CmpE 343 Lecture Notes
11: Regression

Ethem Alpaydın

January 1, 2015

1 Introduction

When we have two random variables X and Y, if they are independent, knowing the value of one has no effect on the other, but when they are dependent, knowing Y has an effect on the probability distribution of X—we use $P(X|Y)$ instead of $P(X)$. Previously we also talked about the covariance, $\text{Cov}(X, Y)$, to measure the relationship between X and Y.

Now we discuss regression where the aim is to estimate the value of one random variable given the value of one or more other random variables. We write

$$Y = f(X|\Theta) + \varepsilon \tag{1}$$

where $f[\cdot)$ is the function that takes X as its argument and is defined in terms of some parameters Θ. We assume that Y is a function of X and ε is the *error term* that makes up for the effect of factors other than X that affect Y. For example X can be height and Y weight of a person and we believe that height affects weight, but there are also other factors, such as, age, gender, and so on. There may be two different people with exactly the same height but with different weight because they may differ in terms of those other factors. We assume that $E[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma^2$; that is, we assume that those other factors do not add any bias and that their effect is constant for everywhere in the X space.

In regression, we are given a sample of paired observations $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^n$, and assuming equation (1), we want to find Θ such that this model has the best approximation ability. To measure this, we define the concept of the *sum of squared error* (SSE):

$$SSE(\Theta|\mathcal{X}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

where

$$y_i = f(x_i|\Theta) \tag{3}$$

The difference $y_i - \hat{y}_i$ is called the *residual* and SSE is the sum of squared residuals. We want to find Θ^* such that SSE is minimized:

$$SSE(\Theta^*|\mathcal{X}) = \min_{\Theta} SSE(\Theta|\mathcal{X}) \tag{4}$$

For every instantiation of Θ, we get a particular $f(\cdot|\Theta)$ and that leads to a particular SSE on the sample $\mathcal{X}$. We can envisage SSE as a function in the space whose dimensions are the elements of Θ and we want to find its minimum. That particular Θ^* is called the *least squares estimator*.

For certain $f(\cdot)$, the optimization of Θ is easy and can be done analytically. Let us see an example.

2 Simple Linear Regression

In linear regression, we assume a linear model:

$$f(x_i|\alpha, \beta) = \alpha + \beta x_i \tag{5}$$

and we are interested in the (a, b) that minimizes

$$SSE(a, b|\mathcal{X}) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - (a + bx_i)]^2 \tag{6}$$

where a and b are the least squares estimators to α and β. Each (a, b) pair defines a line and incurs a certain SSE. To find the minimum of SSE in the (a, b) space, we take its partial derivatives with respect to the two parameters, set them equal to zero, and solve the two equations in two unknowns:

$$\begin{aligned}
\frac{\partial SSE}{\partial a} &= \sum_i (y_i - (a + bx_i)) = 0 \Rightarrow \sum_i y_i = n \cdot a + b \sum_i x_i \\
\frac{\partial SSE}{\partial b} &= \sum_i (y_i - (a + bx_i))x_i = 0 \Rightarrow \sum_i y_i x_i = a \sum_i x_i + b \sum_i (x_i)^2
\end{aligned}$$

These are called the *normal equations* and some manipulation will show that

$$\begin{aligned}
a &= \overline{y} - b\overline{x} \\
b &= \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}
\end{aligned} \tag{7}$$

where $\overline{y} = \sum_i y_i/n$ and $\overline{x} = \sum_i x_i/n$.

SSE is never zero and its minimum value depends on the scale of y_i. To measure the quality of fit in regression, the R^2 statistic is used:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{8}$$

$\hat{y}_i$ is the estimate using regression with x_i as the input. $\overline{y}$ is what we would have used as estimate for y_i if we do not have X and do not use regression. So if regression is useful, the sum of squared error in the numerator is much smaller than the sum of squared error in the denominator and R^2 value will be close to 1; if regression is not useful, if X gives no information, then the two sums of squares would be comparable and R^2 will be close to 0.

3 Sampling Distributions of Least Square Estimators

The regression parameters a and b estimated using equation (8) are a function of the particular sample we use, as does any estimator. Given a different sample from the same population, they will be different, and how different they will be is defined by their sampling distributions.

Let us say A denotes the random variable as estimator to α and one particular instantiation is denoted by a. B and b are similarly defined for β. It is known that $A \sim N(\alpha, \sigma_A^2)$ where

$$\sigma_A^2 = \frac{\sum_i x_i^2}{n \sum_i (x_i - \overline{x})^2} \sigma^2$$

and $B \sim N(\beta, \sigma_B^2)$ where

$$\sigma_B^2 = \frac{1}{\sum_i (x_i - \overline{x})^2} \sigma^2$$

which implies that the least squares estimators are unbiased. An unbiased estimator to σ^2 is

$$s^2 = \frac{SSE}{n - 2} \tag{9}$$

and it can be shown that $(n - 2)s^2/\sigma^2 \sim \chi^2_{n-2}$. Using these, one can derive that

$$\begin{aligned}
\frac{A - \alpha}{s\sqrt{\sum_i x_i^2/n \sum_i (x_i - \overline{x})^2}} &\sim t_{n-2} \\
\frac{B - \beta}{s/\sqrt{\sum_i (x_i - \overline{x})^2}} &\sim t_{n-2}
\end{aligned}$$

These sampling distributions can be used to define confidence intervals or test hypotheses on α or β.

4 Prediction

We are given a sample of n paired observations and we find the least squares estimators, then we are given a new x_0 and asked to estimate y_0. The point estimate is $\hat{y}_0 = a + bx_0$. What about its interval estimate?

We discuss above the sampling distributions of a and b, that is, for different samples of size n, we get different values for a and b, different instantiations of A and B. Now for the new x_0 value, y_0 will be different for different a and b and will itself have a sampling distribution.

Let us start by defining a confidence interval for $E[Y_0|X = x_0]$, also denoted as $\mu_{Y|x_0}$, where $Y_0 = A + Bx_0 + \varepsilon$. It is what we get if we fit many different lines to different samples, evaluate them all at the same x_0, and take their average. Y_0 is normally distributed with

$$E[Y_0|X = x_0] = E[A + Bx_0] = \alpha + \beta x_0$$

We have $\overline{Y} = A + B\overline{X}$, and hence $A = \overline{Y} - B\overline{X}$. Then $Y_0 = A + Bx_0$ can be rewritten as $Y_0 = \overline{Y} + B(x_0 - \overline{X})$ and therefore

$$\text{Var}(\mu_{Y_0}) = \frac{\sigma^2}{n} + (x_0 - \overline{X})^2\sigma_B^2 = \frac{\sigma^2}{n} + \frac{(x_0 - \overline{X})^2}{\sum_i (x_i - \overline{x})^2}$$

where $\text{Cov}(\overline{Y}, B) = 0$. This is the sampling distribution of $E[Y_0|X = x_0]$, which can be used to define confidence intervals or test hypotheses. Note that the variance (and hence the error term) is a function of x; notably, variance is smaller at the center around $\overline{x}$ and higher at the two ends.

Now what about the distribution of Y_0 itself? This case can be viewed as adding ε noise on top of $E[Y_0]$. Because $E[\varepsilon] = 0$, the point estimate is the same and hence the mean of the sampling distribution, but we also add $\text{Var}(\varepsilon) = \sigma^2$ to the variance:

$$\text{Var}(\hat{Y_0} - Y_0) = 1 + \frac{\sigma^2}{n} + \frac{(x_0 - \overline{X})^2}{\sum_i (x_i - \overline{x})^2}$$

5 Generalization

The basic simple linear regression model of equation (5) can be generalized in a number of ways.

In *multiple* linear regression, the response is a function of more than one random variable, for example, two:

$$f(x_i|\alpha, \beta, \gamma) = \alpha + \beta x_i + \gamma z_i \tag{10}$$

In this case, the sample contains observation triples $\{y_i, x_i, z_i\}_{i=1}^n$, but the approach is the same. Again we write the SSE:

$$SSE(a, b, c|\mathcal{X}) = \sum_i [y_i - (a + bx_i + cz_i)]^2 \tag{11}$$

where a, b, c are estimators to α, β, γ, take the derivative with respect to them and set them equal to zero, and solve the three equations in three unknowns. In this case, our fit is a plane.

The same model is also applicable in *polynomial* regression:

$$f(x_i|\alpha, \beta, \gamma) = \alpha + \beta x_i + \gamma x_i^2 \tag{12}$$

where exactly the same approach can be used. Actually if we define $z_i \equiv x_i^2$, equations (10) and (12) are the same—The plane in the two-dimensional (X, Z) space maps to a quadratic in the X space.

Using higher-order terms is one way of implementing nonlinear regression, other nonlinear basis functions can also be used, such as:

$$f(x_i|\alpha, \beta, \gamma) = \alpha + \beta x_i + \gamma \exp(x_i) \tag{13}$$

where again exactly the same method can be used. Note however that it will not work for

$$f(x_i|\alpha, \beta, \gamma) = \alpha + \beta x_i + \exp(\gamma x_i) \tag{14}$$

because the model is no longer linear in its parameters. In such a case, more sophisticated optimization algorithms are needed.