Linear Regression Assumptions

v1.0

Chandler Zachary

The purpose of this document is to provide a quick reference of the classical linear regression assumptions and the consequences of their violation.

The Assumptions of Linear Regression

Functional Form

The outcome is a linear function of the parameters and a stochastic error term:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{b} + \boldsymbol{e}.$$

Nonsingularity of Design Variance Matrix

The matrix

$$(\boldsymbol{x}^T\boldsymbol{x})^{-1}$$

exists and can be calculated.

Mean Zero Stochastic Error

$$\mathbb{E}[\boldsymbol{e}] = 0$$

Stochastic Error Not Correlated with Regressors

$$\mathbb{E}[\boldsymbol{e}|\boldsymbol{x}] = 0$$

Constant Stochastic Error Variance

Errors are homoscedastic:

$$var(\boldsymbol{e}) = \mathbb{E}[\boldsymbol{e}^T\boldsymbol{e}] = \boldsymbol{\sigma}^2\boldsymbol{I}.$$

Stochastic Errors Not Pairwise Correlated

$$cov(e_i, e_j) = \mathbb{E}[e_i, e_j] = 0, \quad \forall i \neq j$$

The assumptions on the error term may be summarized by writing: $\boldsymbol{e} \sim iid(\boldsymbol{0}, \boldsymbol{\sigma}^2\boldsymbol{I})$. When all of these assumptions are satisfied, OLS is an unbiased, minimum variance estimator, often called *BLUE*.

Violations of the Assumptions

Violations of the assumptions can compromise inference either because coefficient estimates will be biased, standard errors will be biased, or both. If regressors are correlated with residuals, then coefficient estimates are biased. If only standard errors are biased, then hypothesis tests and confidence intervals suffer, but coefficient estimates are uncontaminated.

If the functional form is incorrectly specified, then erroneous coefficients may be estimated or important coefficients may be omitted. If the design variance matrix is not invertible, then estimates can not be calculated. If the residuals are not mean zero, then the intercept will be biased, but the slopes are unaffected. If the residuals are correlated with one or more regressors, especially in a way that is correlated with the outcome, then coefficient estimates will be biased.

Non-constant residual variance and pairwise correlation among the residuals will bias standard errors, but they do not necessarily bias coefficient estimates. In either of these scenarios, there is some implied structure or pattern to the error term reflected in the residuals. In this case, OLS is not the most efficient estimator because coefficient variances are incorrectly computed causing standard errors to be inconsistent. This pattern is symptomatic of problems with model specification, sampling design, or experimental design, but these problems have two different sources. Non-constant residual variance is called *heteroscedasticity*, and pairwise residual correlation is sometimes called *autocorrelation*.

The constant error variance assumption implies that all observations have been independently drawn from the same distribution (i.e. - $iid(0, \boldsymbol{\sigma}^2\boldsymbol{I})$). The simplest countermeasure for this problem is robust standard errors, also called Eicker-Huber-White standard errors. This method produces consistent estimates of coefficient variances and attempts to estimate the pattern of variance in the error. It produces wider standard errors, leading to more conservative hypothesis tests and confidence intervals. This is a blunt instrument in the sense that it assumes no knowledge of the pattern of correlation. If the pattern is discernible, a procedure such as generalized least squares can be implemented to accommodate the pattern. It may even be possible to specify a variance-covariance matrix for the error term and estimate the model with this specification.

The pairwise uncorrelated errors assumption implies that the observations have unrelated values of the stochastic component. It means that the observations contribute independent information about the stochastic component. This assumption is violated with observations in the stratum (e.g. - city, demographic sub-population) are similar in *unobserved* ways. When this happens, the treatment or intervention assigned by the hypothesis has heterogeneous effects based on the unobserved strata in the sample, and it has consequences for the efficiency of OLS estimates and the bias of standard errors.

Robust standard errors will remedy this problem *within each stratum* (i.e. - within stratum correlations among unobserved residuals). However, one also must specify the strata, or *clusters*, where errors may be correlated. This is called *clustering standard errors*. It corrects for patterns of correlated error that may exist between observations in the same cluster. Ideally, clusters are assigned at the level of treatment or intervention assumed in the hypothesis, such as the cluster where data are collected. If the data are observational, then clustering is done at the highest level in the data.

Another possibly compromising issue can arise in least squares estimation: multicollinearity. This happens with two or more regressors are correlated with each other. In this case, OLS can not distinguish between the variation in two or more regressors in explaining variation in the outcome. Multicollinearity will cause standard errors to be large, leading to deflated coefficient estimates. Inference remains valid. However, *perfect multicollinearity* occurs when a regressor is a linear combination of one or more other regressors. In this case, coefficients can not be identified.