

EXERCISE 20

Simple, two-variable regression

Before you start

Before proceeding with this Exercise, please read Chapter 12.

Purpose of the project

In this Exercise, we shall look at some of the pitfalls that await the unwary user of regression techniques; in fact, as we shall see, all the cautions and caveats about the **Pearson correlation** apply with equal force to regression.

In Exercise 17, Anscombe's specially devised data set (whose columns were named $X1$, $X2$, $Y1$, $Y2$, $Y3$, $Y4$) was saved in a file named **Anscombe**. Scatterplots and correlation coefficients were obtained for the pairings $(X1, Y1)$, $(X1, Y2)$, $(X1, Y3)$ and $(X2, Y4)$. All sets yielded exactly the same value for the Pearson correlation. When the scatterplots were inspected, however, it was seen that the Pearson correlation was appropriate for only one data set: in the other sets, it would give the unwary user a highly misleading impression of a strong linear association between X and Y . One problem with the Pearson correlation is that it is very vulnerable to the leverage exerted by atypical data points, or **outliers** as they are termed. The Pearson correlation can also have large values with monotonic but non-linear relationships. All this is equally true of the parameters of the regression equation. In this Exercise, we return to Anscombe's data to investigate the statistics of the regression lines for the four sets of paired data.

Opening SPSS

Open SPSS and select the data file **Anscombe** from the opening window. This was the file saved from Exercise 17.

Running the simple regression procedure

Following the procedure described in Section 12.2.3, obtain the regression statistics of $Y1$, $Y2$ and $Y3$ upon $X1$ and of $Y4$ upon $X2$. Remember that the dependent variable is Y , and the independent variable is X . For present purposes, the plotting of the scatterplot of ***ZRESID** (**y-axis** box) against ***ZPRED** (**x-axis** box) should provide illuminating tests of the credibility of the assumption that the data are linear. Full details of preparing the **Linear Regression** dialog box are given in Section 12.2.3 but omit saving predicted values and residuals.

Since we want to carry out regression upon all four (X, Y) data sets, it will be necessary to prepare the **Regression** dialog box for the first pair to include a scatterplot of ***ZRESID** against ***ZPRED**, and then change the variable names on subsequent runs for the remaining three pairs. To return to the **Regression** dialog box after inspecting the scatterplot, click the **Analyze** drop-down menu at the top of the **SPSS Viewer** window and select **Regression** and **Linear** again. Change $Y1$ to $Y2$ in the **Dependent** box and click **OK**. Follow this procedure for each pair of variables (i.e. $Y3$ upon $X1$ and then $Y4$ upon $X2$ – here you need to substitute $X2$

for X1 in the **Independent(s)** box). After each run, you should record the value of **R Squared** and the regression equation, and note the appearance of the scatterplot.

Output for the simple regression analyses

The main features of the Output of a simple regression analysis are fully explained in Chapter 12.

- **Compare the regression statistics and scatterplots for all four bivariate data sets. What do you notice about the values of R Squared and the appearances of the scatterplots? What is the 'take-home' message here?**

Another example

A researcher interested in the relationship between blood alcohol level and road accidents examined the accident rates for various levels of blood alcohol from 5 to 35 mg/100 ml. The data are shown in Table 1:

Alcohol level	5	10	15	20	25	30	35
No of accidents ($\times 10^3$)	10	17	26	30	32	38	42

Find the regression of the number of accidents upon blood alcohol level and predict the number of accidents for a blood alcohol level of 40mg/100ml. Draw the scatterplot with SPSS and fit the regression line.

Preparing the data set

Prepare the data set in the usual manner with two variables.

Running the regression and inspecting the output

Run the regression command as described in Section 12.2.3, but omit the optional extras Descriptives, Casewise diagnostics, and the plot.

- **Write down the regression equation.**

Use either a calculator or SPSS to calculate the number of predicted accidents for an alcohol level of 40 mg/100 ml using the regression equation. In SPSS, insert the value of 40 in the *blood* column of **Data View** and then complete a **Compute** command by entering the appropriate coefficients and variable name to calculate predicted values for a new variable with a name such as *Predicted*.

- **What is the predicted number of accidents for a blood alcohol level of 40 mg/100 ml?**

Drawing the regression line

Use the procedure described in Section 12.2.1 to draw the scatterplot and insert the regression line. To answer the next question, you should edit the scatterplot and extend the abscissa (x axis) to 40 in order to see the position of the regression line for that value. Do this by double-clicking on the scatterplot to open the **Chart Editor** and then click on one of the x axis numbers to open the **Properties** dialog box. Select the **Scale** tab at the top, click off the

tick in the check box for **Maximum**, enter 40 into the **Custom** box alongside, and click on the **Apply** box. The scatterplot and regression line should then extend to include an Alcohol Level value of 40. Grid lines can be added while the scatterplot is still in the **Chart Editor** by clicking **Options**→**Show Grid Lines**. Close the **Properties** dialog box and click off the **Chart Editor** to return the scatterplot to the **Viewer**.

- Does your calculated value for 40 mg/100 ml correspond with what you can see in the scatterplot with its fitted regression line?

Calculating residuals (difference between actual and predicted values)

To calculate the values of residuals, we need to know the predicted value of the number of accidents and then subtract it from the actual number of accidents. This could be done by using a calculator and the regression equation but it is more simply done by using SPSS. Return to the **Linear Regression** dialog box, select the **Save...** button and click on the check boxes for **Unstandardized** in the **Predicted Value** choices and on **Unstandardized** in the **Residuals** choices. After re-running the analysis, the **Data Editor** (not the **SPSS Viewer**) will show the predicted value and residual for each alcohol level.

- What is the value of the largest residual?

Finishing the session

Close down SPSS and any other windows before logging out of the computer.