows you how to use the *R* plugin to do robust regression. You have no excuses now for not using it.

7.10. How to report multiple regression ②

If you follow the American Psychological Association guidelines for reporting multiple regression then the implication seems to be that tabulated results are the way forward. The APA also seem in favour of reporting, as a bare minimum, the standardized betas, their significance value and some general statistics about the model (such as the R^2). If you do decide to do a table then the beta values and their standard errors are also very useful. Personally I'd like to see the constant as well because then readers of your work can construct the full regression model if they need to. Also, if you've done a hierarchical regression you should report these values at each stage of the hierarchy. So, basically, you want to reproduce the table labelled Coefficients from the SPSS output and omit some of the non-essential information. For the example in this chapter we might produce a table like that in Table 7.2.

See if you can look back through the SPSS output in this chapter and work out from where the values came. Things to note are: (1) I've rounded off to 2 decimal places throughout; (2) for the standardized betas there is no zero before the decimal point (because these values cannot exceed 1) but for all other values less than 1 the zero is present; (3) the significance of the variable is denoted by an asterisk with a footnote to indicate the significance level being used (if there's more than one level of significance being used you can denote this with multiple asterisks, such as $*p < .05$, $**p < .01$, and $***p < .001$); and (4) the R^2 for the initial model and the change in R^2 (denoted as ΔR^2) for each subsequent step of the model are reported below the table.

TABLE 7.2 How to report multiple regression

	B	SE B	
Step 1			
Constant	134.14	7.54	
Advertising Budget	0.10	0.01	.58*
Step 2			
Constant	–26.61	17.35	
Advertising Budget	0.09	0.01	.51*
Plays on BBC Radio 1	3.37	0.28	.51*
Attractiveness	11.09	2.44	.19*

Note: $R^2 = .34$ for Step 1, $\Delta R^2 = .33$ for Step 2 ($p < .001$). * $p < .001$.

CHAMORRO-PREMUZIC, T., ET AL. (2008). PERSONALITY AND INDIVIDUAL DIFFERENCES, 44, 965–976.

LABCOAT LENI'S REAL RESEARCH 7.1

Why do you like your lecturers? ①

In the previous chapter we encountered a study by Chamorro-Premuzic et al. in which they measured students' personality characteristics and asked them to rate how much they wanted these same characteristics in their lecturers (see Labcoat Leni's Real Research 6.1 for a full description). In that chapter we correlated these scores; however, we could go a step further and see whether students' personality characteristics predict the characteristics that they would like to see in their lecturers.

The data from this study are in the file **Chamorro-Premuzic.sav**. Labcoat Leni wants you to carry out five multiple regression analyses: the outcome variable in each of the five analyses is the ratings of how much students want to see Neuroticism, Extroversion, Openness to experience, Agreeableness and Conscientiousness. For each of these outcomes, force Age and Gender into the analysis in the first step of the hierarchy, then in the second block force in the five student personality traits (Neuroticism, Extroversion, Openness to experience, Agreeableness and Conscientiousness). For each analysis create a table of the results.

Answers are in the additional material on the companion website (or look at Table 4 in the original article).

7.11. Categorical predictors and multiple regression ③

Often in regression analysis you'll collect data about groups of people (e.g. ethnic group, gender, socio-economic status, diagnostic category). You might want to include these groups as predictors in the regression model; however, we saw from our assumptions that variables need to be continuous or categorical with only two categories. We saw in section 6.5.5 that a point–biserial correlation is Pearson's r between two variables when one is continuous and the other has two categories coded as 0 and 1. We've also learnt that simple regression is based on Pearson's r, so it shouldn't take a great deal of imagination to see that, like the point–biserial correlation, we could construct a regression model with a predictor that has two categories (e.g. gender). Likewise, it shouldn't be too inconceivable that we could then extend this model to incorporate several predictors that had two categories. All that is important is that we code the two categories with the values of 0 and 1. Why is it important that there are only two categories and that they're coded 0 and 1? Actually, I don't want to get into this here because this chapter is already too long, the publishers are going to break my legs if it gets any longer, and I explain it anyway later in the book (sections 9.7 and 10.2.3) so, for the time being, just trust me!

7.11.1. Dummy coding ③

The obvious problem with wanting to use categorical variables as predictors is that often you'll have more than two categories. For example, if you'd measured religiosity you might have categories of Muslim, Jewish, Hindu, Catholic, Buddhist, Protestant, Jedi (for those of you not in the UK, we had a census here a few years back in which a significant portion of people put down Jedi as their religion). Clearly these groups cannot be distinguished using a single variable coded with zeros and ones. In these cases we have to use

what's called **dummy variables**. Dummy coding is a way of representing groups of people using only zeros and ones. To do it, we have to create several variables; in fact, the number of variables we need is one less than the number of groups we're recoding. There are eight basic steps:

1. Count the number of groups you want to recode and subtract 1.
2. Create as many new variables as the value you calculated in step 1. These are your dummy variables.
3. Choose one of your groups as a baseline (i.e. a group against which all other groups should be compared). This should usually be a control group, or, if you don't have a specific hypothesis, it should be the group that represents the majority of people (because it might be interesting to compare other groups against the majority).
4. Having chosen a baseline group, assign that group values of 0 for all of your dummy variables.
5. For your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
6. For the second dummy variable assign the value 1 to the second group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
7. Repeat this until you run out of dummy variables.
8. Place all of your dummy variables into the regression analysis!

Let's try this out using an example. In Chapter 4 we came across an example in which a biologist was worried about the potential health effects of music festivals. She collected some data at the Download Festival, which is a music festival specializing in heavy metal. The biologist was worried that the findings that she had were a function of the fact that she had tested only one type of person: metal fans. Perhaps it's not the festival that makes people smelly, maybe it's only metal fans at festivals who get smellier (as a metal fan, I would at this point sacrifice the biologist to Satan for being so prejudiced). Anyway, to answer this question she went to another festival that had a more eclectic clientele. So, she went to the Glastonbury Music Festival which attracts all sorts of people because it has all styles of music there. Again, she measured the hygiene of concert-goers over the three days of the festival using a technique that results in a score ranging between 0 (you smell like you've bathed in sewage) and 5 (you smell of freshly baked bread). Now, in Chapters 4 and 5, we just looked at the distribution of scores for the three days of the festival, but now the biologist wanted to look at whether the type of music you like (your cultural group) predicts whether hygiene decreases over the festival. The data are in the file called **GlastonburyFestivalRegression.sav**. This file contains the hygiene scores for each of three days of the festival, but it also contains a variable called **change** which is the change in hygiene over the three days of the festival (so it's the change from day 1 to day 3).[13] Finally, the biologist categorized people according to their musical affiliation: if they mainly liked alternative music she called them 'indie kid', if they mainly liked heavy metal she called them a 'metaller' and if they mainly liked sort of hippy/folky/ambient type of stuff then she labelled them a 'crusty'. Anyone not falling into these categories was labelled 'no musical affiliation'. In the data file she coded these groups 1, 2, 3 and 4 respectively.

The first thing we should do is calculate the number of dummy variables. We have four groups, so there will be three dummy variables (one less than the number of groups). Next we need to choose a baseline group. We're interested in comparing those that have different musical affiliations against those that don't, so our baseline category will be 'no musical

[13] Not everyone could be measured on day 3, so there is a change score only for a subset of the original sample.

TABLE 7.3 Dummy coding for the Glastonbury Festival data

	Dummy Variable 1	Dummy Variable 2	Dummy Variable 3
No Affiliation	0	0	0
Indie Kid	0	0	1
Metaller	0	1	0
Crusty	1	0	0

affiliation'. We give this group a code of 0 for all of our dummy variables. For our first dummy variable, we could look at the 'crusty' group, and to do this we give anyone that was a crusty a code of 1, and everyone else a code of 0. For our second dummy variable, we could look at the 'metaller' group, and to do this we give anyone that was a metaller a code of 1, and everyone else a code of 0. We have one dummy variable left and this will have to look at our final category: 'indie kid'. To do this we give anyone that was an indie kid a code of 1, and everyone else a code of 0. The resulting coding scheme is shown in Table 7.3. The thing to note is that each group has a code of 1 on only one of the dummy variables (except the base category that is always coded as 0).

As I said, we'll look at why dummy coding works in sections 9.7 and 10.2.3, but for the time being let's look at how to recode our grouping variable into these dummy variables using SPSS. To recode variables you need to use the *Recode* function. Select Transform > Recode into Different Variables... to access the dialog box in Figure 7.21.

The recode dialog box lists all of the variables in the data editor, and you need to select the one you want to recode (in this case music) and transfer it to the box labelled *Numeric Variable → Output Variable* by clicking on ▸. You then need to name the new variable (the output variable as SPSS calls it), so go to the part that says *Output Variable* and in the box below where it says *Name* write a name for your first dummy variable (you might call it Crusty). You can also give this variable a more descriptive name by typing something in the box labelled *Label* (for this first dummy variable I've called it No Affiliation vs. Crusty). When you've done this click on Change to transfer this new variable to the box labelled *Numeric Variable → Output Variable* (this box should now say *music → Crusty*).

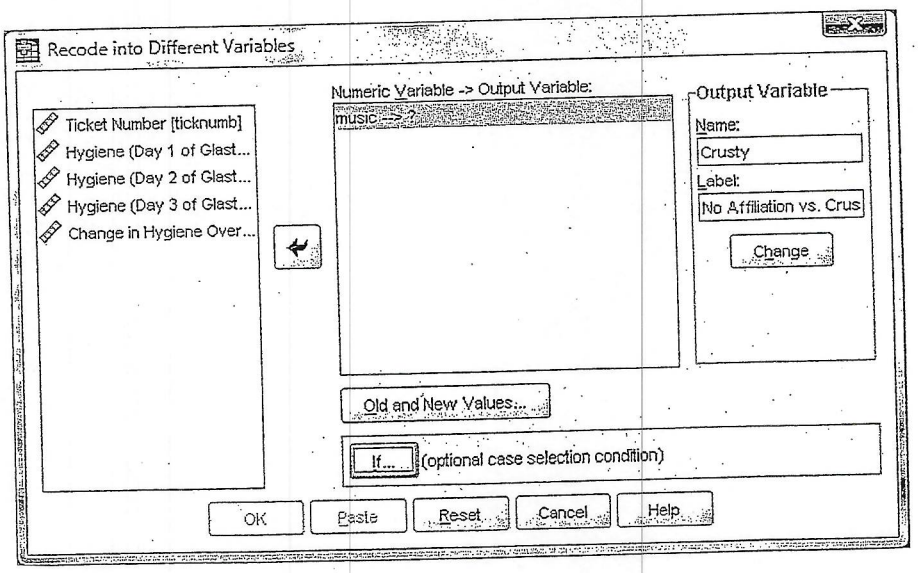


FIGURE 7.21 The recode dialog box

OLIVER TWISTED

Please, Sir, can I have some more ... recode?

'Our data set has missing values,' worries Oliver. 'What do we do if we only want to recode cases for which we have data?' Well, we can set some other options at this point, that's what we can do. This is getting a little bit more involved so if you want to know more, the additional material for this chapter on the companion website will tell you. Stop worrying Oliver, everything will be OK.

Having defined the first dummy variable, we need to tell SPSS how to recode the values of the variable **music** into the values that we want for the new variable, **music1**. To do this click on Old and New Values... to access the dialog box in Figure 7.22. This dialog box is used to change values of the original variable into different values for the new variable. For our first dummy variable, we want anyone who was a crusty to get a code of 1 and everyone else to get a code of 0. Now, crusty was coded with the value 3 in the original variable, so you need to type the value 3 in the section labelled *Old Value* in the box labelled *Value*. The new value we want is 1, so we need to type the value 1 in the section labelled *New Value* in the box labelled *Value*. When you've done this, click on Add to add this change to the list of changes (the list is displayed in the box labelled *Old → New*, which should now say *3 → 1* as in the diagram). The next thing we need to do is to change the remaining groups to have a value of 0 for the first dummy variable. To do this just select All other values and type the value 0 in the section labelled *New Value* in the box labelled *Value*.[14] When you've done this, click on Add to add this change to the list of changes (this list will now also say *ELSE → 0*). When you've done this click on Continue to return to the main dialog box, and then click on OK to create the first dummy variable. This variable will appear as a new column in the data editor, and you should notice that it will have a value of 1 for anyone originally classified as a crusty and a value of 0 for everyone else.

SELF-TEST Try creating the remaining two dummy variables (call them **Metaller** and **Indie_Kid**) using the same principles.

7.11.2 SPSS output for dummy variables ③

Let's assume you've created the three dummy coding variables (if you're stuck there is a data file called **GlastonburyDummy.sav** (the 'Dummy' refers to the fact it has dummy variables in it – I'm not implying that if you need to use this file you're a dummy!)). With dummy variables, you have to enter all related dummy variables in the same block (so use the *Enter* method).

[14] Using this All other values option is fine when you don't have missing values in the data, but just note that when you do (as is the case here) cases with both system-defined and user-defined missing values will be included in the recode. One way around this is to recode only cases for which there is a value (see Oliver Twisted Box). The alternative is specifically to recode missing values using the Range option. It is also a good idea to use the *frequencies* or *crosstabs* commands after a recode and check that you have caught all of these missing values.

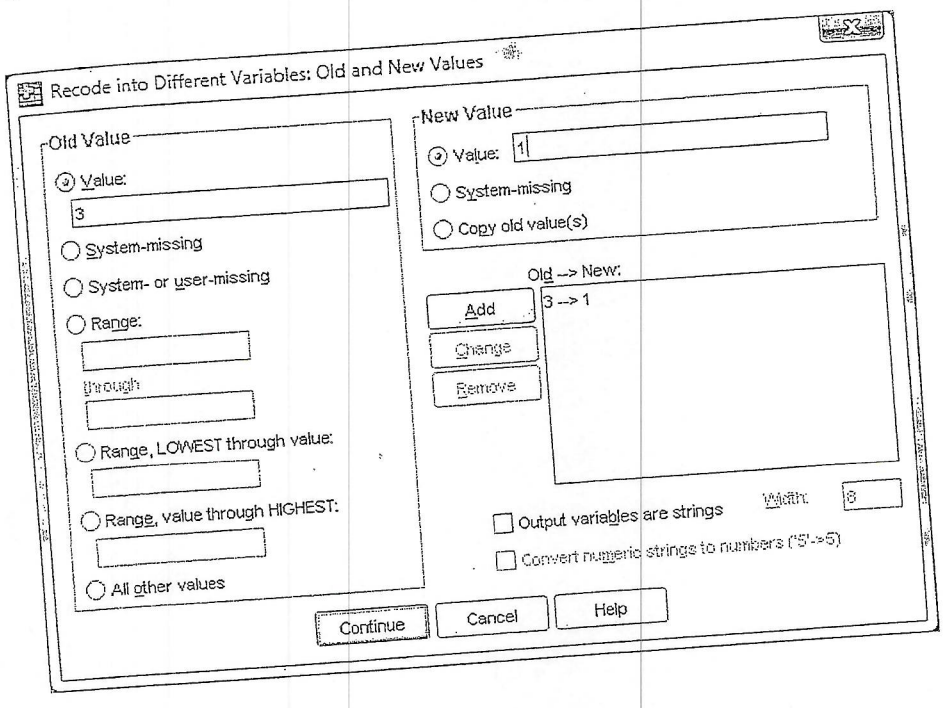


FIGURE 7.22
Dialog boxes for the *Recode* function (see also SPSS Tip 7.1)

SPSS TIP 7.1 Using syntax to recode ③

If you're doing a lot of recoding it soon becomes pretty tedious using the dialog boxes all of the time. I've written the syntax file, **RecodeGlastonburyData.sps**, to create all of the dummy variables we've discussed. Load this file and run the syntax, or open a syntax window (see Section 3.7) and type the following:

```
DO IF  (1-MISSING(change)).
RECODE music (3=1)(ELSE = 0) INTO Crusty.
RECODE music (2=1)(ELSE = 0) INTO Metaller.
RECODE music (1=1)(ELSE = 0) INTO Indie_Kid.
END IF.
VARIABLE LABELS  Crusty 'No Affiliation vs. Crusty'.
VARIABLE LABELS  Metaller 'No Affiliation vs. Metaller'.
VARIABLE LABELS  Indie_Kid 'No Affiliation vs. Indie Kid'.
VARIABLE LEVEL Crusty Metaller Indie_Kid (Nominal).
FORMATS Crusty Metaller Indie_Kid (F1.0).
EXECUTE.
```

Each RECODE command is doing the equivalent of what you'd do using the compute dialog box in Figure 7.22. So, the three lines beginning RECODE ask SPSS to create three new variables (**Crusty, Metaller** and **Indie_Kid**), which are based on the original variable **music**. For the first variable, if **music** is 3 then it becomes 1, and every other value becomes 0. For the second, if **music** is 2 then it becomes 1, and every other value becomes

(Continued)

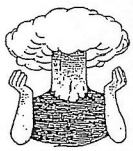
(Continued)

0, and so on for the third dummy variable. Note that all of these RECODE commands are within an IF statement (beginning DO IF and ending with END IF). This tells SPSS to carry out the RECODE commands only if a certain condition is met. The condition we have set is (1- MISSING(change)). MISSING is a built-in command that returns 'true' (i.e. the value 1) for a case that has a system- or user-defined missing value for the specified variable; it returns 'false' (i.e. the value 0) if a case has a value. Hence, MISSING(change) returns a value of 1 for cases that have a missing value for the variable 'change' and 0 for cases that do have values. We want to recode the cases that *do* have a value for the variable change, therefore I have specified '1- MISSING(change)'. This command reverses MISSING(change) so that it returns 1 (true) for cases that have a value for the variable **change** and 0 (false) for system- or user-defined missing values. To sum up, the DO IF (1- MISSING(change)) tells SPSS 'Do the following RECODE commands if the case has a value for the variable **change**.'

The lines beginning VARIABLE LABELS just tells SPSS to assign the text in the quotations as labels for the variables **Crusty**, **Metaller** and **Indie_Kid** respectively. The line beginning VARIABLE LEVEL then sets these three variables to be 'nominal', and the line beginning FORMATS changes the three variables to have a width of 1 and 0 decimal places (hence the 1.0) – in other words, it changes the format to be a binary number.

The final line has the command EXECUTE without which none of the commands beforehand will be executed! Note also that every line ends with a full stop.

So, in this case we have to enter our dummy variables in the same block; however, if we'd had another variable (e.g. socio-economic status) that had been transformed into dummy variables, we could enter these dummy variables in a different block (so, it's only dummy variables that have recoded the same variable that need to be entered in the same block).



SELF-TEST Use what you've learnt in this chapter to run a multiple regression using the change scores as the outcome, and the three dummy variables (entered in the same block) as predictors.

Let's have a look at the output.

SPSS OUTPUT 7.12

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.276[a]	.076	.053	.68818	.076	3.270	3	119	.024	1.893

a. Predictors: (Constant), No Affiliation vs. Indie Kid, No Affiliation vs. Crusty, No Affiliation vs. Metaller

b. Dependent Variable: Change in Hygiene Over The Festival

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.646	3	1.549	3.270	.024[a]
	Residual	56.358	119	.474		
	Total	61.004	122			

a. Predictors: (Constant), No Affiliation vs. Indie Kid, No Affiliation vs. Crusty, No Affiliation vs. Metaller

b. Dependent Variable: Change in Hygiene Over The Festival

SPSS Output 7.12 shows the model statistics. This shows that by entering the three dummy variables we can explain 7.6% of the variance in the change in hygiene scores (the R^2-value × 100). In other words, 7.6% of the variance in the change in hygiene can be explained by the musical affiliation of the person. The ANOVA (which shows the same thing as the R^2 change statistic because there is only one step in this regression) tells us that the model is significantly better at predicting the change in hygiene scores than having no model (or, put another way, the 7.6% of variance that can be explained is a significant amount). Most of this should be clear from what you've read in this chapter already; what's more interesting is how we interpret the individual dummy variables.

SPSS OUTPUT 7.13

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.554	.090		-6.134	.000
	No Affiliation vs. Crusty	-.412	.167	-.232	-2.464	.015
	No Affiliation vs. Metaller	.028	.160	.017	.177	.860
	No Affiliation vs. Indie Kid	-.410	.205	-.185	-2.001	.048

a. Dependent Variable: Change in Hygiene Over The Festival

SPSS Output 7.13 shows a basic Coefficients table for the dummy variables (I've excluded the confidence intervals and collinearity diagnostics). The first thing to notice is that each dummy variable appears in the table with a useful label (such as No Affiliation vs. Crusty). This has happened because when we recoded our variables we gave each variable a label; if we hadn't done this then the table would contain the rather less helpful variable names (crusty, metaller and Indie_Kid). The labels that I suggested giving to each variable give us a hint about what each dummy variable represents. The first dummy variable (No Affiliation vs. Crusty) shows the difference between the change in hygiene scores for the no affiliation group and the crusty group. Remember that the beta value tells us the change in the outcome due to a unit change in the predictor. In this case, a unit change in the predictor is the change from 0 to 1. As such it shows the shift in the change in hygiene scores that results from the dummy variable changing from 0 to 1 (Crusty). By including all three dummy variables at the same time, our baseline category is always zero, so this actually represents the difference in the change in hygiene scores if a person has no musical affiliation, compared to someone who is a crusty. This difference is the difference between the two group means.

To illustrate this fact, I've produced a table (SPSS Output 7.14) of the group means for each of the four groups and also the difference between the means for each group and the no affiliation group. These means represent the average change in hygiene scores for the three groups (i.e. the mean of each group on our outcome variable). If we calculate the difference in these means for the No Affiliation group and the crusty group we get, Crusty – no affiliation = (–0.966) – (–0.554) = –0.412. In other words, the change in hygiene scores is greater for the crusty group than it is for the no affiliation group (crusties' hygiene decreases more over the festival than those with no musical affiliation). This value is the same as the *unstandardized* beta value in SPSS Output 7.13! So, the beta values tell us the relative difference between each group and the group that we chose as a baseline category. This beta value is converted to a *t*-statistic and the significance of this *t* reported. This *t*-statistic is testing, as we've seen before, whether the beta value is 0 and when we have two categories coded with 0 and 1, that means it's testing whether the difference between group means is 0. If it is significant then it means that the group coded with 1 is significantly different from the baseline category – so, it's testing the difference between

two means, which is the context in which students are most familiar with the *t*-statistic (see Chapter 9). For our first dummy variable, the *t*-test is significant, and the beta value has a negative value so we could say that the change in hygiene scores goes down as a person changes from having no affiliation to being a crusty. Bear in mind that a decrease in hygiene scores represents more change (you're becoming smellier) so what this actually means is that hygiene decreased significantly more in crusties compared to those with no musical affiliation!

SPSS OUTPUT 7.14

OLAP Cubes

Variables=Change in Hygiene Over The Festival

Musical Affiliation	Mean	Std. Deviation	N
Indie Kid	-0.964	0.670	14
Metaller	-0.526	0.576	27
Crusty	-0.966	0.760	24
No Musical Affiliation	-0.554	0.708	58
Crusty - No Musical Affiliation	-0.412	0.052	-34
Metaller - No Musical Affiliation	0.028	-0.133	-31
Indie Kid - No Musical Affiliation	-0.410	-0.038	-44
Total	-0.675	0.707	123

Moving on to our next dummy variable, this compares metallers to those that have no musical affiliation. The beta value again represents the shift in the change in hygiene scores if a person has no musical affiliation, compared to someone who is a metaller. If we calculate the difference in the group means for the no affiliation group and the metaller group we get, metaller – no affiliation = (–0.526) – (–0.554) = 0.028. This value is again the same as the unstandardized beta value in SPSS Output 7.13! For this second dummy variable, the *t*-test is not significant, so we could say that the change in hygiene scores is the same if a person changes from having no affiliation to being a metaller. In other words, the change in hygiene scores is not predicted by whether someone is a metaller compared to if they have no musical affiliation.

For the final dummy variable, we're comparing indie kids to those that have no musical affiliation. The beta value again represents the shift in the change in hygiene scores if a person has no musical affiliation, compared to someone who is an indie kid. If we calculate the difference in the group means for the no affiliation group and the indie kid group we get, indie kid – no affiliation = (–0.964) – (–0.554) = –0.410. It should be no surprise to you by now that this is the unstandardized beta value in SPSS Output 7.13! The *t*-test is significant, and the beta value has a negative value so, as with the first dummy variable, we could say that the change in hygiene scores goes down as a person changes from having no affiliation to being an indie kid. Bear in mind that a decrease in hygiene scores represents more change (you're becoming smellier) so what this actually means is that hygiene decreased significantly more in indie kids compared to those with no musical affiliation!

EVERYBODY

So, overall this analysis has shown that compared to having no musical affiliation, crusties and indie kids get significantly smellier across the three days of the festival, but metallers don't. This section has introduced some really complex ideas that I expand upon in Chapters 9 and 10. It might all be a bit much to take in, and so if you're confused or want to know more about why dummy coding works in this way I suggest reading sections 9.7 and 10.2.3 and then coming back here. Alternatively, read Hardy's (1993) excellent monograph!

What have I discovered about statistics? ①

This chapter is possibly the longest book chapter ever written, and if you feel like you aged several years while reading it then, well, you probably have (look around, there are cobwebs in the room, you have a long beard, and when you go outside you'll discover a second ice age has been and gone leaving only you and a few woolly mammoths to populate the planet). However, on the plus side, you now know more or less everything you ever need to know about statistics. Really, it's true; you'll discover in the coming chapters that everything else we discuss is basically a variation on the theme of regression. So, although you may be near death having spent your life reading this chapter (and I'm certainly near death having written it) you are a stats genius – it's official!

We started the chapter by discovering that at 8 years old I could have really done with regression analysis to tell me which variables are important in predicting talent competition success. Unfortunately I didn't have regression, but fortunately I had my dad instead (and he's better than regression). We then looked at how we could use statistical models to make similar predictions by looking at the case of when you have one predictor and one outcome. This allowed us to look at some basic principles such as the equation of a straight line, the method of least squares, and how to assess how well our model fits the data using some important quantities that you'll come across in future chapters: the model sum of squares, SS_M, the residual sum of squares, SS_R, and the total sum of squares, SS_T. We used these values to calculate several important statistics such as R^2 and the F-ratio. We also learnt how to do a regression on SPSS, and how we can plug the resulting beta values into the equation of a straight line to make predictions about our outcome.

Next, we saw that the question of a straight line can be extended to include several predictors and looked at different methods of placing these predictors in the model (hierarchical, forced entry, stepwise). Next, we looked at factors that can affect the accuracy of a model (outliers and influential cases) and ways to identify these factors. We then moved on to look at the assumptions necessary to generalize our model beyond the sample of data we've collected before discovering how to do the analysis on SPSS, and how to interpret the output, create our multiple regression model and test its reliability and generalizability. I finished the chapter by looking at how we can use categorical predictors in regression (and in passing we discovered the *recode* function). In general, multiple regression is a long process and should be done with care and attention to detail. There are a lot of important things to consider and you should approach the analysis in a systematic fashion. I hope this chapter helps you to do that!

So, I was starting to get a taste for the rock-idol lifestyle: I had friends, a fortune (well, two gold-plated winner's medals), fast cars (a bike) and dodgy-looking 8 year olds were giving me suitcases full of lemon sherbet to lick off mirrors. However, my parents and teachers were about to impress reality upon my young mind ...

Key terms that I've discovered

Adjusted predicted value
Adjusted R^2
Autocorrelation
b_i

β_i
Cook's distance
Covariance ratio (CVR)
Cross-validation
Deleted residual
DFBeta
DFFit
Dummy variables
Durbin–Watson test
F-ratio
Generalization
Goodness of fit
Hat values
Heteroscedasticity
Hierarchical regression
Homoscedasticity
Independent errors
Leverage
Mahalanobis distances
Mean squares
Model sum of squares
Multicollinearity
Multiple R
Multiple regression
Outcome variable
Perfect collinearity
Predictor variable
Residual
Residual sum of squares
Shrinkage
Simple regression
Standardized DFBeta
Standardized DFFit
Standardized residuals
Stepwise regression
Studentized deleted residuals
Studentized residuals
Suppressor effects
t-statistic
Tolerance
Total sum of squares
Unstandardized residuals
Variance inflation factor (VIF)

Smart Alex's tasks

- Task 1: A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models. For each model she asked them their salary per day on days when they were working (salary), their age (age), how many years they had worked as a model (years), and then got a panel of experts from modelling agencies to rate the attractiveness of each model as a percentage with 100% being perfectly attractive (beauty). The data are in the file Supermodel.sav. Unfortunately, this fashion student bought some substandard statistics text and so doesn't know how to analyse her data.☺ Can you help her out by conducting a multiple regression to see which variables predict a model's salary? How valid is the regression model? ②

- Task 2: Using the Glastonbury data from this chapter (with the dummy coding in GlastonburyDummy.sav), which you should've already analysed, comment on whether you think the model is reliable and generalizable. ③

- Task 3: A study was carried out to explore the relationship between Aggression and several potential predicting factors in 666 children that had an older sibling. Variables measured were Parenting_Style (high score = bad parenting practices), Computer_Games (high score = more time spent playing computer games), Television (high score = more time spent watching television), Diet (high score = the child has a good diet low in E-numbers), and Sibling_Aggression (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. All other variables were treated in an exploratory fashion. The data are in the file Child Aggression.sav. Analyse them with multiple regression. ②

Answers can be found on the companion website.

Further reading

Bowerman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach* (2nd ed.). Belmont, CA: Duxbury. (This text is only for the mathematically minded or postgraduate students but provides an extremely thorough exposition of regression analysis.)

Hardy, M. A. (1993). *Regression with dummy variables*. Sage university paper series on quantitative applications in the social sciences, 07–093. Newbury Park, CA: Sage.

Howell, D. C. (2006). *Statistical methods for psychology* (6th ed.). Belmont, CA: Duxbury. (Or you might prefer his *Fundamental Statistics for the Behavioral Sciences,* also in its 6th edition, 2007. Both are excellent introductions to the mathematics behind regression analysis.)

Miles, J. N. V., & Shevlin, M. (2001). *Applying regression and correlation: a guide for students and researchers*. London: Sage. (This is an extremely readable text that covers regression in loads of detail but with minimum pain – highly recommended.)

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum. Chapter 3.

Online tutorial

The companion website contains the following Flash movie tutorials to accompany this chapter:

- Regression using SPSS
- Robust Regression

Interesting real research

Chamorro-Premuzic, T., Furnham, A., Christopher, A. N., Garwood, J., & Martin, N. (2008). Birds of a feather: Students' preferences for lecturers' personalities as predicted by their own personality and learning approaches. *Personality and Individual Differences*, *44*, 965–976.