Revision Tutorial Topic 8
Simple Linear Regression

Introduction

In this topic we will be looking at Simple Linear Regression where we attempt to quantify the relationship between variables with one of those variables being the dependant variable. Dependant variable is the variable we are investigating and an Independent variable is an explanatory variable in that it helps explain the variation in the Dependant variable. So simple regression involves a single dependent variable and a single independent variable. In the next topic when we look at multiple regression, will have multiple independent variables that will attempt to explain variations in the dependant variable.

A simple regression business example could be Sales Revenue for a product. So in my example I'd make sales revenue my dependant variable because that is the variable I'm investigating and therefore my independent variable could be advertising expenditure. I could plot the sales revenue for various levels of advertising expenditure on a scatter diagram. From that data, I can create a regression equation. There are some pitfalls here, so I do need to check some assumptions first.

Therefore, the aims of this tutorial are to:

- conduct a simple regression and interpret the meaning of the regression coefficients b_0 and b_1
- use regression analysis to predict the value of a dependent variable based on an independent variable
- assess the adequacy of your estimated model
- evaluate the assumptions of regression analysis
- make inferences about the slope
- make inferences about the correlation coefficient
- estimate mean values and predict individual values

Textbook Questions

12.8   The production of wine is a multibillion-dollar worldwide industry. In an attempt to develop a model of wine quality as judged by wine experts, data was collected from red wine variants of Portuguese vinho verde (data obtained from P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, 'Modeling wine preferences by data mining from physiochemical properties', Decision Support Systems, 47, 2009, 547-553 and <https://bit.ly/9xKlEa>). A sample of 50 wines is stored in **[Dataset: VINHO_VERDE.XLSX]**. Develop a simple linear regression model to predict wine quality, measured on a scale from 0 (very bad) to 10 (excellent), based on alcohol content (%).

   a.   Construct a scatter plot. (For these data, $b_0 = -0.3529$ and $b_1 = 0.5624$.)

   b.   Interpret the meaning of the slope, b_1, in this problem.

   c.   Predict the mean wine quality for wines with a 10% alcohol content.

   d.   What conclusion can you reach based on the results of (a) to (c)?

12.26   In the previous problem,12.8, you used alcohol content to predict wine quality. Perform a residual analysis for these data. Based on these results:

   a.   Determine the adequacy of the fit of the model.

   b.   Evaluate whether the assumptions of regression have been seriously violated.

12.48   Under new workplace agreement structures brought in by the government, an owner of a small chain of clothing stores employing fewer than 100 employees is considering whether to keep on older staff. One element of the decision is an assessment of the staff claim that experienced staff sell more goods. A sample of 12 staff with varying lengths of employment is chosen by the store, and total weekly sales by each collected for one week.

Length of service (years)	Sales ($'00)
5	560
12	634
15	876
6	350
2	532
4	789
7	256
17	822
3	189
20	834
1	366
7	212

   a.   Calculate the coefficient of correlation, r.

   b.   At the 0.05 level of significance, is there a significant linear relationship between the length of service and sales?

   c.   What conclusions can you reach about the relationship between length of service and sales?

12.70   Can you use the annual revenues generated by National Basketball Association (NBA) franchises to predict franchise values? (Franchise values and revenues are stored in [Dataset: NBA_VALUES.XLSX].

   a.   Assuming a linear relationship, use the least-squares method to calculate the regression coefficients b_0 and b_1.

   b.   Interpret the meaning of the Y intercept, b_0, and the slope, b_1, in this problem.

   c.   Predict the mean value of an NBA franchise that generates $150 million of annual revenue.

   d.   Calculate the coefficient of determination, r^2, and interpret its meaning.

   e.   Perform a residual analysis and evaluate the regression assumptions.

   f.   At the 0.05 level of significance, is there evidence of a significant linear relationship between the annual revenues generated and the value of an NBA franchise?

   g.   Construct a 95% confidence interval estimate of the mean value of all NBA franchises that generate $150 million of annual revenue.

   h.   Construct a 95% prediction interval of the value of an individual NBA franchise that generates $150 million of annual revenue.

TEXTBOOK REFERENCE:

Basic Business Statistics: Concepts and Applications. *Berenson, M.L. Levine, D.M. Szabat, K.A. O'Brien, M. Jayne, N. Watson, J.* 5th edition. 2019. Pearson Australia Group Pty Ltd. ISBN 9781488617249. Chapter 12, sections 12 to 12.5 and 12.7 to 12.9.