Statistical Assumptions for Ordinary Least Squares Regression

Introduction

- In Ordinary Least Squares (OLS) regression we wish to model a continuous random variable Y (the *response variable*) given a set of *predictor variables* X_1, X_2, ..., X_k.

- While we will require that the response variable Y be continuous, or approximately continuous, the predictor variables X_1, X_2, ..., X_k can be either continuous or discrete. It is a standard notation to reserve k for the number of predictor variables in the regression model, and p for the number of parameters (regression coefficients or betas) in the regression model. When the model contains an intercept, then $p = k + 1$. When the model does not contain an intercept, then $p = k$.

- When formulating a regression model, we want to explain the variation in the response variable by the variation in the predictor variables.

Statistical Assumptions for OLS Regression

There are two primary assumptions for OLS regression.

1. The regression model can be expressed in the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon. \tag{1}$$

   Notice that the model formulation specifies error term ϵ to be additive, and that the model parameters (the betas) enter the modeling linearly, that is, β_i represents the change in Y for a one unit increase in X_i when X_i is a continuous predictor variable. Any statistical model in which the parameters enter the model linearly is referred to as a *linear model*.

2. The response variable Y is assumed to come from an independent and identically distributed (iid) random sample from a $N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\right)$ distribution where the variance σ^2 is a fixed but unknown quantity. The statistical notation for this assumption is $Y \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\right)$.

Linear Versus NonLinear Regression

Remember that a *linear model* is linear in the parameters, not the predictor variables.

- The following regression models are all linear regression models.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon \tag{2}$$

$$Y = \beta_0 + \beta_1 \ln(X_1) + \epsilon \tag{3}$$

- The following regression models are all nonlinear regression models.

$$Y = \beta_0 \exp(\beta_1 X_1) + \epsilon \tag{4}$$

$$Y = \beta_0 + \beta_2 \sin(\beta_1 X_1) + \epsilon \tag{5}$$

- If you know a little calculus, then there is an easy mathematical definition of a nonlinear regression model. In a nonlinear regression model at least one of the partial derivatives will be dependent on a model parameter.

Distributional Assumptions for OLS Regression

The assumption $Y \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\right)$ can also be presented in terms of the error term ϵ. Most introductory books present the distributional assumption in terms of the error term ϵ, but more advanced books will use the standard Generalized Linear Model (GLM) presentation in terms of the response variable Y.

In terms of the error term ϵ the distributional assumption can also be presented as:

- The error term $\epsilon \sim N\left(0, \sigma^2\right)$. Since $Y \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\right)$, then $\epsilon = Y - \mathbf{X}\boldsymbol{\beta}$ has a $N\left(0, \sigma^2\right)$.

This fact coupled with the fact that we require that our data be a random sample imply the standard assumptions for the error term presentation.

Distributional Assumptions in Terms of the Error

1. The errors are normally distributed.

2. The errors are mean zero.

3. The errors are independent and identically distributed (iid).

4. The errors are *homoscedastic*, i.e. the errors do not have any correlation "in time or space".

When we build statistical models, we will check these assumptions about the errors by assessing the model *residuals*, which are our estimates of the error term.

Further Notation and Details

When we estimate an OLS regression model, we will be working with a random sample of response variables Y_1, Y_2, ... , Y_n, each with a vector of predictor variables $[X_{1i}\ X_{2i}\ \cdots\ X_{ki}]$. In matrix notation we will denote the regression problem by

$$\mathbf{Y}_{(n\times 1)} = \mathbf{X}_{(n\times p)}\boldsymbol{\beta}_{(p\times 1)} + \epsilon_{(n\times 1)} \tag{6}$$

where the matrix size is denoted by the subscript. Note that $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_k]$ and $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2 \cdots \beta_k]$.

- When we want to express the regression in terms of a single observation, then we typically use the i subscript notation

  $$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \tag{7}$$

  or simply

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i. \tag{8}$$