MS&E 228: Inference in Linear Models

Vasilis Syrgkanis

MS&E, Stanford

Linear Regression and the Best Linear Prediction (BLP) Problem

Predictive Modelling

- Let's switch our focus on solving a "predictive" problem
- Simply want to predict an outcome Y
- Having access to a vector of p covariates/features $X = \left(X_1, \ldots, X_p\right)'$

- *Convention:* $X_1 = 1$ (constant covariate)

Predictive Modelling: Mean Squared Error

- Want to construct a function f that "predicts" the value of Y from X
- If a new sample X comes from the same data generating process $f(X)$ is our "best guess" for the corresponding outcome Y
- Goal: minimize Expected or **Mean Squared Error** (MSE)

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

Best Predictive Model

- Goal: minimize Expected or Mean Squared Error (MSE)

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

- If f was allowed to take any "shape" then best function is the Conditional Expectation Function (CEF)

$$f_*(X) \coloneqq E[Y|X]$$

- Simple intuitive proof: **Variance Decomposition**

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

Tower Law of Expectations
$E[U\, h(X)] = E\big[E[Uh(X)|X]\big] = E[E[U|X]h(X)]$

Does not depend on the function f we choose

Is non-negative and takes value zero for $f(X) = E[Y|X]$

Best Linear Prediction (BLP) Problem

- Let's simplify things and just look at the best linear prediction
- Find a linear function of X

$$f(X) = \beta' X := \sum_{j=1}^{p} \beta_j X_j$$

- That minimizes the MSE

$$\min_{b \in \mathbb{R}^p} E\left[(Y - b'X)^2\right]$$

- We call $\beta' X$ the **Best Linear Predictor (BLP)** of Y using X

Best Linear Prediction (BLP) Problem

- The BLP minimizes the MSE

$$\min_{b \in \mathbb{R}^p} E\left[(Y - b'X)^2\right]$$

- Since by the variance decomposition

$$E\left[(Y - b'X)^2\right] = E\left[(Y - E[Y|X])^2\right] + E\left[(E[Y|X] - b'X)^2\right]$$

- First part does not depend on b. The BLP minimizes

$$\min_{b \in \mathbb{R}^p} E\left[(E[Y|X] - b'X)^2\right]$$

- The BLP is the **best linear approximation of the CEF**

Solving for the BLP

- Consider the MSE as a function of the parameter b
$$\text{MSE}(b) \coloneqq E\left[(Y - b'X)^2\right]$$
- Gradient of the MSE with respect to parameter b
$$\nabla_b \text{MSE}(b) \coloneqq E[(Y - b'X)X] = \begin{bmatrix} E[(Y - b'X)X_1] \\ \ldots \\ E\left[(Y - b'X)X_p\right] \end{bmatrix}$$
- The First Order Conditions (FOC) of the BLP problem
$$\nabla_b \text{MSE}(\beta) \coloneqq E[(Y - \beta'X)X] = 0$$
- Many times, referred to as the **Normal Equations**

Numerical Example

- Suppose $X = (1, W)$

$$Y = \gamma\, W^2 + \eta, \qquad \eta \sim N(0,1), W \sim N(0,1)$$

- The *Normal Equations*

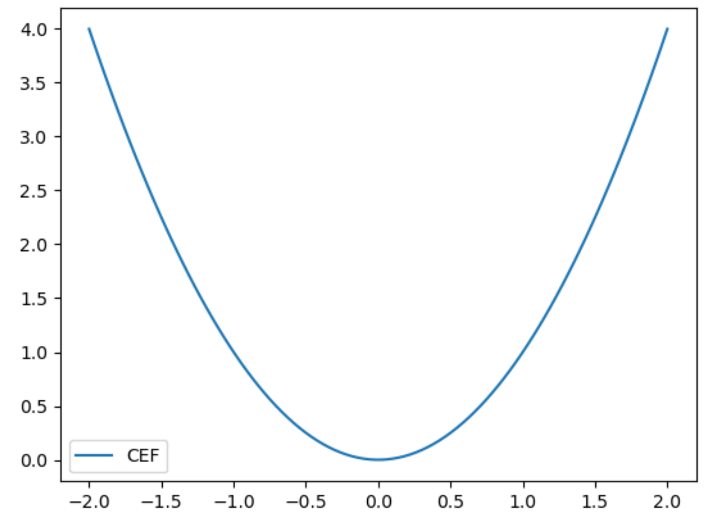
$$E[(Y \; - \beta' X)\; X] = E\left[\left(\gamma W^2 + \eta - \beta' X\right) X\right]$$
$$= E\left[\left(\gamma W^2 - \beta' X\right) X\right] = 0$$

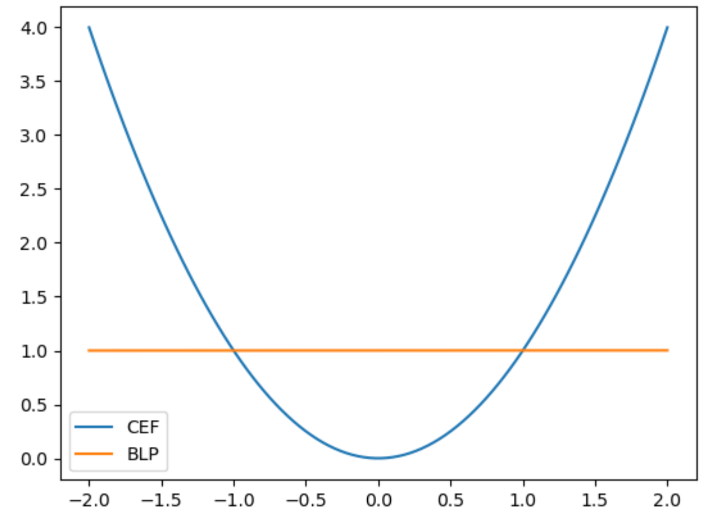
- Since $E[W] = 0, E\left[W^2\right] = 1, E\left[W^3\right] = 0$

$$E\left[\left(\gamma W^2 - \beta' X\right)\, 1\right] = \gamma - \beta_1 = 0$$
$$E\left[\left(\gamma W^2 - \beta' X\right) W\right] = \gamma \cdot 0 - \beta_2 = 0$$

- So $\beta_1 = \gamma, \beta_2 = 0$ and the BLP takes the form:

$$\beta' X = \gamma, \qquad (a\ constant\ prediction)$$





Decomposition of Y

- Define the regression error

$$\epsilon := Y - \beta' X$$

- We can re-write the Normal Equations as

$$E[\epsilon\, X] = 0$$

- We will use the shorthand notation

$$\epsilon \perp X \Leftrightarrow E[\epsilon\, X] = 0$$

- Thus we can decompose Y as

$$Y = \beta' X + \epsilon, \qquad \epsilon \perp X$$

Part of Y that can be **linearly predicted from X**

Remaining un-explained or **residual** part

Numerical Example

- Suppose $X = (1, W)$

$$Y = \gamma W^2 + \eta, \qquad \eta \sim N(0,1), W \sim N(0,1)$$

- Reminder $\beta_1 = \gamma, \beta_2 = 0$ and the BLP takes the form:

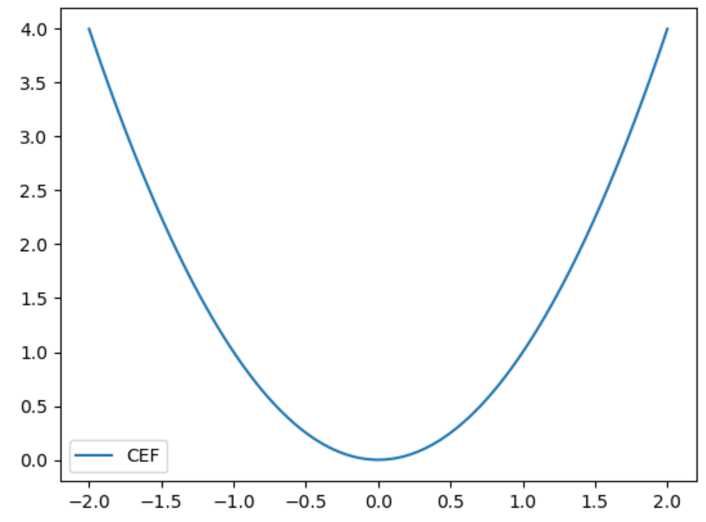
$$\beta' X = \gamma, \qquad (a\ constant\ prediction)$$

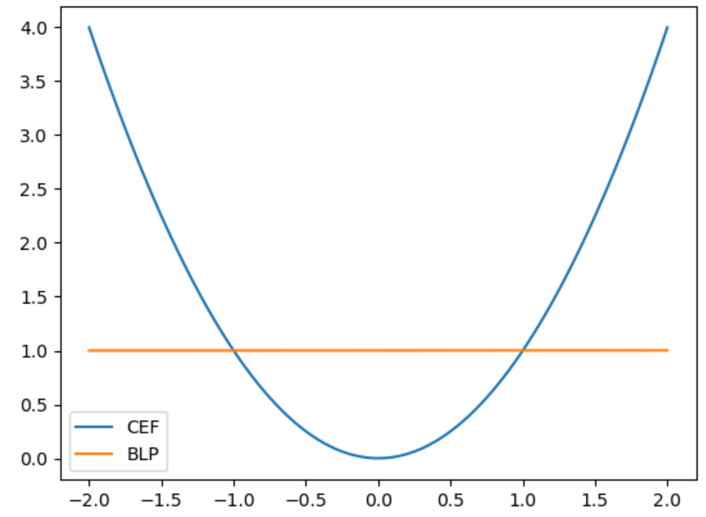
- We can decompose Y as

$$Y = \gamma + \epsilon, \qquad \epsilon \coloneqq \gamma\left(W^2 - 1\right) + \eta$$

- Note that

$$E[\epsilon] = \gamma\left(E\left[W^2\right] - 1\right) = 0$$
$$E[\epsilon W] = \gamma\left(E\left[W^3\right] - E[W]\right) = 0$$





Even if the relationship between outcome Y and covariates X is non-linear, we can always write:

$$Y = \beta' X + \epsilon, \qquad E[\epsilon X] = 0$$

The function $\beta' X$ is the Best Linear Predictor (BLP) or equivalently the best linear approximation to the Conditional Expectation Function (CEF) $E[Y|X]$

Finite Sample Estimation

BLP in Sample

- We have access to n samples
$$(X_1, Y_1), \dots, (X_n, Y_n)$$
- Drawn independent and identically distributed (i.i.d.) according to the distribution F of the random variables (X, Y)

- Consider the empirical analogue of the best linear predictor
- Replace expectations with empirical averages $E_n[Z] = \frac{1}{n}\sum_{i=1}^{n} Z_i$

BLP in Sample: Ordinary Least Squares (OLS)

- Find a linear prediction rule

$$\hat{f}(X) = \hat{\beta}' X$$

- That minimizes the Sample Mean Squared Error

$$\min_{b \in \mathbb{R}^p} E_n\left[(Y - b'X)^2\right] := \frac{1}{n} \sum_{i=1}^{n} (Y_i - b'X_i)^2$$

- Parameters $\hat{\beta}$ are called *sample regression coefficients*

Sample Normal Equations

- The First Order Conditions (FOC) of the Sample BLP problem

$$E_n[(Y - \beta' X)X] = 0$$

- Referred to as the **Sample Normal Equations**

Sample Decomposition of Y

- In sample regression error

$$\hat{\epsilon}_i := Y_i - \hat{\beta}' X_i$$

- We can decompose Y_i as

$$Y_i = \hat{\beta}' X_i + \hat{\epsilon}_i, \qquad E_n[\hat{\epsilon}\, X] = 0$$

Even if the relationship between outcome Y and covariates X is non-linear, we can always write:

$$Y_i = \hat{\beta}' X_i + \hat{\epsilon}_i, \qquad E_n[\hat{\epsilon} X] = 0$$

The function $\hat{\beta}' X$ is the Best Linear Predictor in sample and $\hat{\beta}$ are the sample regression coefficients

How Good is OLS in Recovering BLP

- Is Sample BLP $\hat{\beta}$ (OLS coefficients) close to BLP β?
- The distance between these two quantities depends on the number of parameters we are estimating
- We are estimating p un-constrained parameters from n noisy samples

- We should not expect error to be small if p/n is large
- How does error scale with this ratio?

Approximation of BLP by OLS

Theorem. Under regularity conditions, with probability approaching 1 as $n \to \infty$

$$\sqrt{E_X\left[\left(\beta' X \; - \hat{\beta}' X\right)^2\right]} \le const_F \cdot \sqrt{E[\epsilon^2]}\sqrt{\frac{p}{n}}$$

- E_X expectation with respect to X
- $const_F$ a constant that depends on the distribution F of (X, Y)

Conclusion. If n is large and p is small, for all realizations of data OLS is close to BLP

$$\sqrt{E_X\left[\left(\beta' X \; - \hat{\beta}' X\right)^2\right]} \approx 0$$

$A_n \approx B_n$ means that distance of A_n, B_n concentrates around 0 for some measure of distance d:

$$\forall \epsilon > 0 \colon \lim_{n \to \infty} P(d(A_n, B_n) \le \epsilon) = 1$$

You should expect OLS to produce accurate predictions in the worst-case only if the number of variables is small compared to number of samples.

Its predictions converge to the predictions of the BLP in the population

Interpretable Performance Measures via Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA)

- Reminder: decomposition of Y

$$Y = \beta' X + \epsilon, \qquad E[\epsilon X] = 0$$

- By orthogonality property of residual in the decomposition of Y

$$E\left[Y^2\right] = E\left[(\beta' X + \epsilon)^2\right] = E\left[(\beta' X)^2\right] + E\left[\epsilon^2\right] + 2\beta' E[X\epsilon]$$

- We can decompose the variation in Y, i.e. $E\left[Y^2\right]$, as

$$E\left[Y^2\right] = \underbrace{E\left[(\beta' X)^2\right]}_{\text{explained variation}} + \underbrace{E\left[\epsilon^2\right]}_{\text{residual variation}}$$

Analysis of Variance (ANOVA)

- Reminder: decomposition of Y

$$Y = \beta' X + \epsilon, \qquad E[\epsilon X] = 0$$

- We can decompose the variation in Y, i.e. $E\left[Y^2\right]$, as

$$E\left[Y^2\right] = E\left[(\beta' X)^2\right] + E\left[\epsilon^2\right]$$

- MSE: mean squared prediction error

$$MSE_{pop} = E\left[\epsilon^2\right]$$

- R-squared R^2: Ratio of explained to total variation

$$R^2_{pop} := \frac{explained\ variation}{total\ variation} = \frac{E\left[(\beta' X)^2\right]}{E[Y^2]} = 1 - \frac{E\left[\epsilon^2\right]}{E[Y^2]} \in [0, 1]$$

Standard and advisable definition of R^2 assumes Y is centered (i.e. mean-zero)

Performance Evaluation

In Sample R-squared and MSE

- Decomposition in sample

$$Y_i = \hat{\beta}' X_i + \hat{\epsilon}_i, \qquad E_n[\hat{\epsilon}\, X] = 0$$

- Decomposition of variation in sample

$$E_n[Y^2] = E_n\left[\left(\hat{\beta}' X\right)^2\right] + E_n[\hat{\epsilon}^2]$$

- MSE in sample

$$MSE_{sample} = E_n[\hat{\epsilon}^2]$$

- R-squared in sample

$$R^2_{sample} := \frac{E_n\left[\left(\hat{\beta}' X\right)^2\right]}{E_n[Y^2]} = 1 - \frac{E_n[\hat{\epsilon}^2]}{E_n[Y^2]} \in [0, 1]$$

Standard and advisable definition of R^2 assumes Y is centered (i.e. mean-zero)

When are these good proxies?

- When p/n is small and n is large
- By Law of Large Numbers (LLN) and guarantee theorem for $\hat{\beta}$
- Sample measures are good approximations to population measures

$$R^2_{sample} \approx R^2_{pop}, \qquad MSE_{sample} \approx MSE_{pop}$$

Overfitting: p/n large

- When p/n is large, in-sample BLP performance is mis-leading
- Artificially much smaller than true performance
- Consider case when $n = p$ and variables X linearly independent
- Then we can always find a parameter $\hat{\beta}$ matches Y on samples
$$Y_i = \hat{\beta}' X_i$$
- View it as system of n equations, let $\bar{Y} = (Y_1, \dots, Y_n)$, $\bar{X} = (X_1'; \dots; X_n')$
$$\bar{Y} = \bar{X} b$$
- Since variables are linearly independent, matrix $\bar{X}$ is full rank and invertible
$$MSE_{sample} = 0, \qquad R^2_{sample} = 1$$

An Improvement: Adjusted Measures

- Adjust by factor that relates to ratio p/n
- MSE in sample

$$MSE_{adjusted} = \frac{1}{1 - p/n} E_n\left[\hat{\epsilon}^2\right]$$

- R-squared in sample

$$R^2_{adjusted} \coloneqq 1 - \frac{1}{1 - p/n} \frac{E_n\left[\hat{\epsilon}^2\right]}{E_n[Y^2]} \in [0, 1]$$

- Provably better measures in homoscedastic case (ϵ independent of X)

Sample Splitting: Reliable Performance Measure

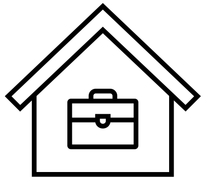
- Use a random subset of $m < n$ of the samples, called the *training set*, to estimate/train the prediction rule $\hat{f}$, e.g. $\hat{f}(X) = \hat{\beta}'X$
- Use the remaining $s = n - m$ samples, called the *test set, denoted as* V, to evaluate the quality of the prediction rule, via R^2 and MSE

$$MSE_{test} = \frac{1}{s} \sum_{k \in V} \left(Y_k - \hat{f}(X_k) \right)^2$$

$$R^2_{adjusted} \coloneqq 1 - \frac{MSE_{test}}{\frac{1}{s} \sum_{k \in V} Y_k^2} \in [0, 1]$$

Stratification

- In moderately sized samples it helps to ensure that train and test set are similar
- In large samples, randomness will guarantee that
- In small samples and with categorical variables, it is advisable to stratify, i.e. split samples in a manner that proportion of samples with each categorical value are similar on each of the two samples



Almost always measure predictive performance of your estimated model on a held-out sample

Inference on Predictive Effects

Predictive Effect

- For some *target* regressor/covariate D of interest
- How do (best linear) predictions change in the population limit, if the value of D changes by a unit, while other regressors are fixed?
- We'll call this the *predictive effect*!
- Partition $X = (D, W)$. We can write:

$$Y = \beta_1 D + \beta_2' W + \epsilon$$

$$\text{Predictive Effect} = \beta_1$$

Partialling-Out

Understanding β_1

- Consider the following partialling out operation

- For any random variable V, let $\tilde{V}$ be the residual of V after subtracting the part of V that is linearly predictable from W

$$\tilde{V} = V - \gamma_V' W, \qquad \gamma_W \in \mathrm{argmin}_\gamma E\left[(V \ - \gamma' W)^2\right]$$

- Note that we can also write the standard decomposition:

$$V = \gamma_V' W + \tilde{V}, \qquad E\left[\tilde{V} W\right] = 0$$

Linearity of Partialling-Out

- Partialling out is a "linear" operation, i.e.

$$Y = V + U \Rightarrow \tilde{Y} = \tilde{V} + \tilde{U}$$

- *Hint:* by decompositions of V, U we have $\gamma_V + \gamma_U$ is BLP of Y using W

$$Y = (\gamma_V + \gamma_U)'W + \tilde{V} + \tilde{U}, \qquad E\left[W\left(\tilde{V} + \tilde{U}\right)\right] = 0$$

Some Magic

- Consider decomposition of Y when considering the BLP using (D, W)

$$Y = \beta_1 D + \beta_2' W + \epsilon, \qquad E[\epsilon\,(D; W)] = 0$$

- Apply linearity of partialling out process

$$\tilde{Y} = \beta_1 \widetilde{D} + \beta_2' \widetilde{W} + \tilde{\epsilon}$$

- Trivially W is fully predictable linearly from W, i.e. $\widetilde{W} = 0$
- Since ϵ is orthogonal to W it is not at all predictable linearly, $\tilde{\epsilon} = \epsilon$

$$\tilde{Y} = \beta_1 \widetilde{D} + \epsilon, \qquad E\left[\epsilon \widetilde{D}\right] = E[\epsilon\,(D - \gamma_D' W)] = 0$$

- β_1 solves the Normal Equations for the regression of $\tilde{Y}$ on $\widetilde{D}$!

Frisch-Waugh-Lovell (FWL) Theorem!

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

$$\beta_1 = \mathrm{argmin}_b E\left[\left(\tilde{Y} - b\,\widetilde{D}\right)^2\right] = \frac{E\left[\tilde{Y}\widetilde{D}\right]}{E\left[\widetilde{D}^2\right]}$$

- We made the assumption that $E\left[\widetilde{D}^2\right] > 0$, i.e. D is not perfectly linearly predictable from W

Predictive effect β_1 of *target variable* is the coefficient in a *simple one variable regression*

$$\begin{pmatrix}\text{part of outcome}\\ \text{un−explained by other}\end{pmatrix} \sim \begin{pmatrix}\text{part of target}\\ \text{un−explained by other}\end{pmatrix}$$

FWL in Sample: Exact same arguments can be repeated in sample

- For any random variable V, let $\check{V}$ be the residual of V after subtracting the part of V that is linearly predictable from W in sample

$$\check{V} = V - \hat{\gamma}_V' W, \qquad \hat{\gamma}_W \in \operatorname{argmin}_\gamma E_n\left[(V \ - \gamma' W)^2\right]$$

- The sample linear regression coefficient $\hat{\beta}_1$ can be recovered from the sample linear regression of $\check{Y}$ on $\check{D}$

$$\hat{\beta}_1 = \operatorname{argmin}_b E_n\left[\left(\check{Y} - b\,\check{D}\right)^2\right] = \frac{E_n\left[\check{Y}\check{D}\right]}{E_n\left[\check{D}^2\right]}$$

- We made the assumption that $E_n\left[\check{D}^2\right] > 0$, i.e. D is not perfectly linearly predictable from W in sample

Coefficient of D in OLS(y~D,W) is mathematically equivalent in samples to

yres = y - OLS(y~W).predict(W)
Dres = D - OLS(D~W).predict(W)

Coefficient of Dres in OLS(yres~Dres)

Asymptotic Distribution and Confidence Intervals

Adaptive Inference

$A_n \overset{\text{a}}{\sim} N(0, V)$ means that as $n \to \infty$:

$$\sup_{R \in \mathcal{R}} |P(A_n \in R) - P(N(0, V) \in R)| \approx 0$$

where $\mathcal{R}$ set of all hyper-rectangles

- Under regularity conditions, if p/n is small, the estimation error in $\check{D}_i, \check{Y}_i$ has no first-order effect on the asymptotic stochastic behavior of $\hat{\beta}_1$

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \approx \sqrt{n} \frac{E_n\left[\epsilon\, \widetilde{D}\right]}{E_n\left[\widetilde{D}^2\right]}$$

- By application of LLN and CLT

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \overset{\text{a}}{\sim} N(0, V)$$

- With asymptotic variance

$$V = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]}$$

- The same statement also holds with estimate of the variance $\hat{V} = \frac{E_n\left[\hat{\epsilon}^2 \check{D}^2\right]}{E_n\left[\check{D}^2\right]}$

Confidence Interval

- $X \overset{a}{\sim} Y \equiv \sup_{[\ell,u]} |P(X \in [\ell,u]) - P(Y \in [\ell,u])| \approx 0$

- If we consider $[\ell,u]$ the $\left(\frac{\alpha}{2}, 1-\frac{\alpha}{2}\right)$ quantile of $N\left(0,\hat{V}\right)$ then

$$P\left(\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \in [\ell,u]\right) \approx 1-\alpha$$

- Equivalently, let z_α the α quantile of $N(0,1)$ and $\boxed{\hat{\sigma}_n = \sqrt{\hat{V}/n}}$ then:

Standard error

$$P\left(\beta_1 \in \left[\hat{\beta}_1 - z_{1-\frac{\alpha}{2}}\hat{\sigma}_n, \hat{\beta}_1 + z_{1-\frac{\alpha}{2}}\hat{\sigma}_n\right]\right) \approx 1-\alpha$$

If we want an interval that roughly contains the predictive effect with probability α, we can use

$$CI(\alpha) := \left[\hat{\beta}_1 - z_{1-\frac{\alpha}{2}}\hat{\sigma}_n, \hat{\beta}_1 + z_{1-\frac{\alpha}{2}}\hat{\sigma}_n\right]$$

$$\hat{\sigma}_n := \frac{1}{\sqrt{n}}\sqrt{\frac{E_n\left[\hat{\epsilon}^2\check{D}^2\right]}{E_n\left[\check{D}^2\right]}}$$

e.g. for 95% confidence interval, $z_{1-\frac{\alpha}{2}} \approx 1.96$

Example: Wage Gap based on Sex Indicator

Revisit Covariate Adjustment for Effect Inference in Experiments

Co-variates for Precision

- Even if we are only interested on ATE covariates can be valuable for precision
- Suppose variance of y is large but can be explained largely by W
- Then we can use W to remove all the explained variation from y
- Then perform our ATE analysis on the remnant variation
- This is oftentimes performed in practice via ordinary linear regression of y on the vector $(1, D, W)$ (after centering W, i.e. $E[W] = 0$)

Is this consistent?

- Suppose that the conditional expectation function (CEF) of the outcome is indeed linear, with $(D, 1, W)$

$$E[\,Y \mid D, W\,] = D\alpha + \alpha_0 + W'\beta$$

- Then note that

$$E[Y(0)] = E\big[E[Y|D=0, W]\big] = \alpha_0$$
$$E[Y(1)] = E\big[E[Y|D=1, W]\big] = \alpha + \alpha_0$$

- Baseline outcome is coefficient associated with the intercept 1
- Average effect is coefficient associated with treatment D
- Next lecture: this does not require the linear CEF assumption

Is this consistent? Beyond Linear CEF

- By the BLP decomposition of Y using $(D, 1, W)$
$$Y = D\alpha + \alpha_0 + \beta' W + \epsilon, \qquad E[\epsilon\,(D; 1; W)] = 0$$
- Note that the quantity:
$$U = \beta' W + \epsilon$$
- Also satisfies
$$\begin{aligned} E[U\,(D;1)] &= \beta' E[W\,(D;1)] + E[\epsilon(D;1)] \\ &= \beta' E[W]\, E[(D;1)] = 0 \end{aligned}$$

$= 0$ by orthogonality of ϵ

By independence of W and D

Since $E[W] = 0$ (de-meaned W)

Is this consistent? Beyond Linear CEF

- By the BLP decomposition of Y using $(D, 1, W)$

$$Y = D\alpha + \alpha_0 + \beta' W + \epsilon, \qquad E[\epsilon\,(D; 1; W)] = 0$$

- So we can write

$$Y = D\alpha + \alpha_0 + U, \qquad E[U\,(D; 1)] = 0$$

- Thus α, α_0 solve the Normal Equations of Y on $D, 1$
- α, α_0 are the BLP of Y using $(D, 1)$

$$E[(Y - \alpha D - \alpha_0)D] = 0 \Rightarrow E[Y|D = 1] = \alpha + \alpha_0$$
$$E[(Y - \alpha D - \alpha_0)\,1] = 0 \Rightarrow E[Y - \alpha D - \alpha_0|D = 0] = 0 \Rightarrow E[Y|D = 0] = \alpha_0$$

- Coefficients α, α_0 are identical to the two means estimate and consistent for ATE

$$\alpha = E[Y|D = 1] - E[Y|D = 0] = E\left[Y^{(1)} - Y^{(0)}\right]$$
$$a_0 = E[Y|D = 0] = E\left[Y^{(0)}\right]$$

The coefficient associated with treatment D in OLS with co-variate adjustment is always consistent for the treatment effect, when run on data from a randomized experiment, as-long-as covariates are de-meaned. The true relationship of outcome with covariates does not need to be linear.