MARNO VERBEEK
Erasmus University, the Netherlands

I Z A
World of Labor
Evidence-based policy making

Using linear regression to establish empirical relationships

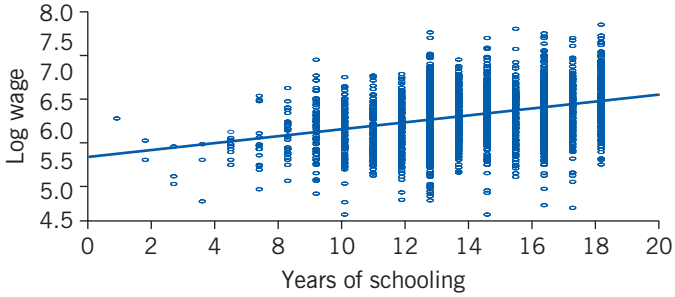
Linear regression is a powerful tool for estimating the relationship between one variable and a set of other variables

Keywords: linear regression, ordinary least squares, model specification, estimation and inference, causality

ELEVATOR PITCH

Linear regression is a powerful tool for investigating the relationships between multiple variables by relating one variable to a set of variables. It can identify the effect of one variable while adjusting for other observable differences. For example, it can analyze how wages relate to gender, after controlling for differences in background characteristics such as education and experience. A linear regression model is typically estimated by ordinary least squares, which minimizes the differences between the observed sample values and the fitted values from the model. Multiple tools are available to evaluate the model.

A simple linear regression can investigate the average relationship between two variables



Source: Author's regression using data from [1] on 3,010 men from the US National Longitudinal Survey of Young Men. Online at: http://www.bls.gov/nls/

IZA World of Labor

KEY FINDINGS

Pros

- Linear regression is a simple and convenient tool to establish an empirical relationship between one variable and a set of other variables.
- Linear regression estimated by ordinary least squares is the "best linear predictor": in a given sample, the estimated linear combination of regressors provides the closest approximation to the actual outcome.
- Ordinary least squares works reasonably well even if the model is not perfectly specified.
- Linear regression with ordinary least squares can provide a quick benchmark for more advanced methods.

Cons

- Causal relationships are most valuable for policy advice and interventions, but interpreting a linear regression model as a causal relationship is challenging and requires strong assumptions.
- Specification of a linear regression model is not always straightforward because there is no simple, hard rule that prescribes how to choose an appropriate specification.
- Specification of a regression model requires care and statistical testing, particularly if estimates of interest appear very sensitive to the specification used or to the set of explanatory variables included.

AUTHOR'S MAIN MESSAGE

Linear regression can be used to empirically establish the relationship between a variable of interest, say a person's wage, and a set of other variables that may be correlated with each other, such as gender, education, and experience. Estimating such relationships is routinely done by ordinary least squares, which tries to make the regression model fit the data as well as possible. Linear regression can predict the outcome variable in cases where it is not observed and thus policymakers can use it to generate predictions for the outcome variable after changing one or more of the explanatory variables to reflect a policy intervention.

Using linear regression to establish empirical relationships. IZA World of Labor 2017: 336
doi: 10.15185/izawol.336 | Marno Verbeek © | February 2017 | wol.iza.org