12: Ordinary Least Square

Naijia Liu

Spring 2025

Partitioned regression

- Partition covariates and coefficients $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$ and $\beta = (\beta_1, \beta_2)'$:
$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{e}$$
- Can we find expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$?
- **Residual regression** or Frisch-Waugh-Lovell theorem to obtain $\hat{\beta}_1$:
  - Use OLS to regress $\mathbf{Y}$ on $\mathbf{X}_2$ and obtain residuals $\tilde{\mathbf{e}}_2$
  - Use OLS to regress each column of $\mathbf{X}_1$ on $\mathbf{X}_2$ and obtain residuals $\tilde{\mathbf{X}}_1$
  - Use OLS to regress $\tilde{\mathbf{e}}_2$ on $\tilde{\mathbf{X}}_1$

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \ldots, X_n)$ and recall inner product:

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{n} X_i Y_i$$

  - Inner products measure how similar two vectors are.
- Slope in this case:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle}$$

- Suppose we add an **orthogonal covariate** $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}$ with $\langle \mathbf{X}, \mathbf{Z} \rangle = 0$:

$$\hat{\beta} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle}, \quad \hat{\gamma} = \frac{\langle \mathbf{Z}, \mathbf{Y} \rangle}{\langle \mathbf{Z}, \mathbf{Z} \rangle}$$

-
  - With exactly orthogonal covariates, multivariate OLS is the same as univariate OLS
  - Only holds in balanced, designed experiments.

Adding the intercept

- Consider the OLS slope with an intercept:

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

- How can we get this?
  1. Regress $\mathbf{X}$ on 1 to get coefficient $\bar{X}$
  2. Regress $\mathbf{Y}$ on residuals from step 1, $\mathbf{X} - \bar{X}\mathbf{1}$
- If we wanted to get the coefficient on added variable Z_i, we could repeat this:
  1. Regress $\mathbf{Z}$ on $\tilde{\mathbf{X}} = \mathbf{X} - \bar{X}\mathbf{1}$ and obtain coefficient $\langle \mathbf{Z}, \tilde{\mathbf{X}}\rangle / \langle \tilde{\mathbf{X}}, \tilde{\mathbf{X}}\rangle$
  2. Regress $\mathbf{Y}$ on residual from ...

Why does residual regression work?

- We can find $\hat{\beta}_1$ by nested minimization:

$$\hat{\beta}_1 = \arg\min_{\beta_1} \left(\min_{\beta_2} \|\mathbf{Y} - \mathbf{X}_1\beta_1 - \mathbf{X}_2\beta_2\|^2 \right)$$

  - First find the minimum of the SSR over β_2, fixing β_1
  - Then find β_1 that minimizes the resulting SSR
- The projection and annihilator matrices are defined only by covariates.

$$\mathbf{M}_2 = I_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$$

  - Creates residuals from a regression on or $\mathbf{X}_2$
- Solving the nested minimization gives:

$$\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{M}_2\mathbf{Y})$$

- When will $\hat{\beta}_1$ be the same regardless of whether $\mathbf{X}_2$ is included?
  - If $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, so $\mathbf{X}_2'\mathbf{X}_1 = 0$ and $\mathbf{M}_2\mathbf{X}_1 = \mathbf{X}_1$

Residual regression

- Define two sets of residuals:
  - $\tilde{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2 =$ residuals from regression of $\mathbf{X}_2$ on $\mathbf{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1\mathbf{Y} =$ residuals from regression of $\mathbf{Y}$ on $\mathbf{X}_1$
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$\begin{aligned}\hat{\beta}_2 &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}) \\ &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{Y}) \\ &= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}(\tilde{\mathbf{X}}_2'\tilde{\mathbf{e}}_1)\end{aligned}$$

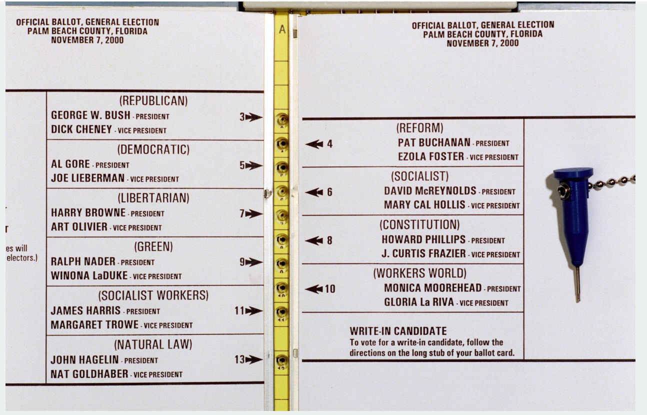
- $\hat{\beta}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\tilde{\mathbf{X}}_2$.
  - Same result applies when using $\mathbf{Y}$ in place of $\tilde{\mathbf{e}}_1$
  - Intuition: residuals are orthogonal
  - Called the **Frisch-Waugh-Lovell Theorem**
  - Sample version of the results we saw for the linear projection

Outliers, leverage points, and influential observations

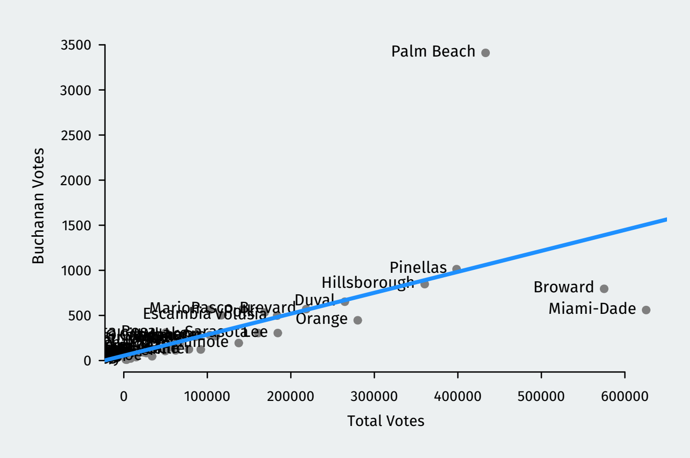
- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
  - Dropping them leads to large changes in the estimated $\hat{\beta}$.
  - Not all "unusual" observations have the same effect, though.
- Useful to categorize:
  1. **Leverage point**: extreme in one X direction
  2. **Outlier**: extreme in the Y direction
  3. **Influence point**: extreme in both directions

Butterfly did it?

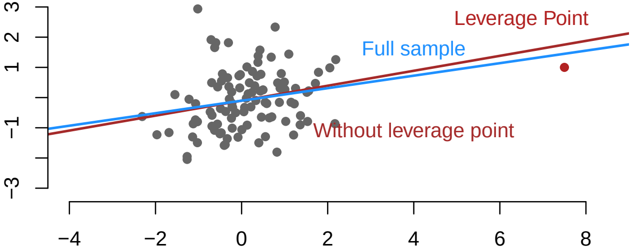
Wand et al, 2001 APSR



Butterfly did it?



Leverage point definition



- Values that are extreme in the X dimension
- That is, values far from the center of the covariate distribution

Leverage values

- Let h_{ij} be the (i, j) entry of $\mathbf{P}$. Then:

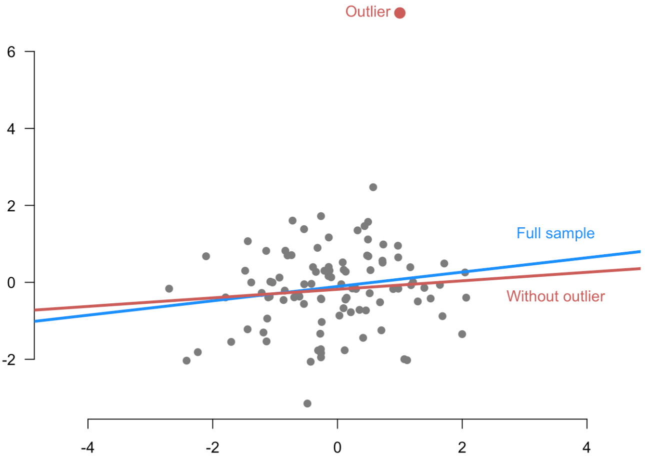
$$\hat{\mathbf{Y}} = \mathbf{PY} \quad \Longrightarrow \quad \hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  - $h_{ij} =$ importance of observation j is for the fitted value $\hat{Y}_i$
- **Leverage/hat values**: h_{ii} diagonal entries of the hat matrix
- With a simple linear regression, we have

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

  - $\rightsquigarrow$ how far i is from the center of the X distribution
- **Rule of thumb**: examine hat values greater than $2(k + 1)/n$

Outlier definition



- An **outlier** is far away from the center of the Y distribution.
- Intuitively: a point that would be poorly predicted by the regression.

Detecting outliers

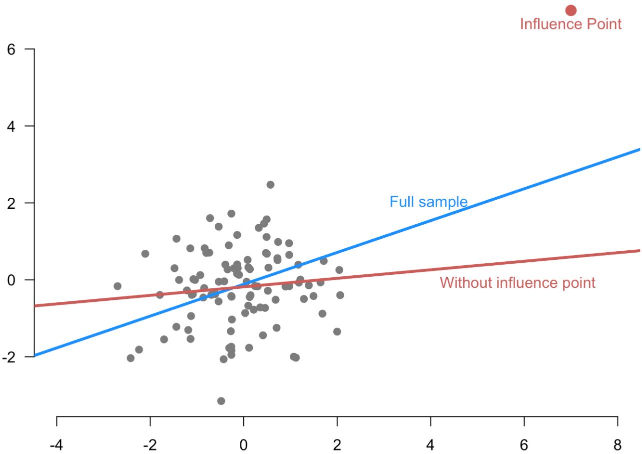
- Want values poorly predicted? Look for big residuals, right?
  - Problem: we use i to estimate $\hat{\beta}$ so $\hat{Y}$ aren't valid predictions.
  - unit might pull the regression line toward itself $\rightsquigarrow$ small residual
- Better: **leave-one-out prediction errors**,
  1. Regress $\mathbf{Y}_{(-i)}$ on $\mathbf{X}_{(-i)}$, where these omit unit i:

$$\hat{\beta}_{(-i)} = \left(\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}\right)^{-1}\mathbf{X}'_{(-i)}\mathbf{Y}_{(-i)}$$

  2. Calculate predicted value of Y_i using that regression: $\tilde{Y}_i = \mathbf{X}'_i\hat{\beta}_{(-i)}$
  3. Calculate prediction error: $\tilde{e}_i = Y_i - \tilde{Y}_i$
- Simple closed-form expressions:

$$\hat{\beta}_{(-i)} = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i\hat{e}_i, \qquad \tilde{e}_i = \frac{\hat{e}_i}{1 - h_{ii}}$$

Influence points



- An **influence point** is one that is both an outlier and a leverage point.
- Extreme in both the X and Y dimensions

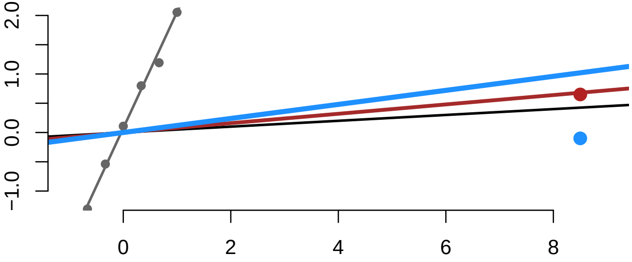
Overall measures of influence

- Influence of i can be measured by change in predictions:

$$\hat{Y}_i - \tilde{Y}_i = h_{ii}\tilde{e}_i$$

  - How much does excluding i from the regression change its predicted value?
  - Equal to "leverage $\times$ outlier-ness"
- Lots of diagnostics exist, but are mostly heuristic.
  - Does removing the point change a coefficient by a lot?

Limitations of the standard tools



- What happens when there are two influence points?
- Red line drops the red influence point
- Blue line drops the blue influence point

What to do about outliers and influential units?

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression, least absolute deviations)