***Linear Regression:***

Sooty suspects that his magic wand is prone to changing its length as the temperature varies: on cold days, he's sure that it's shorter than on warm days. He decides to investigate this, so on each of 50 days he measures the temperature with his little thermometer and measures the length of his wand with his little ruler. He then uses SPSS to perform linear regression on these data, to see if he can predict wand length from temperature. This is how Sooty did it.

(a) The data are in the file ***RegressionData.sav***, on my website. The relevant variables are ***temperature*** and ***wand\_length***. Each row represents a particular day, and shows the ambient temperature (in degrees centigrade) and the length of Sooty's wand (in centimetres - what a big wand Sooty has).

Go to **Analyze..**, then **Regression..**. and then select **Curve Estimation**. (You can also perform this analysis by selecting **Linear...** instead of Curve Estimation, but then you won't automatically get a scatterplot. Life is full of tough choices.)

In the box labelled "Dependent(s):", put the variable that you want to predict (the "Y" variable). In the "Independent" box, put the variable that you want to use as a predictor (the "X" variable). Then click on "OK".



**Independent and dependent variables, and regression:**

The terminology used by SPSS for regression seems a little confusing at first sight. In a basic experimental study, you have an independent variable (IV) and a dependent variable (DV). The independent variable is the one that the experimenter *manipulates*. The dependent variable is the one that is used in order to *measure the effects* of the manipulations of the independent variable. So, for example, you might vary the amount of alcohol that people drink (manipulate the independent variable of "alcohol consumption") and measure the effects of this on how quickly people agree to sing in a karaoke bar (measure the dependent variable of "speed to grab the karaoke microphone when prompted").

In a correlational design (the circumstances in which regression is normally used) the concepts of IV and DV are rather blurred, so don't get too hung up on trying to identify IVs and DVs in this situation. We don't actually *manipulate* any variables in a correlational design, so in a strict sense we have no IV, only two DVs. What we actually have are *measurements* on two variables (in the current example, temperature and wand length). You will see the two variables in a correlational design referred to as IVs, but in a sense, they are both really DVs, because Sooty isn't systematically manipulating either of them - he has no control over either the daily temperature or the length of his wand!

For the purposes of running a regression analysis in SPSS, the variable which is being used as a *predictor* is designated as the independent variable, and the variable whose values we are trying to *predict* is called the dependent variable. Thus, in this instance, Sooty is trying to predict wand length from knowledge of the temperature: wand length is the DV and temperature is the IV, as a far as SPSS is concerned. However, suppose Sooty wanted to predict the temperature from knowledge of his wand's length: in that situation, SPSS would consider wand length to be the IV, and temperature to be the DV.

With this in mind, here's an explanation of the SPSS output.

**The "Model summary" table:**

|  |
| --- |
| **Model Summary and Parameter Estimates** |
| Dependent Variable: wand\_length |
| Equation | Model Summary | Parameter Estimates |
| R Square | F | df1 | df2 | Sig. | Constant | b1 |
| Linear | .139 | 7.770 | 1 | 48 | .008 | 60.952 | .058 |
| The independent variable is temperature. |

The line below the table reminds you which variable you used as a predictor. It designates this as an "independent variable". We used temperature as our predictor, so that's correct.

The line above the table reminds you which variable for which you were trying to predict values. It designates this as a "dependent variable". We used wand length as our predictor, so again that's correct.

There are three boxes in the table itself: "Equation"; "Model Summary"; and "Parameter Estimates".

(a) "Equation" simply reminds you which kind of regression you performed (there are lots of different types). We wanted a "linear" regression, and that's what we've apparently got.

(b) The table section called "Model Summary" contains two things which need a bit of background explanation: the value of "R Squared", and the results of a statistical test called ANOVA.

***Correlation coefficients and R2:***

Pearson's *r* is a measure of the strength of the linear association between two variables. It falls between +1 (a perfect positive correlation between the two variables, such that an increase in one of them is matched by a set amount of increase in the other) and -1 (a perfect negative correlation, such that an increase in one of the variables is always matched by a set amount of *decrease* in the other) An *r* of 0 means that the two variables are not correlated at all. Normally, we work out a value of *r* from our sample data and then assess its *statistical significance:* taking into account how many pairs of scores we have, we can assess how likely it is that we would obtain a correlation coefficient as big as ours merely by chance (i.e., how often we would obtain a given value of Pearson's *r* from our data when in fact there was no relationship between the two variables).

One problem with interpreting correlation coefficients solely in terms of statistical significance is that it tells you whether there is a non-random association between two variables, but it says nothing about how strong this relationship really is. This is where *R* Squared comes in. "R Square" (*R*2) is simply the Pearson's correlation coefficient squared. Thus if you have an *r* of .20, *R*2 is .04 (.20 \* .20 = .04, or 4%). This means that 4% of the variation in one of the variables is accounted for by its relationship with the other one - leaving the remainder (a whopping 96% in this case) wholly unexplained. With a bit of practice, *R*2 enables you to get a feel for how big a correlation coefficient really is, in practical terms. It is an example of a measure of "effect size", designed to give you a feel for how important a test statistic is, rather than merely indicating whether or not it is likely to have arisen by chance.

The advantage of thinking about *R*2 is that it makes the true strength of the correlation more obvious. If you merely considered Pearson's *r*, you might be tempted to think that, say, a .40 correlation was twice as strong as a .20 correlation, and half as good as a .80 correlation. However looking at *R*2 makes it obvious that this is not the case. For predictive purposes, correlations become *much* more useful the closer they are to 1 or -1. A correlation of .20 gives an *R*2 value of .04; a correlation of .40 gives an *R*2 value of .16; while a correlation of .80 gives an *R*2 value of .64! Clearly a .80 correlation is *much* better than a .04 correlation.

Our obtained value of *R*2 is .139, or .14 once rounded up. This means that only 14% of the variation in wand length is accounted for by its relationship with temperature. That leaves 86% of the variation in wand length completely unaccounted for. (Presumably there are other influences on wand length that we are unaware of.) Consequently, even though there is a relationship between wand length and temperature, if Sooty knows how warm it is today, it's not going to give him much insight into how long his wand will be.

***Analysis of Variance results (WARNING - complicated bit, don't get too stressed out by this!):***

In the next part of the table, SPSS shows the results of a statistical test called "Analysis of Variance" ("ANOVA" for short). There's actually a whole family of ANOVA tests, and you'll encounter them in much more detail this year in *Research Skills 2*, and next year in the *Discovering Statistics* module. Without going into too much detail, in the present situation ANOVA is being used to test whether your "model" (your regression line) is significantly better at predicting the values of the variable being predicted than if you simply used the mean of the predicted variable. (If there was no relationship at all between temperature and wand-length, then your best guess at predicting a wand-length, for any given temperature, would be to cite the *average* wand-length. The ANOVA can help you decide whether your regression line does any better at prediction than this.)

 "*F*" is the value of the ANOVA test statistic (after "Fisher", the statistician who devised ANOVA - please address any hate mail to him, not to me). In assessing whether or not the observed value of *F* is so large that it is unlikely to have occurred by chance, you need to take into account two sets of degrees of freedom, and these are shown as "df1" and "df2". For our purposes, "Sig." is the most interesting value. It shows the probability of getting a value of *F* as large as our obtained one, merely by chance. If "Sig." is .05 or smaller, we can conclude that our value of *F* is so large that it is unlikely to have occurred by chance; our regression line is a significantly better fit to the data than a model based on using the mean of the values for the predicted variable.

In the present example, *F* (1, 48) = 7.77, *p* = .008. This *p*-value is much smaller than .05, and so we would conclude that in this case, our regression line is significantly better at predicting the length of Sooty's wand from temperature than if we simply used the mean wand-length every time.

The last part of the table shows two values: "constant", which is the intercept of the regression line, and "b1" which is the slope of the regression line. (It's called "b1" instead of just "b", because the simple linear regression that we have performed can be extended to become **multiple regression**, in which there are a number of different predictor variables instead of just one - e.g. b1, b2, b3, etc.).

The next bit of the output is the scatterplot, with the regression line displayed:



Here we can see that there is a positive relationship between temperature and wand length: as temperature increases, so too does the length of Sooty's wand. However the observed data points (the open circles) are scattered quite widely around the line, reflecting what the statistics have already told us - that there is a relationship between temperature and wand-length, but not a very strong one.

***Reporting the results of a regression analysis:***

You could report these results as follows. "A linear regression was performed. As can be seen from the scatterplot, temperature was a significant predictor of magic wand length. The regression equation was:

*wand length = 60.95 + 0.06 \* temperature*, *R*2 = .14, *F*(1, 48) = 7.77, *p* = .008".

Once we have our regression equation, we can use it to make predictions. In this instance, Sooty can get up in the morning and use the temperature in order to predict what length his magic wand might be that day. (Personally, I'd just measure the wand directly, but it takes all sorts.)

For example, if today's temperature is 90 degrees, then according to the regression equation, Sooty would expect his wand to be:

*wand length = 60.95 + 0.06 \* 90*

*wand length = 60.95 + 5.4*

*wand length = 66.35 cm.*

If the temperature is 60 degrees, Sooty would expect his wand to be:

*wand length = 60.95 + 0.06 \* 60*

*wand length = 60.95 + 3.60*

*wand length = 64.55 cm.*

And so on, for any temperature that Sooty cares to plug into the equation.

***Examples for you to try:***

1. The U.S. army want to make sure that their soldiers are tough enough for active duty. They know that their enemies are now using the most inhumane weapon yet devised - recordings of Gary Barlow songs, played over huge loudspeakers. To ensure that their soldiers can cope with this, the U.S. army take a random group of 60 soldiers and take two measurements from each of them. One is the soldiers' pain threshold (estimated by measuring the length of time in seconds that the soldier can keep their hand in a pail of ice-cold water). The other measure is the length of time that the soldier can listen to Gary Barlow before they scream for mercy (this is their "Barlow-resistance" score). The data are in the variables ***Pain\_threshold*** and ***Barlow\_Resistance***.

Perform a linear regression on these data.

(a) What do you conclude about the relationship between pain threshold and resistance to Gary Barlow?

(b) Here are the pain thresholds for seven marines. You can only send them off for active service if they have a predicted Barlow-resistance score of 50 or above. Use the intercept and slope provided by SPSS to work out the regression formula. Enter the pain thresholds from the recruits into this formula in order to obtain a *predicted* Barlow-resistance score for each man (you need to keep these men at peak fitness, and so you cannot actually expose them to Barlow's singing). Which of these recruits are Barlow-resistant enough to be sent off to the front?

|  |  |  |
| --- | --- | --- |
| **Marine** | **Actual pain threshold** | **Predicted Barlow resistance score** |
| **"Slasher" Malone** | **132** |  |
| **"Noteeth" Magee** | **331** |  |
| **"Crusher" McDermott** | **468** |  |
| **"Bulldog" Grubble** | **329** |  |
| **"Tibbles" Wibble** | **446** |  |
| **"Skull" McCandy** | **395** |  |
| **"Braindeath" Flaherty** | **664** |  |

2. Does the amount of junk mail received predict how angry people become when they receive it? Stichemupp advertising agency investigate this issue by interviewing 40 shoppers. They ask each one to (a) estimate (to the nearest ton) how much junk mail they had received from Virgin media during the past month and (b) how angry they were about it (measured by taking their pulse rate). The data are in the variables **VirginJunk** and **Pulse**. Perform a linear regression on these data. What do you conclude? The data for case 5 have been entered wrongly: his pulse rate should be 120, not 12. Re-run the analysis after correcting this mistake. What happens to the results?

**Answers to the Linear Regression problems:**

***Question 1***

*(a) What do you conclude about the relationship between pain threshold and resistance to Gary Barlow?*

Here are the results of the regression analysis. The scatterplot suggests that there is a fairly strong positive correlation between pain threshold and Gary Barlow-resistance: as pain threshold increases, so too does the length of time that a person can withstand listening to Gary Barlow. The statistical analysis confirms this impression. *R2* is .53, suggesting that about half of the variation in Gary Barlow-resistance can be accounted for by this variable's relationship with pain threshold. The ANOVA shows that this regression model is significantly better than the mean in predicting Gary Barlow- resistance, *F*(1, 58) = 65.79, *p* < .001.

|  |
| --- |
| **Model Summary and Parameter Estimates** |
| Dependent Variable: Barlow\_Resistance |
| Equation | Model Summary | Parameter Estimates |
| R Square | F | df1 | df2 | Sig. | Constant | b1 |
| Linear | .531 | 65.790 | 1 | 58 | .000 | -.556 | .080 |
| The independent variable is Pain\_threshold. |



You could report this as follows. "A linear regression was performed. As can be seen from the scatterplot, pain threshold was a significant predictor of Gary Barlow-resistance. The regression equation was *Gary Barlow resistance = - 0.56 + 0.08 \* pain threshold*, *R2*  = .53, *F*(1, 58) = 65.79, *p* < .001".

*(b) Which of the recruits are Barlow-resistant enough to be sent off to the front?*

In order to decide whether these soldiers should be sent off to active combat, we need to work out what their Gary Barlow-resistance scores are likely to be. To do that, we need to construct the regression equation, and then slot in each soldier's *actual* pain threshold, so that we can find that soldier's *predicted* Barlow-resistance score.

Taking the intercept (constant) and slope (b1) from the SPSS outputm, the formula is:

 *predicted Y* = - 0.56 + 0.08 \* X, where X is the marine's actual pain threshold.

e.g. for "Slasher" Malone, *predicted Y* = - 0.56 + 0.08 \* 132; for "Noteeth" Magee, *predicted Y* = - 0.56 + 0.08 \* 331, and so on.

|  |  |  |
| --- | --- | --- |
| **Marine** | **Actual pain threshold** | **Predicted Barlow resistance score** |
| **"Slasher" Malone** | **132** | **10.00** |
| **"Noteeth" Magee** | **331** | **25.92** |
| **"Crusher" McDermott** | **468** | **36.88** |
| **"Bulldog" Grubble** | **329** | **25.76** |
| **"Tibbles" Wibble** | **446** | **35.12** |
| **"Skull" McCandy** | **395** | **31.04** |
| **"Braindeath" Flaherty** | **664** | **52.56** |

As a check that you've done the sums right, make sure that these points all fall on the regression line. If they don't, you've done something wrong! (As a "quick and dirty" way of finding the predicted Gary Barlow-resistance scores, you could read off the values from the regression line itself; but using the equation is more accurate.)

**Question 2.**

Does the amount of junk mail received predict how angry people become when they receive it?

We want to predict pulse rate from volume of junk mail. Here are the results of the linear regression.

|  |
| --- |
| **Model Summary and Parameter Estimates** |
| Dependent Variable: Pulse |
| Equation | Model Summary | Parameter Estimates |
| R Square | F | df1 | df2 | Sig. | Constant | b1 |
| Linear | .236 | 11.748 | 1 | 38 | .001 | 85.659 | 8.408 |
| The independent variable is VirginJunk. |



You could report this as follows. "A linear regression was performed. As can be seen from the scatterplot, the volume of Virgin junk-mail was a significant predictor of pulse rate. The regression equation was *pulse rate = 85.69 + 8.41 \* junk-mail*, *R*2 = .24, *F*(1, 38) = 11.75, *p* = .001".