More About OLS Regression (Part I)

OLS regression

In many ways, regression analysis lives at the heart of statistics. It is a broad term for a set of methodologies used to predict a `response` variable (also called a `dependent` or `outcome` variable). This prediction is done from one or more `predictor` variables (also called `independent` or `explanatory` variables).

In general, regression analysis can be used to:

- `identify` the explanatory variables that are related to a response variable
- `describe` the form of the relationships involved
- provide an equation for `predicting` the response variable from the explanatory variables

OLS regression predicts the response variable from a set of predictor variables (also called regressing the response variable on the predictor variables - hence the name).

OLS regression fits models of the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}\ , where\ i = 1 \ldots n$$

Our goal is to select model parameters (intercept and slopes) that minimize the difference between actual response values and those predicted by the model. Specifically, model parameters are selected to minimize the sum of squared residuals:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(\hat{Y}_i - (\hat{\beta}_0 + \cdots + \hat{\beta}_k X_{ki}))^2 = \sum_{i=1}^{n}\epsilon_i^2$$

A OLS model is meaningful only if the data satisfies a numbger of statistical assumptions:

a) Normality

For fixed values of independent variables, the dependent variable is normally distributed

b) Independence

The Y_i values are independent of each other

c) Linearity

The dependent variable is linearly related to the independent variable

d) Homoscedasticity

The variance of the dependent variable doesn't vary with the levels of the independent variables.

If we violate these assumptions, our statistical significance tests and confidence intervals stop being valid.

Fitting Regression Models with `lm()`

In R, the basic function for fitting a linear model is `lm()`. The format is

```
myfit <- lm(formula, data)
```

The `formula` is typically written as

```
Y ~ X1 + X2 + . + Xk
```

where the ~ separates the response variable on the left from the predictor variables on the right, and the predictor variables are separated by + signs. Other symbols can be used to modify the formula is various ways.

Symbol	Usage
`~`	Separates response variables on the left from the explanatory variables on the right
`+`	Separates predictor variables
`~`	Separates response variables on the left from the explanatory variables on the right
`+`	Separates predictor variables
`:`	Denotes an interaction between predictor variables.
`*`	A shortcut for denoting all possible interactions
`^`	Denotes interactions up to specified degree
`.`	A placeholder for all variables in the data frame except the dependent variable
`-`	A minus sign removes a variable from the equation
`-1`	Suppress the intercept
`I()`	Elements within the parenthesis are interpreted arithmetically
`function`	Mathematical functions can be used in formulas

In addition to `lm()`, the following table lists several functions that are useful when generating a simple or multiple regression analysis. Each of these functions is applied to the object returned by `lm()` in order to generate additional information based on that fitted model.

Symbol	Usage
`summary()`	Displays detailed results for the fitted model
`plot()`	Generates diagnostics plots for evaluating the fit of a model
`predict()`	Uses a fitted model to predict response values for a new dataset
`coefficients()`	Lists the model parameters (intercept and slopes) for the fitted model
`confint()`	Provides confidence intervals for the model parameters (95% by default)
`fitted()`	Lists the predicted values in a fitted model
`residuals()`	Lists the residual values in a fitted model

Here are some common variations to OLS:

- When the regression model contains one dependent variable and one independent variable, the approach is called *simple linear regression*.
- When there's one predictor variable but power of the variable are included (for example, X, X^2, X^3), it is called *polynomial regression*.
- When there's more than one predictor variable, it is called *multiple linear regression*