Significance Level of Regression Coefficients

Shivank Goel, Email : sg2359@cornell.edu

INTRODUCTION

Consider the regression equation, $y = \alpha + \beta x + u$.[1]. We often say that, the estimate $\hat{\beta}$, obtained from the regression is statistically significant. In this article, we discuss how to test the statistical significance of regression coefficients.

HOMOSKEDASTICITY AND HETEROSKEDASTICITY

We call the OLS estimator as *homoskedastic* if the variance of disturbances, u, does not depend on the covariates x, i.e., $Var(u|x) = \sigma^2$. On the other hand if variance is a function of covariates, i.e., $Var(u|x) = f(x)$, we call that as an *heteroskedastic* estimator. We note that, *Hetereskedasticity* only affects the standard errors, which means the OLS estimate might not be the most efficient, however, it does not affect the consistency of the OLS estimator. Since, the *homoskedastic* assumption might not always be realistic, it is advisable to use *Robust SEs* instead, which assumes *heteroskedastic* disturbances. *Weighted Least Squares (WLS)* is sometimes used to deal with *heteroskedasticity*, however in practice, since we can get consistent estimates from simple OLS, it is common to simply use *Robust SEs* with OLS.

STATISTICAL SIGNIFICANCE OF THE COEFFICIENTS

Consider the regression, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$. The estimates, $\hat{\beta}_1, \hat{\beta}_2$, etc., are functions of random variables and hence they themselves are random variables with variances and covariances with each other. The sampling variance of $\hat{\beta}_j$ (assuming *homoskedasticity*) is given by, $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2(1-R_j^2)}$, where R_j^2 is the R^2 obtained by regressing x_j on all other dependent variables. We note that, a higher variation in x_j, will lead to smaller R_j^2, and hence a smaller $Var(\hat{\beta}_j)$, and smaller standard errors. Intuitively, more variation in x_j helps in identifying it's effect on y. That is why larger samples are always helpful, as they increase the variation. We also note that the SE increases due to the variance of the residuals, σ_u^2, because this variance gives us the variation in y which model could not explain. Hence, to reduce that it is sometimes good to add variables that affect y, to improve the fit, even if they are not necessary for identification. But more variables can be harmful also, in the case if they are highly collinear with other dependent variables, which can increase the value of R_j^2. It becomes difficult to disentangle the effect of variables that are highly collinear, hence we do not add variables that don't affect y, and thus are irrelevant. This is known as the problem of *Multicollinearity*. The *Multicollinearity* does *not* cause a bias or inconsistency, but say x_1 and x_2 are highly correlated, then $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_2)$ may get very large, but it will not have any direct effect on $Var(\hat{\beta}_j)$ for $j \notin \{1, 2\}$.

After obtaining the SEs for the coefficients, we can obtain the $t - stat$, telling us how many standard deviations the estimate $\hat{\beta}$ is from 0, by testing the null hypothesis that $\hat{\beta} = 0$. The p-value gives us the probability to get an estimate, $\hat{\beta}$ standard deviations away from 0, if true $\beta = 0$. Statistical significance is not indicative of economic significance, a statistically insignifacnt coefficient might be economically large and vice-versa.

[1]This pdf has been summarized from *Todd A. Gormley's* notes, Lecture 2, http://www.gormley.info/phd-notes.html