ANESTHESIA & ANALGESIA® Statistical Minute

Related Article, see p 110

OPEN

Linear Regression in Medical Research

Patrick Schober, MD, PhD, MMedStat,* and Thomas R. Vetter, MD, MPH†

Linear Regression Analysis of Exhaled Versus Plasma and Tissue Propofol Concentrations

Exhaled Versus	Slope	95% Confidence Interval	Intercept	95% Confidence Interval	R^2
Plasma	4.6	3.6–5.7	35	9–61	0.71
Brain	2.7	2.1–3.2	0.5	−28 to 29	0.75
Lung	1.9	1.2–2.6	42	6–78	0.52
Liver	0.9	0.7–1.2	36	7–65	0.62
Kidney	2.4	2.0–2.9	17	−4 to 43	0.77
Muscle	1	0.4–1.6	76	37–114	0.25
Fat	0.2	0.1–0.2	14	−22 to 50	0.60

[Exhaled] = Slope × [plasma/tissue] + intercept; R^2 = coefficient of determination.

Slope = simple linear regression coefficient = b_1

Figure. Table 2 given in Müller-Wirtz et al,[1] showing the estimated relationships between tissue (or plasma) propofol concentrations and exhaled propofol concentrations. The authors appropriately report the 95% confidence intervals as a measure of the precision of their estimates, as well as the coefficient of determination (R^2). The presented values indicate, for example, that (1) the exhaled propofol concentrations are estimated to increase on average by 4.6 units, equal to the slope (regression) coefficient, for each 1-unit increase of plasma propofol concentration; (2) the "true" mean increase could plausibly be expected to lie anywhere between 3.6 and 5.7 units as indicated by the slope coefficient's confidence interval; and (3) the R^2 suggests that about 71% of the variability in the exhaled concentration can be explained by its relationship with plasma propofol concentrations.

KEY POINT: Linear regression is used to quantify the relationship between ≥1 independent (predictor) variables and a continuous dependent (outcome) variable.

In this issue of *Anesthesia & Analgesia*, Müller-Wirtz et al[1] report results of a study in which they used linear regression to assess the relationship in a rat model between tissue propofol concentrations and exhaled propofol concentrations (Figure).

Linear regression is used to estimate the association of ≥1 independent (predictor) variables with a continuous dependent (outcome) variable.[2] In the most simple case, thus referred to as "simple linear regression," there is only one independent variable. Simple linear regression fits a straight line to the data points that best characterizes the relationship between the dependent (Y) variable and the independent (X) variable, with the y-axis intercept (b_0), and the regression coefficient being the slope (b_1) of this line:

$$E(Y)\ (\text{expected value of } Y) = b_0 + b_1 X$$

A model that includes several independent variables is referred to as "multiple linear regression" or "multivariable linear regression." Even though the term linear regression suggests otherwise, it can also be used to model curved relationships.

Linear regression is an extremely versatile technique that can be used to address a variety of research questions and study aims. Researchers may want to test whether there is evidence for a relationship between a categorical (grouping) variable (eg, treatment group or patient sex) and a quantitative outcome (eg, blood pressure). The 2-sample t test and analysis of variance,[3] which are commonly used for this purpose, are essentially special cases of linear regression. However, linear regression is more flexible, allowing for >1 independent variable and allowing for continuous independent variables. Moreover, when there is >1 independent variable, researchers can also test for the interaction of variables—in other words, whether the effect of 1 independent variable depends on the value or level of another independent variable.

Linear regression not only tests for relationships but also quantifies their direction and strength. The regression coefficient describes the average (expected) change in the dependent variable for each 1-unit change in the independent variable for continuous

From the *Department of Anesthesiology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands; and †Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Austin, Texas.

Address correspondence to Patrick Schober, MD, PhD, MMedStat, Department of Anesthesiology, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands. Address e-mail to p.schober@amsterdamumc.nl.

independent variables or the expected difference versus a reference category for categorical independent variables. The coefficient of determination, commonly referred to as R^2, describes the proportion of the variability in the outcome variable that can be explained by the independent variables. With simple linear regression, the coefficient of determination is also equal to the square of the Pearson correlation between the x and y values.

When including several independent variables, the regression model estimates the effect of each independent variable while holding the values of all other independent variables constant.[4] Thus, linear regression is useful (1) to distinguish the effects of different variables on the outcome and (2) to control for other variables—like systematic confounding in observational studies or baseline imbalances due to chance in a randomized controlled trial. Ultimately, linear regression can be used to predict the value of the dependent outcome variable based on the value(s) of the independent predictor variable(s).

Valid inferences from linear regression rely on its assumptions being met, including

- the residuals are the differences between the observed values and the values predicted by the regression model, and the residuals must be approximately normally distributed and have approximately the same variance over the range of predicted values;
- the residuals are also assumed to be uncorrelated. In simple language, the observations must be independent of each other; for example, there must not be repeated measurements within the same subjects. Other techniques like linear mixed-effects models are required for correlated data[5]; and
- the model must be correctly specified, as explained in more detail in the next paragraph.

Whereas Müller-Wirtz et al[1] used simple linear regression to address their research question, researchers often need to specify a multivariable model and make choices on which independent variables to include and on how to model the functional relationship between variables (eg, straight line versus curve; inclusion of interaction terms).

Variable selection is a much-debated topic, and the details are beyond the scope of this Statistical Minute. Basically, variable selection depends on whether the purpose of the model is to understand the relationship between variables or to make predictions. This is also predicated on whether there is informed a priori theory to guide variable selection and on whether the model needs to control for variables that are not of primary interest but are confounders that could distort the relationship between other variables.

Omitting important variables or interactions can lead to biased estimates and a model that poorly describes the true underlying relationships, whereas including too many variables leads to modeling the noise (sampling error) in the data and reduces the precision of the estimates. Various statistics and plots, including adjusted R^2, Mallows C_p, and residual plots are available to assess the goodness of fit of the chosen linear regression model.

REFERENCES

1. Müller-Wirtz LM, Maurer F, Brausch T, et al. Exhaled propofol concentrations correlate with plasma and brain tissue concentrations in rats. *Anesth Analg*. 2021;132:110–118.
2. Vetter TR, Schober P. Regression: the apple does not fall far from the tree. *Anesth Analg*. 2018;127:277–283.
3. Schober P, Vetter TR. Analysis of variance in medical research. *Anesth Analg*. 2020;131:508–509.
4. Schober P, Vetter TR. Confounding in observational research. *Anesth Analg*. 2020;130:635.
5. Schober P, Vetter TR. Repeated measures designs and analysis of longitudinal data: if at first you do not succeed-try, try again. *Anesth Analg*. 2018;127:569–575.