Ordinary Least Squares

I discuss ordinary least squares or linear regression when the optimal coefficients minimize the residual sum of squares. I discuss various properties and interpretations of this classic model.

PUBLISHED

04 January 2020

Linear regression

Suppose we have a regression problem with N independent variables $\mathbf{x}_n$ and N dependent variables y_n. Each independent variable is a P-dimensional vector of *predictors*, while each y_n is scalar *response* or *target* variable. In linear regression, we assume our response variables are a linear function of our predictor variables. Let $\boldsymbol{\beta} = [\beta_1, \dots, \beta_P]^{\top}$ denote a P-vector of unknown *parameters* (or "weights" or "coefficients") and let ε_n denote the nth observation's scalar *error* term. Then linear regression is

$$y_n = \beta_1 x_{n,1} + \beta_2 x_{n,2} + \dots + \beta_P x_{n,P} + \varepsilon_n. \tag{1}$$

Written as vectors, Equation 1 is

$$y_n = \boldsymbol{\beta}^{\top} \mathbf{x}_n + \varepsilon_n. \tag{2}$$

If we stack the independent variables into an $N \times P$ matrix $\mathbf{X}$, sometimes called the *design matrix*, and stack the dependent variables and error terms into N vectors $\mathbf{y} = [y_1, \dots, y_N]^{\top}$ and $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_N]^{\top}$, then we can write the model in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

In classical linear regression, $N > P$, and therefore $\mathbf{X}$ is tall and skinny. We can add an intercept to this linear model by introducing a new parameter β_0 and adding a constant predictor as the first column of $\mathbf{X}$. I will discuss the intercept later in this post.

Errors and residuals

Before we discuss how to estimate the model parameters $\boldsymbol{\beta}$, let's explore the linear assumption and introduce some useful notation. Given estimated parameters $\hat{\boldsymbol{\beta}}$, linear regression predicts

$$\hat{\mathbf{y}} \triangleq \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{4}$$

However, most likely, our predictions $\hat{\mathbf{y}}$ will not match the response variables $\mathbf{y}$ exactly. The *residuals* are the differences in the true response variables $\mathbf{y}$ and what we predict or

$$\mathbf{e} \triangleq \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{5}$$

where $\mathbf{e} = [e_1, \dots, e_N]^{\top}$. Note that the residuals are not the error terms $\boldsymbol{\varepsilon}$. The residuals are the differences in the true response variables $\mathbf{y}$ and what we predict, while the errors $\boldsymbol{\varepsilon}$ are the differences between the true data $\mathbf{X}\boldsymbol{\beta}$ and what we observe,

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}. \tag{6}$$

Thus, errors are related to the true data generating process $\mathbf{X}\boldsymbol{\beta}$, while residuals are related to the estimated model $\mathbf{X}\hat{\boldsymbol{\beta}}$ (Figure 1). Clearly, if $\hat{\mathbf{y}} = \mathbf{y}$, then the residuals are zero.

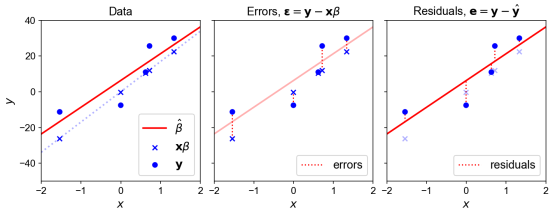

Figure 1. (Left) Ground-truth (un-observed) univariate data $\mathbf{x}\beta$, noisy observations $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$, and estimated $\hat{\beta}$. (Middle) Errors $\boldsymbol{\varepsilon}$ between ground-truth data $\mathbf{x}\beta$ and noisy observations $\mathbf{y}$. (Right) Residuals $\mathbf{e}$ between noisy observations $\mathbf{y}$ and model predictions $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta}$.

Normal equation

Now that we understand the basic model, let's discuss solving for $\boldsymbol{\beta}$. Since $\mathbf{X}$ is a tall and skinny matrix, solving for $\boldsymbol{\beta}$ amounts to solving a linear system of N equations with P unknowns. Such a system is *overdetermined*, and it is unlikely that such a system has an exact solution. Classical linear regression is sometimes called *ordinary least squares* (OLS) because the best-fit coefficients $[\beta_1, \dots, \beta_P]^{\top}$ are defined as those that solve the following minimization problem:

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{n=1}^{N} \left(y_n - \mathbf{x}_n^{\top}\boldsymbol{\beta}\right)^2. \tag{7}$$

Thus, the OLS estimator $\hat{\boldsymbol{\beta}}$ minimizes the *sum of squared residuals* (Figure 2). For a single data point, the squared error is zero if the prediction is exactly correct. Otherwise, the penalty increases quadratically, meaning classical linear regression heavily penalizes outliers. Other loss functions induce other linear models. See my previous post on interpreting these kinds of optimization problems.

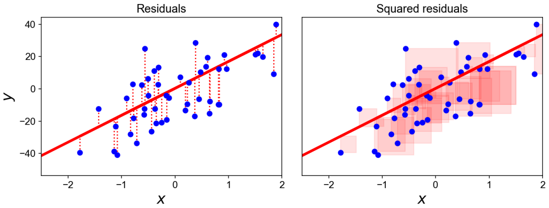

Figure 2. Ordinary least squares linear regression on Scikit-learn's `make_regression` dataset. Data points are in blue. The predicted hyperplane is the red line. **(Left)** Red dashed vertical lines represent the residuals. **(Right)** Light red boxes represent the squared residuals.

In vector form, Equation 7 is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{8}$$

Linear regression has an analytic or closed-form solution known as the *normal equation*,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}. \tag{9}$$

See A1 for a complete derivation of Equation 9.

Hat matrix and residual maker

We'll see in a moment where the name "normal equation" comes from. However, we must first understand some basic properties of OLS. Let's define a matrix $\mathbf{H}$ as

$$\mathbf{H} \triangleq \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}. \tag{10}$$

We'll call this the *hat matrix*, since it "puts a hat" on $\mathbf{y}$, i.e. since it regresses the response variables $\mathbf{y}$ onto the predicted variables $\hat{\mathbf{y}}$:

$$\begin{aligned} &\mathbf{H}\mathbf{y} \\ &= \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y} \\ &\overset{\star}{=} \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \hat{\mathbf{y}} \end{aligned} \tag{11}$$

Step $\star$ is just the normal equation (Equation 8). Note that $\mathbf{H}$ is an orthogonal projection. See A2 for a proof. Furthermore, let's call

$$\mathbf{M} \triangleq \mathbf{I} - \mathbf{H} \tag{12}$$

the *residual maker* since it constructs the residuals, i.e. since it makes the residuals from the response variables $\mathbf{y}$:

$$\begin{aligned} \mathbf{M}\mathbf{y} &= (\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{e}. \end{aligned} \tag{13}$$

The residual maker is also an orthogonal projector. See A3 for a proof.

Finally, note that $\mathbf{H}$ and $\mathbf{M}$ are orthogonal to each other:

$$\begin{aligned} \mathbf{H}\mathbf{M} &= \mathbf{H}(\mathbf{I} - \mathbf{H}) \\ &= \mathbf{H} - \mathbf{H} \\ &= \mathbf{0}. \end{aligned} \tag{14}$$

Geometric view of OLS

There is a nice geometric interpretation to all this. When we multiply the response variables $\mathbf{y}$ by $\mathbf{H}$, we are projecting $\mathbf{y}$ into a space spanned by the columns of $\mathbf{X}$. This makes sense since the model is constrained to live in the space of linear combinations of the columns of $\mathbf{X}$,

$$\mathbf{y} = \beta_1 \begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{N,1} \end{bmatrix} + \beta_2 \begin{bmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{N,2} \end{bmatrix} + \dots + \beta_P \begin{bmatrix} x_{1,P} \\ x_{2,P} \\ \vdots \\ x_{N,P} \end{bmatrix} \tag{15}$$

and an orthogonal projection is the closest to $\mathbf{y}$ in Euclidean distance that we can get while staying in this constrained space. (One can find many nice visualizations of this fact online.) I'm pretty sure this is why the normal equation is so-named, since "normal" is another word for "perpendicular" in geometry.

Thus, we can summarize the predictions made by OLS (Equation 4) by saying that we project our response variables onto a linear hyperplane defined by $\hat{\boldsymbol{\beta}}$ using the orthogonal projection matrix $\mathbf{H}$.

OLS with an intercept

Notice that Equation 1 does not include an intercept. In other words, our linear model is constrained such that the hyperplane $\hat{\boldsymbol{\beta}}$ goes through the origin. However, we often want to model a shift in the response variable, since this can dramatically improve the goodness-of-fit in the model (Figure 3). Thus, we want OLS with an intercept.

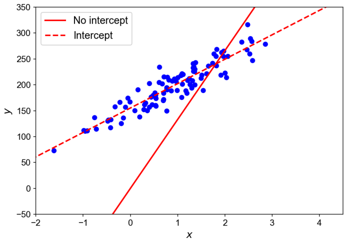

Figure 3. OLS without (red solid line) and with (red dashed line) an intercept. The model's goodness-of-fit changes dramatically depending on this modeling assumption.

In this section, I'll discuss both *partitioned regression* or OLS when the predictors and corresponding coefficients can be logically separated into groups via block matrices and then use that result for the specific case when one of the block matrices in the design matrix $\mathbf{X}$ is a constant, i.e. a constant term for an intercept.

Partitioned regression

Let's rewrite Equation 3 by splitting $\mathbf{X}$ and $\boldsymbol{\beta}$ into block matrices. Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ (we're stacking the columns horizontally) and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]^{\top}$ (we're stacking the rows vertically). Then partitioned regression is,

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon}. \tag{16}$$

The normal equation (Equation 9) can then be written as

$$\begin{bmatrix} \mathbf{X}_1^{\top} \\ \mathbf{X}_2^{\top} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{\top} \\ \mathbf{X}_2^{\top} \end{bmatrix} \mathbf{y}, \tag{17}$$

which in turn can be written as two separate equations:

$$
\begin{aligned}
\mathbf{X}_1^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 &= \mathbf{X}_1^\top \mathbf{y}, \\
\mathbf{X}_2^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 &= \mathbf{X}_2^\top \mathbf{y}.
\end{aligned} \tag{18}
$$

We can then solve for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, since we have two equations and two unknowns. Let's first solve for $\hat{\boldsymbol{\beta}}_1$. We have

$$
\begin{aligned}
\mathbf{X}_1^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 &= \mathbf{X}_1^\top \mathbf{y} - \mathbf{X}_1^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 \\
\hat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2).
\end{aligned} \tag{19}
$$

We can then isolate $\hat{\boldsymbol{\beta}}_2$ by substituting $\hat{\boldsymbol{\beta}}_1$ into the second line of Equation 18:

$$
\begin{aligned}
\mathbf{X}_2^\top \mathbf{y} &= \mathbf{X}_2^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 \\
\mathbf{X}_2^\top \mathbf{y} &= \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2) + \mathbf{X}_2^\top \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 \\
\mathbf{X}_2^\top \mathbf{y} &= \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{y} + \left\{ \mathbf{X}_2^\top \mathbf{X}_2 - \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \right\} \hat{\boldsymbol{\beta}}_2.
\end{aligned} \tag{20}
$$

If we define $\mathbf{H}_1$ and $\mathbf{M}_1$—hat and residual matrices for $\mathbf{X}_1$—as

$$
\begin{aligned}
\mathbf{H}_1 &\triangleq \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top, \\
\mathbf{M}_1 &\triangleq \mathbf{I} - \mathbf{H}_1,
\end{aligned} \tag{21}
$$

then Equation 20 can be rewritten as:

$$
\begin{gathered}
\left\{ \mathbf{X}_2^\top \mathbf{X}_2 - \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \right\} \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2^\top \mathbf{y} - \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{y}, \\
\Downarrow \\
\left\{ \mathbf{X}_2^\top \left(\mathbf{I} - \mathbf{H}_1 \right) \mathbf{X}_2 \right\} \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2^\top \left(\mathbf{I} - \mathbf{H}_1 \right) \mathbf{y}.
\end{gathered} \tag{22}
$$

And we have solved for $\hat{\boldsymbol{\beta}}_2$ in terms of the residual maker for $\mathbf{X}_1$, denoted $\mathbf{M}_1$:

$$
\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}. \tag{23}
$$

As we'll see in the next section, solving for $\hat{\boldsymbol{\beta}}_2$ in terms of $\mathbf{M}_1$ will be make it easier to understand and compute the optimal parameters of OLS with an intercept.

The results in Equations 19 and 23 are part of the Frisch–Waugh–Lovell (FWL) theorem. The basic idea of the FWL theorem is that we can obtain $\hat{\boldsymbol{\beta}}_2$ by regressing $\mathbf{M}_1 \mathbf{y}$ onto $\mathbf{M}_1 \mathbf{X}_2$, where these variables are just the residuals associated with $\mathbf{X}_1$. See (Greene, 2003) for further discussion.

Partitioned regression with a constant predictor

OLS with an intercept is just a special case of partitioned regression. Suppose that we add an intercept to our OLS model. This means we want to estimate a new parameter β_0 such that

$$
y_n = \beta_0 + \mathbf{x}_n^\top \boldsymbol{\beta} + \varepsilon_n. \tag{24}
$$

Notice that if we simply add a constant to our predictor $\mathbf{x}_n$, we can "push" β_0 into the dot product. In matrix notation, we have

$$
\mathbf{y} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1P} \\ 1 & x_{21} & \dots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{NP} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix} + \boldsymbol{\varepsilon}, \tag{25}
$$

which is the same equation as Equation 3 but with a column of ones prepended to $\mathbf{X}$. This means we can write our model as

$$
\mathbf{y} = \mathbf{1} \beta_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{26}
$$

We can use the results of the previous subsection to straightforwardly solve for the scalar β_1 and the P-vector $\boldsymbol{\beta}_2$. First, note that the hat matrix $\mathbf{H}_1$ is just an $N \times N$ matrix filled with the value $1/N$:

$$
\begin{aligned}
\mathbf{H}_1 &= \mathbf{1} (\mathbf{1}^\top \mathbf{1}) \mathbf{1}^\top \\
&= \frac{1}{N} \mathbf{1} \mathbf{1}^\top \\
&= \begin{bmatrix} 1/N & \dots & 1/N \\ \vdots & \ddots & \vdots \\ 1/N & \dots & 1/N \end{bmatrix}.
\end{aligned} \tag{27}
$$

With this in mind, let's compute $\hat{\boldsymbol{\beta}}_2$ using Equation 23. We know that $\mathbf{M}_1 \mathbf{y} = \hat{\mathbf{y}}$. Furthermore, we can simplify $\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2$ as

$$
\begin{aligned}
\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 &= \mathbf{X}_2^\top (\mathbf{I} - \mathbf{H}_1) \mathbf{X}_2 \\
&= \mathbf{X}_2^\top (\mathbf{X}_2 - \mathbf{H}_1 \mathbf{X}_2) \\
&= \mathbf{X}_2^\top (\mathbf{X}_2 - \bar{\mathbf{X}}_2),
\end{aligned} \tag{28}
$$

where $\bar{\mathbf{X}}_2$ is a matrix where each column is an N-vector with the mean of that respective column in $\mathbf{X}_2$ repeated N times since

$$
\bar{\mathbf{X}}_2 \triangleq \begin{bmatrix} 1/N & \dots & 1/N \\ \vdots & \ddots & \vdots \\ 1/N & \dots & 1/N \end{bmatrix} \begin{bmatrix} x_{11} & \dots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NP} \end{bmatrix} = \begin{bmatrix} \bar{x}_{:,1} & \dots & \bar{x}_{:,P} \\ \vdots & \ddots & \vdots \\ \bar{x}_{:,1} & \dots & \bar{x}_{:,P} \end{bmatrix}, \tag{29}
$$

where $\bar{x}_{:,p}$ is defined as

$$
\bar{x}_{:,p} = \frac{1}{N} \sum_{n=1}^{N} x_{n,p}, \tag{30}
$$

or the mean of a column of $\mathbf{X}$. Thus, we are just mean centering $\mathbf{X}_2$ in the calculation $\mathbf{X}_2 - \bar{\mathbf{X}}_2$. This gives us

$$
\hat{\boldsymbol{\beta}}_2 = \left\{ \mathbf{X}_2^\top (\mathbf{X}_2 - \bar{\mathbf{X}}_2) \right\}^{-1} \mathbf{X}_2^\top (\mathbf{y} - \hat{\mathbf{y}}). \tag{31}
$$

In other words, when OLS has an intercept, the optimal coefficients $\hat{\boldsymbol{\beta}}_2$—which are the parameters associated with the predictors in our design matrix—are just the result of the normal equation after mean-centering our targets and predictors. Intuitively, this means that the hyperplane defined by just $\hat{\boldsymbol{\beta}}_2$ goes through the origin (Figure 4).

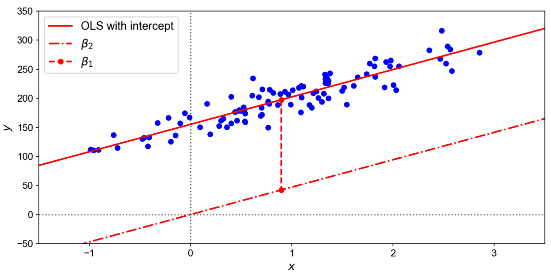

Figure 4. OLS with an intercept (solid line) can be decomposed into OLS without an intercept (dashed-and-dotted line) and a bias term (dashed line). Without an intercept, OLS goes through the origin. With an intercept, the hyperplane is shifted by the distance between the original hyperplane and the mean of the data.

Finally, we can solve for the scalar $\hat{\beta}_1$—which is really the intercept parameter β_0 in Equation 24—as

$$\begin{aligned} \hat{\beta}_1 &= (\mathbf{1}^{\top}\mathbf{1})^{-1}\mathbf{1}^{\top}(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) \\ &= \frac{1}{N}\mathbf{1}^{\top}(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) \\ &= \bar{y} - \bar{\mathbf{x}}_2\hat{\boldsymbol{\beta}}_2, \end{aligned} \tag{32}$$

where $\bar{\mathbf{x}}_2$ is a row of $\bar{\mathbf{X}}_2$, i.e. a vector of means for each predictor. The interpretation of this bias parameter (Equation 32) becomes more clear if we diagram these quantities (Figure 4). As we can see, OLS with an intercept can be decomposed into OLS without an intercept, where the hyperplane defined by $\hat{\boldsymbol{\beta}}_2$ passes through the origin, and the bias parameter β_1 (again elsewhere denoted β_0), which shifts the hyperplane from the origin to the mean of the response variables.

A probabilistic perspective

OLS can be viewed from a probabilistic perspective. Recall the linear model

$$y_n = \mathbf{x}_n^{\top}\boldsymbol{\beta} + \varepsilon_n. \tag{33}$$

If we assume our error ε_n is additive Gaussian noise, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, then this induces a conditionally Gaussian assumption on our response variables,

$$p(y_n \mid \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}(y_n \mid \mathbf{x}_n^{\top}\boldsymbol{\beta}, \sigma^2). \tag{34}$$

If our data is i.i.d., we can write

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = \prod_{n=1}^{N}\mathcal{N}(y_n \mid \mathbf{x}_n^{\top}\boldsymbol{\beta}, \sigma^2). \tag{35}$$

In this statistical framework, maximum likelihood (ML) estimation gives us the same optimal parameters as the normal equation. To compute the ML estimate, we first take derivative with respect to the parameter of the log likelihood function and then solve for $\boldsymbol{\beta}$. We can represent the log likelihood compactly using a multivariate normal distribution,

$$\log p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{36}$$

See A4 for a complete derivation of Equation 36. If we take the derivative of this log likelihood function with respect to the parameters, the first term is zero and the constant $1/2\sigma^2$ does not effect our optimization. Thus, we are looking for

$$\hat{\boldsymbol{\beta}}_{\texttt{MLE}} = \operatorname*{argmax}_{\boldsymbol{\beta}} \big\{ -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \big\}. \tag{37}$$

Of course, maximizing the negation of a function is the same as minimizing the function directly. Thus, this is the same optimization problem as Equations 7 and 8.

Furthermore, let $\boldsymbol{\beta}_0$ and σ_0^2 be the true generative parameters. Then

$$\begin{aligned} \mathbb{E}[\mathbf{y} \mid \mathbf{X}] &= \mathbf{X}\boldsymbol{\beta}_0 \\ \mathbb{V}[\mathbf{y} \mid \mathbf{X}] &= \sigma_0^2\mathbf{I}. \end{aligned} \tag{38}$$

See A5 for a derivation of Equation 38. Since we know that the conditional expectation is the minimizer of the mean squared loss—see my previous post if needed—, we know that $\mathbf{X}\boldsymbol{\beta}_0$ would be the best we can do given our model. An interpretation of the conditional variance in this context is that it is the smallest expected squared prediction error.

Conclusion

Classical linear regression or ordinary least squares is a linear model in which the estimated parameters minimize the sum of squared residuals. Geometrically, we can interpret OLS as orthogonally projecting our response variables onto a hyperplane defined by these linear coefficients. OLS typically includes an intercept, which shifts the hyperplane so that it goes through the targets' mean, rather than through the origin. In a probabilistic view of OLS, the maximum likelihood estimator is equivalent to the solution to the normal equation.

Acknowledgements

I thank Andrei Margeloiu for correcting an error in an earlier version of this post. When $\hat{\mathbf{y}} = \mathbf{y}$, the residuals are zero, but the true errors are still unknown.

Appendix

A1. Normal equation

We want to find the parameters or coefficients $\boldsymbol{\beta}$ that minimize the sum of squared residuals,

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{A1.1}$$

Note that we can write

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{A1.2}$$

This can be easily seen by writing out the vectorization explicitly. Let $\mathbf{v}$ be a vector such that

$$\mathbf{v} = \begin{bmatrix} y_1 - \mathbf{x}_1^{\top}\boldsymbol{\beta} \\ \vdots \\ y_N - \mathbf{x}_N^{\top}\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} x_{11} & \dots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NP} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_P \end{bmatrix}. \tag{A1.3}$$

The squared L2-norm $\|\mathbf{v}\|_2^2$ is the sums the squared components of $\mathbf{v}$. This is equivalent to taking the dot product $\mathbf{v}^{\top}\mathbf{v}$. Now define the function $J(\cdot)$ such that

$$J(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{A1.4}$$

To minimize $J(\cdot)$, we take its derivative with respect to $\boldsymbol{\beta}$, set it equal to zero, and solve for $\boldsymbol{\beta}$,

$$\begin{aligned} \nabla_{\boldsymbol{\beta}}J(\boldsymbol{\beta}) &\stackrel{1}{=} \nabla_{\boldsymbol{\beta}}\Big[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big] \\ &\stackrel{2}{=} \nabla_{\boldsymbol{\beta}}\Big[(\mathbf{y}^{\top} - \boldsymbol{\beta}^{\top}\mathbf{X}^{\top})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big] \\ &\stackrel{3}{=} \nabla_{\boldsymbol{\beta}}\Big[\boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta} - \mathbf{y}^{\top}\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^{\top}\mathbf{y} - \boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{y}\Big] \\ &\stackrel{4}{=} \nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta} - \mathbf{y}^{\top}\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^{\top}\mathbf{y} - \boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{y}\big) \\ &\stackrel{5}{=} \nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta}\big) - \nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\mathbf{y}^{\top}\mathbf{X}\boldsymbol{\beta}\big) + \cancel{\nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\mathbf{y}^{\top}\mathbf{y}\big)} - \nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{y}\big) \\ &\stackrel{6}{=} \nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta}\big) - 2\nabla_{\boldsymbol{\beta}}\operatorname{tr}\big(\boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{y}\big) \\ &\stackrel{7}{=} 2\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^{\top}\mathbf{y} \end{aligned} \tag{A1.5}$$

In step 4, we use the fact that the trace of a scalar is the scalar. In step 5, we use the linearity of differentiation and the trace operator. In step 6, we use the fact that $\operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\mathbf{A}^{\top})$. In step 7, we take the derivatives of the left and right terms using identities 108 and 103 from (Petersen et al., 2008), respectively.

If we set line 7 equal to zero and divide both sides of the equation by two, we get the normal equation:

$$\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\top}\mathbf{y}. \tag{A1.6}$$

A2. Hat matrix is an orthogonal projection

A square matrix is a projection if $\mathbf{A} = \mathbf{A}^2$. Thus, the hat matrix $\mathbf{H}$ is a projection since

$$\begin{aligned} \mathbf{H}^2 &= \mathbf{H}\mathbf{H} \\ &= \left(\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\right)\left(\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\right) \\ &= \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top} \\ &= \mathbf{H}. \end{aligned} \tag{A2.1}$$

A real-valued projection is orthogonal if $\mathbf{A} = \mathbf{A}^{\top}$. Thus, the hat matrix $\mathbf{H}$ is an orthogonal projection since

$$\begin{aligned} \mathbf{H}^{\top} &= (\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top})^{\top} \\ &= (\mathbf{X}^{\top})^{\top}[(\mathbf{X}^{\top}\mathbf{X})^{-1}]^{\top}\mathbf{X}^{\top} \\ &= \mathbf{H}. \end{aligned} \tag{A2.2}$$

A3. Residual maker is an orthogonal projector

A square matrix $\mathbf{A}$ is a projection if $\mathbf{A} = \mathbf{A}^2$. For the residual maker $\mathbf{M}$, we have:

$$\begin{aligned} \mathbf{M}^2 &= (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) \\ &= \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}^2 \\ &\overset{\star}{=} \mathbf{I} - \mathbf{H} \\ &= \mathbf{M}. \end{aligned} \tag{A3.1}$$

Step $\star$ holds because we know that the hat matrix $\mathbf{H}$ is an orthogonal projector. A real-valued projection $\mathbf{A}$ is orthogonal if $\mathbf{A} = \mathbf{A}^{\top}$. For the residual maker $\mathbf{M}$, we have:

$$\begin{aligned} \mathbf{M}^{\top} &= (\mathbf{I} - \mathbf{H})^{\top} \\ &= \mathbf{I}^{\top} - \mathbf{H}^{\top} \\ &\overset{\dagger}{=} \mathbf{I} - \mathbf{H}. \end{aligned} \tag{A3.2}$$

Again, step $\dagger$ holds because $\mathbf{H}$ is an orthogonal projector. Therefore, $\mathbf{M}$ is an orthogonal projector.

A4. Multivariate normal representation of the log likelihood

The probability density function for a D-dimensional multivariate normal distribution is

$$p(\mathbf{z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D \det(\boldsymbol{\Sigma})}} \exp\left\{ -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right\}. \tag{A4.1}$$

The mean parameter $\boldsymbol{\mu}$ is a D-vector, and the covariance matrix $\boldsymbol{\Sigma}$ is a $D \times D$ positive definite matrix. In the probabilistic view of classical linear regression, the data are i.i.d. Therefore, we can represent the likelihood function as

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{A4.2}$$

The above formulation leverages two properties from linear alegbra. First, if the dimensions of the covariance matrix are independent (in our case, each dimension is a sample), then $\boldsymbol{\Sigma}$ is diagonal, and its matrix inverse is just a diagonal matrix with each value replaced by its reciprocal. Second, the determinant of a diagonal matrix is just the product of the diagonal elements.

The log likelihood is then

$$\log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{A4.3}$$

as desired.

A5. Conditional expectation and variance

$$\begin{aligned} \mathbb{E}[\mathbf{y} \mid \mathbf{X}] &= \mathbb{E}[\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \mid \mathbf{X}] \\ &= \mathbb{E}[\mathbf{X}\boldsymbol{\beta}_0 \mid \mathbf{X}] + \mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}] \\ &= \mathbf{X}\boldsymbol{\beta}_0 + \mathbb{E}[\boldsymbol{\varepsilon}] \\ &= \mathbf{X}\boldsymbol{\beta}_0, \\ \\ \mathbb{V}[\mathbf{y} \mid \mathbf{X}] &= \mathbb{V}[\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \mid \mathbf{X}] \\ &= \mathbb{V}[\mathbf{X}\boldsymbol{\beta}_0 \mid \mathbf{X}] + \mathbb{V}[\boldsymbol{\varepsilon} \mid \mathbf{X}] \\ &= \mathbb{V}[\mathbf{X} \mid \mathbf{X}] + \mathbb{V}[\boldsymbol{\varepsilon}] \\ &= \sigma_0^2\mathbf{I}. \end{aligned} \tag{A5.1}$$

1. Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.
2. Petersen, K. B., Pedersen, M. S., & others. (2008). The matrix cookbook. *Technical University of Denmark*, *7*(15), 510.