# Minimum-norm OLS estimator with intercept

**Olivier Grisel**

## 1 Context and notation

Consider the minimum norm OLS estimator in the underdetermined case:

$$\begin{aligned} &\min_{w,w_0} \frac{1}{2}\left(w^T \cdot w + \gamma w_0^2\right) \\ &\text{s.t. } X \cdot w + w_0 1_n = y \end{aligned} \tag{1}$$

- $X$ holds the input feature values and has shape $(n, p)$;
- $y$ is the column vector has shape $(n, 1)$;
- $n$ is the number of samples;
- $p$ is the number of features;
- $w$ is a column vector of trainable parameter shape $(p, 1)$;
- $w_0$ is an extra scalar trainable parameter ("the intercept");
- $1_n$, $1_p$ are column vectors of ones of shape $(n, 1)$ and $(p, 1)$;
- $\overline{X}$ is the column vector of the mean of each column of $X$;
- $\overline{y}$ is the mean of $y$;
- $X_c$ is the centered version of $X$ such that $X_c = X - 1_n \cdot \overline{X}^T$;
- $y_c$ is the centered version of $y$ such that $y_c = y - \overline{y} 1_n$;
- $\gamma \in \{0, 1\}$ makes it possible decide whether we want to include the intercept in the computation of the norm or not.

Setting $\gamma = 1$ would yield the standard formulation which is equivalent to concatenating a column of 1 to $X$ to avoid having to handle a separate intercept coefficient. In this case we solve as in the standard presentations that omit the intercept such as [1].

However here are interested in the $\gamma = 0$ to compute the minimum norm OLS estimator where the magnitude of the intercept does not participate in the computation of the norm, to be consistent with the choice to not penalize the intercept in ridge regression for instance, and ensure the continuity of the solutions when $\alpha \to 0$.

## 2 Solving for $\gamma = 0$ with the method of Lagrange multipliers

Consider the centered data:

$$X = X_c + 1_n \cdot \overline{X}^T \tag{2}$$

$$y = y_c + \overline{y} 1_n \tag{3}$$

We can rewrite the generic formulation of the problem in Equation 1 as:

$$\begin{aligned} &\min_{w,w_0} \frac{1}{2}\left(w^T \cdot w + \gamma w_0^2\right) \\ &\text{s.t. } X_c \cdot w + 1_n \cdot \overline{X}^T \cdot w + w_0 \cdot 1_n = y_c + \overline{y} 1_n \end{aligned} \tag{4}$$

Let's introduce Langrange multipliers $\lambda$ to define our unconstrained objective function:

$$
\begin{aligned}
L(w, w_0, \lambda) = & \frac{1}{2} w^T \cdot w + \frac{\gamma}{2} w_0^2 \\
& + \lambda^T \cdot X_c \cdot w + \left(\lambda^T \cdot 1_n\right)\left(\overline{X}^T \cdot w\right) + w_0 \lambda^T \cdot 1_n \\
& - \lambda^T \cdot y_c - \overline{y} \lambda^T \cdot 1_n
\end{aligned} \tag{5}
$$

The minimizer of this objective function is a critical point:

- $\nabla L_w(w, w_0, \lambda) = 0_p$ yields:

$$
w + X_c^T \cdot \lambda + \left(\lambda^T \cdot 1_n\right) \overline{X} = 0_p \tag{6}
$$

- $\nabla L_{w_0}(w, w_0, \lambda) = 0$ yields:

$$
\gamma w_0 + \lambda^T \cdot 1_n = 0 \tag{7}
$$

- $\nabla L_\lambda(w, w_0, \lambda) = 0_n$ yields:

$$
X_c \cdot w + \left(\overline{X}^T \cdot w\right) 1_n + w_0 1_n = y_c + \overline{y} 1_n \tag{8}
$$

Right-multiplying Equation 8 by $1_n^T$ yields:

$$
1_n^T \cdot X_c \cdot w + \left(1_n^T \cdot 1_n\right)\left(\overline{X}^T \cdot w\right) + w_0 \left(1_n^T \cdot 1_n\right) = 1_n^T \cdot y_c + \overline{y} 1_n^T \cdot 1_n \tag{9}
$$

Since $1_n^T \cdot 1_n = n$, $1_n^T \cdot X_c = 0_p$ and $1_n^T \cdot y_c = 0$ we recover the usual:

$$
w_0 = \overline{y} - \overline{X}^T \cdot w \tag{10}
$$

Note that Equation 10 holds for any value of $\gamma$.

For the case where $\gamma = 0$, then Equation 7 becomes:

$$
\lambda^T \cdot 1_n = 0 \tag{11}
$$

and Equation 6 yields:

$$
w = - X_c^T \cdot \lambda \tag{12}
$$

and therefore:

$$
w_0 = \overline{y} + \overline{X}^T \cdot X_c^T \cdot \lambda \tag{13}
$$

Let's subtitute in $w_0$ and $w$ in Equation 8:

$$
-X_c \cdot X_c^T \cdot \lambda - \left(\overline{X}^T \cdot X_c^T \cdot \lambda\right) 1_n + \left(\overline{y} + \overline{X}^T \cdot X_c^T \cdot \lambda\right) 1_n = y_c + \overline{y} 1_n \tag{14}
$$

Hence, after simplification, and assuming $X_c \cdot X_c^T$ is invertible:

$$
\lambda = - \left(X_c \cdot X_c^T\right)^{-1} y_c \tag{15}
$$

and therefore the solution is:

$$
\begin{aligned}
\hat{w} \; & = - X_c^T \cdot \left(X_c \cdot X_c^T\right)^{-1} y_c \\
\hat{w}_0 & = \overline{y} - \overline{X}^T \cdot \hat{w}
\end{aligned} \tag{16}
$$

The minimum norm solution for the centered problem without intercept is also the minimum norm solution for the original problem (with intercept).

## Bibliography


[1] Stephen Boyd, "Least-norm solutions of undetermined equations," 2007. [Online]. Available: https://see.stanford.edu/materials/lsoeldsee263/08-min-norm.pdf