Deriving the Ordinary Least Squares Linear Regression Solution

Colan F. Biemer

I've found that most resources on linear regression give the solution, $\beta = (X^TX)^{-1}X^TY$, but do not show the required steps. If they do show the steps, there is a moment in the solution where something non-obvious is presented as obvious—well, it isn't obvious to me. My goal is to show how the solution is come to in all of its details, within reason. If you find that a step is missing or could use more details, feel free to email me at bi3mer93@gmail.com. Our process starts with the residual sum of squares or RSS.

$$RSS = \sum_{i=1}^{N}(Y - \widehat{Y})^2$$

Y is the true value we want to predict and $\widehat{Y}$ is the value our model has predicted. Both of these are $N \times 1$ matrices—N rows by 1 column. The question that arises is, how do we find $\widehat{Y}$?

This problem assumes we are given a matrix X with dimensions $N \times P$. Notice that the number of rows in X is the same Y. Every row of X is associated with the same numbered row in Y. So we are going to take a row in X and use that as input to a model. That model will output a continuous value—since this is in the realm of regression—which will create $\widehat{Y}$.

The title gives away that we will be assuming that a linear relationship can model the data. This is a *bias* of our model. We have no guarantee that a line is the true trend of the dataset. Regardless, going further is out of scope and so we'll assume a linear model fits the data. A linear model assigns weights to each input value and has an extra parameter for the intercept. These weights, β, are what we find to solve the problem.

$$\widehat{Y_i} = \beta_0 + \beta_1 X_{i,p} + ... + \beta_N X_{i,p} = \beta_0 + \sum_{k=1}^{P}\beta_k X_{i,k}$$

To simplify notation, we'll assume that X has been modified with an additional column of all ones.

$$\widehat{Y_i} = \sum_{k=0}^{N}\beta_k X_{i,k}$$

We can also express $\widehat{Y}$ completely with linear algebra.

$$\widehat{Y} = X\beta$$

First, let's check that the dimensions are correct. $X\beta$ needs to result in dimensions $N \times 1$. So X is $N \times P$ and β is $P \times 1$. This results in a $N \times 1$ vector which matches what we expected. Some readers may think that they can stop here. We have a system of equations that can be solved. This, though, is incorrect if we have more equations than variables. This case is referred to as an overdetermined system and is our use case for this problem. We need a different approach, and that approach is using RSS to find β.

$$RSS_{OLS} = \sum_{i=1}^{N}(Y_i - X_i^T\beta)^2$$

The above equation now expresses RSS in terms of our model for $\widehat{Y}$. The ith index of $\widehat{Y}$ is expressed as $X_i^T\beta$. Examining the dimensions we find $(1 \times P)(P \times 1) \rightarrow 1$, a single value.

We can view RSS as an error function. Therefore, the best β should minimize RSS. We can find the best β by taking the partial derivative with respect to β and setting the result equal to 0. The question is why? To do that, we have to understand convexity. I actually think wikipedia [1] does a really good job of explaining the idea, and I won't be covering it here. The important thing to note is that RSS is a convex function. A convex function by definition has a global minima that can be found by following the gradient, which is why we take the partial derivative with respect to β.

Now before we do that, I want to re-express RSS in terms of matrices to remove the sum and the square.

$$RSS_{OLS} = (Y - X\beta)^T(Y - X\beta)$$

To understand why the two equations are equivalent, I think it's best to simplify: $RSS_{OLS} = (Y - \widehat{Y})^T(Y - \widehat{Y})$ since $\widehat{Y} = X\beta$. From this perspective it is clear that we are squaring the result of the subtraction for each row which is why the sum and linear algebra equations are equivalent.

Before taking the derivative, we expand out the equation. We use the identities $(AB)^T = B^TA^T$ and $(A+B)^T = A^T + B^T$

$$\begin{aligned}
&RSS_{OLS} = (Y - X\beta)^T(Y - X\beta)\\
&RSS_{OLS} = (Y^T - \beta^TX^T)(Y - X\beta)\\
&RSS_{OLS} = Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta
\end{aligned}$$

This step is the part that confused me the most: $Y^TX\beta = \beta^TX^TY$. The easiest way to understand why this is true is to do the matrix multiplication yourself with generic variables. Looking at the dimensions, you'll find that they both reduce to a singular values. When you finish simplifying both to a sum, the two will be equivalent. If someone asks, I'm happy to write this out in more detail.

$$RSS_{OLS} = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$$

Now we take the derivative.

$$\frac{\partial}{\partial\beta}\ Y^TY - 2B^TX^TY + \beta^TX^TX\beta = -2X^TY + 2X^TX\beta$$

To calculate the derivative, we only look at terms with β. The first term has a β^T and that is easy to remove. The second, $\beta^TX^TX\beta$ is going to be β^2 which is why we follow the power rule to simplify. With this we can set the partial derivative equal to 0 to find the minimum β. As a reminder, the minimum value for a convex function is where the slope is 0. Therefore, we set the derivative equal to 0.

$$\begin{aligned}
&-X^TY + 2X^TX\beta = 0\\
&-X^TY + X^TX\beta = 0\\
&X^TX\beta = X^TY\\
&\beta = (X^TX)^{-1}X^TY
\end{aligned}$$

Te last step—multiplying both sides by $(X^TX)^{-1}$—is important to understand. A matrix multiplied by its inverse is the identity matrix, which is why the left-hand-side is just β. But this assumes that the inversion exists in the first place. For a matrix to be invertable, it must be square (i.e. number of columns equals the number of rows) and

[1] https://en.wikipedia.org/wiki/Convex_function

its determinant [2] must not equal 0. A determinant of 0 means that there are linearly dependent basis vectors in the matrix. More simply, you have two or more columns in the matrix $X^T X$ which are duplicates or one column can be expressed as a linear combination of another column. In practice, the number of rows of data tend to make this not a problem.

With that, we're done. I'll leave you with this:

$$\begin{aligned}
&X\beta = Y \quad \text{(multiply both sides by } X^T\text{)}\\
&X^T X\beta = X^T Y\\
&\beta = (X^T X)^{-1} X^T Y
\end{aligned}$$

We get the same exact result as all the work above. Why?

[2] `https://en.wikipedia.org/wiki/Determinant`