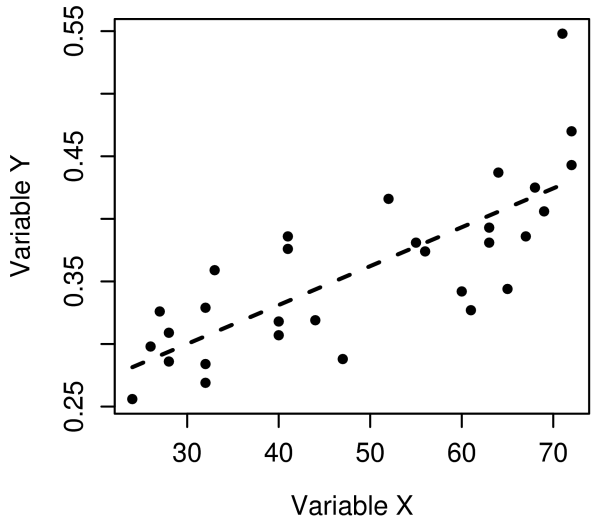
Hutcheson, G. D. (2011). Ordinary Least-Squares Regression. In L. Moutinho and G. D. Hutcheson, The SAGE Dictionary of Quantitative Management Research. Pages 224-228.

Ordinary Least-Squares Regression

Introduction

Ordinary least-squares (OLS) regression is a generalized linear modelling technique that may be used to model a single response variable which has been recorded on at least an interval scale. The technique may be applied to single or multiple explanatory variables and also categorical explanatory variables that have been appropriately coded.

Figure 1
least squares regression line



Key Features

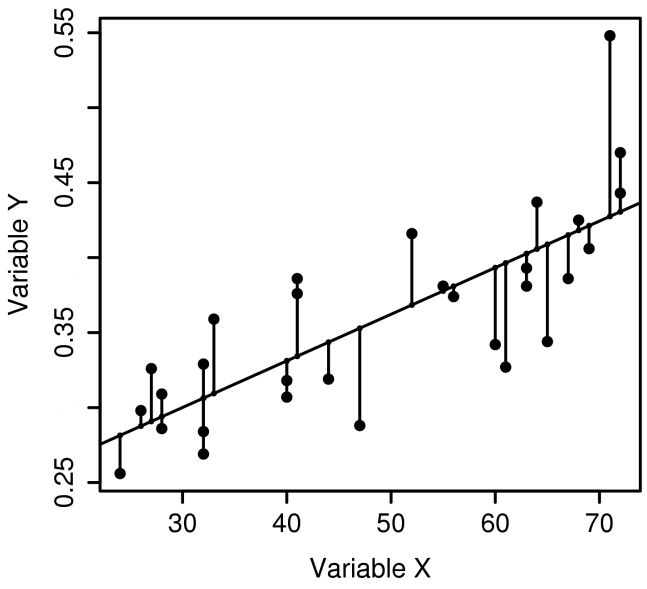
At a very basic level, the relationship between a continuous response variable (Y) and a continuous explanatory variable (X) may be represented using a line of best-fit, where Y is predicted, at least to some extent, by X. If this relationship is linear, it may be appropriately represented mathematically using the straight line equation 'Y = α + βx', as shown in Figure 1 (this line was computed using the least-squares procedure; see Ryan, 1997).

The relationship between variables Y and X is described using the equation of the line of best fit with α indicating the value of Y when X is equal to zero (also known as the intercept) and β indicating the slope of the line (also known as the regression coefficient). The regression coefficient β describes the change in Y that is associated with a unit change in X. As can be seen from Figure 1, β only provides an indication of the average expected change (the observed data are scattered around the line), making it important to also interpret the confidence intervals for the estimate (the large sample 95% two-tailed approximation of the confidence intervals can be calculated as β ± 1.96 s.e. β).

In addition to the model parameters and confidence intervals for β, it is useful to also have an indication of how well the model fits the data. Model fit can be determined by comparing the observed scores of Y (the values of Y from the sample of data) with the expected values of Y (the values of Y predicted by the regression equation). The difference between these two values (the deviation, or residual as it is also called) provides an indication of how well the model predicts each data point. Adding up the deviances for all the data points after they have been squared (this basically removes negative deviations) provides a simple measure of the degree to which the data deviates from the model overall. The sum of all the squared residuals is known as the residual sum of squares (RSS) and provides a measure of model-fit for an OLS regression model. A poorly fitting model will deviate markedly from the data and will consequently have a relatively large RSS, whereas a good-fitting model will not deviate markedly from the data and will consequently have a relatively small RSS (a perfectly fitting model will have an RSS equal to zero, as there will be no deviation between observed and expected values of Y). It is important to understand how the RSS statistic (or the ***deviance*** as it is also known; see Agresti,1996, *pages 96-97)* operates as it is used to determine the significance of individual and groups of variables in a regression model. A graphical illustration of the residuals for a simple regression model is provided in Figure 2. Detailed examples of calculating deviances from residuals for null and simple regression models can be found in Hutcheson and Moutinho, 2008.

The deviance is an important statistic as it enables the contribution made by explanatory variables to the prediction of the response variable to be determined. If by adding a variable to the model, the deviance is greatly reduced, the added variable can be said to have had a large effect on the prediction of Y for that model. If, on the other hand, the deviance is not greatly reduced, the added variable can be said to have had a small effect on the prediction of Y for that model. The change in the deviance that results from the explanatory variable being added to the model is used to determine the significance of that variable's effect on the prediction of Y in that model. To assess the effect that a single explanatory variable has on the prediction of Y, one simply compares the deviance statistics before and after the variable has been added to the model. For a simple OLS regression model, the effect of the explanatory variable can be assessed by comparing the RSS statistic for the full regression model ($Y = \alpha + \beta x$) with that for the null model ($Y = \alpha$). The difference in deviance between the nested models can then be tested for significance using an F-test computed from the following equation.

Figure 2
OLS regression model residuals



$$F_{df_p - df_{p+q}, df_{p+q}} = \frac{RSS_p - RSS_{p+q}}{\left(df_p - df_{p+q}\right)\left(RSS_{p+q} \;/\; df_{p+q}\right)}$$

where p represents the null model, $Y = \alpha$, $p+q$ represents the model $Y = \alpha + \beta x$, and df are the degrees of freedom associated with the designated model. It can be seen from this equation that the F-statistic is simply based on the difference in the deviances between the two models as a fraction of the deviance of the full model, whilst taking account of the number of parameters.

In addition to the model-fit statistics, the R-square statistic is also commonly quoted and provides a measure that indicates the percentage of variation in the response variable that is `explained' by the model. R-square, which is also known as the coefficient of multiple determination, is defined as

$$R^2 = \frac{RSS \;\; after \;\; regression}{total \;\; RSS}$$

and basically gives the percentage of the deviance in the response variable that can be accounted for by adding the explanatory variable into the model. Although R-square is widely used, it will always increase as variables are added to the model (the deviance can only go down when additional variables are added to a model). One solution to this problem is to calculate an adjusted R-square statistic (R^2_a) which takes into account the number of terms entered into the model and does not necessarily increase as more terms are added. Adjusted R-square can be derived using the following equation

$$R^2_a = R^2 - \frac{k\left(1 - R^2\right)}{n - k - 1}$$

where n is the number of cases used to construct the model and k is the number of terms in the model (not including the constant).

An example of simple OLS regression

A simple OLS regression model with a single explanatory variable can be illustrated using the example of predicting ice cream sales given outdoor temperature (Koteswara, 1970). The model for this relationship

(calculated using software) is

$$\text{Ice cream consumption} = 0.207 + 0.003 \text{ temperature}.$$

The parameter for α (0.207) indicates the predicted consumption when temperature is equal to zero. It should be noted that although the parameter α is required to make predictions of ice cream consumption at any given temperature, the prediction of consumption at a temperature of zero might be of limited usefulness, particularly when the observed data does not include a temperature of zero in it's range (predictions should only be made within the limits of the sampled values). The parameter β indicates that for each unit increase in temperature, ice cream consumption increases by 0.003 units. The significance of the relationship between temperature and ice cream consumption can be estimated by comparing the deviance statistics for the two nested models in the table below; one that includes temperature and one that does not. This difference in deviance can be assessed for significance using the F-statistic.

Model	**deviance (RSS)**	**df**	**change in deviance**	**F-statistic**	**P-value**
consumption = a	0.1255	29	0.0755	42.28	<.0001
consumption = α + β temperature	0.0500	28			

On the basis of this analysis, outdoor temperature would appear to be significantly related to ice cream consumption with each unit increase in temperature being associated with an increase of 0.003 units in ice cream consumption. Using these statistics it is a simple matter to also compute the R-square statistic for this model, which is 0.0755/0.1255, or 0.60. Temperature “explains” 60% of the deviance in ice cream consumption (i.e., when temperature is added to the model, the deviance in the Y variable is reduced by 60%).

OLS regression with multiple explanatory variables

The OLS regression model can be extended to include multiple explanatory variables by simply adding additional variables to the equation. The form of the model is the same as above with a single response variable (Y), but this time Y is predicted by multiple explanatory variables (X_1 to X_3).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The interpretation of the parameters (α and β) from the above model is basically the same as for the simple regression model above, but the relationship cannot now be graphed on a single scatter plot. α indicates the value of Y when all vales of the explanatory variables are zero. Each β parameter indicates the average change in Y that is associated with a unit change in X, whilst controlling for the other explanatory variables in the model. Model-fit can be assessed through comparing deviance measures of nested models. For example, the effect of variable X_3 on Y in the model above can be calculated by comparing the nested models

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

The change in deviance between these models indicates the effect that X_3 has on the prediction of Y when the effects of X_1 and X_2 have been accounted for (it is, therefore, the unique effect that X_3 has on Y after taking into account X_1 and X_2). The overall effect of all three explanatory variables on Y can be assessed by comparing the models

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Y = \alpha.$$

The significance of the change in the deviance scores can be assessed through the calculation of the F-statistic using the equation provided above (these are, however, provided as a matter of course by most software packages). As with the simple OLS regression, it is a simple matter to compute the R-square statistics.

An example of multiple OLS regression

A multiple OLS regression model with three explanatory variables can be illustrated using the example from the simple regression model given above. In this example, the price of the ice cream and the average income of the neighbourhood are also entered into the model. This model is calculated as

$$\text{Ice cream consumption} = 0.197 - 1.044\ \text{price} + 0.033\ \text{income} + 0.003\ \text{temperature}.$$

The parameter for α (0.197) indicates the predicted consumption when all explanatory variables are equal to zero. The β parameters indicate the average change in consumption that is associated with each unit increase in the explanatory variable. For example, for each unit increase in price, consumption goes down by 1.044 units. The significance of the relationship between each explanatory variable and ice cream consumption can be estimated by comparing the deviance statistics for nested models. The table below shows the significance of each of the explanatory variables (shown by the change in deviance when that variable is removed from the model) in a form typically used by software (when only one parameter is assessed, the F-statistic is equivalent to the t-statistic ($F = \sqrt{t}$) which is often quoted in statistical output).

	deviance change	**df**	**F-value**	**P-value**
coefficient				
price	0.002	1	$F_{1,26}$= 1.567	0.222
income	0.011	1	$F_{1,26}$= 7.973	0.009
temperature	0.082	1	$F_{1,26}$= 60.252	<0.0001
residuals	0.035	26		

Within the range of the data collected in this study, temperature and income appear to be significantly related to ice cream consumption.

Conclusion

OLS regression is one of the major techniques used to analyse data and forms the basis of many other techniques (for example ANOVA and the Generalised linear models, see Rutherford, 2001). The usefulness of the technique can be greatly extended with the use of dummy variable coding to include grouped explanatory variables (see Hutcheson and Moutinho, 2008, for a discussion of the analysis of experimental designs using regression) and data transformation methods (see, for example, Fox, 2002). OLS regression is particularly powerful as it relatively easy to also check the model asumption such as linearity, constant variance and the effect of outliers using simple graphical methods (see Hutcheson and Sofroniou, 1999).

Further Reading

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley and Sons, Inc.

Fox, J. (2002). *An R and S-Plus Companion to Applied Regression.* London: Sage Publications.

Hutcheson, G. D. and Moutinho, L. (2008). Statistical Modeling for Management. Sage Publications.

Hutcheson, G. D. and Sofroniou, N. (1999). *The Multivariate Social Scientist.* London: Sage Publications.

Koteswara, R. K. (1970). Testing for the Independence of Regression Disturbances. *Econometrica*, 38:,97-117.

Rutherford, A. (2001). Introducing ANOVA and ANCOVA: a GLM approach. London: Sage Publications.

Ryan, T. P. (1997). *Modern Regression Methods.* Chichester: John Wiley and Sons.

Graeme Hutcheson
Manchester University