Review of OLS Regression

This version: October 25, 2023

1 Ordinary Least Squares (OLS) Regression

OLS aims at summarizing the dependency between $\boldsymbol{y} \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ by finding a linear combination of $\boldsymbol{X}$ that minimizes the mean squared error (MSE)

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2,$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip}, \end{pmatrix}$$

and

$$y_i - \boldsymbol{x_i'}\boldsymbol{\beta} = y_i - (x_{i1}\beta_1 + \cdots + x_{ip}\beta_p).$$

The coefficients $\hat{\boldsymbol{\beta}}$ that solve the optimization problem is the OLS estimator. One common way to specify the OLS estimator is to write down the OLS equation

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \epsilon_i,$$

where ϵ_i is the unobserved error term.

Remark 1. *When seeing a regression equation like*

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \epsilon_i,$$

it is important to distinguish whether it is a specification of the model (the data generation process) or a specification of the OLS estimator. Throughout the note, we will view such an equation as a specification of the OLS estimator only, i.e., we do not intend to assume that y_i is a linear function of $\boldsymbol{x}_i$.

Example 1 (simple regression)**.** *A simple regression is the case When the regression only contains an intercept and a regressor:*

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

One can show that

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \bar{y} - \widehat{\beta}\bar{x},$$

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

and $\hat{\beta}_1 \xrightarrow{p} \frac{Cov(X,Y)}{Var(X)}$. So, in this case, the OLS coefficient is just the covariance between Y and X scaled by the variance of X.

The primary usage of OLS regression are

1. Make prediction: how to predict y_{n+1} based on x_{n+1}? (Ans $\hat{y}_{n+1} = x_{n+1}$)
2. Estimate marginal effect: how does a 1 unit increase of if x_j change y? (Ans: β_j)

Example 2 (CLTV prediction)**.** *Recall that in customer management, one important variable is the customer lifetime value (CLTV), which can be measured by the consumption in the next year. We can use the recency-frequency-monetary (RFM) framework:*

- *Recency: when was the last time a customer make transaction?*

- *Frequency: how often does a customer make transaction?*
- *Monetary: how much does a customer usually spend?*

and can make predictions on CLTV by considering the regression of CLTV on RFM variables.

Example 3 (marketing mixing)**.** *A common challenge marketers face is the marketing mix problem. Suppose that a marketer has to allocate the advertisement budget on Google, Facebook, and Instagram. Let:*

- y_t *= sales on day* t
- $Google_t$ *= advertisement spending on Google on day* t
- FB_t *= advertisement spending on Facebook on day* t
- IG_t *= advertisement spending on Instagram on day* t

Then a regression of y_t *on* $Google_t$, FB_t, IG_t *can tell us which platform has higher impact on sales. The regression is a simple example of market mix regression.*

Quiz 1. *Suppose that Table 1 is our data.*

Table 1: Sample data

ID	Height	Male	Female
1	180	1	0
2	170	1	0
3	163	0	1
4	157	0	1

Consider the following three OLS estimators

$$Height_i = \alpha + \epsilon_i,$$
$$Height_i = \gamma_0 Female_i + \gamma_1 Male_i + \epsilon_i,$$
$$Height_i = \beta_0 + \beta_1 Male_i + \epsilon_i,$$

If you were to run a regression using the sample data, what would be $\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1$*?*

Notice that, in the three regressions specifications in Quiz 1, the OLS estimator are simply (differences of) sample means.[1] In fact, we can generalize this observation to more general situations as OLS regression is the approximation to conditional mean, which we will explore in the next part.

Remark 2. *A variable is called a dummy variable if its value only takes 0 or 1.*

2 Statistical Interpretation of OLS Regression

In the examples in Quiz 1, OLS estimators are simply (linear combinations of) conditional mean. This is not a coincidence as OLS is the "best linear approximation" of the conditional mean, and OLS is the conditional mean when the conditional mean is linear.

First, recall that conditional mean of Y given $X = x$ is

$$\mathbb{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\mathrm{d}y,$$

where

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \frac{f_{X,Y}(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

is the conditional p.d.f./ p.m.f. of Y given $\boldsymbol{X}$. For example, $\mathbb{E}\left[Height|Male = 1\right]$ is the average height of males, and $\mathbb{E}\left[Height|Male = 0\right]$ is the average height of females.

Remark 3. *It is often useful to think of condition mean as a function, that is,*

$$g(\boldsymbol{x}) = \mathbb{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right].$$

So the function $f(x)$ would give the average of Y given $\boldsymbol{X} = \boldsymbol{x}$. Moreover, you can think of

$$\mathbb{E}\left[Y|\boldsymbol{X}\right] = g(\boldsymbol{X})$$

as a transformation of the random variable $\boldsymbol{X}$, which is itself a random variable. For example, if the conditional mean $\mathbb{E}\left[Height|Male\right]$ will be equal to the average height of men if the random variable $Male = 1$ and average height of women if $Male = 0$.

[1] Actually, the last two regression would give the same prediction.

One useful property of the conditional mean is the **Law of Iterated Expectation**.

Proposition 1 (Law of Iterated Expectation)**.** *The Law of Iterated Expectation states that*

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|\boldsymbol{X}]].$$

The Law of Iterated Expectation is like an accounting equation. To calculate the average of Y, you can first calculate the average of Y within subgroups, and then take average across the subgroups again. For example, the average height, is the average of the average height among man and women.

Another important property of conditional mean is that, it is the best predictor under the L_2-loss. See the proposition below.

Proposition 2 (conditional mean is the best predictor)**.** *Consider all the possible functions f*

$$f : \mathcal{X} \to \mathcal{Y}$$

that maps from the support of X, $\mathcal{X}$, to the support of Y, $\mathcal{Y}$. One can think of these functions $f(\cdot)$ as predictors of Y based on the information X. We are interested in finding the best predictor

$$\min_f \mathbb{E}\left[(Y - f(X))^2\right]$$

that minimizers the L_2-loss (the mean squared error). Then

$$f^*(x) = \mathbb{E}[Y|X = x].$$

is the solution to the minimization problem. That is, the conditional mean is the best predictor of Y under the L_2-loss.

Proof. Before we proceed to the actual proof, let's consider first show that the quantity

$$\mathbb{E}\left[(Y - c)^2\right]$$

is minimized by $c^* = \mathbb{E}[Y]$. To show this, note that, for any constant $c \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\left[(Y-c)^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y] + \mathbb{E}[Y] - c)^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] - 2\mathbb{E}\left[(Y - \mathbb{E}[Y])\right](\mathbb{E}[Y] - c) + (\mathbb{E}[Y] - c)^2 \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] + (\mathbb{E}[Y] - c)^2
\end{aligned}$$

as the middle term

$$-2\mathbb{E}\left[(Y - \mathbb{E}[Y])\right](\mathbb{E}[Y] - c) = -2(\mathbb{E}[Y] - \mathbb{E}[Y])(\mathbb{E}[Y] - c) = 0.$$

Now, we can see that,

$$\begin{aligned}
&\mathbb{E}\left[(Y-c)^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] + (\mathbb{E}[Y] - c)^2 \\
&\geq \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] \\
&\geq Var(Y),
\end{aligned}$$

which is achieved by setting $c^* = \mathbb{E}[Y]$. The proof of the proposition follows an almost exact structure. For any predictor $f(\cdot)$,

$$\begin{aligned}
&\mathbb{E}\left[(Y - f(X))^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - f(X))^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] - 2\mathbb{E}\left[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f(X))\right] + \mathbb{E}\left[(\mathbb{E}[Y|X] - f(X))^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y|X] - f(X))^2\right],
\end{aligned}$$

since, by the law of iterated expectation,

$$\begin{aligned}
&\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X\right])(\mathbb{E}\left[Y|X\right]-f(X))\right] \\
&=\mathbb{E}\left[\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X\right])(\mathbb{E}\left[Y|X\right]-f(X))|X\right]\right] \\
&=\mathbb{E}\left[(\mathbb{E}\left[Y|X\right]-f(X))\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X\right])|X\right]\right] \\
&=\mathbb{E}\left[(\mathbb{E}\left[Y|X\right]-f(X))(\mathbb{E}\left[Y|X\right]-\mathbb{E}\left[Y|X\right])\right] \\
&=0.
\end{aligned}$$

Notice that, to minimize the objective function, one should set $f=\mathbb{E}\left[Y|X\right]$ as

$$\begin{aligned}
&\mathbb{E}\left[(Y-f(X))^2\right] \\
&=\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X\right])^2\right]+\mathbb{E}\left[(\mathbb{E}\left[Y|X\right]-f(X))^2\right] \\
&\geq\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X\right])^2\right],
\end{aligned}$$

which is obtained by setting $f(x)=\mathbb{E}\left[Y|X\right]$. □

Remark 4. *Proposition 2 only shows that the conditional mean is the predictor for the L_2-loss. However, there are loss functions such as the L_1-loss*

$$\mathbb{E}\left[|Y-f(X)|\right],$$

in which the conditional mean is actually ***not*** *the best predictor for this norm.*

Remark 5. *While the conditional mean is the best predictor, we still need to estimate it from the data.*

Now, we are ready to show that, the OLS regression is the "best linear approximation" to the conditional expectation function, and, therefore, the OLS regression is the best linear predictor. Surprisingly (or not surprisingly?), we can use the proof to show this. First, notice that the objective function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\boldsymbol{x_i'\beta})^2$$

is the sample analogue of

$$\mathbb{E}\left[(y_i - \boldsymbol{x_i'\beta})^2\right],$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 \xrightarrow{p} \mathbb{E}\left[(y_i - \boldsymbol{x_i'\beta})^2\right].$$

when $n \to \infty$.[2] So we can view OLS as solving the best linear predictor $\boldsymbol{x_i'\beta}$ that minimizes the average L_2 prediction error,

$$\mathbb{E}\left[(y_i - \boldsymbol{x_i'\beta})^2\right].$$

Recall earlier we show that

$$\mathbb{E}\left[(Y - f(X))^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y|X] - f(X))^2\right],$$

which holds for any function $f(\cdot)$. Now, if we plug in $f(x) = \boldsymbol{x'\beta}$, we get

$$\mathbb{E}\left[(Y - \boldsymbol{x'\beta})^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y|X] - \boldsymbol{x'\beta})^2\right].$$

Since the first part of expression does not involve on $\boldsymbol{\beta}$, minimizes

$$\mathbb{E}\left[(Y - \boldsymbol{x'\beta})^2\right]$$

is equivalent to

$$\mathbb{E}\left[(\mathbb{E}[Y|X] - \boldsymbol{x'\beta})^2\right].$$

Therefore, OLS regression amounts to finding best linear approximation of the conditional mean, and this is why we saw OLS estimators are (differences of) sample means in Quiz 1.

Example 4 (predicting CLTV with recency)**.** *Given what we learned so far, how would*

[2] In case you are wondering, the convergence is uniform in β, and this is called the *Uniform Law of Large Number*.

you predict a customer's LTV based on the recency, $R = 1, 2, 3, ...$(e.g., $R = 10$ means the last transaction made by the customer is 10 days ago.)? For one thing, we know $\mathbb{E}[CLTV|R]$ is the predictor, so that would be a reasonable choice. Is there any reason why we would prefer the linear regression

$$CLTV_i = \beta_0 + \beta_1 CLTV_i + \epsilon_i$$

over the conditional mean? Could you find an OLS regression that is equivalent to the conditional mean?

Quiz 2. *Is the regression model in Example 4 causal? Or, consider the regression:*

$$Covid_i = \beta_0 + \beta_1 Vaccine_i + \epsilon_i,$$

where $Covid_i$ is a dummy for whether a person gets Covid-19 and $Vaccine_i$ is a dummy for whether a person was vaccinated or not. Is this regression causal? Is the regression still useful even if it's not causal? Hint: what is $\hat{beta}_1$?

3 Identification problems in OLS

Identification may fail in OLS regression for two primary reasons. First reason is the lack of variations among the variables. Consider the following example.

Example 5. *Suppose Table 2 is our data. Consider the following regression*

Table 2: Failure of identification due to the lack of variations

ID	Height	Male	Female	Taipei
1	180	1	0	0
2	170	1	0	0
3	163	0	1	1
4	157	0	1	1

$$Height_i = \beta_0 + \beta_1 Female + \beta_2 Taipei + \epsilon_i.$$

Can you find $\hat{\boldsymbol{\beta}}$? No. It is impossible to separate the effect of Taipei from gender because

all people who were born in Taipei are female.

For another example, think of 徐乃麟 and 曾國城. [3] If you are interested estimate each of their effect on the box office, but, in your data, all the shows are co-host by them. Can you estimate their effects separately? No.

The second common reason why OLS is not identified is because you have too many independent variables. Consider the following example.

Example 6. *Suppose that $n = 3 < p = 4$, and the data is Table 3*

Table 3: Failure of identification due to high dimensionality

ID	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	10	1	0	0	0
2	12	0	1	0	0
3	2	0	0	1	0

As the data shown in Table, it is the third group that does not have any observation, then

$$\sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})^2$$

does not depend on β_3, so β_3 can be any number. The OLS estimator is not well-defined. Notice that, regardless of the actual values of $\boldsymbol{x}$, the model is bound to not identified as $n < p$.

Notice that in both Example 5 and 6, we have regression models that are **observationally equivalent**.

[3] They are two famous TV show hosts who frequently partner up with each other.