CHEST®

Official publication of the American College of Chest Physicians

Documenting Research in Scientific Articles: Guidelines for Authors*

Tom Lang

Chest 2007;131;628-632
DOI 10.1378/chest.06-2088

CHEST Postgraduate Education Corner

MEDICAL WRITING TIPS

Documenting Research in Scientific Articles: Guidelines for Authors*

3. Reporting Multivariate Analyses

Tom Lang, MA

(CHEST 2007; 131:628–632)

Multivariate analyses include two broad statistical techniques, regression analysis and analysis of variance (ANOVA). The reporting guidelines for each are similar and here have been condensed from the book *How To Report Statistics in Medicine*.[1]

Reporting Regression Analysis

Regression analysis attempts to predict or estimate the value of a response variable or outcome from the known values of one or more explanatory variables or predictors. The type of regression analysis is determined by the number of explanatory (or independent) variables and of the response (or dependent) variables, as well as by the "level of measurement" of these variables.

The phrase *level of measurement* refers to the kind of information collected about a variable. Nominal data are categorical data with no inherent ranking, such as blood type (*eg*, A, B, AB, and O); ordinal data are categorical data that do have an inherent ranking, such as severity categories (*eg*, mild, moderate, and severe); and continuous data are measurements made on a continuous scale of equal intervals. The level of measurement can also be set by the researcher. For example, data on BP can be collected as a nominal variable (hypertensive or not hypertensive), an ordinal variable (hypotensive, normotensive, or hypertensive) or a continuous variable (systolic BP measured in millimeters of mercury.)

The most common types of regression analyses are as follows:

- Simple linear regression is used to assess the relationship between a single continuous explanatory variable and a single continuous response variable that varies linearly over a range of values.
- Multiple linear regression is used to assess the linear relationship between two or more continuous or categorical explanatory variables and a single continuous response variable.
- Simple logistic regression is used to assess the relationship between a single continuous or categorical explanatory variable and a single categorical response variable, usually a binary variable, such as whether or not a heart attack has occurred.
- Multiple logistic regression is used to assess the relationship between two or more continuous or categorical explanatory variables and a single categorical response variable.
- Nonlinear regression is used to assess variables that are not linearly related and that cannot be transformed into a linear relationship. These equations model more complex relationships than the other forms of regression analysis.
- Polynomial regression can be used for any of the above combinations of explanatory and response variables when the relationship among the variables is curvilinear, which requires, say, squaring or cubing one or more explanatory variables in the model.
- Cox proportional hazards regression, an aspect of time-to-event (survival) analysis, is used to assess the relationship between two or more continuous or categorical explanatory variables and a single continuous response variable (the time to the event). Typically, the event (usually death) has not yet occurred for all participants in the sample, which creates censored observations.

*From Tom Lang Communications and Training, Davis, CA.
The author receives royalties from the sale of *How To Report Statistics in Medicine*, from which this article is taken. He has no other conflicts of interest with the publication of this article.

Manuscript received August 21, 2006; revision accepted August 24, 2006.

Correspondence to: Tom Lang, MA, 1925 Donner Ave, No. 3, Davis, CA 95618; e-mail: tomlangcom@aol.com

DOI: 10.1378/chest.06-2088

Guideline: Describe the Relationship of Interest or the Purpose of the Analysis

In addition to predicting one value from one or more others, regression analysis can be used to "control for" the potential confounding effects of explanatory variables that are associated with the response variable. Regression analysis can separate the effects of, say, age and sex on survival after surgery, for example.

Regression analysis can also be used to create risk scores. Here, the variables of the risk score are those of the regression equation, and the score itself is the value predicted by the regression model.

Guideline: Identify the Variables Used in the Analysis and Summarize Each With Descriptive Statistics

Continuous variables should be summarized with medians and ranges or interquartile ranges (or means and SDs if the data are normally distributed), and categorical data can be summarized with counts or percentages.

Guideline: Confirm That the Assumptions of the Analysis Were Met and State How Each Was Checked

A statement that the assumptions were verified and by which methods is all that need be included. There are both formal checks (*eg*, hypothesis tests) and informal checks (*eg*, inspection of graphs of residuals) for these assumptions. Sometimes, data that violate the assumptions can be adjusted (*eg*, with data transformations) to meet the assumptions. If such adjustments were made, they should be identified.

Guideline: Report How Any Missing Data Were Treated in the Analyses

Missing data can be a problem in multivariate analysis because it reduces the sample size unless corrective measures are taken. To create a model for predicting weight from age and height, for example, values for each of these variables must be collected for each patient. If age is missing from one patient, the patient is excluded from the analysis, and the sample size is reduced by one. In regression models with several variables, losses to missing data can be common.

$$\boxed{Y = 40.8 + 3.98X_1 + 1.22X_2 - 2.09X_3}$$

Figure 1. A multiple linear regression equation. In this example, the model predicts overall function score, Y, for patients with multiple sclerosis based on: disease severity, X_1; ambulatory ability (measured as the rate of walking in laps per minute), X_2; and number of lesions, X_3. Here, X_1, X_2, and X_3 are explanatory variables (sometimes called *risk factors*); the numbers in front of the X values are called regression *coefficients* or β-*weights*. (40.8 is the Y intercept point, where the line crosses the Y axis.) Coefficients are interpreted as follows: if X_1 and X_3 are held constant (or "controlling for" disease severity and number of lesions), then mean functional score increases by about 1.25 times (1.22, the coefficient for X_2) for each additional lap per minute. The final model had a coefficient of multiple determination, R^2, of 0.58, indicating that the three variables in the model explain 58% of the variation in the response variable.

However, missing data can be replaced in a process called *imputation*. Simple imputation methods include using the mean of all observed values for all people in place of the missing value; using the mean observed value for the same person in other time periods; using the mean of the previous and following values for the person, if they exist; or using the most recent observed value for the person (called *the last-observation-carried-forward method*, which is commonly used in pharmaceutical research). Other methods of imputing data are possible, but they should be based on sound judgment.

Guideline: Report How Any Outlying Values Were Treated in the Analysis

Outliers are extreme values that appear to be anomalies. Outliers cannot be ignored: even a single outlier can have a profound effect on the relationship derived from the regression line.[2,3] All outliers must be reported, but it is permissible to report the results with and without the outliers to indicate their effect on the results.

Guideline: Report the Regression Model

A simple linear regression equation can be reported in the text or in a scatter plot of the data. Multiple linear regression models can be reported as equations (Fig 1) or in tables (Table 1); logistic regression models are typically reported in tables because the equations are so complex (Table 2).

Guideline: Report the actual p Value and the 95% Confidence Interval for the Regression Coefficient(s) of the Explanatory Variable(s), and in Logistic Regression, Report the Odds Ratio and the Associated 95% Confidence Interval

In regression analysis, the regression coefficient for an explanatory variable indicates how much the

Table 1—*A Table for Reporting a Multiple Linear Regression Model With Three Explanatory Variables**

Variables	Coefficient (β)	SE	95% CI	Wald χ^2	p Value†
Intercept	40.79	2.55			
X_1	3.98	2.37	− 0.67 to 8.63	1.68	0.10
X_2	1.23	0.29	0.66 to 1.80	4.20	< 0.001
X_3	− 2.09	0.28	− 2.64 to − 1.54	− 7.34	< 0.001

*Intercept = a mathematical constant (no clinical interpretation); X_1 to X_3 = the explanatory variables; Coefficient = the mathematical weightings of the explanatory variables in the equation (the regression coefficient or β -weight); SE = estimated precision of the coefficients; 95% CI = 95% confidence intervals for the coefficients; Wald χ^2 = the Wald test statistic calculated from the data to be compared with the χ^2 distribution with 1 degree of freedom.

†Variables X_2 and X_3 are statistically significant independent predictors of the response variable.

average value of the response variable, Y, varies with each unit change in the explanatory variable, X. The coefficient, or β-weight, is an estimate and so should be accompanied by a confidence interval that indicates its precision.

Odds ratios are widely used in logistic regression analysis. For a binary explanatory variable, the odds ratio is the ratio of the odds that an event will occur in one group to the odds that the event will occur in the other group. An odds ratio of 1 means that both groups have a similar likelihood of having a heart attack. The larger the odds ratio, the more likely the event is expected to occur in the group used in the numerator.

Guideline: Specify How the Explanatory Variables That Appear in the Final Regression Model Were Chosen

One of the first steps in building a multiple regression model is to identify the explanatory variables that are significantly related to the response variable.[4] Several dozens of variables may be considered one at a time in this process, called *univariate analysis*. Often, a less-restrictive α-level, such as 0.1, is used in the univariate analysis to identify a broad range of explanatory variables that might be associated with the response variable. That is, variables with p values less than 0.1 on univariate analysis are considered for inclusion in the model.

The second step in building a regression model is to identify the best combination of explanatory variables to include in the model. In simultaneous regression, all of the explanatory variables are included in the model and are tested as a group. In hierarchical regression, the investigator defines the number and order in which the explanatory variables are entered into the model. Common procedures are forward, backward, stepwise, and best-subset techniques.

Guideline: In Multiple Regression Models, Specify Whether All Potential Explanatory Variables Were Assessed for Collinearity (Nonindependence)

The explanatory variables in a multiple linear regression equation should be independent of one another.[4] If two or more explanatory variables are correlated, that is, if their regression lines are parallel or “collinear,” then they are not independent. Collinear variables add much the same information to the model, so only one is needed. The variable with the strongest relationship with the response variable should be considered for inclusion in the final model.

Table 2—*A Table for Reporting a Multiple Logistic Regression Model With Four Explanatory Variables**

Variable	Coefficient (β)	SE	Wald χ^2	p Value	Odds Ratio	95% CI
Intercept	− 1.88	0.48				
X_1	1.435	0.589	5.93	0.02	4.2	1.32–13.33
X_2	− 0.847	0.690	1.51	0.22	0.43	0.11–1.66
X_3	3.045	1.260	5.84	0.02	21.01	1.78–248.29
X_4	2.200	0.990	4.94	0.03	9.03	1.30–62.83

*Odds Ratio = controlling for other variables in the model, for every unit increase in, for example, variable 1, the odds of having the event of interest increase by 4.2 (likewise, controlling for other variables in the model, for every unit increase in, for example, variable 2, the odds of having the event decrease by 0.43); 95% CI = the 95% confidence interval for the estimated odds ratio. See Table 1 for other abbreviations or explanations not used in the text.

Guideline: In Multiple Regression Models, Specify Whether the Explanatory Variables Were Tested for Interaction

Two explanatory variables are said to interact if the effect of one explanatory variable on the response variable depends on the level of the second explanatory variable. Interaction implies that the variables should be considered together, not separately. So, for example, if alcohol interacts with antibiotics in the blood, the model should have a variable for blood alcohol level, one for blood antibiotic level, and an interaction term that expresses the relationship between serum alcohol and antibiotic level.

Guideline: Provide a Measure of the "Goodness of Fit" of the Model to the Data

The predictive value of a regression model is affected by how well it "fits" the data.[5,6] Thus, a measure of goodness of fit is useful because it reveals how well the model reflects the data on which it was created.

Simple linear regression analysis can be thought of as an extension of correlation analysis, except that now one variable is being used to predict the other with the addition of a regression line. As in correlation analysis, scatter plots can be useful for showing this relationship. The correlation coefficient itself can indicate indirectly how well the model can predict. Correlations have to be high, say, above 0.7, as well as statistically significant, if a simple linear regression model is to predict with any degree of accuracy.

In simple linear regression analysis, the correlation coefficient associated with the scatter plot is also useful in the form of the coefficient of determination (r^2). This coefficient indicates how much of the variability in the response variable is explained by the explanatory variable. For example, if the correlation between skin-fold thickness and body fat is 0.8, then $r^2 = 0.64$, or 64%. That is, 64% of the variability in body fat can be accounted for by skin-fold thickness. In multiple linear regression analysis, the coefficient of multiple determination (R^2) has the same function.

A residual is the difference between the value predicted by the model and the actual value of the data point as collected. The smaller the residual, the better the prediction. Residuals can also be graphed to determine how well the assumption of linearity was met. Thus, a graph of residuals (one kind of "model diagnostic plot") in which the values are small for all values of X, meaning that they stay close to an average difference of zero, indicates that the assumption of linearity was met and that the model predicts reasonably well. Outlier assessments work the same way as residual assessments, in that they and their associated residuals are apparent on the graph as data points to investigate.

Formal goodness-of-fit tests calculate a p value. If the p value is statistically significant, the model does not appropriately fit the data.

Guideline: Specify Whether the Model Was Validated

Regression models can be validated or tested against a similar set of data to show that they explain what they seek to explain. One method used when the sample is large is to develop the model on, say, 75% of the data, then to create another model on the remaining 25% of the data, and determine whether the models are similar. Another method involves removing the data from one subject at a time and recalculating the model. The coefficients and the predictive validity of all the models can then be assessed. Such methods are called *jack-knife procedures*. A third method involves developing another model on a separate set of similar data and determining whether the models differ.

Guideline: Name the Statistical Package or Program Used in the Analysis

Although commercial statistical programs generally are validated and updated, and have met the test of time, the performance characteristics of privately developed programs are often unknown.

Reporting ANOVA

ANOVA is a form of hypothesis testing for studies involving two or more variables. It is closely related to regression analysis and should be reported according to the same general guidelines. Usually, ANOVA is used to assess categorical explanatory variables, whereas regression analysis is used to assess continuous explanatory variables. When a study includes both continuous and categorical explanatory variables, the analysis may be called multiple regression or analysis of covariance.

ANOVA is a "group comparison" that determines whether a statistically significant difference exists somewhere among the groups studied. If a significant difference is indicated, ANOVA is usually followed by a multiple comparison procedure that compares combinations of groups to examine further any differences among them.

Table 3—*A Table for Presenting the Results of a Two-Way ANOVA for Analyzing the Two Factors Group and Age**

Source of Variation	df†	Sums of Squares	Mean Square	F Statistic	p Value
Group	1	0.64	0.64	2.24	0.16
Age	3	3.92	1.31	4.57	0.02
Group × age	3	4.91	1.64	5.72	0.01
Error	12	3.43	0.29		

*ANOVA = includes the two factors: group (two levels or categories) and age (four categories or levels), and the levels of each category should be stated in the description of the study (group and age significantly interact and so must be considered together); Source of variation = identification of the sources of variability in the response variable as the factors in the model (group, age, and the interaction between group and age) and as random error (the variability not explained by the factors); df = the degrees of freedom, a mathematical concept; Sums of squares = unlike one-way ANOVA, the sums of squares in multiway ANOVA are not easily explained and are best regarded as simply steps in the calculation of the mean squares; Mean square = the sums of squares divided by the degrees of freedom (essentially, estimates of the variation in the data); F statistic = the test statistic for the F distribution, for testing for interaction effects and main effects, equals the mean square for each factor divided by the mean square of the error; p Value = the probability values indicating the statistical significance of the effect of each factor on the response variable (*eg*, age and group interact [p = 0.01] in affecting the response variable and should be further investigated together; *ie*, the main effect of group or the main effect of age should not be investigated alone).

†For two groups, the df is 2 − 1, or 1. For four age categories, the df is 4 − 1, or 3. For the interaction effect between group and age (*ie*, group × age), the df values for each factor are multiplied (3 × 1 = 3).

The most common ANOVA procedures used in biomedical research are as follows:

- One-way ANOVA assesses the effect of a single (hence the "one-way" designation) categorical explanatory variable (sometimes called a *factor*) on a single continuous response variable. Note, too, that the factor (category) has three or more alternatives (or "levels" or "values"; *eg*, blood type is A, B, AB, or O). When there are only two alternatives (two groups), this analysis reduces to Student *t* test.
- Two-way ANOVA assesses the effect of two categorical explanatory variables (again, sometimes called *factors*) on a single continuous response variable.
- Multiway ANOVA assesses the effect of three or more categorical explanatory variables (still called *factors*) on a single continuous response variable.
- Analysis of covariance assesses the effect of one or more categorical explanatory variables while controlling for the effects of some other (possibly continuous) explanatory variables (now called *covariates*) on a single continuous response variable.
- Repeated-measures ANOVA is used to assess several, or repeated, measurements of the same participants under different conditions (such as BP measurements taken while the patient is supine, sitting, or standing) or at different points over time (such as muscle strength measured 1, 5, 10, and 20 days after surgery).

ANOVA is typically used to compare three or more group means on a certain response variable. It can also be expanded to include additional explanatory variables and can assess their simultaneous effects on the response variable. Whereas the purpose of regression analyses is usually to predict the value of the response variable, the purpose of ANOVA is usually to compare groups for differences in the means of the response variable. ANOVA models are also usually reported in tables (Table 3).

ACKNOWLEDGMENT: This article draws heavily from *How To Report Statistics in Medicine*, by Tom Lang and Michelle Secic.[1]

References

1 Lang T, Secic M. How to report statistics in medicine. 2nd ed. Philadelphia, PA: American College of Physicians, 2006
2 Godfrey K. Simple linear regression in medical research. In: Bailar JC, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, MA: NEJM Books, 1992; 201–232
3 Altman DG, Gore SM, Gardner MJ, et al. Statistical guidelines for contributors to medical journals. BMJ 1983; 286: 1489–1493
4 Shutty M. Guidelines for presenting multivariate statistical analyses in rehabilitation psychology. Rehabil Psych 1994; 39:141–144
5 Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 2001; 54:979–985
6 Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. Am J Public Health 1991; 81:1630–1635

Documenting Research in Scientific Articles: Guidelines for Authors*

Tom Lang

Chest 2007;131; 628-632
DOI 10.1378/chest.06-2088