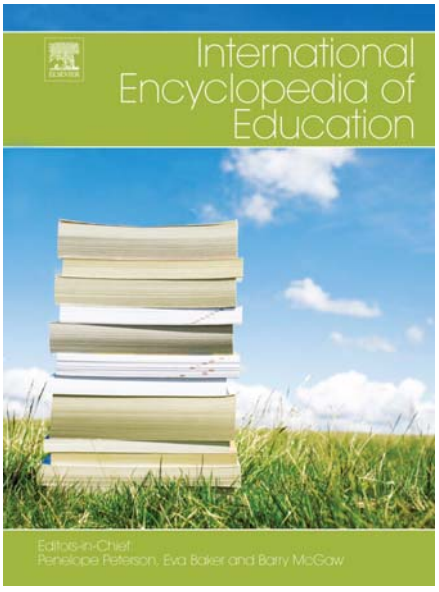

Hazelton M L (2010), Univariate Linear Regression. In: Penelope Peterson, Eva Baker, Barry McGaw, (Editors), International Encyclopedia of Education. volume 7, pp. 482-488. Oxford: Elsevier.

Univariate Linear Regression

M L Hazelton, Massey University, Palmerston North, New Zealand

Glossary

Akaike information criterion – A measure of the overall quality of a model which takes account of both the complexity of the model and how well it fits the data.
Heteroscedasticity – Refers to the presence of different variances between two or more random variables. The complementary concept is homoscedasticity, which is usually assumed to apply to the error terms in a linear regression model.
Method of least squares – A well-founded technique for estimating the unknown parameters in a linear regression model, based on minimization of the sum of squared differences between observed and modeled responses.
Ordinary least squares – The method of least squares is also sometimes referred to as ordinary least squares to distinguish it from related techniques such as weighted least squares.
Raw residual – The value of an observed response minus the corresponding predicted response based on the regression model.
Residual analysis – Refers to the examination of residuals from a regression model in order to assess the validity of the assumptions underlying the model.
Standardized residual – A scaled version of a raw residual.

The problem of modeling the behavior of one random response variable in terms of one or more other explanatory variables is one of the most important aspects of modeling in statistics. When the variables are quantitative (as we shall assume for the most part in this article), regression modeling can be used. Regression models are wide ranging and have been used extensively in the education literature (see, e.g., Hsu, 2005).

In principle, we could attempt to describe multiple features of the distribution of the response in terms of the explanatory variables. However, in many applications it is sufficient to assume that only the mean (or expected) value of the response varies with the predictors, and that it does so in a linear fashion. Under these assumptions, we obtain the classical (multiple) linear regression model. One of the reasons for the widespread use of linear regression is that such models have a highly tractable mathematical structure (e.g., Jørgensen, 1993). As a result, it has been possible to develop an extensive body of methods and theory for linear regression. This in turn provides the statistical practitioner with a comprehensive and well-understood set of tools for fitting, examining, comparing, and interpreting regression models.

In this article, a mathematical formulation of the linear regression model is provided. Model fitting by the method of least squares is described, and the application of regression models illustrated through an example. The underlying assumptions for regression modeling are discussed, and methods for examining their validity are examined. This article also covers the interpretation of regression models; selection of explanatory variables and comparison of models; and connections with other methods including analysis of variance (ANOVA) and hierarchical linear models.

Formulation of the Linear Regression Model

Suppose we observe a response variable Y and p explanatory (or predictor or regressor) variables $x_1, \ldots, x_p$ on n individuals or entities. In classifying one of them as the response, we are not treating the variables in a symmetric manner. Rather, we are focusing on the (stochastic) behavior of the response in terms of the other variables. Therefore, we will consider the explanatory variables to be fixed for the purposes of the regression model. In some applications, the explanatory variables will be fixed by an experimenter; for example, the length of time allocated to a student to complete some task for which we wish to measure a response. In other cases, the predictor variables will have arisen as the result of some random process, but the regression model will represent the behavior of the response conditional on the values of the predictor variables.

As a concrete example (to which we shall refer repeatedly), Guber (1999) describes data from 1994–95 for each of the $n = 50$ US states including the following variables: mean total SAT score (SAT); mean expenditure per pupil in thousands of dollars (EXP); mean pupil/teacher ratio (PTR); estimated mean annual salary of teachers in thousands of dollars (SAL); and percentage of all eligible students taking the SAT (PER). It is quite natural to think of modeling SAT as the response variable so as to understand its relationship with the other variables. While these

explanatory variables were not fixed by an experimenter, it is equally natural to think of modeling SAT conditional on their values so that one may use the model to predict SAT for any given set of values EXP, PTR, PER, and SAL.

The linear regression model is defined by the following equation:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad [1]$$

where Y_i is the response for the ith individual (e.g., SAT for the ith state); x_{ij} is the value of the jth explanatory variable for the ith individual (e.g., x_{i1} is the value of EXP in the ith state); and ε_i is the error term for that individual. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are model parameters whose true values are unknown and hence must be estimated from the data. When there is just one explanatory variable, the (multiple) linear regression model in eqn [1] reduces to the simple linear regression model defined by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad [2]$$

where x_i now denotes the value of the single explanatory variable for the ith individual.

It is usual to make the following four assumptions about a linear regression model, and more specifically, about the error terms in the model:

(A1) The mean error is zero, that is, $\mathrm{E}[\varepsilon_i] = 0$. This assumption is equivalent to writing that

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad [3]$$

which is a statement that we have correctly specified the form of the mean value of the response variable.

(A2) The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are statistically independent.

(A3) $\mathrm{Var}(\varepsilon_i) = \sigma^2$ is constant for all observations, $i = 1, \ldots, n$.

(A4) The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed.

We examine the consequences of failure of these assumptions, and methods for checking their validity, when we discuss model diagnostics.

The linear regression model can be expressed compactly and conveniently using matrix notation. Model [1] becomes simply

$$y = X\beta + \varepsilon \quad [4]$$

where $y = (Y_1, \ldots, Y_n)^T$ is the vector of responses (with superscript T denoting matrix transposition); $\beta = (\beta_0, \ldots, \beta_p)^T$ is the vector of regression parameters; and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ the vector of random error terms. In [4]

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad [5]$$

is the design matrix for the regression. The rows of X correspond to given individuals while the columns correspond to given variables (with the initial column of ones corresponding to the intercept term β_0).

Fitting Regression Models

Application of a linear regression model to any given data set requires that the model parameters (including the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ and the error variance, σ^2) be estimated. The regression coefficients can be estimated using the method of least squares (sometimes also referred to as ordinary least squares (OLS) to distinguish it from related techniques such as weighted least squares). If the observed responses are $y_1, \ldots, y_n$, then the least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ of the regression coefficients are derived by minimizing the sum of squared discrepancies

$$SS(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2. \quad [6]$$

Using matrix notation, the vector of least-square estimates can be expressed explicitly as

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad [7]$$

In practice, least-squares estimates are almost invariably calculated using statistical software packages. These use stable numerical algorithms (based on matrix decomposition techniques) to compute $\hat{\beta}$ rather than use a direct implementation of eqn [7].

The least-squares method is not only an intuitively plausible approach to computing estimates of the regression parameters, but is also well supported from a theoretical perspective. It can be shown under assumptions (A1)–(A4) that the least-squares estimates are also maximum likelihood estimates, and hence are optimal in a number of important statistical senses. Even if assumption (A4) fails, least-squares estimation still produces the so-called best linear unbiased estimates (BLUEs), courtesy of Gauss–Markov theory (see, e.g., Plackett, 1950).

Having obtained the least-squares estimates of the regression parameters, we may compute the fitted values,

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip} \quad [8]$$

which estimate the mean responses. We may also calculate the residual sum of squares,

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 \quad [9]$$

which provides an overall measure of the discrepancy between the observed and fitted responses, and consequently forms the basis for estimating the error variance σ^2.

Specifically, an unbiased estimator of the error variance is given by

$$s^2 = \frac{1}{n-p-1} RSS. \qquad [10]$$

Inference

The fitted regression model (i.e., the model with the regression parameters replaced by the least-squares estimates thereof) can be used for a number of purposes. These include estimation of the effect of each of the explanatory variables on the response; testing of hypotheses regarding the relationship between the response and one or more of the explanatory variables; and prediction of the response for a given set of explanatory variables.

To illustrate these ideas **Table 1** provides estimated regression coefficients (typical output from any standard statistics software package) from fitting the SAT data that were introduced earlier. The fitted model equation is hence

$$\widehat{\text{SAT}} = 1045.97 + 4.46\text{EXP} - 3.62\text{PTR} + 1.64\text{SAL} - 2.90\text{PER}. \qquad [11]$$

As an example interpretation of the parameter estimates, the model indicates a 4.46 point increase in expected SAT for a unit increase in EXP (i.e., for a \$1000 increase in mean expenditure per pupil), but it is important to recognize that this assumes that all other variables are held constant (cf. Courville and Thompson, 2001).

The standard error associated with each variable is a measure of the precision with which the coefficient of that variable is estimated. Standard errors for regression coefficients are given by the square roots of the diagonal elements of the (estimated) variance–covariance matrix, $\widehat{\text{Var}}(\hat{\beta}) = s^2(X^T X)^{-1}$. The t value for the ith variable (which appears in standard computer output for a regression analysis, like **Table 1**) is the ratio of the variable's estimated coefficient to its standard error: $t = \hat{\beta}_1 / SE(\hat{\beta}_i)$. This is an appropriate statistic for testing whether or not the true underlying parameter value is zero. A large t statistic provides evidence against the null hypothesis that the true value of the parameter is zero, and will produce a small p-value (derived from the t-distribution with $n-p-1$ degrees of freedom). For example, the p-value for EXP of $P = 0.674$ (from **Table 1**) provides no evidence to suggest that the coefficient of EXP is different from zero, and hence no statistically significant evidence that expenditure has an effect of SAT, having adjusted for the other variables in the model. Nonetheless, a simple linear regression of SAT on just EXP alone returns an estimated coefficient for EXP of $\hat{\beta}_1 = -20.9$ with corresponding p-value of $P = 0.006$, indicating that SAT score is related to expenditure but in a manner that is interrelated with the variables PTR, SAL, and PER.

Table 1 Table of regression coefficient for the SAT data

	Estimate	*Std. error*	*t value*	*Pr(>|t|)*
(Intercept)	1045.97	52.87	19.78	0.000
EXP	4.46	10.55	0.42	0.674
PTR	–3.62	3.21	–1.13	0.266
SAL	1.64	2.39	0.69	0.496
PER	–2.90	0.23	–12.56	0.000

Confidence intervals for regression parameters provide guidance on the size of the effect of an explanatory variable on the response. A $100(1-\alpha)\%$ confidence interval for the coefficient of the ith variable is given by:

$$\left(\hat{\beta}_i - t_{\alpha/2} SE(\hat{\beta}_i), \hat{\beta}_i + t_{\alpha/2} SE(\hat{\beta}_i)\right), \qquad [12]$$

where $t_{\alpha/2}$ is the appropriate critical point (i.e., the $(1-\alpha/2)$ quantile) of the t-distribution on $n-p-1$ degrees of freedom. For example, the 95% confidence interval for the coefficient of SAL is given by $(1.638 \pm 2.014 \times 2.387) = (-3.17, 6.45)$ using $t_{0.025} = 2.014$ for a t-distribution on 45 degrees of freedom. This interval includes zero, indicating that we cannot be sure (based on this analysis) that higher teacher salaries are associated with higher SAT scores (having adjusted for the other explanatory variables).

Point prediction from a regression model proceeds by substituting the requisite values of the explanatory variables into the fitted model equation. For example, if we wished to predict SAT results for EXP = 9, PTR = 14, SAL = 50, and PER = 80, then by substitution into eqn [11] we get $\widehat{\text{SAT}} = 885$ (to the nearest whole number) which is reasonably similar to the observed mean SAT score of 908 in Connecticut for which the values of the explanatory variables are similar to those chosen above. A prediction interval (which should be supplied as a matter of course) to accompany the point prediction $\hat{y}$ can be computed by

$$\left(\hat{y} - t_{\alpha/2} PE(\hat{y}), \hat{y} + t_{\alpha/2} PE(\hat{y})\right) \qquad [13]$$

where $PE(\hat{Y})$ is the prediction error. For the point prediction above the corresponding 95% prediction error is $PE(\widehat{\text{SAT}}) = 36.1$ (and $t_{0.025} = 2.014$ as earlier) so that the prediction interval is (812,958).

Regardless of the other purposes to which a regression model is developed, it is almost always of interest to characterize the extent to which the explanatory variables can describe the behavior of the response variable. This can be quantified using the coefficient of determination, R^2, which is equal to the square of the correlation coefficient between the observed and fitted values. It can be interpreted as the proportion of the variation in the

response which is attributable to the explanatory variables. For instance, for the SAT data $R^2 = 0.825$ indicating that over 80% of the variability in SAT scores can be explained by the variables EXP, PTR, SAL, and PER. It should be noted that all these comments on R^2 assume that the regression model in question includes an intercept term. For models without an intercept term (i.e., regression through the origin), R^2 does not have a particularly useful interpretation.

Model Diagnostics

The validity of any conclusions that we draw from a fitted regression model will be contingent on the correctness of the assumptions (A1–A4) listed earlier. Failure of A1 is particularly serious, since this indicates that the form of the regression function (as given in eqn [3]) is misspecified. In such circumstances, estimates of regression parameters will be biased and the model will be unreliable.

Failure of A2 occurs when the responses are dependent (e.g., because of clustering effects which may occur because of spatial proximity, or because the data include repeated observations on a given set of individuals). In this case, the least-squares parameter estimates will be unbiased (though somewhat inefficient) but standard errors computed in the usual way will be incorrect, leading to unreliable test statistics and confidence intervals. The same type of problems will occur if A3 is violated, that is, the errors are heteroscedastic. Failure of A4, that is errors which are not normally distributed, is of less concern. While in theory this invalidates t and F tests based on normality, versions of the central limit theorem and related asymptotic theory imply that the results of such standard tests will be quite reliable unless the sample size is very small.

A variety of diagnostic tools are available for assessing the validity of the model assumptions. These are typically based on the model residuals, and hence the application and interpretation of such diagnostics is often referred to as residual analysis. The raw residuals for the linear regression model are defined by:

$$e_i = y_i - \hat{\mu}_i \qquad [14]$$

while the standardized residuals are given by

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} \qquad [15]$$

where h_{ii} is the ith diagonal element of the hat matrix, $H = X(X^TX)^{-1}X^T$. The raw residuals act as substitutes for the (unobserved) error terms, $\varepsilon_1, \ldots, \varepsilon_n$, but these residuals have different variances as a result of the model fitting process. The standardized residuals are scaled to have unit variance and are therefore often preferred for diagnostic purposes.

One of the most commonly applied diagnostic tools is a plot of the residuals (either raw or standardized) against the fitted values. Such a plot can be used to spot a number of problems with the fitted model, including detection of:

- outlying observations, which will appear as points with extreme residuals;
- misspecification of the regression function (i.e., failure of A1), which may manifest itself through a discernible trend (e.g., curvature) in the plot; and
- heteroscedasticity in the error terms (i.e., failure of A3), which will typically lead to systematic variation (e.g., showing a funnel-shaped residual plot) in the vertical spread of the residuals from left to right in the plot.

A plot of the raw residuals against fitted values for the SAT data is given in **Figure 1**. It appears from this that West Virginia is an outlier, since the residual for this state is of a markedly larger size than the residuals for any other state. There is also a hint of curvature in the plot (with a downward trend at the left-hand side and an upward trend at the right-hand side of the plot) suggesting that the relationship between the SAT score and each of the explanatory variables may not be linear in every case. A plot of the residuals against each explanatory variable in turn can provide further insight. For instance, plots of the residuals for the SAT model against each of EXP, PER, PTR, and SAL (not shown) reveal clear curvature only for PER, indicating that the assumption of a linear relationship between SAT and this variable is questionable.

Assessment of the independence assumption (A2) can be difficult. A relatively common reason for failure of this assumption is the existence of serial correlation between observations with a clear time dimension (e.g., data collected on a monthly basis). In such cases, a plot of the residuals against the order in which the data were collected can sometimes be revealing although it may be more enlightening to apply standard time series methods to the residuals.

The standard tool for assessing normality of the error terms (A4) is a normal Q–Q (or normal probability) plot of the standardized residuals. This should appear as an

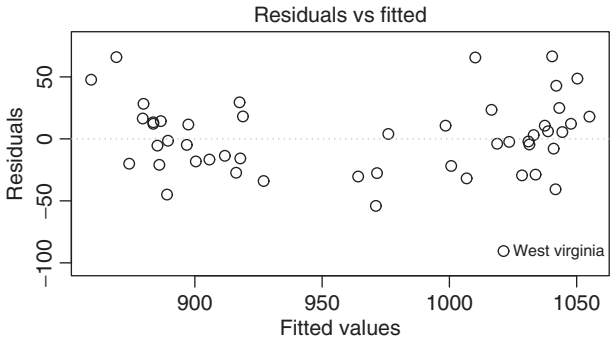

Figure 1 Plot of residuals vs. fitted values for SAT data.

approximately straight line if the assumption is reasonable. The *Q–Q* plot for the standardized residuals for the SAT data is given in **Figure 2**. Leaving aside the previously identified outlier (West Virginia), this plot does not suggest any major departure from normality in the distribution of the error terms.

As we have seen, the analysis of the SAT data may be complicated because of the existence of an apparent outlying observation – West Virginia. The extent to which this is likely to have a major impact on the model fit (and hence on inferences drawn from the model) depends on the influence that this data point exerts. At a general level, a data point with extreme values for one or more of its explanatory variables has the potential (depending on the value of the response) to have a more pronounced impact on the fit of the model than a data point for which the predictors sit near the center of the data. This potential can be measured by leverage, which is given by h_{ii} (as defined above) for the ith observation. A plot of the standardized residuals against leverage can therefore be illuminating. In addition, Cook's distance provides a measure of the influence for each data point which takes account of both the values of the predictors and the response variable. In essence, Cook's distance works by examining how the model fit would change were the data point under consideration excluded when estimating the regression parameters. Data points with a Cook's distance larger than 1 can be considered very influential and warrant close attention, while those with a Cook's distance between 0.5 and 1 might be described as moderately influential. A number of alternative measures of influence exist, with DFFITS (Belsley *et al.*, 1980) being arguably the most important.

To illustrate these methods, we plot the standardized residuals against leverage for the SAT data in **Figure 3**. It can be seen that West Virginia has rather low leverage and so has only a modest influence on the fitted model despite its extreme value for SAT (relative to its explanatory variables). This is reflected in the value of 0.11 for the Cook's distance for that state. The observation from Utah appears of greater concern, but its Cook's distance of 0.47 (although the largest of all the observations) is not excessive.

When problems are found with a regression model, then it is necessary to take remedial action. When misspecification of the regression function is suspected, then one approach is to transform the response variable (e.g., by a logarithmic transformation), while another is to consider inclusion of polynomial terms in the explanatory variables. For example, for the SAT data it transpires that addition of the quadratic term in PER leads to a statistically significant improvement in the model fit ($P = 0.000$ when testing the coefficient of PER^2). An alternative is to employ nonparametric or semiparametric regression methods (e.g., Takezawa, 2006; Ruppert *et al.*, 2003). When there is evidence of heteroscedasticity, then the method of weighted least squares should be preferred to ordinary least squares, while failure of the independence assumption can be countered by explicit modeling of a more general covariance structure for the errors. These extensions to the standard linear regression model can both be described using the unifying theory of generalized least squares (e.g., Kariya and Kurata, 2004).

The handling of outliers can be a delicate matter. The first step is to check for transcription errors in the data and correct as necessary. One can consider discarding any remaining outliers (especially those with high influence) and then refitt the model, but it is important to document that this has been done. Mindless and automated outlier deletion is to be avoided, since outliers are often of considerable interest. For example, outliers provided an early indication of the depletion of the ozone layer (Christie, 2004). An alternative to deletion is the use of robust statistical methods.

Model Comparison

When there are multiple explanatory variables available, then one can define a variety of alternative regression

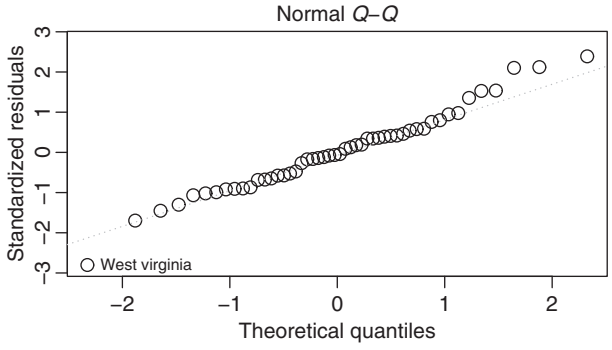

Figure 2 Normal *Q–Q* plot for standardized residuals for SAT data.

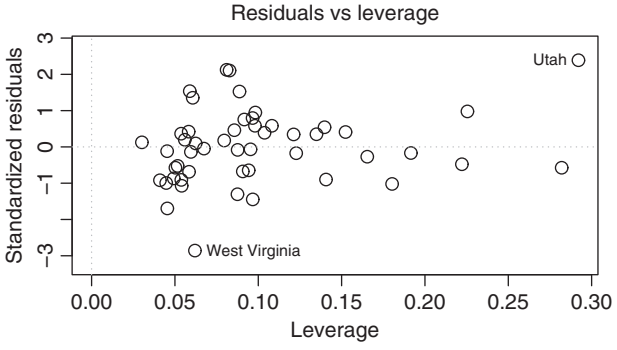

Figure 3 Plot of standardized residuals against leverage for SAT data.

models based on different subsets of these predictors. The question then arises as to which of these models should be preferred, and hence we have need for methods for comparing models. A classical approach to the comparison of two nested models (where a small model contains a subset of the variables in a large model) is to employ F tests based on ANOVA. This reduces to t-testing for a coefficient (as described above) when the models differ by only a single variable. For example, comparison of a large model containing all four explanatory variables (EXP, PTR, SAL, and PER) with a small model containing just the three variables EXP, SAL, and PER can be achieved using the t-statistic for PTR. The value $t = -1.13$ from **Table 1** gives a p-value 0.266, and hence provides no significant evidence against the null hypothesis that the small model is adequate.

An alternative approach is to define some measure of overall quality for each model, and then choose the model for which this is optimum. In quantifying model quality, one should keep in mind two competing goals in model selection. First, we have the principle of parsimony, which in essence states that simple models are preferable to more complex models (other things being equal). Second, we want the model to fit as closely to the data as possible. For linear regression models, complexity can be measured by the number of regression parameters or explanatory variables, while the closeness of the model fit can be measured by residual sum of squares. This motivates measures of quality, Q, of the following form:

$$Q = \log(RSS)/\lambda + \gamma p + c \qquad [16]$$

where RSS is the model residual sum of squares, p the number of explanatory variables, and λ, γ are tuning constants and c is an arbitrary constant. When conducting such a comparison, a smaller value for Q indicates a better model.

A number of so-called information criteria can be written in the form of eqn [16] with suitable values for the tuning parameters. The Akaike Information Criterion (AIC) is perhaps best known of these and is obtained by setting $\lambda = 1/n$ and $\gamma = 2$, while the Bayes (or Schwarz) Information Criterion (BIC) is obtained when $\lambda = 1/n$ and $\gamma = \log(n)$ (see Akaike (1974) and Schwarz (1978)). A related quantity is Mallows Cp (Mallows, 1973) in which $\gamma = 2$ and the first term on the right-hand side of eqn [16] is replaced by $RSS/\hat{\sigma}^2$, where $\hat{\sigma}^2$ is an estimate of variance which is assumed fixed for all models under comparison.

As an illustration of the application of the AIC, **Table 2** lists the values of AIC for five models for the SAT data, based on different subsets of explanatory variables. The best model (i.e., with lowest AIC) of those listed is M2, which regresses SAT score on the three variables PTR, SAL, and PER. However, models M1 and M3 are only a few units worse on the AIC scale, so there is little to choose between these and M2.

Table 2 Values of AIC for five regression models for the SAT data

Model	*Explanatory variables in model*	*AIC*
M1	EXP + PTR + SAL + PER	353.5
M2	PTR + SAL + PER	351.7
M3	SAL + PER	354.6
M4	PTR + PER	356.2
M5	SAL + PTR	425.3

When a large number of predictor variables are available, a direct comparison of all possible models will be highly computer intensive. A computationally cheaper methodology is then desirable. One approach is to employ a stepwise variable selection algorithm, where we consider systematically adding and/or removing variables in an iterative manner in order to improve the model quality. More specifically, at each step of this type of algorithm we compare the current best model with all alternative models which can be obtained by removing single terms from this model, or adding in single variables which are currently not part of the model. These comparisons may be conducted with F tests, or using AIC scores.

To illustrate the the AIC approach, we conduct variable selection starting with M1 (from **Table 2**) as the initial model. At the first step, we compare M1 with each of the four models defined by deleting one of the variables in M1. It transpires that out of these models and M1 itself, the lowest AIC belongs to model M2 (obtained by removing EXP from M1). Hence M2 becomes the current best model. At the next step, we consider all models that can be obtained by adding or deleting a single variable from M2. These models are M1 (obtained by adding EXP) and M3, M4, and M5 (obtained by deleting the variables PTR, SAL, and PER respectively). All these models have higher AIC values than M2, indicating that this model cannot be improved by a single addition or deletion. The algorithm then terminates, returning M2 as the model of choice.

Summary

Linear regression models are among the most commonly used statistical methods. They combine widespread applicability with a highly developed theoretical basis. Functionality for fitting linear regression models exists in every significant statistics software package (including Minitab, SAS, SPSS, and R) and also in some spreadsheet packages (e.g., Excel).

This article has focused on regression modeling with quantitative explanatory variables. The analysis of models with categorical explanatory variables is covered in the article on ANOVA in this volume, although both regression and ANOVA type models can be unified with the

general framework of the linear model (e.g., Cohen, 1968). Regression models for clustered data and other hierarchical data structures are described in the article on hierarchical linear model in this volume. Regression models for highly non-normal response variables (e.g., binary responses) are covered in the article on generalized linear models in this volume.

See also: Analysis of Variance; Generalized Linear Models; Hierarchical Linear Models; Time Series Analysis.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Belsley, D. A., Edwin, K., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

Christie, M. (2004). Data collection and the ozone hole: Too much of a good thing? *Proceedings of the International Commission on History of Meteorology* **1.1**, 99–105.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin* **70**, 426–443.

Courville, T. and Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement* **61**(2), 229–248.

Guber, D. L. (1999). Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education* **7**(2). http://www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm (accessed May 2009).

Hsu, T.-C. (2005). Research methods and data analysis procedures used by educational researchers. *International Journal of Research and Methods in Education* **28**(2), 109–133.

Jørgensen, B. (1993). *The Theory of Linear Models.* London: Chapman and Hall.

Kariya, T. and Kurata, H. (2004). *Generalized Least Squares.* Chichester: Wiley.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics* **15**, 661–675.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika* **37**, 149–157.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**(2), 461–464.

Takezawa, K. (2006). *Introduction to Nonparametric Regression.* Hoboken, NJ: Wiley-InterScience.

Further Reading

Chatterjee, S., Hadi, A., and Price, B. (2000). *Regression Analysis by Example.* New York: Wiley.

Chatterjee, S. and Yilmaz, M. (1992). A review of regression diagnostics for behavioral research. *Applied Psychological Measurement* **16**(3), 209–227.

Chen, X., Ender, P., Mitchell, M., and Wells, C. (2003). *Regression with SPSS* . http://www.ats.ucla.edu/stat/spss/webbooks/reg/default.htm (accessed May 2009).

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B* **45**, 311–354.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304–1308.

Kim, K. and Timm, N. (2006). *Univariate and Multivariate General Linear Models Theory and Applications with SAS,* 2nd edn. New York: CRC Press.

Miller, A. (2002). *Subset Selection in Regression,* 2nd edn. London: CRC Press.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis,* 3rd edn. Hoboken, NJ: Wiley-Interscience.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research,* 3rd edn. Ft.Worth, TX: Harcourt Brace.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

StatSoft Inc. (2007). General linear models (GLM). *Electronic Statistics Textbook.* http://www.statsoft.com/textbook/stglm.html (accessed May 2009).

Weisberg, S. (2005). *Applied Linear Regression,* 3rd edn. Hoboken, NJ: Wiley-Interscience.

Wikipedia Contributors (2005). Ordinary least squares. *Wikipedia, the Free Encyclopedia.* http://en.wikipedia.org/w/index.php?title=Ordinary_least_squares&oldid=17317645 (accessed May 2009).

Wikipedia Contributors (2008). Linear regression. *Wikipedia, the Free Encyclopedia.* http://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=216828588 (accessed May 2009).