

Practical 1: Correlation and regression

What you need to do

Before starting the Minitab practical, it would be a good idea (if you've not already done so) to bookmark the webpage containing all of the information for Semester 2 of ACE2013:

www.mas.ncl.ac.uk/~nlf8/teaching/ace2013/

To add this page to your favourites, select Favourites → Add to Favourites from the menu bar at the top. You will now be able to get to this page easily by selecting Favourites and selecting the page from the list.

In this practical you will be using Minitab to perform correlation and regression analyses similar to those shown in lectures over the past two weeks. You should work through *all* the questions given below, although only the starred questions will be assessed.

Along with assessed work from the other practical sessions, solutions to the starred questions should be typed up and submitted (with a cover sheet – see the course website) to the homework submission postbox outside the Maths & Statistics General Office, on the third floor of the Herschel Building, by no later than **4pm, Monday 12th May 2014**.

Log-in to a PC and open Minitab by clicking on Start → All programs → Statistical Software → Minitab 15 Statistical Software English. Now work through the following questions.

1. The following data comes from a recent British Gas investigation that examined factors which influenced the length of time it takes customers to pay their utility bills (see the British Gas website for details about this investigation and a fully downloadable report). Shown below are the quarterly gas bills for ten randomly selected British Gas customers (X , in pounds), and the corresponding time it took these customers to pay their bills (Y , in days).

Customer	1	2	3	4	5	6	7	8	9	10
X	400	105	205	150	460	250	315	420	100	300
Y	35	15	18	20	30	22	25	34	10	20

- (a) Type these data into the first two columns of a blank Minitab worksheet, making sure you label your columns appropriately.
- (b) Produce a scatterplot of these data by clicking on Graphs → Scatterplot → Simple; enter your X variables and Y variables, and then click OK. How would you describe, in words, the relationship that you see? To include this in a Microsoft WORD document, you would right-click on the graph and then click Copy graph; you can then paste this into WORD directly from the clipboard.
- (c) Compute the *sample* correlation coefficient r between X and Y by selecting Stat → Basic Statistics → Correlation, entering your Variables and clicking OK. Does this match up with what you see in the your scatterplot produced in part (b)?

- (d) Use the **Minitab** output from part (c) to perform a hypothesis test for the *population* correlation coefficient ρ . Clearly state your hypotheses, and use the p -value to
- state the strength of evidence against H_0 ;
 - make your decision (reject/retain H_0);
 - explain what this means in non-statistical terms.

- (e) Now use **Minitab** to perform a simple linear regression analysis for these data. Select **Stat**→**Regression**→**Regression**; enter your **Response** (Y) and **Predictor** (X) variables, and click **OK**.

- (i) Write down the full regression equation, including your estimate of σ , in the following form:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

- (ii) Using the **Minitab** output, perform a hypothesis test for the slope term β_1 .
- (iii) Interpret the value **R-Sq** in your **Minitab** output. Can you see how this is related to the sample correlation coefficient obtained in part (c)?
- (f) Re-produce the scatterplot from part (b) but now with the regression line superimposed. This can be done by clicking on **Stat**→**Regression**→**Fitted line plot**; entering your **Response** and **Predictor** variables; then clicking on **OK**.
- (g) British Gas apply a late payment fee to customers who take longer than twenty five days to pay their bill. Daniel Ballantyne's bill this quarter is £327. Can we expect Mr Ballantyne to be subject to the British Gas late payment fee?

2. Holly Gale works as an advertising manager for a company which markets do-it-yourself products through a nationwide chain of retail outlets. The company advertises mainly on a local basis, via newspapers, hoardings and the exterior of buses. Holly has been asked to assess the relationship between the amount spent on advertising per quarter and the corresponding level of sales achieved.

After giving it some thought, Holly realises that many factors could influence sales: as well as the amount spent on advertising, the number of households in the catchment area of a particular shop could affect sales; so too could the number of competitor shops within this catchment area; even the weather in the region may have an impact on do-it-yourself activity! Clearly it is impossible to include every variable which may show some relationship with sales, so she decides as a starting point to include

1. Advertising expenditure
2. Number of households within a five mile radius
3. Number of competitor shops within a five mile radius

These data can be opened by clicking on the **DIY** filename on the **ACE2013** course website (in the "Practicals" section). Open this datafile from the course website, and then answer the following questions:

- (a) Produce scatterplots of sales against each of the three candidate predictor variables by clicking on **Graph**→**Scatterplot**→**Simple**; enter **Sales** against **Advertising**, **Sales**

against Households and Sales against Competing stores; click on Multiple graphs, and then select In separate panels of the same graph; click OK, and OK again. Describe, in words, the relationship between sales and the three predictor variables.

- (b) Obtain the sample correlation coefficients for sales with each of the three predictor variables. Which predictor variable seems to show the strongest (linear) association with sales?
- (c) Now perform a multiple linear regression analysis on these data and obtain the estimated multiple linear regression equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

where Y represents sales and X_1 , X_2 and X_3 represent advertising expenditure, number of households and number of competitors (respectively). To do this, click on Stat → Regression → Regression, enter the Response variable (Y) and the three Predictor variables (X_1 , X_2 and X_3) and click OK. Write down your full estimated multiple linear regression equation.

- (d) Look at the output you obtained in part (c). You should notice that the coefficients for Advertising, Households and Competing stores ($\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$) have p -values of 0.420, 0.308 and 0.092, respectively. Hence, since all three p -values are greater than 0.05, we would retain all three null hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = 0, \\ H_0 &: \beta_2 = 0 \quad \text{and} \\ H_0 &: \beta_3 = 0, \end{aligned}$$

though β_1 is the *least significant* coefficient in the model since it has the highest p -value. Thus, you should now re-fit the multiple linear regression model using the menu options given in part (c), but exclude the advertising expenditure variable.

- (e) Look at the output you obtained in part (d). Having removed Advertising, you should now notice that the p -values for the coefficients of Households and Competing stores are 0.175 and 0.088. Again, since both are greater than 0.05, this implies that both Households and Competing stores *could* be removed from the model, since we would retain the null hypotheses:

$$\begin{aligned} H_0 &: \beta_2 = 0 \quad \text{and} \\ H_0 &: \beta_3 = 0. \end{aligned}$$

However, β_2 is the *least significant* coefficient since it has the highest p -value. Thus, you should now re-fit the linear regression model but exclude the Households variable.

- (f) Look at the output you obtained in part (e).
 - (i) Write down the estimated regression equation. Is this a *simple* linear regression model or a *multiple* linear regression model?
 - (ii) You should see that the p -value for the only remaining predictor variable (Competing stores) is 0.000 to 3 decimal places. What does this imply?
 - (iii) Interpret the R^2 statistic.

***3.** *Noakes Service Corporation* is trying to see how the number of shifts worked, bonus rates paid to employees, average hours of overtime and staff morale affect production. They have collected the data given in the file **Noakes**, available via the course website. Load up this worksheet, and then answer the following questions.

- (a) Perform a multiple linear regression analysis on these data and copy-and-paste the resulting **Minitab** output into your solutions.
- (b) Look at the output in part (a). Write down the full estimated regression equation in the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

where Y represents production and X_1 , X_2 , X_3 and X_4 represent the number of shifts, bonus rates paid, average hours of overtime and staff morale, respectively.

- (c) It might be that not all four predictor variables are necessary in your model. Look at the output in part (a) and, if there are any insignificant parameters in your model, remove the variable associated with the *least significant* parameter and re-fit the model. Do this until all remaining parameters are significant. Include, in your solutions, the **Minitab** output for the final model.
- (d) Interpret the R^2 statistic.
- (e) Perform an overall test for your final model using the F -test.
- (f) *Howard Enterprises* runs a similar company to *Noakes Service Corporation*. What can we expect production to be at *Howard Enterprises* if, this week, 6 shifts were worked, the bonus rate was 22%, the average hours of overtime was 15 and staff morale was rated as 9?