OLS Regression

POST 8000 – Foundations of Social Science Research for Public Policy

Steven V. Miller

Department of Political Science

Goal for Today

Discuss OLS regression, the backbone of quantitative analysis.

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.
– Gujarati (1998)

Population Regression Function (PRF)

The PRF is a linear function that hypothesizes a theoretical relationship between a DV (Y) and a set of IVs (i.e. X_i).

- A stochastic error term rounds it out as well.

$$Y_i = \beta_0 + \beta_i * X_i + u_i$$

Sample Regression Function (SRF)

The PRF is not directly observable, but we can estimate it from the SRF.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_i * X_i + \hat{u}_i$$

In terms of the SRF:

- $Y_i = \hat{Y}_i + \hat{u}_i$

In terms of the PRF:

- $Y_i = E(Y|X_i) + u_i$

Toward Ordinary Least Squares (OLS)

But how is SRF itself determined? First express:

$$\hat{Y}_i + \hat{u}_i$$

as:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 * X_i$$

Meaning that: the residuals $(\hat{u}_i)$ are the differences between actual and estimated Y values.

The Least Squared Principle:

The sum of squared residuals should be as small possible.

$$\Sigma \hat{u_i}^2 = \Sigma(Y_i - \hat{Y}_i) = \Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 * X_i)$$

The regression parameters that minimize the sum of squared residual errors will be the best fit for the model.

- Residuals are squared because unsquared residuals will always sum to zero.

Statistical Properties of OLS Estimators

1. OLS estimators are expressed solely in terms of the observable quantities (i.e. x and y)
2. They are *point* estimators. Given the sample, each estimator emerges as only a point value for the relevant population parameter.

Once OLS estimates are obtained, the sample regression line emerges as well.

Properties of SRF from OLS Estimators.

1. It passes through the sample means of y and x.
2. The mean value of $\hat{Y_i}$ is equal to the mean value of the actual Y.
3. The mean of the residuals ($\hat{u_i}$) is zero.
4. The residuals ($\hat{u_i}$) are uncorrelated with the predicted/fitted values of Y.
5. The residuals are unccorrelated with X_i.

The "10 Commandments" of OLS (i.e. Assumptions)

1. The regression is linear in its parameters.
2. X is assumed fixed in repeated sampling (i.e. it's nonstochastic)
3. Zero mean value of the disturbance/error term of u_i. I.e. $E(u_i|X_i) = 0$.
4. Variance of u_i is constant/"homoscedastic".
5. No *autocorrelation* between disturbances.

The "10 Commandments" of OLS (continued)

6. Zero covariance between u_i and X_i (i.e. residuals are uncorrelated with X_i)
7. Can't estimate more parameters than observations.
8. X can't be a constant.
9. The regression model is correctly specified.
10. There is no perfect linear relationship among IVs (i.e. no multicollinearity).

Conclusion/Discussion

1. What's in the disturbance/error term?
2. How do you check the assumptions?

Table of Contents