Introducing the Linear Model

What is Correlational Research?

Correlational designs are when many variables are measured simultaneously but unlike in an experiment none of them are manipulated. When we use correlational designs we can't look for cause-effect relationships because we haven't manipulated any of the variables, and also because all of the variables have been measured at the same point in time (if you're really bored, Field, 2013, Chapter 1 explains why experiments allow us to make causal inferences but correlational research does not). In psychology, the most common correlational research consists of the researcher administering several questionnaires that measure different aspects of behaviour to see which aspects of behaviour are related. Many of you will do this sort of research for your final year research project (so pay attention!).

The linear model

In the first few lectures we saw that the only equation we ever really need is this one:

$$\text{outcome}_i = (\text{Model}_i) + \text{error}_i$$

We also saw that we often fit a linear model, which in its simplest form can be written as:

$$\text{outcome}_i = (b_0 + b_1 X_i) + \text{error}_i$$
$$\text{y}_i = (b_0 + b_1 X_i) + \varepsilon_i \qquad \text{Eq. 1}$$

The fundamental idea is that an outcome for an entity can be predicted from a model and some error associated with that prediction (ε_i). We are predicting an outcome variable (y_i) from a predictor variable (X_i) and a parameter, b_1, associated with the predictor variable that quantifies the relationship it has with the outcome variable. We also need a parameter that tells us the value of the outcome when the predictor is zero; this parameter is b_0.

You might recognize this model as 'the equation of a straight line'. I have talked about fitting 'linear models', and linear simply means 'straight line'. Any straight line can be defined by two things: (1) the slope (or gradient) of the line (usually denoted by b_1); and (2) the point at which the line crosses the vertical axis of the graph (known as the *intercept* of the line, b_0). These parameters b_1 and b_0 are known as the **regression coefficients** and will crop up time and time again, where you may see them referred to generally as b (without any subscript) or b_n (meaning the b associated with variable n). A particular line (i.e., model) will have has a specific intercept and gradient.

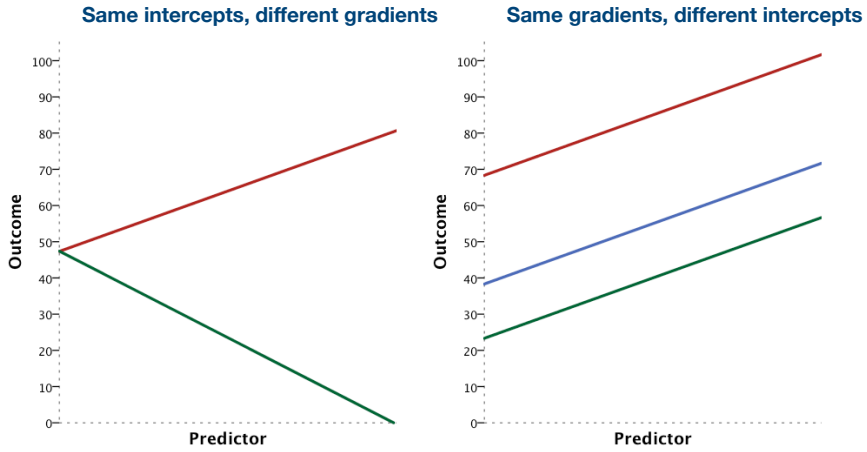


Figure 1: Shows lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

Figure 1 shows a set of lines that have the same intercept but different gradients. For these three models, b_0 will be the same in each but the values of b_1 will differ in each; this figure also shows models that have the same gradients (b_1 is the same in each model) but different intercepts (the b_0 is different in each model). I've mentioned already that b_1 quantifies the relationship between the predictor variable and the outcome, and Figure 1 illustrates this point. A model with a positive b_1 describes a positive relationship, whereas a line with a negative b_1 describes a negative relationship. Looking at Figure 1 (left) the red line describes a positive relationship whereas the green line describes a negative relationship. As such, we can use a linear model (i.e., a straight line) to summarize the relationship between two variables: the gradient (b_1) tells us what the model looks like (its shape) and the intercept (b_0) tells us where the model is (its location in geometric space).

This is all quite abstract so let's look at an example. Imagine that I was interested in predicting physical and downloaded album sales (outcome) from the amount of money spent advertising that album (predictor). We could summarize this relationship using a linear model by replacing the names of our variables into Eq. 1.

$$\begin{aligned} y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{album sales}_i &= b_0 + b_1 \text{advertising budget}_i + \varepsilon_i \end{aligned} \qquad \text{Eq. 2}$$

Once we have estimated the values of the *b*s we would be able to make a prediction about album sales by replacing 'advertising' with a number representing how much we wanted to spend advertising an album. For example, imagine that b_0 turned out to be 50 and b_1 turned out to be 100, our model would be:

$$\text{album sales}_i = 50 + (100 \times \text{advertising budget}_i) + \varepsilon_i \qquad \text{Eq. 3}$$

Note that I have replaced the betas with their numeric values. Now, we can make a prediction. Imagine we wanted to spend £5 on advertising, we can replace the variable 'advertising budget' with this value and solve the equation to discover how many album sales we will get:

$$\begin{aligned} \text{album sales}_i &= 50 + (100 \times 5) + \varepsilon_i \\ &= 550 + \varepsilon_i \end{aligned}$$

So, based on our model we can predict that if we spend £5 on advertising, we'll sell 550 albums. I've left the error term in there to remind you that this prediction will probably not be perfectly accurate. This value of 550 album sales is known as a **predicted value**.

The linear model with several predictors

We have seen that we can use a straight line to 'model' the relationship between two variables. However, life is usually more complicated than that: there are often numerous variables that might be related to the outcome of interest. To take our album sales example, we might expect variables other than simply advertising to have an effect. For example, how much someone hears songs from the album on the radio, or the 'look' of the band might have an influence. One of the beautiful things about the linear model is that it can be expanded to include as many predictors as you like. To add a predictor all we need to do is place it into the model and give it a *b* that estimates the relationship in the population between that predictor and the outcome. For example, if we wanted to add the number of plays of the band on the radio per week (airplay), we could add this second predictor in general as:

$$Y_i = (b_0 + b_1 X_{1i} + b_2 X_{2i}) + \varepsilon_i \qquad \text{Eq. 4}$$

Note that all that has changed is the addition of a second predictor (X_2) and an associated parameter (b_2). To make things more concrete, let's use the variable names instead:

$$\text{album sales}_i = b_0 + b_1 \text{advertising budget}_i + b_2 \text{airplay}_i + \varepsilon_i \qquad \text{Eq. 5}$$

The new model includes a *b*-value for both predictors (and, of course, the constant, b_0). If we estimate the *b*-values, we could make predictions about album sales based not only on the amount spent on advertising but also in terms of radio play.

Multiple regression can be used with three, four or even ten or more predictors. In general we can add as many predictors as we like, and the linear model will expand accordingly:

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i} \cdots b_nX_{ni}) + \varepsilon_i \quad \text{Eq. 6}$$

In which, Y is the outcome variable, b_1 is the coefficient of the first predictor (X_1), b_2 is the coefficient of the second predictor (X_2), b_n is the coefficient of the nth predictor (X_{ni}), and ε_i is the error for the ith participant. (The brackets aren't necessary, they're just to make the connection to Eq. 1). This equation illustrates that we can add in as many predictors as we like until we reach the final one (X_n), but each time we do, we assign it a regression coefficient (b).

Estimating the model

Linear models can be described entirely by a constant (b_0) and by parameters associated with each predictor (bs). These parameters are estimated using the method of least squares (described in your lecture). This method is known as **ordinary least squares (OLS)** regression. In other words, SPSS finds the values of the parameters that have the least amount of error (relative to any other value) for the data you have.

Assessing the goodness of fit, sums of squares, R and R^2

Once Nephwick and Clungglewad have found the model of best fit it is important that we assess how well this model fits the actual data (we assess the goodness of fit of the model). We do this because even though the model is the best one available, it can still be a lousy fit to the data. One measure of the adequacy of a model is the sum of squared differences (think back to lecture 2, or Field, 2013, Chapter 2). There are several sums of squares that can be calculated to help us gauge the contribution of our model to predicting the outcome. Let's go back to our example of predicting album sales (Y) from the amount of money spent advertising that album (X). One day my boss came in to my office and said 'Andy, I know you wanted to be a rock star and you've ended up working as my stats-monkey, but how many albums will we sell if we spend £100,000 on advertising?' In the absence of any data probably the best answer I could give would be the mean number of album sales (say, 200,000) because on average that's how many albums we expect to sell. However, what if he the asks 'How many albums will we sell if we spend £1 on advertising?' Again, in the absence of any accurate information, my best guess would be the mean. There is a problem: whatever amount of money is spent on advertising I always predict the same levels of sales. As such, the mean is a fairly useless model of a relationship between two variables.—but it is the simplest model available.

Using the mean as a model, we can calculate the difference between the observed values, and the values predicted by the mean. We saw in lecture 1 that we square all of these differences to give us the sum of squared differences. This sum of squared differences is known as the **total sum of squares** (denoted SS_T) because it is the total amount of error present when the most basic model is applied to the data (Figure 2). Now, if we fit the more sophisticated model to the data, such as a line of best fit, we can work out the differences between this new model and the observed data. Even if an optimal model is fitted to the data there is still some inaccuracy, which is represented by the differences between each observed data point and the value predicted by the regression line. These differences are squared before they are added up so that the directions of the differences do not cancel out. The result is known as the **sum of squared residuals** (SS_R). This value represents the degree of inaccuracy when the best model is fitted to the data. We can use these two values to calculate how much better the regression line (the line of best fit) is than just using the mean as a model (i.e. how much better is the best possible model than the worst model?). This improvement in prediction is the difference between SS_T and SS_R. This difference shows us the reduction in the inaccuracy of the model resulting from fitting the regression model to the data. This improvement is the **model sum of squares** (SS_M).

If the value of SS_M is large then the regression model is very different from using the mean to predict the outcome variable. This implies that the regression model has made a big improvement to how well the outcome variable can be predicted. However, if SS_M is small then using the regression model is little better than using the mean (i.e. the regression model is no better than taking our 'best guess'). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is easily calculated by dividing the sum of squares for the model by the total sum of squares. The resulting value is called R^2 and to express this value as a percentage you should multiply it by 100. So, R^2 represents the amount of variance in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T).

$$R^2 = \frac{SS_M}{SS_T}$$ Eq. 7

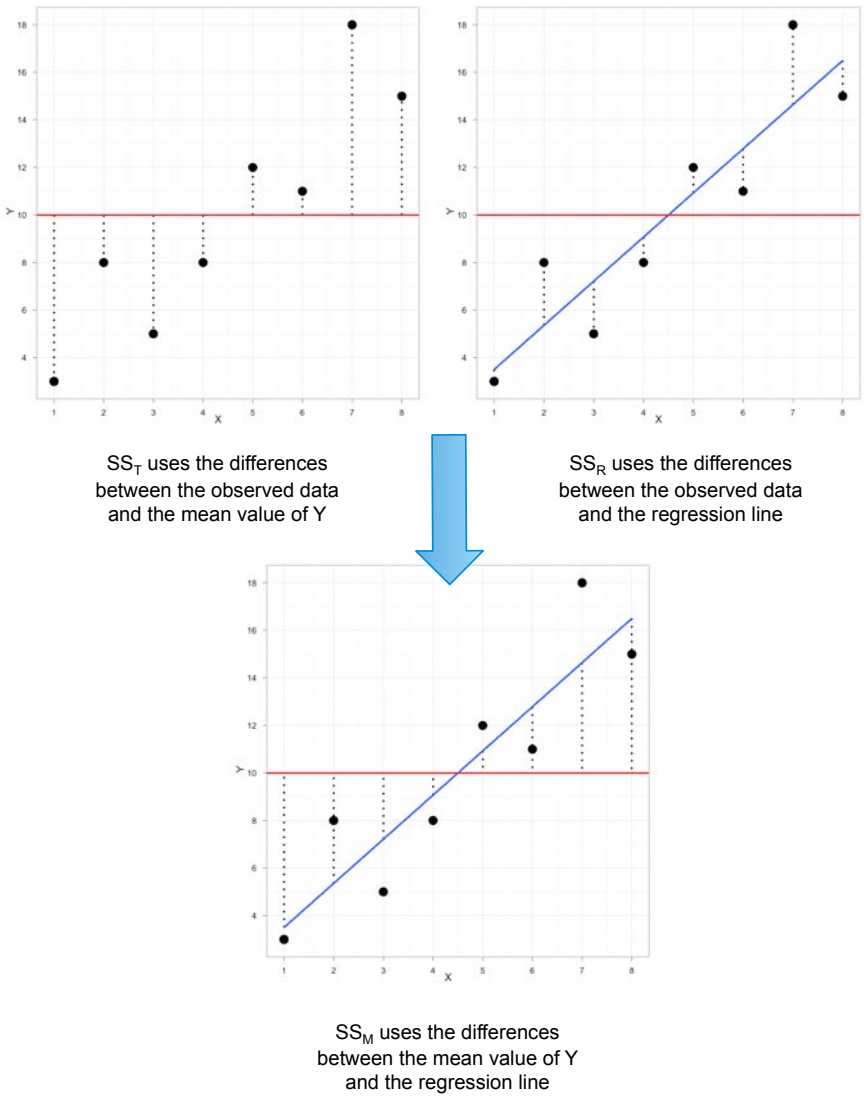


Figure 2: Diagram showing from where the regression sums of squares derive

A second use of the sums of squares in assessing the model is the *F*-test. This test is based upon the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R). Rather than using the sums of squares, it uses the mean sums of squares (referred to as the **mean squares** or MS). The result is the mean squares for the model (MS_M) and the residual mean squares (MS_R) — see Field 2013 for more detail. At this stage it isn't essential that you understand how the mean squares are derived (it is explained in Field, 2013). However, it is important that you understand that the ***F*-ratio** (Eq. 8) is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model. If a model is good, then the improvement in prediction due to the model (MS_M) to be large and the difference between the model and the observed data (MS_R) to be small. In short, a good model should have a large *F*-ratio.

$$F = \frac{MS_M}{MS_R}$$ Eq. 8

Assessing individual predictors

The value of *b* represents the change in the outcome resulting from a unit change in the predictor. If the model is very bad then we would expect the change in the outcome to be zero. A regression coefficient of 0 means: (1) a unit change in the predictor variable results in no change in the predicted value of the outcome (the predicted value of the outcome does not change at all). Logically if a variable significantly predicts an outcome, then it should have a *b*-value that is different from zero. The ***t*-statistic** tests the null hypothesis that the value of *b* is 0: therefore, if it is significant we gain confidence in the hypothesis that the *b*-value is significantly different from 0 and that the predictor variable contributes significantly to our ability to estimate values of the outcome.

Like *F*, the *t*-statistic is based on the ratio of explained variance against unexplained variance or error. What we're interested in here is not so much variance but whether the *b* is big compared to the amount of error in that estimate. To estimate how much error we could expect to find in *b* we use the standard error because it tells us about how different *b*-values would be across different samples. Eq. 9 shows how the *t*-test is calculated: The b_{expected} is the value of *b* that we would expect to obtain if the null hypothesis were true (i.e., zero) and so this value can be replaced by 0. The equation simplifies to become the observed value of *b* divided by the standard error with which it is associated:

$$t = \frac{b_{\text{observed}} - b_{\text{expected}}}{SE_b} = \frac{b_{\text{observed}}}{SE_b} \qquad \text{Eq. 9}$$

The values of *t* have a special distribution that differs according to the degrees of freedom for the test. In this context, the degrees of freedom are *N*–*p*–*1*, where *N* is the total sample size and *p* is the number of predictors. In simple regression when we have only one predictor, this reduces down to *N*–2. SPSS provides the exact probability that the observed value (or a larger one) of *t* would occur if the value of *b* was, in fact, 0. As a general rule, if this observed significance is less than .05, then scientists assume that *b* is significantly different from 0; put another way, the predictor makes a significant contribution to predicting the outcome.

Generalization and Bootstrapping

Remember from your lecture on bias that linear models assume:

- Linearity and additivity: the relationship you're trying to model is, in fact, linear and with several predictors, they combine additively.
- Normality: For *b* estimates to be optimal the residuals should be normally distributed. For CIs and confidence intervals to be accurate, the sampling distribution of *b*s should be normal.
- Homoscedasticity: necessary for *b* estimates to be optimal and significance tests and CIs of the parameters to be accurate.

If these assumptions are met then we can trust the estimates of our *b*s, which means that we can generalize our model (i.e. assume that it works in samples other than the one from which we collected data). If we have concerns about these assumptions we can use bootstrapping to compute robust estimates of *b*s and their confidence intervals. Lack of normality prevents us from knowing the shape of the sampling distribution unless we have big samples. **Bootstrapping** Efron & Tibshirani, 1993) gets around this problem by estimating the properties of the sampling distribution from the sample data. In effect, the sample data are treated as a population from which smaller samples (called bootstrap samples) are taken (putting each score back before a new one is drawn from the sample). The parameter of interest (e.g., the regression parameter) is calculated in each bootstrap sample. This process is repeated perhaps 2000 times. The end result is that we have 2000 parameter estimates, one from each bootstrap sample. There are two things we can do with these estimates: the first is to order them and work out the limits within which 95% of them fall. We can use these values as an estimate of the limits of the 95% confidence interval of the parameter. The result is known as a *percentile bootstrap* confidence interval (because it is based on the values between which 95% of bootstrap sample estimates fall). The second thing we can do is to calculate the standard deviation of the parameter estimates from the bootstrap samples and use it as the standard error of parameter estimates. An important point to remember is that because bootstrapping is based on taking random samples from the data you've collected the estimates you get will be slightly different every time. This is nothing to worry about. For a fairly gentle introduction to the concept of bootstrapping see Wright, London and Field (2011).

Fitting a linear model

Figure 3 shows the general process of conducting regression analysis. First, we should produce scatterplots to get some idea of whether the assumption of linearity is met, and also to look for any outliers or obvious unusual cases. At this stage we might transform the data to correct problems. Having done this initial screen for problems we fit a model and save the various diagnostic statistics that we will discuss next week. If we want to generalize our model beyond the sample, or we are interested in interpreting significance tests and confidence intervals then we examine these residuals to check for homoscedasticity, normality, independence and linearity. If we find problems then we take corrective action and re-estimate the model. Also, it's probably wise to use bootstrapped confidence intervals when we first estimate the model because then we can basically forget about things like normality.

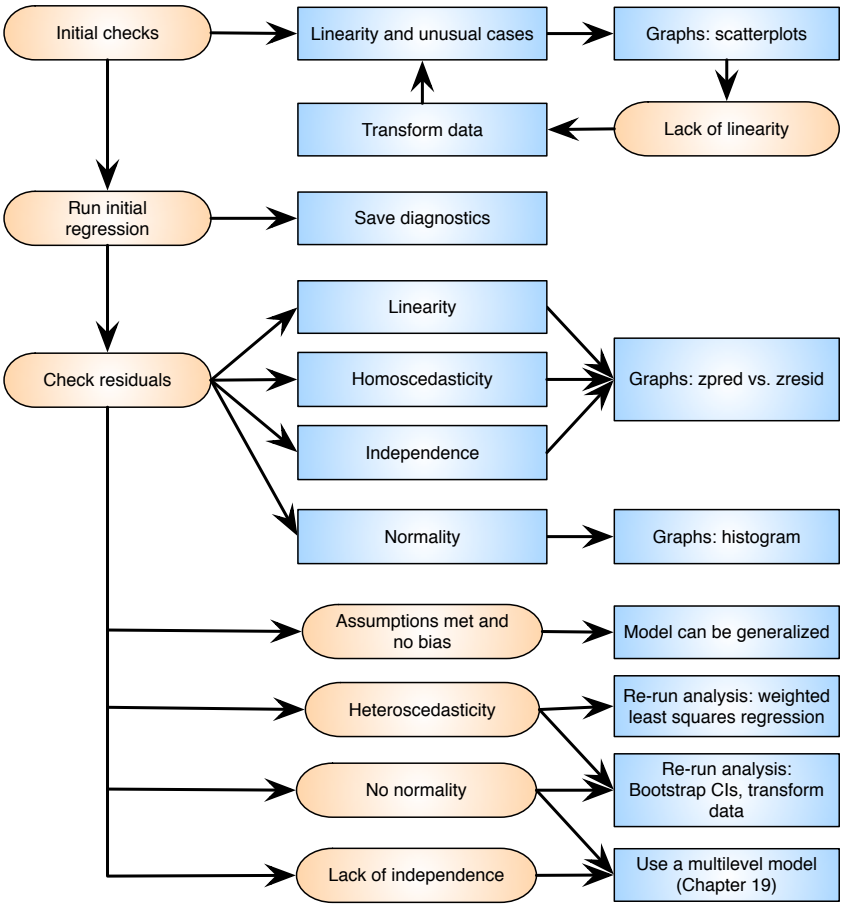


Figure 3: The process of fitting a regression model

Regression using SPSS

There are some data from Field 2013 in the file **Album Sales.sav**. This data file has 200 rows, each one representing a different album. There are also two columns, one representing the sales of each album in the week after release and the other representing the amount (in pounds) spent promoting the album before release. This is the format for entering regression data: the outcome variable and any predictors should be entered in different columns, and each row should represent independent values of those variables.

The pattern of the data is shown in Figure 4 and it should be clear that a positive relationship exists: so, the more money spent advertising the album, the more it is likely to sell. Of course there are some albums that sell well regardless of advertising (top left of scatterplot), but there are none that sell badly when advertising levels are high (bottom right of scatterplot). The scatterplot also shows the line of best fit for these data: bearing in mind that the mean would be represented by a flat line at around the 200,000 sales mark, the regression line is noticeable different.

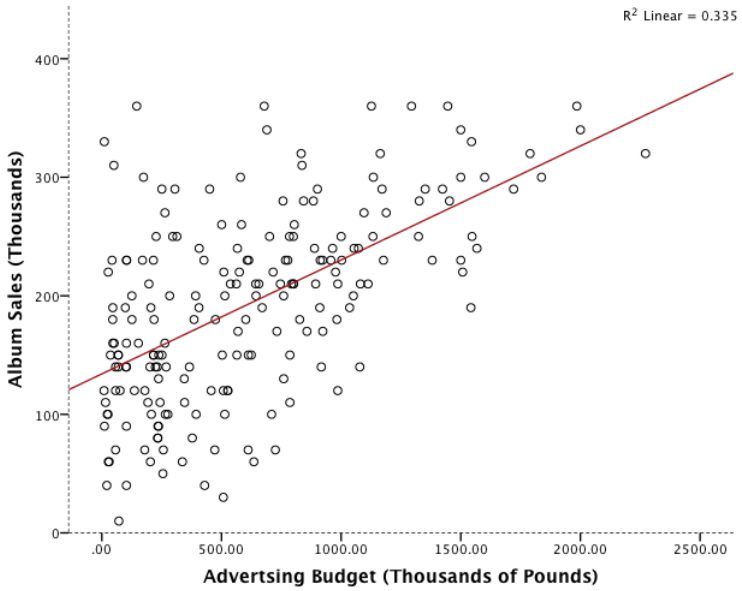


Figure 4: Scatterplot showing the relationship between album sales and the amount spent promoting the album.

Running a basic Analysis

To find out the parameters that describe the regression line, and to see whether this line is a useful model, we need to run a regression analysis. To do the analysis you need to access the main dialog box by selecting Analyze Regression Linear... Figure 4 shows the resulting dialog box. There is a space labelled *Dependent* in which you should place the outcome variable (in this example **sales**). So, select **sales** from the list on the left-hand side, and transfer it by dragging it or clicking on . There is another space labelled *Independent(s)* in which any predictor variable should be placed. In simple regression we use only one predictor (in this example, **adverts**) and so you should select **adverts** from the list and click on to transfer it to the list of predictors. There are a variety of options available, but these will be explored within the context of multiple regression.

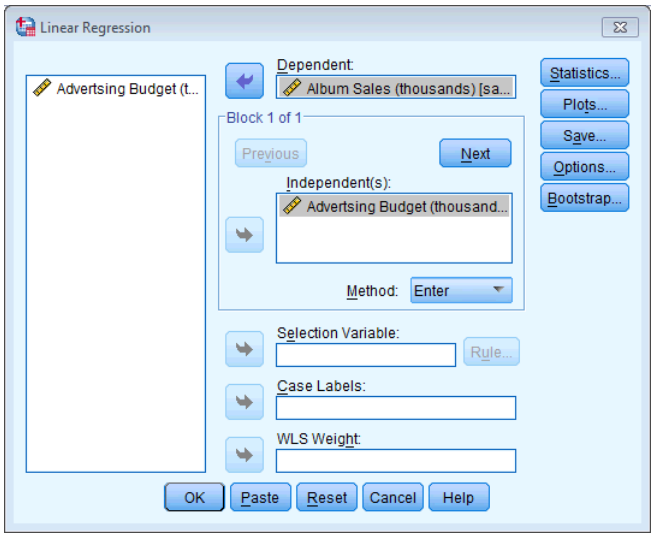


Figure 5: Main dialog box for regression

If we are worried about assumptions then we can get bootstrapped confidence intervals for the regression coefficients by clicking Bootstrap... Select Perform bootstrapping to activate bootstrapping, and to get a 95% confidence interval click Bias corrected accelerated (BCa). Click on OK in the main dialog box to run the basic analysis.

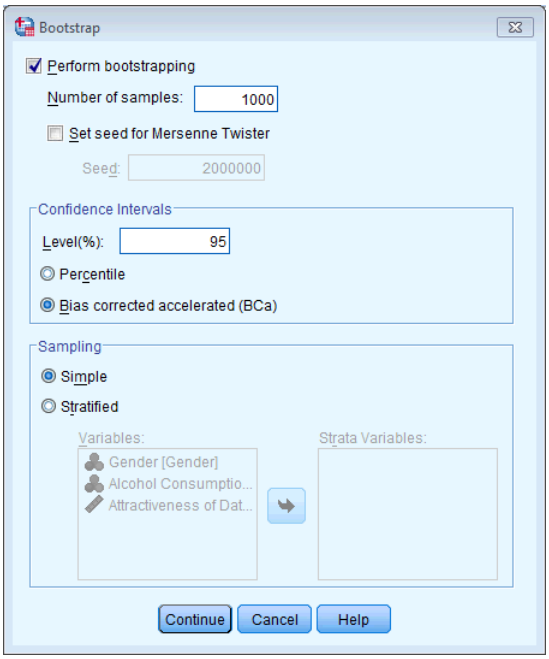


Figure 6: Bootstrap dialog box

Output from SPSS

Overall Fit of the Model

Model Summary

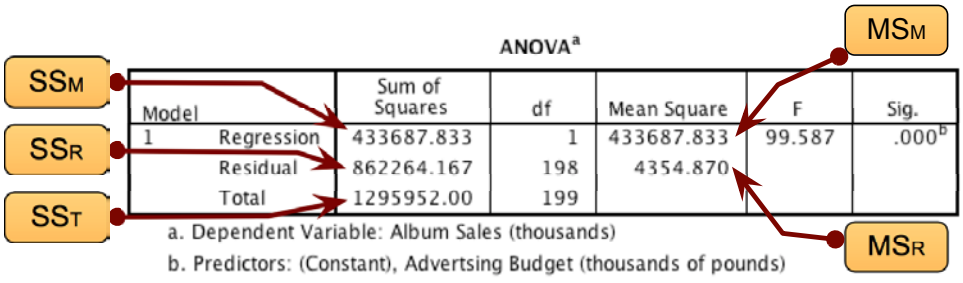
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.991

a. Predictors: (Constant), Advertsing Budget (thousands of pounds)

Output 1

The first table provided by SPSS is a summary of the model that gives the value of R and R^2 for the model. For these data, R is 0.578 and because there is only one predictor, this value represents the simple correlation between advertising and album sales (you can confirm this by running a correlation). The value of R^2 is 0.335, which tells us that advertising expenditure can account for 33.5% of the variation in album sales. There might be many factors that can explain this variation, but our model, which includes only advertising expenditure explains 33%: 66% of the variation in album sales is unexplained. Therefore, there must be other variables that have an influence also

The next part of the output reports an analysis of variance (ANOVA—see Field, 2013, Chapter 11). The most important part of the table is the *F*-ratio, which is calculated using Eq. 8, and the associated significance value. For these data, *F* is 99.59, which is significant at p <.001 (because the value in the column labelled *Sig.* is less than.001). This result tells us that there is less than a 0.1% chance that an *F*-ratio this large would happen if there were no effect. Therefore, we can conclude that our regression model results in significantly better prediction of album sales than if we used the mean value of album sales. In short, the regression model overall predicts album sales significantly well.



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[b]
	Residual	862264.167	198	4354.870		
	Total	1295952.00	199			

a. Dependent Variable: Album Sales (thousands)

b. Predictors: (Constant), Advertsing Budget (thousands of pounds)

Output 2

Model Parameters

The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However, the ANOVA doesn't tell us about the individual contribution of variables in the model (although in this simple case there is only one variable in the model and so we can infer that this variable is a good predictor). The table in Output 3 provides details of the model parameters (the beta values) and the significance of these values. We saw in Eq. 1 that b_0 was the *Y* intercept and this value is the value *B* for the constant. So, from the table, we can say that b_0 is 134.14, and this can be interpreted as meaning that when no money is spent on advertising (when $X = 0$), the model predicts that 134,140 albums will be sold (remember that our unit of measurement was thousands of albums). We can also read off the value of b_1 from the table and this value represents the gradient of the regression line. It is 0.096.[1] Although this value is the slope of the regression line, it is more useful to think of this value as representing *the change in the outcome associated with a unit change in the predictor*. Therefore, if our predictor variable is increased by 1 unit (if the advertising budget is increased by 1), then our model predicts that 0.096 extra albums will be sold. Our units of measurement were thousands of pounds and thousands of albums sold, so we can say that for an increase in advertising of £1000 the model predicts 96 (0.096 × 1000 = 96) extra album sales. As you might imagine, this investment is pretty bad for the record company: they invest £1000 and get only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of album sales.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Album Sales (thousands)

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	134.140	.356	8.214	.001	117.993	151.258
	Advertsing Budget (thousands of pounds)	.096	.000	.009	.001	.080	.113

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 3

We saw earlier that, in general, values of the regression coefficient *b* represent the change in the outcome resulting from a unit change in the predictor and that if a predictor is having a significant impact on our ability to predict the outcome then this *b* should be different from 0 (and big relative to its standard error). We also saw that the *t*-test tells us whether the *b*-value is different from 0. SPSS provides the exact probability that the observed value of *t* would occur if the value of *b* in the population were zero. If this observed significance is less than .05, then the result reflects a genuine effect. For both *t*s, the probabilities are .000 (zero to 3 decimal places) and so we can say that the probability of these *t* values (or larger) occurring if the values of *b* in the population were zero is less than .001. Therefore, the *b*s are significantly different from 0. In the case of the *b* for advertising budget this result means that the advertising budget makes a significant contribution ($p < .001$) to predicting album sales.

The bootstrap confidence interval tells us that the population value of *b* for advertising budget is likely to fall between .08 and .11 and because this interval doesn't include zero we would conclude that there is a genuine positive relationship between advertising budget and album sales in the population. Also, the significance associated with this confidence interval is $p = .001$, which is highly significant. Also, note that the bootstrap process involves re-estimating the standard error (it changes from .01 in the original table to a bootstrap estimate of .009). This is a very small change. For the constant, the standard error is 7.537 compared to the bootstrap estimate of 8.214, which is a difference of

[1] Sometimes small values are reported by SPSS as things like 9.612 E–02 and many students find this notation confusing. Well, think of E–02 as meaning 'move the decimal place 2 steps to the left', so 9.612 E–02 becomes 0.09612.

0.677. The bootstrap confidence intervals and significance values are useful to report and interpret because they do not rely on assumptions of normality or homoscedasticity.

Using the Model

So far, we have discovered that we have a useful model, one that significantly improves our ability to predict album sales. However, the next stage is often to use that model to make some predictions. The first stage is to define the model by replacing the *b*-values in Eq. 2 with the values from the output. In addition, we can replace the *X* and *Y* with the variable names so that the model becomes:

$$\begin{aligned}\text{album sales}_i &= b_0 + b_1\text{advertising budget}_i \\ &= 134.14 + (0.096 \times \text{advertising budget}_i)\end{aligned} \qquad \text{Eq. 10}$$

It is now possible to make a prediction about album sales, by replacing the advertising budget with a value of interest. For example, imagine a recording company executive wanted to spend £100,000 on advertising a new album. Remembering that our units are already in thousands of pounds, we can simply replace the advertising budget with 100. He would discover that album sales should be around 144,000 for the first week of sales:

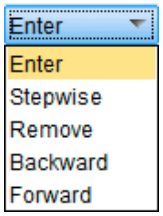
$$\begin{aligned}\text{album sales}_i &= 134.14 + (0.096 \times \text{advertising budget}_i) \\ &= 134.14 + (0.096 \times 100) \\ &= 143.74\end{aligned} \qquad \text{Eq. 11}$$

Regression with several predictors using SPSS

SELF-TEST: Produce a matrix scatterplot of **Sales**, **Adverts**, **Airplay** and **Attract** including the regression line.

Main options

The executive has past research indicating that advertising budget is a significant predictor of album sales, and so he should include this variable in the model first. His new variables (**airplay** and **attract**) should, therefore, be entered into the model *after* advertising budget. This method is hierarchical (the researcher decides in which order to enter variables into the model based on past research). To do a hierarchical regression in SPSS we have to enter the variables in blocks (each block representing one step in the hierarchy). To get to the main *regression* dialog box select Analyze Regression Linear... . To set up the first block we do exactly what we did before. Select the outcome variable (album sales) and drag it to the box labelled *Dependent* (or click on). We also need to specify the predictor variable for the first block. We've decided that advertising budget should be entered into the model first, so select this variable in the list and drag it to the box labelled *Independent(s)* (or click on). Underneath the *Independent(s)* box, there is a drop-down menu for specifying the *Method* of regression. You can select a different method of variable entry for each block by clicking on Enter, next to where it says *Method* (see Figure 5). The default option is forced entry, and this is the option we want, but next week we'll look at other approaches.



Having specified the first block in the hierarchy, we need to move onto to the second. To tell the computer that you want to specify a new block of predictors you must click on Next. This process clears the *Independent(s)* box so that you can enter the new predictors (you should also note that above this box it now reads *Block 2 of 2* indicating that you are in the second block of the two that you have so far specified). We decided that the second block would contain both of the new predictors and so you should click on **Airplay** and **Attract** (while holding down *Ctrl*, or *Cmd* if you use a Mac) in the variables list and drag them to the *Independent(s)* box or click on . The dialog box should now look like Figure 7. To move between blocks use the Previous and Next buttons (so for example, to move back to block 1, click on Previous). We can get bootstrapped confidence intervals for the regression coefficients by clicking Bootstrap... as we did before.

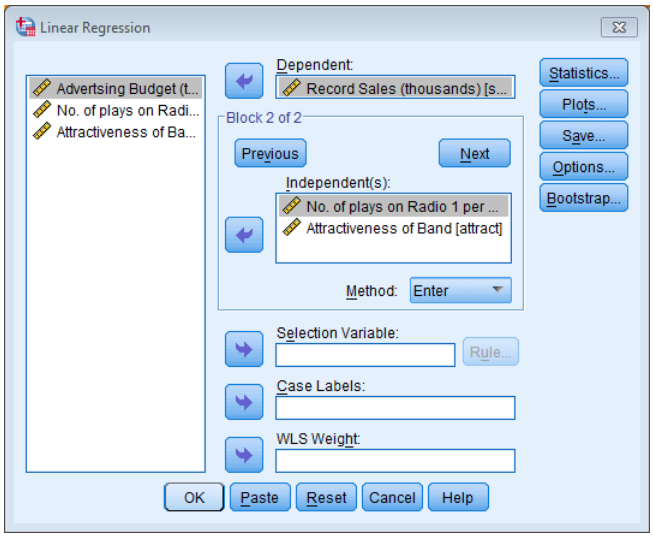


Figure 7: Main dialog box for block 2 of the multiple regression

Output from SPSS

Overall Fit of the Model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.991
2	.815[b]	.665	.660	47.087

a. Predictors: (Constant), Advertsing Budget (Thousands of Pounds)

b. Predictors: (Constant), Advertsing Budget (Thousands of Pounds), Attractiveness of Band, No. of plays on Radio

c. Predictors: (Constant), Advertsing Budget (Thousands of Pounds), No. of plays on Radio, Attractiveness of Band

Output 4

Note that there are two models. Model 1 refers to the first stage in the hierarchy when only advertising budget is used as a predictor. Model 2 refers to when all three predictors are used. Under this table SPSS tells us what the dependent variable (outcome) was and what the predictors were in each of the two models. The column labelled *R* contains the values of the multiple correlation coefficient between the predictors and the outcome. When only advertising budget is used as a predictor, this is the simple correlation between advertising and album sales (0.578). In fact, all of the statistics for model 1 are the same as the regression model earlier (Output 1).

The next column gives us a value of R^2, which we already know is a measure of how much of the variability in the outcome is accounted for by the predictors. For the first model its value is .335, which means that advertising budget accounts for 33.5% of the variation in album sales. However, when the other two predictors are included as well (model 2), this value increases to .665 or 66.5% of the variance in album sales. Therefore, if advertising accounts for 33.5%, we can tell that attractiveness and radio play account for an additional 33%.[2] So, the inclusion of the two new predictors has explained quite a large amount of the variation in album sales

Output 5 contains an ANOVA that tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'. For the initial model the *F*-ratio is 99.59, $p < .001$. For the second model the value of *F* is 129.498, which is also highly significant ($p < .001$). We can interpret these results as meaning that both models significantly improved our ability to predict the outcome variable compared to not fitting the model.

[2] That is, 33% = 66.5% – 33.5%.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[b]
	Residual	862264.167	198	4354.870		
	Total	1295952.00	199			
2	Regression	861377.418	3	287125.806	129.498	.000[c]
	Residual	434574.582	196	2217.217		
	Total	1295952.00	199			

a. Dependent Variable: Album Sales (Thousands)

b. Predictors: (Constant), Advertsing Budget (Thousands of Pounds)

c. Predictors: (Constant), Advertsing Budget (Thousands of Pounds), Attractiveness of Band, No. of plays on Radio

Output 5

Model Parameters

Remember that in multiple regression the model takes the form of Eq. 6 and in that equation there are several unknown parameters (the *b*-values). The first part of Output 6 gives us estimates for these *b*-values and these values indicate the individual contribution of each predictor to the model. We can use these estimates to define our model:

$$\begin{aligned}\text{sales}_i &= b_0 + b_1\text{advertising}_i + b_2\text{airplay}_i + b_3\text{attractiveness}_i \\ &= -26.61 + (0.08\ \text{advertising}_i) + (3.37\ \text{airplay}_i) + (11.09\ \text{attractiveness}_i)\end{aligned} \quad \text{Eq. 12}$$

The *b*-values tell us about the relationship between album sales and each predictor. All three predictors have positive *b*-values indicating positive relationships. So, as advertising budget increases, album sales increase; as plays on the radio increase, so do album sales; and finally more attractive bands will sell more albums. The *b*-values tell us more than this, though. They tell us to what degree each predictor affects the outcome *if the effects of all other predictors are held constant*:

- **Advertising budget** (*b* = 0.085): as advertising budget increases by one unit, album sales increase by 0.085 units.
- **Airplay** (*b* = 3.367): as the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units.
- **Attractiveness** (*b* = 11.086): a band rated one unit higher on the attractiveness scale can expect additional album sales of 11.086 units.

For this model, the advertising budget, $t(196) = 12.26$, $p < .001$, the amount of radio play prior to release, $t(196) = 12.12$, $p < .001$ and attractiveness of the band, $t(196) = 4.55$, $p < .001$, are all significant predictors of album sales.[3] Remember that these significance tests are accurate only if the assumptions discussed in your lecture are met.

If these assumptions aren't met, or you want to ignore them, you could look at the table of bootstrap confidence intervals for each predictor and their significance value[4]. These tell us that advertising, *b* = 0.09 [0.07, 0.10], *p* = .001, airplay, *b* = 3.37 [2.74, 4.02], *p* = .001, and attractiveness of the band, *b* = 11.09 [6.46, 15.01], *p* = .001, all significantly predict album sales. The confidence intervals are constructed such that in 95% of samples the boundaries contain the population value of *b*. Therefore, it's likely that the confidence interval we have constructed for this sample will contain the true value of *b* in the population. Therefore, we can use the confidence intervals to tell us the likely size of the parameter in the population (i.e., the true value). If the confidence interval contains zero then this means that the true value might be zero (i.e., no effect at all) or opposite to what we observed in the sample (e.g., a negative *b* instead of

[3] For all of these predictors I wrote $t(196)$. The number in brackets is the degrees of freedom. In regression the degrees of freedom are $N–p–1$, where N is the total sample size (in this case 200) and p is the number of predictors (in this case 3). For these data we get 200–3–1 = 196.

[4] Remember that because of how bootstrapping works the values in your output will be slightly different to mine, and different if you re-run the analysis.

the positive one that we observed). Therefore, because zero does not fall within the boundaries of any of our bootstrap confidence intervals, we can conclude very confidently that the population values of *b* are positive— in other words, all of the predictor variables are genuine predictors of album sales.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	134.140	7.537		17.799	.000	119.278	149.002
	Advertsing Budget (Thousands of Pounds)	.096	.010	.578	9.979	.000	.077	.115
2	(Constant)	-26.613	17.350		-1.534	.127	-60.830	7.604
	Advertsing Budget (Thousands of Pounds)	.085	.007	.511	12.261	.000	.071	.099
	No. of plays on Radio	3.367	.278	.512	12.123	.000	2.820	3.915
	Attractiveness of Band	11.086	2.438	.192	4.548	.000	6.279	15.894

a. Dependent Variable: Album Sales (Thousands)

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	134.140	-.116	7.952	.001	120.108	148.793
	Advertsing Budget (Thousands of Pounds)	.096	.000	.008	.001	.079	.112
2	(Constant)	-26.613	.489	16.295	.097	-55.403	8.595
	Advertsing Budget (Thousands of Pounds)	.085	.000	.007	.001	.072	.098
	No. of plays on Radio	3.367	.010	.321	.001	2.735	4.022
	Attractiveness of Band	11.086	-.119	2.221	.001	6.458	15.013

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 6

Tasks

Task 1

Lacourse et al. (2001) conducted a study to see whether suicide risk was related to listening to heavy metal music. They devised a scale to measure preference for bands falling into the category of heavy metal. This scale included heavy metal bands (Black Sabbath, Iron Maiden), speed metal bands (Slayer, Metallica), death/black metal bands (Obituary, Burzum) and gothic bands (Marilyn Manson, Sistsers of Mercy). They then used this (and other variables) as predictors of suicide risk based on a scale measuring suicidal ideation etc. devised by Tousignant et al., (1988).

- Lacourse, E., Claes, M., & Villeneuve, M. (2001). Heavy Metal Music and Adolescent Suicidal Risk. Journal of Youth and Adolescence, 30 (3), 321-332. [Available through the Sussex Electronic Library].

Let's imagine we replicated this study. The data file **HMSuicide.sav** (on the course website) contains the data from such a replication. There are two variables representing scores on the scales described above: hm (the extent to which the person listens to heavy metal music) and suicide (the extent to which someone has suicidal ideation and so on). Using these data carry out a regression analysis to see whether listening to heavy metal predicts suicide risk.

How much variance does the final model explain?

Your Answers:

Does listening to heavy metal significantly predict suicide risk (quote relevant statistics)?

Your Answers:

What is the nature of the relationship between listening to heavy metal and suicide risk? (sketch a scatterplot if it helps you to explain).

Your Answers:

Write out the regression equation for the model

Your Answers:

As listening to heavy metal increases by 1 unit, how much does suicide risk increase?

Your Answers:

Task 2

One of my favourite activities, especially when trying to do brain-melting things like writing statistics books, is to drink tea. I am English after all. Fortunately, tea improves your cognitive function, well, in old Chinese people at any rate (Feng, Gwee, Kua, & Ng, 2010). I may not be Chinese and I'm not *that* old, but I nevertheless enjoy the idea that tea might help me think. There are some data (**Tea Makes You Brainy 716.sav**) based on Feng et al.'s study that measured the number of cups of tea drunk and cognitive functioning. Use regression to construct a model that predicts cognitive functioning from tea drinking, what would cognitive functioning be if someone drank 10 cups of tea? Is there a significant effect?

Task 3

In the lecture we saw an example of outliers vs. influential cases in which we predicted mortality rates form the number of pubs in the area. Run a regression analysis for the **pubs.sav** data predicting **mortality** from the number of **pubs**. Try repeating the analysis but bootstrapping the confidence intervals.

Task 4

The Honesty Lab (www.honestylab.com) looked at how people evaluated dishonest acts. Participants evaluated the dishonesty of acts based on watching videos of people confessing to those acts. The media would have us believe that the more likeable the perpetrator was, the more positively their dishonest acts were viewed. Imagine we took 100 people and gave them a random dishonest act, described by the perpetrator. We asked them to evaluate the honesty of the act (0 = appalling behaviour to 10 = it's Ok really) and how much they liked the person (0 = not at all, 10 = a lot). I've fabricated some data (**HonestyLab.sav**) relating to people's ratings of dishonest acts and the likeableness of the perpetrator. Run a regression using bootstrapping to predict ratings of dishonesty from the likeableness of the perpetrator.

References

Feng, L., Gwee, X., Kua, E. H., & Ng, T. P. (2010). Cognitive function and tea consumption in community dwelling older Chinese in Singapore. Journal of Nutrition Health & Aging, 14(6), 433-438.

Wright, D. B., London, K., & Field, A. P. (2011). Using Bootstrap Estimation and the Plug-in Principle for Clinical Psychology Data. Journal of Experimental Psychopathology., 2(2), 252–270. doi: doi:10.5127/jep.013611

Terms of Use