OLS (Linear) Regression

Forthcoming in:

J. C. Barnes & D. R. Forde (Eds.), The encyclopedia of research methods and statistical techniques in criminology and criminal justice. New York, NJ: Wiley.

Alexander L. Burton

University of Cincinnati

A common statistical approach used by criminal justice researchers to examine relationships between variables is the regression framework. Regression modeling allows researchers to examine the specific effects variables have on one another, while simultaneously controlling for the effects that other variables may also have. Although many types of regression frameworks exist, the most frequently used in criminal justice research are logistic regression techniques (e.g., binary logistic regression and ordinal logistic regression) and Ordinary Least Squares (OLS) regression. The latter, OLS, is the focus of this essay.

Thus, this paper is structured in the following way. First, I define what the ordinary least squares method is. Second, I provide a practical guide with special attention paid to the data assumptions that must be met to conduct OLS linear regression. Finally, I conclude with the statistics that should be interpreted in an OLS regression model output.

THE OLS LINEAR REGRESSION FRAMEWORK

The Least Squares Method

The least squares method used in OLS regression is relatively straightforward. Imagine a scatterplot of datapoints that form a linear trend. An OLS linear regression procedure builds a *line of best fit* that would serve as the most accurate way of depicting the spread of the data points with a single line. The least squares property states that the line fit in the OLS method will have the smallest value of the summed squared deviations of each data point from the line.

OLS Regression Procedure

The OLS regression method of analysis fits a regression plane onto a “cloud” of data that is assumed to have a linear trend (Fox, 2015). Although the regression plane does not touch every point in the data cloud, it does model the partial relationships between each slope (i.e., each regression coefficient “*b*”) and the outcome variable, while holding constant the effects of the remaining variables (Fox, 2015). Thus, regression coefficients in OLS are estimated by minimizing the sum of squares of the differences between values fitted into the regression plane and the observed values in the data. For a handful of reasons (discussed below), OLS regression has many data assumptions that a researcher must check before conducting the analysis.

OLS Regression Assumptions

Linearity. The linearity assumption states that a model cannot be correctly specified if the independent variables in the model do not share, collectively, a linear relationship with the dependent variable (Fox, 2015). Also, each independent variable (excluding binary variables) must share a linear relationship with the dependent variable (Fox, 2015). This assumption is salient, as a non-linear model fails to explain the systematic pattern of the relationship between the dependent and independent variables (Fox, 2015). Given that the units (or levels) of the independent variable are inconsequential (i.e., each unit change in the independent variables always corresponds to the same resultant change in the dependent variable) in OLS, nonlinearity biases the interpretability of the least squares estimators (Weisburd & Britt, 2014).

This assumption can be assessed in a handful of ways—both graphically and statistically. One graphical method of assessing the linearity assumption is to examine a scatterplot of the studentized residuals plotted against the unstandardized predicted values. This scatterplot can reveal whether a linear relationship exists between the independent variables and the dependent variable, collectively. Further, partial regression plots should be generated to show whether each independent variable and the dependent variable share a linear relationship. In this approach, it is fine to exclude binary variables as the partial regression plots will not be substantively useful.

To statistically assess for linearity, incremental *F*-tests ("lack-of-fit" tests) can be used to assess whether any independent variables in the model cannot be specified linearly with the outcome variable. These tests can depict whether variables in the model significantly contribute to a deviation from linearity for the full model. If no significant deviations exist, the linearity assumption is met and it can be concluded that a linear model fits well with the data.

Independence (Non-autocorrelation). The independence assumption of OLS is met if error terms in the regression model are uncorrelated (i.e., independent of each other) (Fox, 2015). This assumption is largely a result of how the data were collected. Thus, if data were randomly sampled from a large population, there is likely no correlation between the error terms. However, if there is reason to believe there may be autocorrelation between error terms (e.g., multilevel data), this assumption can be assessed statistically by consulting the Durbin-Watson statistic. With a possible range between zero and four, a Durbin-Watson test statistic of approximately *two* is indicative of weak correlation between residuals, thereby implying that the error terms in the model are independent (Rutledge & Barros, 2002).

Normality. The normality assumption states that the distribution of errors (residuals) must be normally distributed around the multiple regression plane (Fox, 2015). Fox (2015) offers three reasons why it is important to test for this assumption, even with large samples. First, OLS estimators are less efficient if the error distribution has "heavy tails" (Fox, 2015), which occurs when outlying data points create non-normal error distributions. Second, skewed error distributions can adversely affect the interpretation of the least squares fit (Fox, 2015). This is due to the conditional mean of the dependent variable, given the predictors, being affected by the skewed distribution (Fox, 2015). Finally, multimodal error distributions may cause data to become dichotomized into groups, which causes non-normality in the error distribution (often caused by the inclusion of binary variables in the model) (Fox, 2015).

Normality is best assessed graphically with quantile-comparison plots and kernel-density estimates. These graphical assessments test for normality by plotting studentized residuals onto a line representing the normal distribution quantile (quantile-comparison plot) and a normal curve (kernel-density plot).

Constant Error Variance. This assumption states that the variation of the dependent variable around the regression plane (i.e., the error variance) is constant (Fox, 2015). Heteroskedasticity, or nonconstant error variance, is problematic in regression models because it lessens the efficiency of least squares estimators and can lead to miscalculations of coefficient standard errors (Fox, 2015). This assumption can be assessed by examining a scatterplot of the studentized residuals plotted against the unstandardized predicted values ($\hat{Y}$). A visual inspection of this plot can reveal whether this assumption has been met with the model. If heteroskedasticity is presumed to exist, and visual inspection does not provide conclusive evidence of homoscedasticity, this assumption can be further assessed statistically.

One option to statistically assess this assumption is to conduct a Breusch-Pagan test. Testing the null hypothesis that the model has constant error variance (i.e., homoscedasticity), a significant result ($p < .05$) with this test implies that the variation of the dependent variable around the regression plane is nonconstant (i.e., heteroskedastic). Conversely, a nonsignificant p-value indicates that the assumption of homoscedasticity has been met with the model.

Multicollinearity

Multicollinearity exists in an OLS multiple regression model when two or more independent variables share a *near perfect* linear relationship (Fox, 2015). Multicollinearity can cause values of least squares estimators to be unstable (i.e., subject to change with slight variation in the data) (Fox, 2015). Further, multicollinearity makes it difficult to establish the unique effects of each independent variable on the dependent variable (Weisburd & Britt, 2014).

Two common approaches to assess for multicollinearity involve examining the bivariate correlations between each pair of independent variables and checking the Variance Inflation Factors (VIF) of each predictor in the model. In the first approach, check for correlations at or above .80 because that is considered a strong indicator of multicollinearity in the model (Weisburd & Britt, 2014). In the latter approach, VIF values are examined because they estimate how much of the variance in regression coefficients is inflated due to multicollinearity (Fox, 2015). A recommended cutoff value for VIFs is 10, where VIFs less than 10 indicate that the model does not suffer from multicollinearity (Fox, 2015).

A Note on Data Transformations

When conducting parametric tests (such as OLS linear regression), it is assumed that the data is approximately normally distributed across every category of the dependent variable. Commonly, however, this is not the case—especially in the social sciences. To remedy a violation of this assumption of parametric tests, transformations to the data can be made. Transformations can come from the families of logarithms, roots, and powers. Transformations may be necessary for independent variables, the dependent variable, and in some instances, both. Understanding when a data transformation is necessary will generally come as a result of checking the data assumptions mentioned above. A useful tool for deciding which type of data transformation should be performed is Tukey and Mosteller's bulging rule (1977).

To provide a substantive example of when a transformation would be necessary, it is instructive to examine a recent article published by Haner, Cullen, Jonson, Burton, and Kulig (2019). In one of their OLS regression models, it was found that the dependent variable (support for banning firearms from risky people) was strongly positively skewed (which conveyed a majority of the sample agreed that access to firearms should be banned from risky people). As a result, the OLS assumption of constant error variance was initially violated. To remedy this violation, a logarithmic transformation was made to the dependent variable, which resulted in all assumptions of the model being met.

Unusual Data

Outliers. As mentioned earlier when describing the OLS regression framework, the regression plane fitted onto the "data cloud" is an estimation. Thus, this line can be highly influenced by *odd* cases in the data cloud. One salient consideration when performing an OLS regression is to check the data for significant outliers. An outlier is defined as an observation whose dependent variable value is conditionally unusual given the value of the independent variable(s) (Fox, 2015). Two statistics are generally used to assess for outliers—standardized and studentized deleted residuals. Cases with standardized residuals of ± 3 standard deviations and studentized deleted residuals of ± 2 standard deviations are considered outliers (Fox, 2015).

Leverage and influence. In addition to checking data for outliers, it is also essential to examine whether cases in the data exhibit significant influence and leverage on the regression plane. In OLS regression, cases that have unusual combinations of independent variable values have high leverage (Fox, 2015). Observations that have both high leverage and large studentized residual exhibit robust influence on the regression coefficients (Fox, 2015). Common methods used to check the model for cases exhibiting high leverage and influence are: examine hat-values ($\bar{h}$) (cutoff $2(\bar{h})$), Cook's D (see Fox, 2015, pp. 282, for the cutoff), DFITTS (see Chatterjee & Hadi, 1988, for cutoff), and COVRATIO (See Belsey, Kuh, & Welsch, 1980, for cutoff).

INTERPRETING OLS REGRESSION RESULTS

Once the assumptions of the OLS regression framework have been met, a researcher can interpret their results with confidence. In a standard OLS linear regression output table, there are approximately six statistics that should be examined by the researcher. These are the: p-value, 95% confidence interval, unstandardized regression coefficient (b), standardized regression coefficient (β), standard error (SE), R^2, and the F-statistic.

p-value

In the regression output table, the p-value denotes the results of hypothesis tests of the independent variables' slopes. In these tests, the null hypothesis states that the slope of the independent variable is no different than zero (implying it has no effect on predicting the dependent variable). Thus, regression coefficients with p-values less than .05 are significant predictors of the dependent variable.

95% Confidence Interval

In most OLS regression output tables, the 95% confidence interval is provided for each unstandardized regression coefficient. This column in the table provides the lower and upper bound values for the confidence interval. One can be 95% certain that the true value found within the population lies within the lower and upper bound interval.

Standardized Regression Coefficients (β)

The standardized regression coefficient, often referred to as Beta, allows for comparisons to be made between independent variables in the same model. This is due to variables in the model not being measured in the same units (e.g., number of prior arrests and hours spent in treatment). Standardized regression coefficients are calculated using beta weights. Thus, to calculate Beta values for each variable, all unstandardized regression coefficients are placed into an equation and standardized according to the ratio of the standard deviation of the variable examined to the standard deviation of the dependent variable (Weisburd & Britt, 2014).

Unstandardized Regression Coefficients (*b*)

Unstandardized regression coefficients in an OLS linear regression represent the slope of the line between the independent variables and the dependent variable, while holding constant the effects of the other variables in the model. Thus, coefficients interpret as *a one unit increase in X* (independent variable) *corresponds to and expected increase/decrease in Y* (dependent variable), *independent of the effects of the remaining variables in the model*. It is important to note that only significant regression coefficients should be interpreted (i.e., when $p < .05$), as insignificant slopes (unstandardized regression coefficients) are—statistically speaking—no different from zero. Unstandardized regression coefficients should be used when examining results across studies.

Standard Error (SE)

The standard error shown in the regression output table depicts the estimated amount of error in the estimate compared to the actual value in the population. Standard error for the estimate is reduced by increasing the sample size.

R^2

To assess how well the overall model fits the data, the coefficient of determination (R^2) can be consulted. This statistic represents the percentage of variation explained in the dependent variable by the independent variables in the model. Typically, in the social sciences, R^2 values of .40 or larger are considered robust. A consideration to note when interpreting the R^2 is that the value will increase simply due to the inclusion of more independent variables.

Researchers may, instead, choose to consult the adjusted version of R^2. In addition to being adjusted for the number of predictors in the model, the adjusted R^2 is a measure of the proportion of variance in the dependent variable explained by the independent variables, versus the mean model (i.e., a model with no independent variables included using only the mean of the dependent variable) (Fox, 2015). This adjusted statistic is not necessarily based on the sample like the R^2 (which leads to a positively biased estimation), and instead, is and the value expected to exist in the true population (Fox, 2015).

F-statistic

To assess whether the overall regression model is significant, the *F*-test statistic should be interpreted. The statistic is generated from an *F*-test (an omnibus test such as that used in ANOVA) and denotes whether the addition of the independent variables lead to a model that is significantly more efficient at predicting the dependent variable compared to the mean model. Further, a significant *F*-statistic implies that the full model (with the inclusion of all independent variables) is a significantly better fit for the data than the mean model. A significant *F* value also conveys that at least one regression slope (b) in the model significantly differs from zero.

References

Belsley, D. A., Kuh, E., & Weisch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: John Wiley.

Chatterjee, S., & Hadi, A. S. (1988). Impact of simultaneous omission of a variable and an observation on a linear regression equation. *Computational Statistics & Data Analysis*, *6*(2), 129-144.

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage publications.

Haner, M, Cullen, F. T., Jonson, C. L., Burton, A. L., & Kulig, T. C. (2019). Price of liberty or never again: Americans' views on preventing mass murder. *Justice Evaluation Journal*. Advance online publication. DOI: 10.1080/24751979.2019.1569474

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression: A second course in regression.* Reading, MA: Addison-Wesley.

Rutledge, D. N., & Barros, A. S. (2002). Durbin–Watson statistic as a morphological estimator of information content. *Analytica Chimica Acta*, *454*(2), 277-295.

Weisburd, D., & Britt, C. L. (2014). *Statistics in criminal justice*. New York, NY: Springer.