JOURNAL SERIES:

# Statistics Review Part 11: Linear and Logistic Regression

by Russell Findlay, PharmD, Jordan Rush, PharmD, Kevin Look, PharmD, PhD, and Amanda Margolis, PharmD, MS, BCACP

## Objectives

1. Recognize and explain the differences between linear and logistic regression.
2. Identify how to interpret linear and logistic regression results in clinical studies.
3. Explain the role of regression in minimizing bias due to confounding variables.

Determining the strength and direction of the relationship between independent (i.e., exposure) and dependent (i.e., outcome) variables is vital when interpreting the results of study data. There are a number of statistical tools that can quantify the strength of this correlation and guide the reader to appropriate interpretation and application of results. This article builds off of the previous article in this series (Part 10 on causality and confounding). Regression is the standard statistical technique used to understand how independent variables affect outcome variables, and is used widely in published literature. This article reviews two main types of regression, linear and logistic regression, the interpretation of regression results in clinical studies, and gives examples that illustrate the appropriate application of each approach. It is important to reiterate that a statistically significant correlation between the independent and dependent variables does not imply causation.

At the outset, it is important to be able to determine when it is appropriate to use linear or logistic regression. The key is to correctly identify the type of dependent variable present. Linear regression is used in situations where the dependent variable is continuous and logistic regression is used in situations where the dependent variable is binary.

Both approaches help researchers control and account for the impact of confounding variables. Estimates obtained in a study may be biased because of potential measured and unmeasured factors. These extraneous variables can obscure the clear relationship between variables and cause variability in responses.[1] Regression techniques enable researchers to account for measured potential confounders. Of note, unmeasured confounders cannot be accounted for using simple regression techniques. Researchers can account for unmeasured confounders effects through randomization, literature evaluation, or by using advanced statistical methods that are beyond the scope of this article.

## Linear Regression

Linear regression attempts to fit a straight line to the data that best approximates the relationship between the independent and dependent variables. Linear regression analysis can be used to quantify the strength of the relationship between two variables to help understand how changes in the independent (or input) variable(s) affect the dependent (or output) variable. In other words, for every "x unit" increase in the independent variable, there is a "y unit" change (decrease or increase) in the dependent variable. This relationship is linear and has a defined, constant rate of change and is often identified as a regression line, estimate, or fit line.

The $r^2$ is a dimensionless measure (i.e., it lacks units such as milligrams) that provides a sense of how well the data fit a statistical model (also called "goodness of fit"). That is, it tells you how well the regression line fits the data points on a scale from 0 to 1. An $r^2$ of 1 indicates that the regression line perfectly fits the data.

Consider an example where there is one independent variable and one dependent variable, such as the relationship between salt intake and blood pressure. It is well known that excessive salt intake can increase blood pressure.[2] In simple linear regression, the amount of salt consumed would positively correlate with a rise in blood pressure, meaning individual blood pressure readings would increase in proportion to salt intake. The data points would fit to a straight line with a positive slope. If you estimated the association between salt intake and blood pressure and found an $r^2$ value of 0.81, this would mean the regression model explained 81% of the

**TABLE 1. Multiple Linear Regression Results of Factors Related to LDL Levels**

| Variable | Effect Size | 95% Confidence Interval | P-value |
|---|---|---|---|
| **Age** | -2.5 | (-5.5, 0.5) | 0.08 |
| **BMI** | | | |
| <30 | -12.5 | (-22.5, -2.5) | 0.04 |
| 30-40 | reference | reference | reference |
| >40 | 30 | (37, 23) | 0.002 |
| **Moderate Intensity Statin** | -53 | (-44, -62) | 0.000 |
| **High Intensity Statin** | -82 | (-74, -90) | 0.000 |

total variance and represents an excellent fit.[3] However, often times the $r^2$ values seen in clinical studies are considerably lower, as there are many different factors that can influence a patient's health or behavior.

Linear regression can also be used when there is more than one independent variable that may affect a single dependent variable. This is one approach that can be used to control for potential confounding variables that may bias the results. Referred to as multiple linear regression, it is able to account for the effects of multiple independent variables at one time by measuring the separate effect of each independent variable while holding all other independent variables constant. Multiple linear regression attempts to fit a linear equation to the observed data (as in single variable regression) but this time there are multiple relationships that may explain the variability of the data.[4]

Table 1 provides an example of how results may be presented in a theoretical study looking at the effect of age, body mass index (BMI), use of a moderate intensity statin, and use of a high intensity statin on low-density lipoprotein (LDL) levels. In this fictitious example, there are continuous variables (i.e., age) and categorical variables (i.e., type of statin used, categories of BMI). The coefficient or effect size estimate of -2.5 for age means that for each additional year of age, LDL decreases by 2.5mg/dL, holding BMI and statin use constant. This effect is not statistically significant based on the *a priori* significance value of 5%. [For more on significance testing, refer to Part 4 of this series.] In contrast, the coefficient on BMI >40 suggest that a BMI greater than is 40 is associated with an LDL increase of 30 mg/dL, holding the other variables constant, and constitutes a statistically significant result. Of note, researchers rarely, if ever, publish the results of a whole continuous regression model. Instead, authors routinely present only the main findings of interest and a list of control variables; full regression models may be included only in appendices.

## Logistic Regression

Logistic regression is used to predict the effect of one or more independent variables on a binary or dichotomous dependent variable (i.e., having one of two possible outcomes, such as 0=no, 1=yes). Although logistic regression uses a different statistical process than linear regression, it can still identify the effect of an independent variable on a dependent variable while controlling for potential confounders. One major difference from linear regression is that the effect sizes in logistic regression are reported in the form of an odds ratio (OR). An odds ratio is defined as the odds that an outcome will occur given an exposure, compared to the odds the outcome will occur in the absence of that exposure.[5]

For example, a research study by Farland et al. assessed the impact of β-blocker use on the incidence of exacerbations in patients with chronic obstructive pulmonary disease (COPD).[6] These factors include beta-blocker use (OR: 0.61, 95% CI: 0.40-0.93) and beta-blocker cardioselectivity (OR: 0.84, 95% CI: 0.38-1.83). An OR value less than one indicates a reduced odds compared to placebo, meaning there is 39% reduced odds of a patient on a beta-blocker experiencing a COPD exacerbation controlling for covariates. Other covariates controlled for in the study were having diabetes, use of an aldosterone antagonist, use of inhaled corticosteroids, and provider type. This is a statistically significant difference because the OR does not include 1. However, it seems there is no statistically significant relationship between cardioselectivity and the odds of a patient experiencing a COPD exacerbation controlling for covariates. Although the researchers found a 16% reduction in the odds of a COPD exacerbation, this relationship was not statistically significant because the confidence interval included 1.

## Summary

Regression is a statistical tool used to understand the relationships between independent variables and their impact on dependent variables. Two commonly used types of regression models are linear and logistic regression. Linear regression is used for continuous dependent variables, whereas logistic regression is used when the dependent variable is dichotomous. Regardless of which type of regression is used, it is a statistical technique that enables researchers to account for measured potential confounders during data analysis.

## Practice Questions

1. What type of analysis would be most appropriate to analyze the relationship between the amount of time a person studies for an exam and whether or not the individual passes the exam?
   a. Linear regression
   b. Logistic regression
   c. T-test
   d. Chi-square test

2. In a study examining the relationship between the number of medications a patient is taking and the likelihood of medication adherence, the $r^2$ value for the regression line is 1.0. What does this indicate?
   a. For each additional medication a patient is taking, adherence

increases by 1.0.

b. Since $r^2$ equals 1.0, this indicates there is no relationship between the number of medications a person is taking and medication adherence.

c. Since $r^2$ equals 1.0, the regression line perfectly fits the data.

d. For each additional medication a patient is taking, there is no effect on adherence.

3. Suppose a study examined the relationship between working in a coal mine and the risk of developing asthma. If the study found an increased risk of asthma in the group who worked in the coal mines (OR: 1.6, 95% CI: 1.3-1.8). What does this mean?

a. Employees who worked in the coal mines were 1.60 times less likely to develop asthma than those who did not.

b. Employees who worked in the coal mines had a 60% decreased odds of developing asthma than those who did not.

c. The results are significant because the confidence interval does not contain 1.0.

d. The results are not significant because the confidence interval is >1.0.

### *Answers:*

1. **b** The amount of time the individual studies is the continuous independent variable. The dependent variable, whether or not the person passes the exam (yes or no), is dichotomous. Therefore, logistic regression is most appropriate.

2. **c** The $r^2$ value gives you the "goodness of fit" between the regression line and data points. An $r^2$ value of 1.0 indicates a perfect fit and that all data points fall on the regression line.

3. **c** The only time the result becomes non-significant is if the CI contains 1.0.

---

Russel Findlay and Jordan Rush are Senior Administrative Residents at the University of Wisconsin Hospital and Clinics, Madison, WI. Kevin Look is an Assistant Professor in the Social & Administrative Sciences Division at the University of Wisconsin School of Pharmacy, Madison, WI. Amanda Margolis is a Lecturer at the UW-Madison School of Pharmacy and a Clinical Pharmacist at the William S. Middleton Memorial Veterans Hospital, Madison, WI.

### *References and suggestions for further review*

1. Gaddis ML, Gaddis GM. Introduction to Biostatistics: Part 6, Correlation and Regression. Ann Emerg Med. 1990;19(12):1462-1468.
2. He FJ, Li J, Macgregor GA. Effect of longer-term modest salt reduction on blood pressure. Cochrane Database Syst Rev. 2013;4:CD004937. doi: 10.1002/14651858.CD004937.pub2.
3. Agresti A, Franklin CA. Statistics: The Art and Science of Learning from Data. 2nd ed. Upper Saddle River, NJ: Pearson Education; 2007.
4. Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med. 2003;138(8):644-650.
5. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010;19(3):227-229.
6. Farland MZ, Peters CJ, Williams JD, et al. β-Blocker use and incidence of chronic obstructive pulmonary disease exacerbations. Ann Pharmacother. 2013;47(5):651-656.

## New Members

***Welcome to the newest members of PSW (2/03/2015 to 4/13/2015)***

**Michael G. Barbiaux,** RPh, New London
**Nicholas Baulch,** Madison
**Stacy Bernd,** HP, Madison
**Karen L. Brimer,** RPh, Madison
**Jaime Cadotte,** Washburn
**Michael H. Cain,** RPh, Jim Falls
**David E. Call,** RPh, Bay Area Medical Center Pharmacy, Marinette
**Lauren Darling,** Meijer Pharmacy, Oak Creek
**Caitlin Carberry Drayna,** PharmD, Froedtert Health Hometown Pharmacy, West Bend
**Tina Durrant,** Meijer Pharmacy
**Diane Erdman,** PharmD, BCPS, BCACP, CDE, CPPS, Wheaton Franciscan, Milwaukee
**Megan M. Erickson,** PharmD, Mundelein
**Jodie Shawn Evans,** Union Grove
**Rachael Kathryn Gang,** Jefferson
**Timothy Gardner,** RPh, Bay Area Medical Center Pharmacy, Marinette
**Mary S. Genisot,** RPh, Marinette
**Matthew Glaser,** PharmD, Walgreens, Sun Prairie
**JoAnna L. Gollmer**, Cambridge
**Sara Jane Guth**, Birnamwood
**Lauren Elaine Hansen,** Racine
**Michael Hapka,** Glendale
**Hannah Hendrickson,** Ontario
**Andrew R. Hochradel,** Prescriptions Plus, West Allis
**Andrea Hudson,** Ashland
**Wendy Densie Humphrey,** Burlington
**Daniel Janke,** Walgreens, Deerfield
**Don Johnson,** Meijer Pharmacy, Grand Rapids
**Mary Johnson,** Meijer Pharmacy, Grand Rapids
**Chayani Kaenkumchorn,** Meijer Pharmacy, Milwaukee
**Jessica K. Kannemeier,** RPh, Community Pharmacy, Madison
**Kathleen King,** Village Pharmacy Inc., Coleman
**Laura Labreche,** Good Value Pharmacy, Kenosha
**Sherie Latva,** St Vincents Pharmacy, DePere
**Andrew Roger Lindloff,** New Richmond
**Adam Noah Lubin,** PharmD, Green Bay
**Cary B. Luedtke**, RPh, Manitowoc
**Man Ly,** Bay Area Medical Center Pharmacy, Marinette
**Ben Mak,** Bay Area Medical Center Pharmacy, Marinette
**Jess A. Mangold**, Combined Locks
**Ryan McFadden**, Waunakee
**Lisa Ann Miller,** Mequon
**Michelle N. Mitchell,** Brookfield
**Lisa Monks,** Meijer Pharmacy, Grafton
**Chelsey Novak,** Eannelli Pharmacy, Prairie Du Sac
**Bolaji K. Olasusi,** PharmD, Walgreens Pharmacy, Weston
**Kurt Ostmann,** RPh, Walgreens Pharmacy, Wausau
**H. Allan Peterson,** RPh, Marinette
**Restina Polovic,** RPh, Franklin
**Stephanie K. Roehrig,** PharmD, Waupaca
**Douglas J. Schara,** Davenport
**Geoffrey A. Schnelle,** Medford
**Linda Shaver,** Bay Area Medical Center Pharmacy, Marinette
**Craig R. Sherman,** Waukesha
**Mary Siegler,** Brookfield
**Sarah M. Stade,** Pleasant Prairie
**Effie Steele,** Meijer Pharmacy, Grand Rapdis
**Diane Valeo-Seabright,** Highland Park
**Anuja Vallabh,** Clement J. Zablocki VA Medical Center, Milwaukee
**Jennifer A. Wagner,** Janssen, Middleton
**Jason Watt,** Cardinal Health, Naperville
**Syreeta Williams,** Racine