Gov 2001 Section 9, 2025

Regression (Predictive): Defining $\boldsymbol{\beta}$

1. Goal: Study (estimate) the relationship between a covariate (or predictor variables) $\mathbf{X} = \{X_1, X_2, \ldots, X_K\}$ and an outcome variable Y. This week, we only discuss how to properly derive our estimand and do not care about the estimator.

2. We would like to know the predicted outcome (expected value of the outcome) given the predictors with value $\mathbf{X} = \mathbf{x}$. It is thus in the form of a CEF:

$$\mu(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$$

   For discrete $\mathbf{X}$, we only need to estimate parameters $\mu(\mathbf{x})$ for each possible $\mathbf{x}$, and an easy way is to use subclassification, constructing a plug-in estimator:

$$\hat{\mu(\mathbf{x})} = \frac{\sum_{i=1}^{n} Y_i 1(\mathbf{X}_i = \mathbf{x})}{\sum_{i=1}^{n} 1(\mathbf{X}_i = \mathbf{x})}$$

   This is the average of outcomes when $\mathbf{X} = \mathbf{x}$.

3. However, if $\mathbf{X}$ is continuous, subclassification will not work, as we need an infinite number of parameters. We can limit our focus and study the linear relations between the predictors and outcome, assuming that the CEF follows a simple linear function:

$$\mu(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_K$$

   We often add an intercept for mathematical necessity. Intuitively, the outcome may not be 0 if all $x_1, \ldots, x_k$ are set to 0. In coding, this is sometimes called the "bias-merging trick": For example, we have a two-dimensional covariate data $\mathbf{x} = (x_1, x_2)$. We add an $x_0 = 1$ in the first position of the data, making it $(1, x_1, x_2)$. Then we will have

$$\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_K$$

   Now, we only need to estimate $\boldsymbol{\beta}$, with $k + 1$ estimands. Denote the estimator as $\hat{\boldsymbol{\beta}}$, and $\hat{\mu(\mathbf{x})} = \mathbf{x}'\hat{\boldsymbol{\beta}}$ is the fitted regression. We will discuss how to construct $\hat{\boldsymbol{\beta}}$ in detail next week.

4. The true value of outcome Y almost always differs from the conditional expectation, regardless of the model we use. We can write $Y = \mu(\mathbf{X}) + U(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}] + \varepsilon$, where ε is the regression error (see the last section), or $Y = \mathbf{x}'\boldsymbol{\beta} + \epsilon$, where ϵ is the projection error (residual, see the next section). Notice that in general $\varepsilon \neq \epsilon$.

Regression (Descriptive): Deriving $\boldsymbol{\beta}$

1. To derive the exact form of $\boldsymbol{\beta}$, we can interpret the regression in another way. Suppose we are interested in the relationship between two r.v.s X, Y (X is now set to one-dimensional

for simplicity), and the theoretical regression coefficient β_1 can be perceived as a summary (or description) of this joint distribution, defined as

$$\beta_{Y\sim X} = \frac{\text{Cov}(Y,X)}{\text{Var}(X)}$$

One way to show this is to find the β_1 that minimizes the expected difference between Y and $\beta_0 + \beta_1 X$, so that the latter is the best linear mimic of Y. We are solving:

$$(\beta_0, \beta_1) = \arg\min_{(a,b\in\mathbb{R}^2)} E[(Y - a - bX)^2]$$

One important theorem here is that we can differentiate under the integral sign (DUThIS):

$$\frac{d}{dt}\left(\int_a^b f(x,t)\,dx\right) = \int_a^b \frac{\partial}{\partial t} f(x,t)\,dx$$

Recall that expectation is essentially an integral $\int\int (y-a-bx)^2 f(x,y)\,dx\,dy$, where the joint PDF $f(x,y)$ is clearly free of a, b, so we can directly apply DUThIS to take derivatives and get the minimal expected difference (let it be D):

$$\frac{\partial D}{\partial a} = -2E[(Y - a - bX)] = 0, \quad \frac{\partial D}{\partial b} = -2E[X(Y - a - bX)] = 0$$

And $\beta_0 = a^*, \beta_1 = b^*$. Reorganize, we get

$$\beta_0 = E[Y] - \beta_1 E[X]$$

Plug in back to a^*, we have

$$E[XY - E[X]E[Y] + \beta_1 E[X]E[X] - \beta_1 E[X^2]] = 0$$

Reorganize, it is just

$$(E[XY] - E[X]E[Y]) = (E[X^2] - (E[X])^2)\beta_1$$

So we get $\beta_{Y\sim X} = \beta_1 = \frac{\text{Cov}(Y,X)}{\text{Var}(X)}$. If we center (X, Y) to $(\tilde{X} = X - E[X], \tilde{Y} = Y - E[Y])$ so that they have mean zero, then we have

$$\beta_{\tilde{Y}\sim\tilde{X}} = \frac{E[\tilde{X}\tilde{Y}]}{E[\tilde{X}^2]}$$

The linear regression can also be written as

$$\mu(x) = E[Y] + \beta_{Y\sim X}(x - E[X])$$

We also call it the linear projection of Y on X at $X = x$. It is the best linear function of x for approximating Y.

2. To estimate $\beta_{Y\sim X}$ with data $(X_{1:n}, Y_{1:n})$, we can simply use a plug-in estimator:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

We will discuss its properties next week.

3. Now consider the generalized case for a K-dimensional $\mathbf{X}$ (*Note*: the intercept term is now included as the first term X_1). We solve

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}\in\mathbb{R}^k} E[(Y - \mathbf{X}'\mathbf{b})^2]$$

For this one, we apply matrix calculus conclusions:

$$\frac{d\mathbf{x}^\top \mathbf{a}}{d\mathbf{x}} = \frac{d\mathbf{a}^\top \mathbf{x}}{d\mathbf{x}} = \mathbf{a}, \quad \frac{d\mathbf{x}^\top \mathbf{x}}{d\mathbf{x}} = 2\mathbf{x}$$

We use the chain rule (similar to the scalar case) and by DUThIS:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{b}} E[(Y - \mathbf{X}'\mathbf{b})^2] &= 2E\left[(Y - \mathbf{X}'\mathbf{b})\frac{\partial}{\partial \mathbf{b}}(Y - \mathbf{X}'\mathbf{b})\right] \\
&= -2E[(Y - \mathbf{X}'\mathbf{b})\mathbf{X}] \\
&= -2E[\mathbf{X}Y - \mathbf{X}\mathbf{X}'\mathbf{b}] = 0
\end{aligned}$$

Therefore, we have

$$\boldsymbol{\beta} = \frac{E[\mathbf{X}Y]}{E[\mathbf{X}\mathbf{X}']}$$

We can thus write the linear projection as

$$\mu(\mathbf{X}) = \mathbf{X}'(E[\mathbf{X}\mathbf{X}'])^{-1}E[\mathbf{X}Y]$$

4. The *projection error (residual)* of a regression is the difference between the true value and its predicted (fitted) value:

$$\epsilon = Y - \mathbf{X}'\boldsymbol{\beta}$$

Important property of projection error:

$$\begin{aligned}
E[\mathbf{X}\epsilon] &= E[\mathbf{X}(Y - \mathbf{X}'\boldsymbol{\beta})] \\
&= E[\mathbf{X}Y] - E[\mathbf{X}\mathbf{X}']\boldsymbol{\beta} \\
&= E[\mathbf{X}Y] - E[\mathbf{X}\mathbf{X}'](E[\mathbf{X}\mathbf{X}'])^{-1}E[\mathbf{X}Y] \\
&= E[\mathbf{X}Y] - E[\mathbf{X}Y] = 0
\end{aligned}$$

So for all $j \in \{1, 2, \ldots, K\}$, we have $E[X_j\epsilon] = 0$. When including the intercept term $X_1 = 1$ in the regression, we have $E[\epsilon] = E[X_1\epsilon] = 0$. Then the projection error is uncorrelated with covariates:

$$\mathrm{Cov}(\mathbf{X}, \epsilon) = E[\mathbf{X}\epsilon] - E[\mathbf{X}]E[\epsilon] = 0 - 0 = 0$$

Interpretation of Regression Coefficients

1. Simple linear case: $\mu(x_{1:K}) = \beta_0 + \sum_{j=1}^{K} \beta_K x_K$. We have for any $j \in \{1, \ldots, K\}$,

$$\beta_j = \mu(x_1, \ldots, x_j + 1, \ldots, x_k) - \mu(x_1, \ldots, x_j, \ldots, x_k)$$

   Interpretation: the change in the predicted outcome for increasing X_j by one unit.

2. Polynomial regression: for simplification, consider covariates (x_1, x_2) with quadratic term x_1^2. We have

$$\mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$$
$$\frac{\partial \mu}{\partial x_1} = \beta_1 + 2\beta_2 x_1$$

   β_1 or β_2 alone cannot capture x_1's contribution to the outcome. The slope of the predicted outcome to x_1 is now $\beta_1 + 2\beta_2 x_1$, so it depends on x_1. Therefore, $\beta_2 > 0$ implies that the relationship is convex, i.e., the effect of x_1 increases as x_1 grows; $\beta_2 < 0$ implies that the relationship is concave, i.e., the effect of x_1 decreases as x_1 grows.

3. Interaction term: for simplification, consider covariates (x_1, x_2) with interaction term $x_1 x_2$. We have

$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$
$$\frac{\partial \mu}{\partial x_1} = \beta_1 + \beta_3 x_2, \quad \frac{\partial \mu}{\partial x_2} = \beta_2 + \beta_3 x_1$$

   Therefore, $\beta_3 > 0$ means that the marginal effect of one of (x_1, x_2) on the predicted outcome increases as the other grows (they amplify each others' effects); $\beta_3 < 0$ means that the marginal effect of one of (x_1, x_2) on the predicted outcome decreases as the other grows (they dampen each others' effects).

Omitted Variable Bias

1. *Omitted variable bias* occurs when the model to estimate $E[Y|\mathbf{X} = \mathbf{x}]$ is incorrectly specified. Intuitively, let Z be an unobserved variable that also explains Y, so that the true relationship between $\mathbf{X}, Y, Z$ is $Y = \mathbf{X}'\boldsymbol{\beta} + \theta Z + U$. Z causes bias when it is also correlated with X, and let their true relationship be $Z = \mathbf{X}'\boldsymbol{\gamma} + V$. Therefore, the true relationship between X and Y is $Y = \mathbf{X}'(\boldsymbol{\beta} + \theta\boldsymbol{\gamma}) + \theta V + U$. So when we fit $\mu(\hat{\mathbf{X}}) = \mathbf{X}'\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta} + \theta\boldsymbol{\gamma}$ rather than $\boldsymbol{\beta}$.

2. One consequence of omitted variable bias is that the projection error's conditional expectation is non-zero:

$$E[\epsilon|\mathbf{X}] = E[Y - \mathbf{X}'\boldsymbol{\beta}|\mathbf{X}] = E[\theta Z + U|\mathbf{X}] = \theta E[Z|\mathbf{X}]$$

   which equals to 0 only when $Z \perp \mathbf{X}$. Notice that even though the projection and covariates have zero covariance, they are not necessarily independent.

3. In many situations, we may need to choose whether to include a variable in a regression, so it can be helpful to understand how this choice might affect the population coefficients on the other variables in the regression. Suppose we have a variable Z_i that we may add to our regression, which currently has $\mathbf{X}_i$ as the covariates. We can write the original regression as $\mu_{Y\sim X}(\mathbf{X}_i) = \mathbf{X}_i'\delta$ and the new projection as $\mu_{Y\sim X,Z}(\mathbf{X}_i, Z_i) = \mathbf{X}_i'\beta + Z_i\gamma$. By projection we have

$$\delta = \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_i Y_i].$$

Let $\epsilon_i = Y_i - \mu_{Y\sim X,Z}(\mathbf{X}_i, Z_i)$ be the projection errors from the long regression, we can plug in and rewrite this as

$$\begin{aligned}
\delta &= \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_i(\mathbf{X}_i'\beta + Z_i\gamma + \epsilon_i)] \\
&= \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\beta + \mathbb{E}[\mathbf{X}_i Z_i]\gamma + \mathbb{E}[\mathbf{X}_i\epsilon_i]\right) \\
&= \beta + \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_i Z_i]\gamma
\end{aligned}$$

Note that the vector in the second term is the vector of linear projection coefficients of a population linear regression of Z_i on the $\mathbf{X}_i$. If we call these coefficients π, then the short coefficients are

$$\delta = \beta + \pi\gamma.$$

We can rewrite this to show that the difference between the coefficients in these two projections is

$$\delta - \beta = \pi\gamma$$

or the product of the coefficient on the "excluded" Z_i and the coefficient of the included $\mathbf{X}_i$ on the excluded.

4. We introduce *Partitioned Regression* in general (you will learn more about this in the Frisch-Waugh-Lovell theorem). With a regression of an outcome on two covariates, understanding how the coefficients of one variable relate to the other is helpful. Consider the following best linear projection:

$$(\alpha, \beta, \gamma) = \arg\min_{(a,b,c)\in\mathbb{R}^3} \mathbb{E}\left[(Y_i - (a + bX_i + cZ_i))^2\right]$$

From the above results, we know that the intercept has a simple form:

$$\alpha = \mathbb{E}[Y_i] - \beta\mathbb{E}[X_i] - \gamma\mathbb{E}[Z_i].$$

Let's investigate the first order condition for β:

$$\begin{aligned}
0 &= \mathbb{E}[Y_i X_i] - \alpha\mathbb{E}[X_i] - \beta\mathbb{E}[X_i^2] - \gamma\mathbb{E}[X_i Z_i] \\
&= \mathbb{E}[Y_i X_i] - \mathbb{E}[Y_i]\mathbb{E}[X_i] + \beta\mathbb{E}[X_i]^2 + \gamma\mathbb{E}[X_i]\mathbb{E}[Z_i] - \beta\mathbb{E}[X_i^2] - \gamma\mathbb{E}[X_i Z_i] \\
&= \mathrm{Cov}(Y_i, X_i) - \beta\mathbb{V}[X_i] - \gamma\mathrm{Cov}(X_i, Z_i)
\end{aligned}$$

We can see from this that if $\mathrm{Cov}(X_i, Z_i) = 0$, then the coefficient on X_i will be the same as in the simple regression case, $\mathrm{Cov}(Y_i, X_i)/\mathrm{Var}(X_i)$. When X_i and Z_i are uncorrelated,

we call them **orthogonal**. To write a simple formula for β when the covariates are not orthogonal, we **orthogonalize** X_i by obtaining the prediction errors from a population linear regression of X_i on Z_i:

$$\tilde{X}_i = X_i - (\delta_0 + \delta_1 Z_i) \quad \text{where} \quad (\delta_0, \delta_1) = \arg\min_{(d_0,d_1)\in\mathbb{R}^2} \mathbb{E}\left[(X_i - (d_0 + d_1 Z_i))^2\right]$$

Given the properties of projection errors, we know $\text{Cov}(\tilde{X}_i, Z_i) = E[\tilde{X}_i Z_i] = 0$. This time, we can project $Y_i \sim \tilde{X}_i$, and

$$\beta_{Y_i \sim \tilde{X}_i} = \text{Cov}(Y_i, \tilde{X}_i)/\text{Var}(\tilde{X}_i)$$

You can substitute $X_i = \tilde{X}_i + (\delta_0 + \delta_1 Z_i)$ into the original regression and verify that $\beta_{Y_i \sim \tilde{X}_i}$ is the same as the coefficient β in multivariate regression. With $\mathbf{X}$, this holds for all X_k.

Regression Error and R^2 (Optional)

1. *Regression Error*: The difference between random and predicted outcome given predictors,

$$U(\mathbf{x}) = Y - E[Y|\mathbf{X} = \mathbf{x}] = Y - \mu(\mathbf{x})$$

It almost always exists and is *the part of Y that cannot be explained by X*, even if we knew the true conditional expectation of Y given X. Sometimes we call $\mu(\mathbf{x})$ the "signal term" and $U(\mathbf{x})$ the "noise term".

Note: Both the residual (projection error) and regression error are random variables. But we can observe and calculate the residual in real life, using the crystallized data $(\mathbf{X}, Y) = (\mathbf{x}, y)$, whereas the regression error is a theoretical noise that can never be observed.

Properties of the regression error (writing it for $\mathbf{X}$ random):

- $E[U(\mathbf{X})|\mathbf{X} = \mathbf{x}] = E[Y|\mathbf{X} = \mathbf{x}] - E[E[Y|\mathbf{X} = \mathbf{x}]] = 0$.
- By law of iterated expectation, $E[U(\mathbf{X})] = E[E[U(\mathbf{X})|\mathbf{X} = \mathbf{x}]] = 0$.
- For each $j \in \{1, \ldots, K\}$,

$$\text{Cov}(U(\mathbf{X}), X_j) = E[X_j U(\mathbf{X})] = E[E[X_j U(\mathbf{X})|\mathbf{X}]] = E[X_j E[U(\mathbf{X})|\mathbf{X}]] = E[0 X_j] = 0$$

Note: The above properties are always true for the regression error but not necessarily for the projection error (particularly when there is omitted variable bias).

2. R^2 Statistic: By EVVE, we have

$$\text{Var}(Y) = \text{Var}(\mu(X)) + \text{Var}(U(X))$$

We can decompose variation in outcome to variation in prediction (1st term) and variation in random noise (2nd term). We thereby define

$$R^2 = \frac{\text{Var}(\mu(X))}{\text{Var}(Y)} = 1 - \frac{\text{Var}(U(X))}{\text{Var}(Y)}$$

to be the proportion of variation in Y that is accounted for by total variation in prediction. Hence, R^2 is between 0 and 1, and having an R^2 statistic closer to 1 means less variation in outcome is due to random noise, which means that the model can explain our data better.