Predictive Inference 1: Refresh Linear Regression in Moderately High Dimensions

Alexander Quispe[1]

[1]PUCP

September 3, 2021

Best Linear Predictor: In Population

1. Regression as Best Linear Prediction Problem. For theoretical purposes we work with the population first.
   We consider a scalar random variable Y, an outcome of interest, and a vector of covariates

$$\begin{aligned} & X = (X_1, ..., X_p)' \\ & EY, EYC, EXX'; \textit{ Theoretical expected values} \end{aligned} \tag{1}$$

2. Best linear prediction rule for Y using X

$$\sum_{j=1}^{p} \beta_j X_j = \beta' X, \quad \textit{for} \;\; \beta = (\beta_1, ..., \beta_p) \tag{2}$$

   We define now β as any solution to the Best Linear Prediction (BLP) Problem:

$$\min_{b \in \mathbb{R}^p} E(Y - b'X)^2 \tag{3}$$

Best Linear Predictor: In Population

1. We can compute an optimal β by solving the First Order Conditions (FOC) for the BLP problem, called Normal Equations:

$$E(Y - \beta' X)X = 0 \tag{4}$$

Any optimal $b = \beta$ satisfies the Normal Equations. Defining the regression error as

$$\varepsilon := Y - b' X \tag{5}$$

we have the simple decomposition of Y:

$$Y = \beta' X + \varepsilon, \quad EX\varepsilon = 0 \tag{6}$$

2. $\beta' X$ is the part of Y that can be predicted and
3. ε is the unexplained or residual part.

Best Linear Predictor: In Sample

1. In applications the researcher does not have access to the population in total, but observes only a sample

$$(Y_i, X_i)_1^n = ((Y_1, X_1), ..., (Y_n, X_n)) \tag{7}$$

2. Best Linear Prediction Problem in the Sample:

$$\min_{b \epsilon \mathbb{R}^p} \mathbb{E}_n (Y_i - b' X_i)^2 \tag{8}$$

where β is any solution to the BLP problem in the sample. The βs are called the sample regression coefficients.

3. Again from FOC we have

$$\mathbb{E}_n X_i (Y_i - X_i' \hat{\beta}) = 0 \tag{9}$$

Best Linear Predictor: In Sample

1. defining the in-sample regression error as

$$\hat{\varepsilon}_i := (Y_i - \hat{\beta}' X_i) \tag{10}$$

we have the simple decomposition of Y:

$$Y_i = X_i'\hat{\beta} + \hat{\varepsilon}_i, \;\; \mathbb{E}_n X_i \hat{\varepsilon}_i = 0 \tag{11}$$

2. $X_i'\hat{\beta}$ is the part of Y that can be predicted and
3. $\hat{\varepsilon}_i$ is the unexplained or residual part.

Understanding β_1 via "Partialling-Out"

1. Lets imagine we have the next equation

$$Y = \underbrace{[\beta_1 D + \beta_2' W]}_{Predicted \quad value} + \underbrace{\varepsilon}_{error} \tag{12}$$

How does the predicted value of Y change if D increases by a unit while W remains unchanged?

2. **Partialling-out operation:** procedure that takes a random variable V and creates a "residual" $\widetilde{V}$ by subtracting the part of V that is linearly predicted by:

$$\widetilde{V} = V - \gamma_{VW}' W, \quad \gamma_{VW} = \arg\min_{\gamma} E(V - \gamma' W)^2 \tag{13}$$

3. We can show that

$$Y = V + U \implies \widetilde{Y} = \widetilde{V} + \widetilde{U} \tag{14}$$

Understanding β_1 via "Partialling-Out"

1. Partialling-out to both sides of our regression equation $Y = \underbrace{[\beta_1 D + \beta_2' W]}_{Predicted \quad value} + \underbrace{\varepsilon}_{error}$, we get:

$$\widetilde{Y} = \beta_1 \widetilde{D} + \beta_2' \widetilde{W} + \widetilde{\varepsilon} \tag{15}$$

which simplifies to :

$$\widetilde{Y} = \beta_1 \widetilde{D} + \varepsilon, \quad E\varepsilon\widetilde{D} = 0 \tag{16}$$

- $\beta_2' W \quad \widetilde{W} = 0$
- $\widetilde{\varepsilon} = \varepsilon$
- $E\varepsilon\widetilde{D} = 0$; since $\widetilde{D}$ is a linear function of $X = (D, W)$

2. What we found in (16) are the Normal Equations for the population regression of $\widetilde{Y}$ on $\widetilde{D}$.

Understanding β_1 via "Partialling-Out"

Theorem (Frisch-Waugh-Lovell, FWL)

The population linear regression coefficient β_1 can be recovered from the population linear regression of $\widetilde{Y}$ on $\widetilde{D}$.

$$\beta_1 = \arg\min_{b_1} E(\widetilde{Y} - \beta_1 \widetilde{D})^2 = (E\widetilde{D}^2)^{-1} E\widetilde{D}\widetilde{Y}, \quad E\widetilde{D}^2 > 0 \tag{17}$$

Theorem (Frisch-Waugh-Lovell, FWL)

Follow the next algorithm:

- *Regress Y on W, obtain residuals ε_1*
- *Regress D on W, obtain residuals ε_2*
- *Regress ε_1 on ε_2, obtain OLS estimates β_1*

Understanding β_1 via "Partialling-Out"

Theorem (Frisch-Waugh-Lovell, FWL)

In otherwords, β_1 can be interpreted as a (univariate) linear regression coefficient in the linear regression of ***residualized*** *Y* ***on residualized*** *D, where the residuals are defined by* ***partialling-out the linear effect of*** *W* ***from*** *Y* ***and*** *D.*

Understanding $\hat{\beta}_1$ via "Partialling-Out"

Theorem (Frisch-Waugh-Lovell, FWL-IN SAMPLE)

When we work with the sample, we mimic the partiallingout in the population.

$$\hat{\beta}_1 = \arg\min_{b_1} \mathbb{E}(\check{Y}_i - \beta_1 \check{D}_i)^2 = (\mathbb{E}\check{D}_i^2)^{-1}\mathbb{E}\check{D}_i\check{Y}_i \tag{18}$$

where $\check{V}_i$ denote the residual left after predicting V_i with controls W_i in the sample:

$$\check{V}_i = V_i - \hat{\gamma}'_{VW} W_i, \;\; \hat{\gamma}_{VW} = \arg\min_{\gamma} \mathbb{E}_n (V_i - \gamma' W_i)^2 \tag{19}$$