Answers to Exercises:

Chapter 2: Prediction and Analysis of Variance

NB It is a good idea to learn how to use SPSS syntax as it will enable you to keep a record of all the commands you have used to create your final data and output. These commands, if kept together in a logical, well labelled, syntax file, can allow you to replicate, in a matter of minutes, results and analysis that originally took months to create. Because syntax files are basically text files they take up very little hard-disk space and so you can easily keep multiple backup copies of them on a variety of devices so as to minimise the risk of losing your work. This is particularly useful for large projects, such as dissertations and PhD research.

Exercise 2.1.1

1. What is the "residual" and how is it related to how OLS draws the line of best fit? What other names are given to the residual?

   The residual, u, is the difference between the predicted values, $\hat{y}$, and the observed values of the dependent variable in our sample, y. That is,

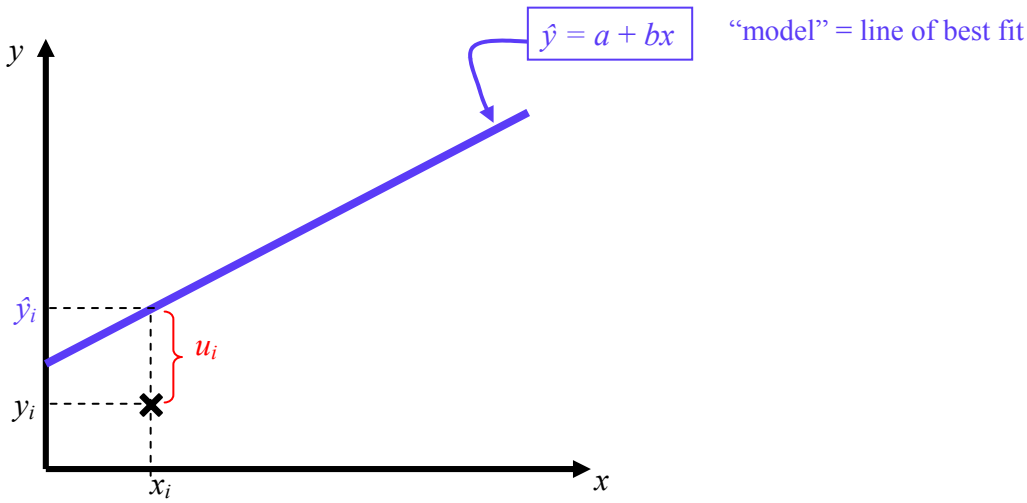
   $$\begin{aligned} u_i &= y_i - \hat{y}_i \\ &= y_i - (a + b_1 x_1 + b_2 x_2) \end{aligned}$$

   where a, b_1 and b_2 are our sample estimates of the population parameters, α, β_1 and β_2. The residual measures the unexplained variation in the dependent variable. The residual is also called "the error term" or "disturbance term", and is usually assumed to be white noise – i.e. to vary in a purely randomly in a way that is unrelated (not contingent upon) the explanatory variables, x_1, x_2 etc, included in the model.

2. How is the error term different to the "standard errors" reported in the table of coefficients?

   The error term is a measure of the prediction accuracy of the model within our sample. It is the difference between the observations in our sample on y and the predicted values of y obtained by plugging into the estimated model the values of the explanatory variables x_1, x_2, … x_k etc. The standard error, SE, on the other hand, is a measure of how much a sample estimate varies from sample to sample. The standard errors reported in the coefficients table relate to an individual slope or intercept term, such as b_1, the slope coefficient for x_1. So, the standard error of b_1, $SE(b_1)$, is the estimated standard deviation of the slope coefficients b_1 across repeated samples. Usually we only have one sample, however, so we cannot be sure what the true value of the standard error is (we can only estimate its likely value using probability theory). In contrast, we can measure the error term exactly because it is simply the difference between the predicted values, $\hat{y}$, and the actual values of the dependent variable, y, in our sample.

3. Draw a hypothetical scatter plot with line of best fit depicting the residual for a particular observation.



4. The formula for the Residual Sum of Squares is given above. What do you think the formulas for the Total Sum of Squares and Regression Sum of Squares will look like?

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

RSS measures how much variation there is in *u;* i.e. how much it deviates from zero.

$$REGSS = \sum_i (\hat{y}_i - \bar{y})^2$$

REGSS measures how much variation there is in $\hat{y}$, the predicted values of *y;* i.e. how much $\hat{y}$ deviates from the average value of *y*.

$$TSS = \sum_i (y_i - \bar{y})^2$$

REGSS measures how much variation there is in *y*, the observed values of the dependent variable*;* i.e. how much *y* deviates from the average value *y*.

Exercise 2.1.1

1. Enter the following data on y, x_1 and x_2 into a new SPSS data sheet. Label y as income from employment measured in £000s, x_1 as years of post-school education, and x_2 as years of work experience. Then save your data as "income_education_exp.sav".

y	x_1	x_2
35	5	10
22	2	9
31	7	10
21	3	9
42	9	13

   Click *File, New, Data* on the menu bar. Enter the data in Data View (notice the tab labelled Data View in the bottom left of the screen). Then go to Variable View by clicking on the tab of that name in the bottom left of the screen and enter variable labels in the Labels column of Variable View. Alternatively, you can label the variables using the following syntax:

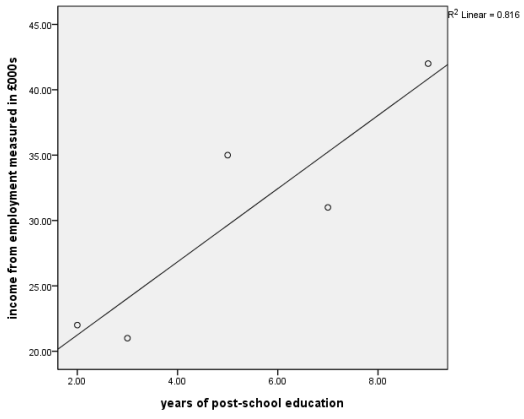
```
VARIABLE LABELS y "income from employment measured in £000s" .
VARIABLE LABELS x1 "years of post-school education".
VARIABLE LABELS x2 "years of work experience".
```

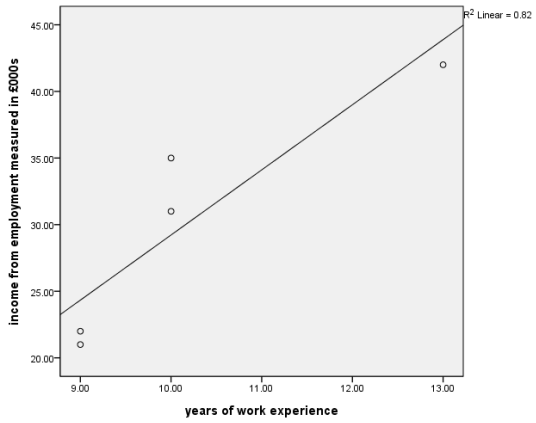
2. Run a scatter plot of y on x_1 and of y on x_2. Include a line of best fit in each graph. Comment on your results.

```
GRAPH    /SCATTERPLOT(BIVAR)=x1 WITH y.
GRAPH    /SCATTERPLOT(BIVAR)=x2 WITH y.
```

   (To add a line of best fit, double click on the graph, right-click on the data points on the chart, and select "Add Fit Line at Total" then select Linear, and click Close. Then close the graph editor to return to the Statistics Viewer output window).

   The graphs should look like those below. There are too few observations to derive strong conclusions, but the observations available do suggest an upward, linear relationship.





3. Run a regression of y on x_1 and x_2. Based on the estimated coefficients from the regression output, use the COMPUTE command to create a new variable called y_hat, computed as the predicted income for each of the 5 individuals in your sample.

```
REGRESSION  /DEPENDENT y  /METHOD=ENTER x1 x2  .
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.200	23.951		-.175	.877
	years of post-school education	1.450	1.789	.468	.811	.503
	years of work experience	2.633	3.117	.488	.845	.487

a. Dependent Variable: income from employment measured in £000s

Note that if you double click on the coefficient table, you can highlight particular results, right click on the item, copy and then paste into your syntax file. This allows you to enter the parameters in your syntax command to the maximum number of decimal places. Here we have used the first 5 decimal places:

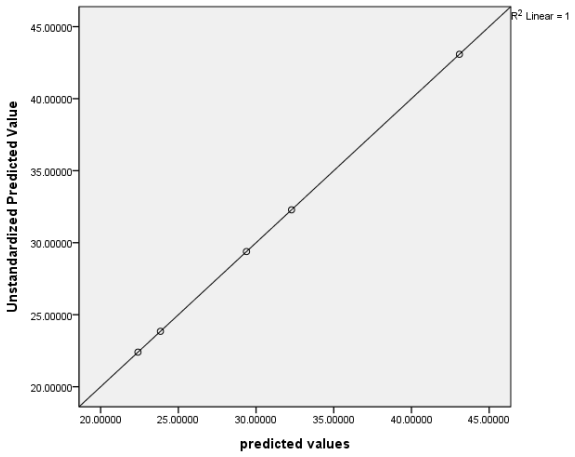
```
COMPUTE y_hat = -4.20000 + 1.45000*x1 + 2.63333*x2.
EXECUTE.
```

4. Now use the "\SAVE PRED" option in the SPSS regression function to calculate the predicted values (SPSS will create a new variable in the next available column in your data set with an automatically generated name such as PRE_1). Look at the variables in the Data View window to confirm that the values are the same as y_hat (there might be small differences due to rounding). Run a scatter plot of y_hat against the predicted values from the "Save" option and include a line of best fit. Does the plot look as you would have expected?

```
REGRESSION /DEPENDENT y /METHOD=ENTER x1 x2 /SAVE
PRED.
```

	y	x1	x2	y_hat	PRE_1
1	35.00	5.00	10.00	29.38330	29.38333
2	22.00	2.00	9.00	22.39997	22.40000
3	31.00	7.00	10.00	32.28330	32.28333
4	21.00	3.00	9.00	23.84997	23.85000
5	42.00	9.00	13.00	43.08329	43.08333
6					

The values of y_hat, which we computed manually using the COMPUTE function, are pretty much identical to those of PRE_1, which SPSS created automatically as part of the "/SAVE" regression option. We can see this in the scatter plot below with a line that passes through all the points and an R^2 of 100%.



Save your amended data set under a new name to keep the changes you have made.

```
SAVE OUTFILE='C:\STATISTICS\income_education_exp1.sav' /COMPRESSED.
```

Exercise 2.3.1

1. Open up your revised income_education_exp.sav dataset and use the COMPUTE command to create a new variable called u, equal to y minus $\hat{y}$. Are the errors large or small?

   ```
   COMPUTE u = y - y_hat.
   EXECUTE.
   ```

   Compared to the corresponding value of y, the errors appear to be reasonably small.

2. Now use the "`/SAVE RESID`" option in the regression function to create the residual. Open the Data View window and compare this variable with u, the variable you created manually. Are they the same? Is the mean value of the residual as you would have expected?

   ```
   REGRESSION  /DEPENDENT y  /METHOD=ENTER x1 x2   /SAVE
   RESID.
   ```

   Yes, the two variables, u and RES_1 have almost exactly the same values, the differences being due to rounding errors. If you use the SPSS generated predicted values, PRE_1, to compute u, and then run descriptive statistics on u and RES_1, the saved residuals automatically created by SPSS, you will see that the variables are identical (make sure the u variable has the same number of decimal places in the Variable View window):

   ```
   COMPUTE u = y - PRE_1.
   EXECUTE.
   REGRESSION  /DEPENDENT y  /METHOD=ENTER x1 x2   /SAVE RESID.
   DESCRIPTIVES VARIABLES=RES_1 u /STATISTICS=MEAN STDDEV MIN MAX.
   ```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Unstandardized Residual	5	-2.85000	5.61667	.0000000	3.26534837
u	5	-2.85000	5.61667	.0000000	3.26534837
Valid N (listwise)	5				

Note that *u* has zero mean, which is what we would expect since the OLS algorithm ensures this is the case. Essentially the intercept term is computed in such a way as to guarantee the mean value of the error term equals zero.

Further Practice:

3. Load up Nationwideextract data. Run a regression of purchase price on floor area. Calculate a predicted values variable first using the COMPUTE syntax, then by selecting it as an option in the Linear Regression Window, and then by using the "/SAVE PRED" syntax (give a different name to each version of the variable). Check that they all produce equivalent results by running Descriptive Stats on each of the variables.

4. Re-run the above regression but this time include number of bedrooms. Calculate the predicted values using the same three methods and compare.

5. Run a scatter plot of the residuals against the dependent variable and comment on your results.

Exercise 2.4.1

1. Comment on the Regression, Residual, and Total sum of squares from the output from the last two questions using the Nationwideextract.sav data (i.e. whether relatively large or small, and the meaning of each).

```
REGRESSION  /DEPENDENT purchase   /METHOD=ENTER floorare.
```

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	434054196416.331	1	434054196416.331	598.724	.000(a)
	Residual	401630697844.870	554	724965158.565		
	Total	835684894261.200	555			

a Predictors: (Constant), Floor Area (sq meters)
b Dependent Variable: Purchase Price

```
REGRESSION /DEPENDENT purchase /METHOD=ENTER floorare bedrooms
/SAVE PRED RESID.
```

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	434054208017.860	2	217027104008.930	298.822	.000(a)
	Residual	401630686243.340	553	726276105.323		
	Total	835684894261.200	555			

a Predictors: (Constant), Number of Bedrooms , Floor Area (sq meters)
b Dependent Variable: Purchase Price

The regression sum of squares tells us about the amount of variation in *y* which is successfully explained by the independent variables. It measures this variation using the difference between the predicted values and the mean value of *y*, therefore the ***higher*** the regression sum of squares figure and the nearer this figure is to the total sum of squares, the better the model is at successfully predicting the value of the dependent variable. For both regression analyses, the regression sum of squares is slightly larger than the residual sum of squares. However, the difference between the REGSS and RSS is relatively small thus the models explained just over 50% of the variation in purchase price (found by dividng the total sum of squares by the regression sum of squares).

2. Confirm that TSS = REGSS + RSS for both regressions.

3. Confirm that R^2 = REGSS / TSS for both regressions.

4. Open up your revised income_education_exp.save dataset. Attempt to create the RSS manually (HINT: first use the COMPUTE command to create the square of the error term, then run the DESCRIPTIVES command with `/STATISTICS= SUM` option to obtain the sum of squared residuals). Compare your results with those obtained from the ANOVA regression table.

```
COMPUTE u_sq = u*u.
EXECUTE.
DESCRIPTIVES VARIABLES  = u_sq  /STATISTICS= SUM.
```

Descriptive Statistics

	N	Sum
u_sq	5	42.65
Valid N (listwise)	5	

```
REGRESSION  /DEPENDENT y  /METHOD=ENTER x1 x2  .
```

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	272.150	2	136.075	6.381	.135[a]
	Residual	42.650	2	21.325		
	Total	314.800	4			

a. Predictors: (Constant), years of work experience, years of post-school education

b. Dependent Variable: income from employment measured in £000s

Comparing the value for u_sq under the Sum column of the Descriptive Statistics table with the Residual Sum of Squares value in the ANOVA table, we can confirm that they are the same = 42.65.

5. Use the COMPUTE and DESCRIPTIVES commands to calculate the TSS and REGSS. Compare your results with those produced in the ANOVA regression table.

First run descriptives on *y* to obtain the mean value of 30.2:

```
DESCRIPTIVES VARIABLES=y.
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
income from employment measured in £000s	5	21.00	42.00	30.2000	8.87130
Valid N (listwise)	5				

Then use this output to create a variable called ybar equal to the mean of y for all its values. Then compute `y_ybar = (y - ybar);` and `y_ybar__sq` = (y - ybar)^2. You need this to calculate TSS.

Then compute `yhat_ybar = (y_hat – ybar)` and `y_yhat__sq` = (y_hat – ybar)^2. You need this to calculate REGSS. Finally, run descriptives with the `/STATISTICS= SUM` option to obtain the sum of (y_hat –ybar)^2 and sum of (y_hat –ybar)^2 for all observations in the sample:

```
COMPUTE ybar             = 30.2.
COMPUTE y_ybar           = y - ybar.
COMPUTE y_ybar__sq       = y_ybar**2.
COMPUTE yhat_ybar        = y_hat - ybar.
COMPUTE yhat_ybar__sq    = yhat_ybar **2.
DESCRIPTIVES VARIABLES   = y_ybar__sq   yhat_ybar__sq
  /STATISTICS= SUM.
```

Descriptive Statistics

	N	Sum
y_ybar__sq	5	314.80
yhat_ybar__sq	5	272.15
Valid N (listwise)	5	

Comparing the Descriptive Statistics output we can see that the sum of `y_ybar__sq` equals the Total Sum of Squares (TSS) = 314.8. We can also see that the sum of `yhat_ybar__sq` equals the Regression Sum of Squares (REGSS) = 272.15.

Exercise 2.5.1

1. Calculate the F statistic for the two regressions you ran earlier using the general formula:

$$F = \frac{(RSS_R - RSS_U)/r}{RSS_U/(n-k-1)}$$

2. Now calculate the F statistics using the $(R^2/k)/((1-R^2)/(n-k-1))$ formula.

3. Now calculate the F statistic using the ratio of mean squares formula:

F = regression mean squares / residual mean squares