Least Squares

EPsy 8264

2018-05-02

A whole class of statistical estimators is based on the concept of least squares. Given a random variable represented by Y_i and predicted values for this random variable (via some model) represented by $\hat{Y}_i$, least squares estimation lead to $\hat{Y}_i$ (predicted values) such that $\sum(Y_i - \hat{Y}_i)^2$ is a minimum. Thus, given the constraints of your model (What constraints?), there is no other set of $\hat{Y}_i$ scores which will lead to smaller squared discrepancies between the actual and predicted values.

Note that there are other methods of obtaining estimators—e.g. maximum likelihood, Bayesian.

Note that we can express Y_i as a function of the predicted value plus error,

$$Y_i = \hat{Y}_i + \epsilon_i$$

For estimation purposes we wish this error to be as small as possible. (One would not need ϵ_i if we always had perfect prediction.) Note too that ϵ_i is assumed random with no systematic relationship to the $\hat{Y}_i$ values and that the mean of the ϵ_i values is 0. (Thus, the mean for Y_i and $\hat{Y}_i$ are the same.)

Suppose we wish to model (predict) the midterm scores for everyone in this class. Further suppose that we wish to do this with a constant, say β_0. What is the best constant (value of β_0) we can choose whereby $\sum(Y_i - \beta_0)^2$ is a minimum?

The way we solve a minimization or maxumumization problem is with calculus. In calculus, we would take the derivative of the equation, with respect to β_0, set the resulting equation equal to 0, and then simplify/solve the equation. Note that below we drop the i subscript for ease of writing the equations, but it could easily be added back on.

Calculus is not a prerequisite for this course, so I present the result for information only.

$$\begin{aligned} f &= \sum(Y - \beta_0)^2 \\ &= \sum(Y^2 - 2\beta_0 Y + \beta_0^2) \end{aligned}$$

Now we want to find the derivative with respect to β_0. We will use the fact that the derivative of a sum equals the sum of the derivatives. Thus

$$\begin{aligned} f' &= \sum(0 - 2Y + 2\beta_0) \\ &= \sum -2(Y - \beta_0) \end{aligned}$$

We can now set this equal to zero and solve for A. We do this using rules of summation and algebra.

$$
\begin{aligned}
0 &= \sum -2(Y - \beta_0) \\
0 &= -2\sum (Y - \beta_0) \\
0 &= \sum (Y - \beta_0) \\
0 &= \sum Y - \sum \beta_0 \\
\sum \beta_0 &= \sum Y \\
n\beta_0 &= \sum Y \\
\beta_0 &= \frac{\sum Y}{n}
\end{aligned}
$$

This implies that the value for β_0 which minimizes the sum of squared errors (make the best predictions) is the mean of Y. From that we can conclude that if we are to make *one single prediction* for everyone, our best guess is to assume everyone is average.

One could also find the second derivative to ascertain whether the value for β_0 we found is a minimum or a maximum. I do not perform this check here.

However, what if we have additional information? Suppose we know everyone's score on a regression pretest. Since this is an advanced regression course, we may assume that knowing a student's score on a regression pretest may be predictive of their score on the midterm. Those who did extremely well on the pretest should also do better on the midterm. We will now use a model that will predict midterm scores in EPSY 8264 for those with different pretest scores. We will add a further restriction to our model that will force the relationship between pretest scores and midterm scores to be linearly related. Mathematically, we write the model as,

$$Y_i = \beta_0 + \beta_1(X_i) + \epsilon$$

where Y_i is again the midterm score of interest for Student i, X_i is the pretest score for Student i, β_0 is the intercept of the line relating pretest to midterm scores, and β_1 is the slope of that line.

Here we use Greek letters (β_0 and β_1) to represent population parameters (truth). Thus, we ASSUME that there is a linear relationship between the midterm scores and the pretest scores for the POPULATION and, that even in the population, the relationship is not perfect (there are errors between the line and the actual observed midterm scores).

Once we fit the model to a set of data, we will obtain estimates for the parameters in the model. To indicate they are estimates we will use either a "hat" on the parameter (e.g., $\hat{\beta}_0$) or use Roman letters (e.g., B_0).

Using least squares to estimate the parameter values, again will try to minimize the sum of squared errors,

$$f = \sum (Y_i - \hat{Y}_i)^2$$

Now, $\hat{Y}_i$ is just a more complicated equation than when we only included a constant value.

$$\hat{Y}_i = \beta_0 + \beta_1(X_i)$$

Substituting this into f,

$$f = \sum \Big(Y_i - \beta_0 - \beta_1(X_i) \Big)$$

Once again we find the derivative, but this time, since there are two parameters we need to minimize across, we find the partial derivative with respect to β_0 and the partial derivative with respect to β_1.

KEY POINT: It is the variances which are unbiased (not the standard deviations)! This is why statisticians like to estimate variances rather than standard deviations; despite their unwieldy (squared) metric.

References