METHODOLOGY

Understanding linear and logistic regression analyses

Andrew Worster, MD, MSc;[*†] Jerome Fan, MD;[*] Afisi Ismaila, MSc[†]

SEE RELATED ARTICLE PAGE 105

Regression analysis, also termed regression modeling, is an increasingly common statistical method used to describe and quantify the relation between a clinical outcome of interest and one or more other variables. In this issue of *CJEM*, Cummings and Mayes used linear and logistic regression to determine whether the type of trauma team leader (TTL) impacts emergency department (ED) length-of-stay or survival.[1] The purpose of this educational primer is to provide an easily understood overview of these methods of statistical analysis. We hope that this primer will not only help readers interpret the Cummings and Mayes study, but also other research that uses similar methodology.

Linear regression

The most common type of regression analysis is linear regression which, as its name implies, assumes that a linear relation exists between the dependent variable (i.e., the variable that the researchers are trying to predict, also known as the response or outcome variable) and the independent variable(s) that the researchers choose to evaluate (i.e., the known or hypothesized predictor variable[s]). Linear regression produces a mathematical equation (or "model") for a "best fit" line to describe the relation. This can be portrayed visually by a scatter plot with a line running through it as shown in Figure 1.[2]

To be suitable for linear regression, the outcome of interest must be a "continuous" variable (that is, a variable with a continuous numerical range rather than a categorical outcome). For example, a researcher could evaluate the potential for injury severity score (ISS) to predict ED length-of-stay by first producing a scatter plot of ISS graphed against ED length-of-stay to determine whether an apparent linear relation exists, and then by deriving the best fit straight line for the data set using linear regression carried out by statistical software. The mathematical formula for this relation would be: ED length-of-stay = k(ISS) + c. In this equation, k (the slope of the line) indicates the factor by which length-of-stay changes as ISS changes and c (the "constant") is the value of length-of-stay when ISS equals zero and crosses the vertical axis.[2] In this hypothetical scenario, the overlap of the line and the data points in Figure 2

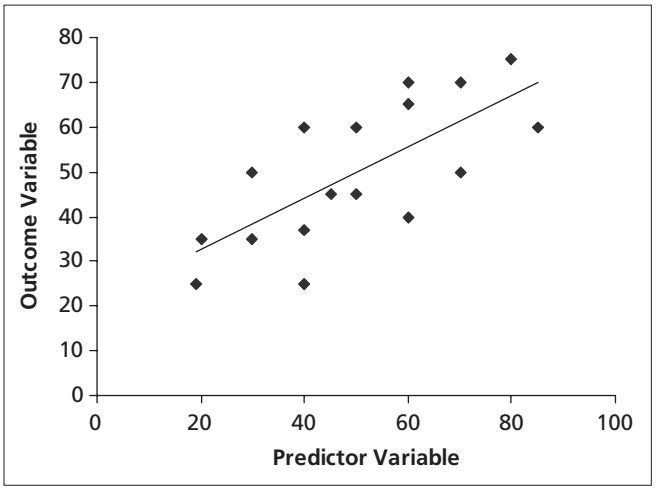

Fig. 1. A scatter plot depicting the linear relation between the outcome and predictor variables.

From the *Division of Emergency Medicine, McMaster University, Hamilton, Ont. and the †Department of Clinical Epidemiology, McMaster University, Hamilton, Ont.

Received: Feb. 1, 2007; accepted: Feb. 2, 2007

This article has not been peer reviewed.

Can J Emerg Med 2007;9(2):111-3

demonstrate a perfect positive relation between the 2 variables, indicating that ISS can be used as a single predictor of ED length-of-stay. Suppose, however, that ISS has no impact on ED length-of-stay. The line would then have a slope of zero, as shown in Figure 3, and the intersection of the line itself with the Y axis would represent the mean length-of-stay value (c = 14.3). Figure 4 represents a situation in between these 2 extremes and suggests that while ISS likely impacts ED length-of-stay, the scatter of data points around the best fit line makes it is difficult to know if the apparent magnitude of the relations is real. Linear regression analysis addresses this by measuring the proportion of the variability in ED length-of-stay that the model, represented by the fitted line, explains.[3]

Simple linear regression is used when there is only a single continuous predictor variable and a single continuous outcome variable. Multivariate (or multiple) linear regression is used to produce a model with 2 or more continuous or categorical predictor variables and a continuous outcome variable. For example, a researcher might also want to determine whether sex (a categorical variable) and age (a continuous variable) are predictors of ED length-of-stay. By incorporating these variables, along with ISS, multivariate linear regression can produce a more complete model and, thus, a better understanding of the independent impact of different predictor variables on the outcome, as well as any potential interaction between the predictor variables themselves.[4]

Some of the important terms in the statistical outputs of a linear regression analysis include:

- coefficient of correlation (R), a dimensionless quantity ranging from –1 to +1 that describes the strength of the association between 2 variables (R = 1 in Figure 2, R = 0 in Figure 3, and R = 0.28 in Figure 4);
- coefficient of determination (R^2), another dimensionless quantity ranging from 0% to 100%, that measures the proportion of variation in the outcome variable explained by the regression model (R^2 = 100% in Figure 2, R^2 = 0% in Figure 3, and R^2 = 8% in Figure 4) and;
- analysis of variance (ANOVA), a global test of significance of linear association in which $p < 0.05$ generally implies a linear association between outcome and predictor variables.

Linear regression analysis also provides estimates of regression variables and test of significance for each variable.

Logistic regression

Logistic regression is similar to multivariate linear regression in that it creates a model to describe the impact of multiple predictors on a single response variable. However, in logistic regression, the outcome variable must be categorical (usually dichotomous, i.e., with 2 possible out-

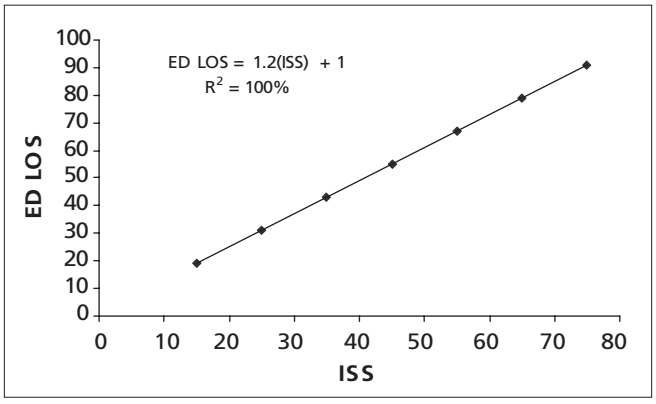

Fig. 2. A perfect positive relation between the Injury Severity Score (ISS) and emergency department (ED) length-of-stay (LOS) (i.e., the ISS can be used as a single predictor of ED LOS).

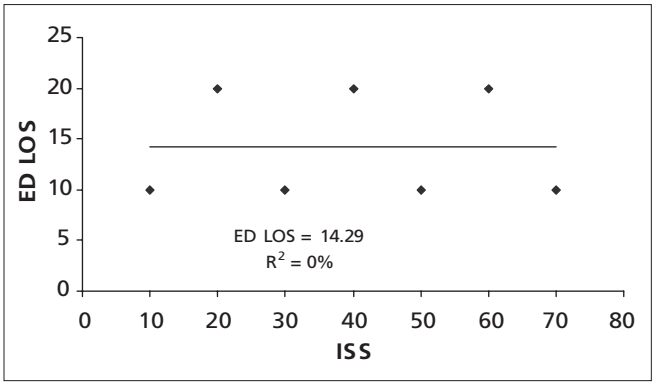

Fig. 3. With a slope of 0, the Injury Severity Score has no impact on emergency department length-of-stay (LOS).

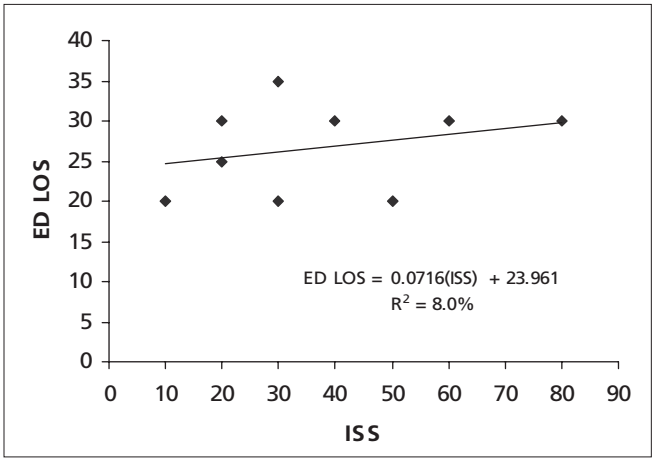

Fig. 4. This data set suggests that while Injury Severity Score likely has an impact on emergency department length-of-stay (LOS), the apparent magnitude of the relation is uncertain.

https://doi.org/10.1017/S1481803500014883 Published online by Cambridge University Press

comes, such as death or survival, although special techniques allow further categorized data to be modelled). A continuous outcome variable can be converted to a categorical one in order for logistic regression to be used, but collapsing continuous variables in this manner is generally discouraged as it reduces precision. If this is done, the categories should not be arbitrary but rather must make clinical sense. Unlike linear regression, the predictor variables in logistic regression do not need to be linearly related, normally distributed or have equal variance within each group. Because the relation between the predictor and outcome variables is not presumed to be a linear function, the measure of association between the outcome of interest and the predictor variable is represented by an odds ratio (OR) instead of a multiplicative factor. Comparisons of ORs between predictor variables help determine the factors of greatest importance, while their confidence intervals indicate their statistical significance.

Limitations

Readers should be cautious when interpreting the results of regression analyses, and several issues must be considered. First, researchers occasionally mistake a predictor variable for the outcome variable or vice versa (and so the model must make clinical sense).[5] Second, the determination of a statistically significant relation between the predictor and outcome variables does not necessarily mean that a causal relation exists.[5] Third, there may be other predictor variables that were not included in the model that have equal or greater impact on the outcome variable. Finally, the predictive ability of a model does not apply to data outside the range of the data from which the model was derived.[5] It should also be noted that there are several important statistical assumptions essential for the correct application of regression techniques that are beyond the scope of this review.

Cummings and Mayes study

Since ED length-of-stay, the first outcome in the Cummings and Mayes study, is a continuous variable, the authors used multivariate linear regression analyses to determine whether it is impacted by any of the studied predictor variables: age, sex, ISS, revised trauma score (RTS) and TTL. However, because age, sex and RTS were not found to be significant factors ($p > 0.05$), they were removed from the final model. The researchers were then able to determine that TTL has no significant impact on ED length-of-stay ($p = 0.37$). They assessed whether the same predictor variables had any impact on survival, a categorical variable. To answer this they used logistic regression. Again, the researchers found that TTL has no significant impact ($p = 0.58$). This was further demonstrated by the ORs for each of the TTL categories, the confidence intervals of which included a value of one.

In considering the Cummings and Mayes study, the authors appear to have correctly identified which variables were predictors and which were outcomes. However, it remains conceivable that there might be other predictor variables, not included in the model, that have a statistically significant impact on the response variables. For example, hospital overcrowding, resource availability and quality of care indicators could all be influential. As previously mentioned, failure to include all possible predictors in a study can weaken a regression model.

Summary

While the mathematical concepts and details of statistical tests can appear complex, most of the issues involved are straightforward. By understanding some basic statistical concepts and the role of the different statistical tests, one can usually determine whether the method of analysis for any given study was appropriate. Regression analysis is a powerful statistical method to determine which variables are predictors of an outcome and the magnitude of that relation. Linear and logistic regression simply differ in the type of outcome and predictor variables they employ, and the type of results they produce.

Competing interests: None declared

Key words: regression analysis, linear regression, logistic regression

References

1. Cummings GE, Mayes DC. A comparative study of designated trauma team leaders on trauma patient survival and emergency department length of stay. Can J Emerg Med 2007;9:105-10.
2. Marill KA. Advanced statistics: linear regression, part I: simple linear regression. Acad Emerg Med 2004;11:87-93.
3. Glover T, Mitchell K. An introduction to biostatistics. 1st ed. New York (NY): McGraw-Hill; 2002.
4. Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. Acad Emerg Med 2004;11:94-102.
5. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology 2003;227:617-22.

Correspondence to: Dr. Andrew Worster, Division of Emergency Medicine, McMaster University Medical Centre, 1200 Main Street West, Hamilton ON L8N 3Z5; aworster@rogers.com