Lecture 3: Agnostic Regression and Approximation Inference

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

January 31, 2022

Today:

- Continuing today with the theme of estimation and inference.
- Properties of regression estimator from an "agnostic" perspective.
- "Approximation inference."

Next time:

- Connecting back to causal inference.
- Regression as a tool for causal effect estimation.
- Bringing potential outcomes back in.

Overview: the "agnostic" mindset

- Experiment example from last class was "model free."
  - Binary treatment
  - Randomization
  - Easy to be agnostic
- Sometimes the inference needs more modeling:
  - Effects of continuous treatments, interactions with continuous moderators, etc.
  - Identification that requires a model (e.g., regression discontinuity)
- But models at best only approximate.
- How do we do "honest" inference when working with approximations?

"Dishonest": linear regression as a parametric model

- Suppose Y_i iid with $Y_i = X_i'\beta + \varepsilon_i$ (X_i length K, incl. constant).

"Dishonest": linear regression as a parametric model

- Suppose Y_i iid with $Y_i = X_i'\beta + \varepsilon_i$ (X_i length K, incl. constant).
- Further, let $\varepsilon_i \sim N(0, \sigma^2)$.
- Then $Y_i \sim N(X_i'\beta, \sigma^2)$.

"Dishonest": linear regression as a parametric model

- Suppose Y_i iid with $Y_i = X_i'\beta + \varepsilon_i$ (X_i length K, incl. constant).
- Further, let $\varepsilon_i \sim N(0, \sigma^2)$.
- Then $Y_i \sim N(X_i'\beta, \sigma^2)$.
- In this case, we can compute the probability of the data given different values of β and σ, and choose the β and σ that maximizes this probability. For normal data, this would imply,

$$\max_{\beta,\sigma} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right]$$

- Take the natural log and then solve. Solution is OLS:

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

- Inference proceeds from what the model entertains (iid, correct linear specification, homogenous effects, homoskedasticity).

Rather“dishonest”: linear regression as a semi-parametric model

- The model can be recast in “semi-parametric” terms.

Rather“dishonest”: linear regression as a semi-parametric model

- The model can be recast in “semi-parametric” terms.
- Condition on observed $\mathbf{X}$ and suppose $Y = \mathbf{X}\beta + \varepsilon$.
- Moment assumptions: $\mathrm{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$ and $\mathrm{Cov}[\varepsilon|\mathbf{X}] = \sigma^2 I_{n\times n}$.

Rather“dishonest”: linear regression as a semi-parametric model

- The model can be recast in “semi-parametric” terms.
- Condition on observed $\mathbf{X}$ and suppose $Y = \mathbf{X}\beta + \varepsilon$.
- Moment assumptions: $\mathrm{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$ and $\mathrm{Cov}[\varepsilon|\mathbf{X}] = \sigma^2 I_{n\times n}$.
- Target a β estimator (i) unbiased with (ii) minimum variance.

Rather“dishonest”: linear regression as a semi-parametric model

- The model can be recast in “semi-parametric” terms.
- Condition on observed $\mathbf{X}$ and suppose $Y = \mathbf{X}\beta + \varepsilon$.
- Moment assumptions: $\mathrm{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$ and $\mathrm{Cov}[\varepsilon|\mathbf{X}] = \sigma^2 I_{n\times n}$.
- Target a β estimator (i) unbiased with (ii) minimum variance.
- Gauss-Markov theorem: OLS is it (BLUE).
- Inference proceeds from these assumptions (iid, correct linear specification, homogenous effects, homoskedasticity).
- Can weaken the moment restrictions to have, $\mathrm{Cov}[\varepsilon|\mathbf{X}] = \Omega$, not necessarily homoskedastic nor diagonal.
- Then, the GLS estimator,

  $$\hat{\beta}_{GLS} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}Y,$$

  is BLUE.
- Still motivated by a homogenous effects assumption.

- In classical regression, inference (intervals, p-values, etc.) is not robust to departures from assumptions (correct specification of linear predictor and error and homogenous effects).
- But we *know* that these assumptions are only approximations.
- Motivates a desire to be honest about such approximation and have inferential approaches that account for them.

“Agnostic” regression and approximation inference

"Agnostic" regression and approximation inference

- Suppose (Y_i, X_i) from an arbitrary population member.

“Agnostic” regression and approximation inference

- Suppose (Y_i, X_i) from an arbitrary population member.
- By *definition* of the CEF (not an assumption),

$$Y_i = \underbrace{\mathrm{E}[Y_i|X_i]}_{\text{explained by } X_i} + \underbrace{\varepsilon_i}_{\text{unexplained and \textit{orthogonal} to } X_i}$$

with $\mathrm{E}[\varepsilon_i|X_i] = 0$ and $\mathrm{E}[h(X_i)\varepsilon_i] = 0$ for arbitrary $h(.)$.

"Agnostic" regression and approximation inference

- Suppose (Y_i, X_i) from an arbitrary population member.
- By *definition* of the CEF (not an assumption),

$$Y_i = \underbrace{\mathrm{E}[Y_i|X_i]}_{\text{explained by } X_i} + \underbrace{\varepsilon_i}_{\text{unexplained and } \textit{orthogonal} \text{ to } X_i}$$

  with $\mathrm{E}[\varepsilon_i|X_i] = 0$ and $\mathrm{E}[h(X_i)\varepsilon_i] = 0$ for arbitrary $h(.)$.
- *If* linearity holds, then $\mathrm{E}[Y_i|X_i] = X_i'\beta$ for some β, and,

$$\begin{aligned} \mathrm{E}[X_i\varepsilon_i] = \mathrm{E}[X_i(Y_i - X_i'\beta)] &= 0 \\ \mathrm{E}[X_iY_i] - \mathrm{E}[X_iX_i']\beta &= 0 \\ \beta &= \mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iY_i] \end{aligned}$$

- When X_i is all dummy variables, linearity holds by construction.
- When X_i includes continuous variables (perhaps transformed), then linearity is a substantive assumption about the CEF.

“Agnostic” regression: approximation inference

- Now suppose linearity *does not* necessarily hold.

“Agnostic” regression: approximation inference

- Now suppose linearity *does not* necessarily hold.
  - I.e., our specification is missing higher order terms, interactions, or discontinuities.

“Agnostic” regression: approximation inference

- Now suppose linearity *does not* necessarily hold.
  - I.e., our specification is missing higher order terms, interactions, or discontinuities.
- The linear specification may still serve a pragmatic function:
  - Lower-order, interpretable approximation.
  - But we want to be honest about this.

“Agnostic” regression: approximation inference

- Now suppose linearity *does not* necessarily hold.
  - I.e., our specification is missing higher order terms, interactions, or discontinuities.
- The linear specification may still serve a pragmatic function:
  - Lower-order, interpretable approximation.
  - But we want to be honest about this.
- If we use mean squared error as our target criterion for prediction accuracy, we have

$$\min_b \mathrm{E}\left[(Y_i - X_i' b)^2\right]$$

$$\text{FOC: } \mathrm{E}\left[X_i(Y_i - X_i' b^*)\right] = 0$$

$$b^* = \mathrm{E}\left[X_i X_i'\right]^{-1} \mathrm{E}\left[X_i Y_i\right].$$

- (Whether MSE is a good criterion depends on whether the conditional mean is appropriate for the problem at hand.)

"Agnostic" regression: approximation inference

- Now suppose linearity *does not* necessarily hold.
  - I.e., our specification is missing higher order terms, interactions, or discontinuities.
- The linear specification may still serve a pragmatic function:
  - Lower-order, interpretable approximation.
  - But we want to be honest about this.
- If we use mean squared error as our target criterion for prediction accuracy, we have

$$\min_b \mathrm{E}\left[(Y_i - X_i'b)^2\right]$$

$$\text{FOC: } \mathrm{E}\left[X_i(Y_i - X_i'b^*)\right] = 0$$

$$b^* = \mathrm{E}\left[X_i X_i'\right]^{-1}\mathrm{E}\left[X_i Y_i\right].$$

- (Whether MSE is a good criterion depends on whether the conditional mean is appropriate for the problem at hand.)
- Goal is to translate this into an estimator and then do honest inference.

“Agnostic” regression: approximation inference

- Define the *population* regression coefficient for our linear approximation as,

$$\beta \equiv \mathrm{E}\left[X_i X_i'\right]^{-1} \mathrm{E}\left[X_i Y_i\right]$$

- (This is well defined whether or not $\mathrm{E}\left[Y_i | X_i\right]$ is linear.)

"Agnostic" regression: approximation inference

- Define the *population* regression coefficient for our linear approximation as,

$$\beta \equiv \mathrm{E}[X_i X_i']^{-1}\mathrm{E}[X_i Y_i]$$

- (This is well defined whether or not $\mathrm{E}[Y_i|X_i]$ is linear.)
- Suppose a random sample, S, of size N from a large population.
- Then, for arbitrary $i \in S$, $W_i = \begin{pmatrix} Y_i & X_i' \end{pmatrix}'$ is an iid vector.
- The sample mean $\frac{1}{N}\sum_{i=1}^{N} W_i \xrightarrow{p} \mathrm{E}[W_i]$ by the WLLN.

“Agnostic” regression: approximation inference

- Define the *population* regression coefficient for our linear approximation as,

$$\beta \equiv \mathrm{E}[X_i X_i']^{-1} \mathrm{E}[X_i Y_i]$$

- (This is well defined whether or not $\mathrm{E}[Y_i|X_i]$ is linear.)
- Suppose a random sample, S, of size N from a large population.
- Then, for arbitrary $i \in S$, $W_i = \begin{pmatrix} Y_i & X_i' \end{pmatrix}'$ is an iid vector.
- The sample mean $\frac{1}{N}\sum_{i=1}^{N} W_i \xrightarrow{p} \mathrm{E}[W_i]$ by the WLLN.
- Holds for higher moments, e.g., $\frac{1}{N}\sum_{i=1}^{N} W_i W_i' \xrightarrow{p} \mathrm{E}[W_i W_i']$.
- As such,

$$\begin{aligned}\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y &= \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i Y_i \\ &= \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right) \xrightarrow{p} \beta\end{aligned}$$

"Agnostic" regression: approximation inference

- Define the population residual,

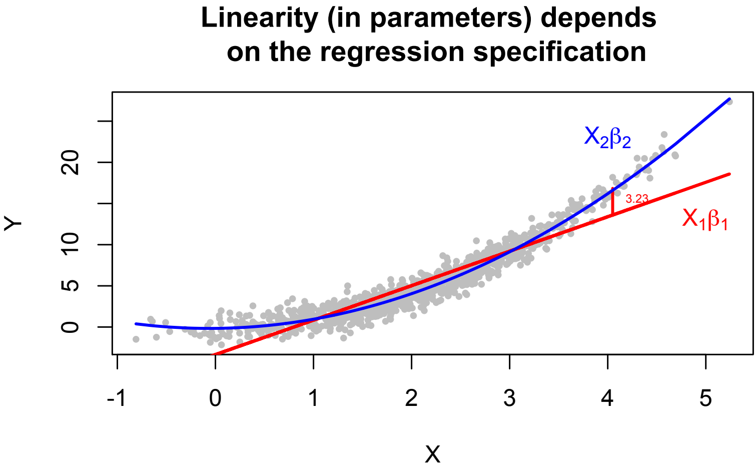
$$Y_i = X_i'\beta + (Y_i - X_i'\beta) \equiv X_i'\beta + e_i.$$

- Then, we have orthogonality,

$$\mathrm{E}[X_i e_i] = \mathrm{E}[X_i(Y_i - X_i'\beta)] = \mathrm{E}[X_i(Y_i - X_i'(X_i X_i')^{-1}(X_i Y_i)] = 0,$$

  by construction of the regression coefficient.
- However, $\mathrm{E}[e_i|X_i] = \mathrm{E}[Y_i|X_i] - X_i'\beta$ is zero *only if* linearity holds, but we have not asserted this.
- (Example 1...)

Example 1



Linearity (in parameters) depends on the regression specification
$X_2\beta_2$
3.23
$X_1\beta_1$
Y
0
5
10
20
-1
0
1
2
3
4
5
X

"Agnostic" regression: approximation inference

Whether or not the CEF is linear, we can get a *consistent* estimate of the *linear approximation* with OLS:

“Agnostic” regression: approximation inference

Whether or not the CEF is linear, we can get a *consistent* estimate of the *linear approximation* with OLS:

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i Y_i = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i (X_i'\beta + e_i)$$

$$= \beta + \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i e_i$$

$$\Rightarrow \sqrt{N}(\hat{\beta} - \beta) = \frac{N}{\sqrt{N}} \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i e_i = \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i'\right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i e_i,$$

which, by Slutsky, has the same asymptotic distribution as

$$\mathrm{E}\left[X_i X_i'\right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i e_i.$$

"Agnostic" regression: approximation inference

Whether or not the CEF is linear, we can get a *consistent* estimate of the *linear approximation* with OLS:

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i Y_i = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i (X_i'\beta + e_i)$$

$$= \beta + \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i e_i$$

$$\Rightarrow \sqrt{N}(\hat{\beta} - \beta) = \frac{N}{\sqrt{N}} \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i e_i = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i e_i,$$

which, by Slutsky, has the same asymptotic distribution as

$$\mathrm{E}\left[X_i X_i'\right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i e_i.$$

By CLT, $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i e_i$ is distributed normal with mean $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathrm{E}\left[X_i e_i\right] = 0$ and variance,

$$\mathrm{Var}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i e_i\right] = \frac{1}{N} N \mathrm{Var}\left[X_i e_i\right] = \mathrm{E}\left[(X_i e_i - \mathrm{E}\left[X_i e_i\right])(X_i e_i - \mathrm{E}\left[X_i e_i\right])'\right] = \mathrm{E}\left[X_i X_i' e_i^2\right].$$

"Agnostic" regression: approximation inference

- Putting it all together, we have,

  $$\sqrt{N}(\hat{\beta}-\beta) \overset{a}{\sim} \mathrm{N}(0,\Omega), \text{ with } \Omega = \mathrm{E}[X_i X_i']^{-1}\mathrm{E}[X_i X_i' e_i^2]\mathrm{E}[X_i X_i']^{-1},$$

  or, letting $\mathbf{V} = \Omega/N$,

  $$\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta, \mathbf{V})$$

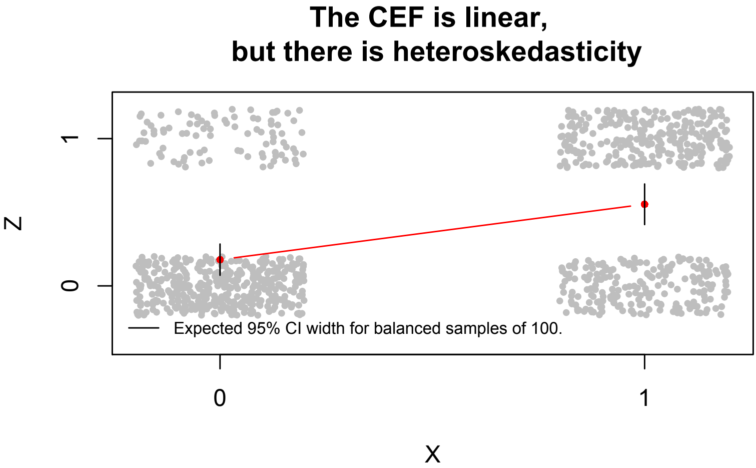
- A consistent estimator for $\mathbf{V}$ is given by,

  $$\begin{aligned}\hat{\mathbf{V}} &\equiv \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \hat{e}_i^2\right)\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N} X_i X_i' \hat{e}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1},\end{aligned}$$

- **V** is the "Huber-Eicker-White" covariance estimator.
- Square roots of diagonals are "het. robust" s.e.'s, used for C.I.'s and tests based on normal (or t_{N-K}) approximation.
- Refinements for finite samples have been proposed (cf. Imbens & Kolesar, 2016).

- $\mathbf{V}$ is the "Huber-Eicker-White" covariance estimator.
- Square roots of diagonals are "het. robust" s.e.'s, used for C.I.'s and tests based on normal (or t_{N-K}) approximation.
- Refinements for finite samples have been proposed (cf. Imbens & Kolesar, 2016).
- Random sampling from a large population implies the vectors, (Y_i, X_i), are iid. *This is not the same thing* as assuming homoskedasticity, which is an assumption on *the population residual*, e_i, which itself is a function of (i) the approximation that we use for $\mathrm{E}[Y_i|X_i]$ and (ii) the scale of Y_i.
- Heteroskedasticity of the population residual arises naturally when we only approximate $\mathrm{E}[Y_i|X_i]$.
- Heteroskedasticity of the population residual also arises when $\mathrm{E}[Y_i|X_i]$ is linear, but the conditional variance is not constant (example 2...).

Example 2



The CEF is linear,
but there is heteroskedasticity
Expected 95% CI width for balanced samples of 100.
Z
X
0
1
0
1

Take a moment and think:

What have we **assumed**
to obtain these results on
consistency
and the large sample **distribution**?

Properties of the OLS solution

Partial regression, or "Frisch-Waugh-Lovell"

Properties of the OLS solution

- Suppose a sample regression solution given by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$.
- Let $\mathbf{X}_1$ refers to the first $K-1$ columns, while X_K is the Kth.

Properties of the OLS solution

- Suppose a sample regression solution given by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$.
- Let $\mathbf{X}_1$ refers to the first $K-1$ columns, while X_K is the Kth.
- Two partial regressions and associated residuals:
  - Let $\hat{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'Y$ and $f = Y - \mathbf{X}_1\hat{\gamma}_1$.
  - Let $\hat{\gamma}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'X_K$ and $g = X_K - \mathbf{X}_1\hat{\gamma}_2$.

Properties of the OLS solution

- Suppose a sample regression solution given by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$.
- Let $\mathbf{X}_1$ refers to the first $K-1$ columns, while X_K is the Kth.
- Two partial regressions and associated residuals:
  - Let $\hat{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'Y$ and $f = Y - \mathbf{X}_1\hat{\gamma}_1$.
  - Let $\hat{\gamma}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'X_K$ and $g = X_K - \mathbf{X}_1\hat{\gamma}_2$.
- Residual-residual regression: let $\hat{\gamma}_3 = g'f/g'g$ and $e = f - g\hat{\gamma}_3$.

Properties of the OLS solution

- Suppose a sample regression solution given by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$.
- Let $\mathbf{X}_1$ refers to the first $K-1$ columns, while X_K is the Kth.
- Two partial regressions and associated residuals:
  - Let $\hat{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'Y$ and $f = Y - \mathbf{X}_1\hat{\gamma}_1$.
  - Let $\hat{\gamma}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'X_K$ and $g = X_K - \mathbf{X}_1\hat{\gamma}_2$.
- Residual-residual regression: let $\hat{\gamma}_3 = g'f/g'g$ and $e = f - g\hat{\gamma}_3$.
- By the construction of the residual, $\mathrm{E}[X_{1i}f_i] = 0$, $\mathrm{E}[X_{1i}g_i] = 0$, and $\mathrm{E}[g_ie_i] = 0$.
- In that case, $\mathrm{E}[X_{1i}e_i] = \mathrm{E}[X_{1i}f_i] - \mathrm{E}[X_{1i}g_i]\hat{\gamma}_3 = 0$.
- Also $\mathrm{E}[(g_i + X_{1i}\hat{\gamma}_2)e_i] = \mathrm{E}[X_{Ki}e_i] = 0$, so $\mathrm{E}[X_ie_i] = 0$.

Properties of the OLS solution

- Suppose a sample regression solution given by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$.
- Let $\mathbf{X}_1$ refers to the first $K-1$ columns, while X_K is the Kth.
- Two partial regressions and associated residuals:
  - Let $\hat{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'Y$ and $f = Y - \mathbf{X}_1\hat{\gamma}_1$.
  - Let $\hat{\gamma}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'X_K$ and $g = X_K - \mathbf{X}_1\hat{\gamma}_2$.
- Residual-residual regression: let $\hat{\gamma}_3 = g'f/g'g$ and $e = f - g\hat{\gamma}_3$.
- By the construction of the residual, $\mathrm{E}[X_{1i}f_i] = 0$, $\mathrm{E}[X_{1i}g_i] = 0$, and $\mathrm{E}[g_ie_i] = 0$.
- In that case, $\mathrm{E}[X_{1i}e_i] = \mathrm{E}[X_{1i}f_i] - \mathrm{E}[X_{1i}g_i]\hat{\gamma}_3 = 0$.
- Also $\mathrm{E}[(g_i + X_{1i}\hat{\gamma}_2)e_i] = \mathrm{E}[X_{Ki}e_i] = 0$, so $\mathrm{E}[X_ie_i] = 0$.
- Then,

$$
\begin{aligned}
Y &= \mathbf{X}_1\hat{\gamma}_1 + f = \mathbf{X}_1\hat{\gamma}_1 + g\hat{\gamma}_3 + e = \mathbf{X}_1\hat{\gamma}_1 + (X_K - \mathbf{X}_1\hat{\gamma}_2)\hat{\gamma}_3 + e \\
&= \mathbf{X}_1(\hat{\gamma}_1 - \hat{\gamma}_2\hat{\gamma}_3) + \mathbf{X}_K\hat{\gamma}_3 + e \text{ with } \mathrm{E}[\mathbf{X}e] = \mathbf{0} \\
&\Rightarrow \hat{\beta} = \begin{pmatrix} \hat{\gamma}_1 - \hat{\gamma}_2\hat{\gamma}_3 \\ \hat{\gamma}_3 \end{pmatrix}
\end{aligned}
$$

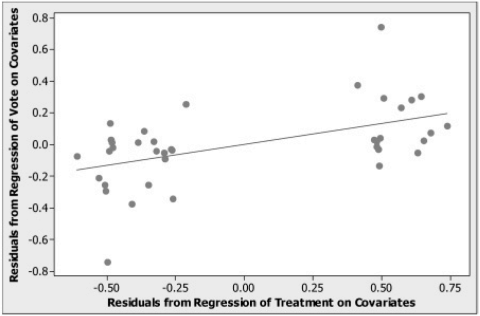
Properties of the OLS solution

- Frisch-Waugh-Lovell theorem, also known as "partial regression."
- Allows us to express any multiple regression solution in terms of a bivariate regression solutions—e.g.,

$$\hat{\beta}_k = \frac{\mathrm{Cov}\,(\tilde{Y}_i, \tilde{X}_{ki})}{\mathrm{Var}\,(\tilde{X}_{ki})},$$

  where $\tilde{Y}_i$ and $\tilde{X}_{ik}$ are Y_i and X_{ik} "residualized" on other regressors.

$$
\begin{aligned}
LN\frac{Turnout_{t,i}}{100 - Turnout_{t,i}} &= \beta_0 + \beta_1 Treatment_i \\
&\quad + \beta_2 San\ Francisco_i \\
&\quad + \beta_3 Portland_i + \beta_4 Lewiston_i \\
&\quad + \beta_5 Austin_i + \beta_6 Pittsburgh_i \\
&\quad + \beta_7 Hartford_i + \beta_8 Stockton_i \\
&\quad + \beta_9 Green\ Bay_i + \beta_{10} St.\ Paul_i \\
&\quad + \beta_{11} Oakland_i + \beta_{12} Tallahassee_i \\
&\quad + \beta_{13} New\ Haven\ Municipal_i \\
&\quad + \beta_{14} New\ Hampshire_i \\
&\quad + \beta_{15}\left(LN\frac{Turnout_{t-1,i}}{100 - Turnout_{t-1,i}}\right) \\
&\quad + \varepsilon_{t,i}.
\end{aligned}
$$



(Addonizio et al., 2007)

Properties of the OLS solution

Implications of sampling distribution for testing

Properties of the OLS solution

- Recall under random sampling of individual units, $\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta, \mathbf{V})$.
- Hypothesis testing for β applies this result.

Properties of the OLS solution

- Recall under random sampling of individual units, $\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta, \mathbf{V})$.
- Hypothesis testing for β applies this result.
- Construct linear restrictions matrix, $R_{(q\times k)}$, to test $H_0 : R\beta = r_{(q\times 1)}$, that is, q linear restrictions.
- E.g., suppose an intercept and 3 coefficients, and we want to test that $\beta_2 = \beta_3 = 0$. Then we can write,

$$R\beta = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = r$$

Properties of the OLS solution

- Recall under random sampling of individual units, $\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta, \mathbf{V})$.
- Hypothesis testing for β applies this result.
- Construct linear restrictions matrix, $R_{(q\times k)}$, to test $H_0 : R\beta = r_{(q\times 1)}$, that is, q linear restrictions.
- E.g., suppose an intercept and 3 coefficients, and we want to test that $\beta_2 = \beta_3 = 0$. Then we can write,

$$R\beta = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = r$$

- Because $R\hat{\beta}$ is a sum of $\overset{a}{MVN}$ variables, $R\hat{\beta}$ is $\overset{a}{MVN}$, and under H_0, $R\hat{\beta} - r$ is mean zero $\overset{a}{MVN}$.
- Therefore under H_0,

$$W \equiv (R\hat{\beta} - r)'(R\hat{\mathbf{V}}R')^{-1}(R\hat{\beta} - r) \overset{a}{\sim} \chi^2_q \text{ (Wald test statistic)}$$

- Finite sample refinement for normal errors tests W/q on $F_{q,N-K}$.

Properties of the OLS solution

Measuring the influence of observations via the leverage

Properties of the OLS solution

- Start again with $Y_i = X_i\beta + e_i$.

Properties of the OLS solution

- Start again with $Y_i = X_i\beta + e_i$.
- $\hat{\beta}_k$ equals kth row of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ times Y: "weighted avg" over Y.

Properties of the OLS solution

- Start again with $Y_i = X_i\beta + e_i$.
- $\hat{\beta}_k$ equals kth row of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ times Y: "weighted avg" over Y.
- Define ι^j, a dummy variable equal to 1 for j but 0 for others, and consider the regression, $Y_i = X_i\beta^j + \alpha\iota_i^j + u_i$.
- By FWL, estimate β^j with $\tilde{Y}$ and $\tilde{\mathbf{X}}$, residualized on ι^j. $\tilde{Y}$ and $\tilde{\mathbf{X}}$ equal Y and $\mathbf{X}$ respectively but with the jth position zeroed out. So $\hat{\beta}^j$ is $\hat{\beta}$ if j were simply omitted.
- By FWL, we can estimate α with $\tilde{Y}$ and $\tilde{\iota}^j$, residualized on X.
- Now, $\tilde{Y} = Y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')Y = \mathbf{M}_X Y = \hat{e}$, with $\mathbf{M}_X$ idempotent, while $\tilde{\iota}^j = \mathbf{M}_X\iota^j$.
- Then, $\hat{\alpha} = \frac{\iota^{j\prime}\mathbf{M}_X Y}{\iota^{j\prime}\mathbf{M}_X\iota^j} = \frac{\hat{e}_j}{1-h_j}$, where h_j is jth diag. of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Properties of the OLS solution

- Start again with $Y_i = X_i\beta + e_i$.
- $\hat{\beta}_k$ equals kth row of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ times Y: "weighted avg" over Y.
- Define ι^j, a dummy variable equal to 1 for j but 0 for others, and consider the regression, $Y_i = X_i\beta^j + \alpha\iota_i^j + u_i$.
- By FWL, estimate β^j with $\tilde{Y}$ and $\tilde{\mathbf{X}}$, residualized on ι^j. $\tilde{Y}$ and $\tilde{\mathbf{X}}$ equal Y and $\mathbf{X}$ respectively but with the jth position zeroed out. So $\hat{\beta}^j$ is $\hat{\beta}$ if j were simply omitted.
- By FWL, we can estimate α with $\tilde{Y}$ and $\tilde{\iota}^j$, residualized on X.
- Now, $\tilde{Y} = Y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')Y = \mathbf{M}_X Y = \hat{e}$, with $\mathbf{M}_X$ idempotent, while $\tilde{\iota}^j = \mathbf{M}_X\iota^j$.
- Then, $\hat{\alpha} = \frac{\iota^{j\prime}\mathbf{M}_X Y}{\iota^{j\prime}\mathbf{M}_X\iota^j} = \frac{\hat{e}_j}{1-h_j}$, where h_j is jth diag. of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Putting it all together yields, $\hat{\beta}^j - \hat{\beta} = -\frac{\hat{e}_j}{1-h_j}(\mathbf{X}'\mathbf{X})^{-1}X_j$
- h_j is leverage. As it gets bigger, so does j's potential influence.
- Influence also depends on $\hat{e}_j$.

Properties of the OLS solution

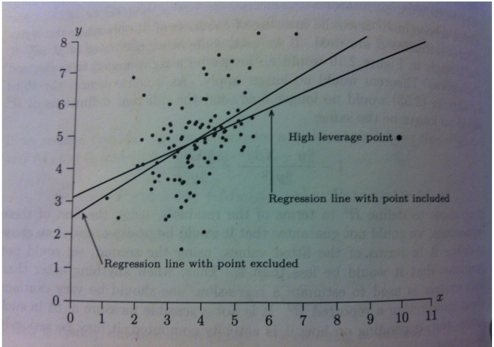


Figure 2.14 An influential observation

- In bivariate regression, leverage measures deviation from $\bar{X}$.

$$h_j = \frac{1}{n} + \frac{(X_j - \bar{X})^2}{\sum_{i=1}^{N}(X_j - \bar{X})^2},$$

- In multiple regression, h_j measures distance from $(\bar{X}_1, \ldots, \bar{X}_K)$.

The “agnostic” mindset

- Recall our goal: machinery for modeling that is honest about approximation.
- New mindset: “honest” approximation inference.

The “agnostic” mindset

- Recall our goal: machinery for modeling that is honest about approximation.
- New mindset: “honest” approximation inference.
- Even if we don’t get the CEF *exactly correct*, we can still do robust inference on the approximation. Classical regression inference was dishonest because of specification assumptions.

The "agnostic" mindset

- Recall our goal: machinery for modeling that is honest about approximation.
- New mindset: "honest" approximation inference.
- Even if we don't get the CEF *exactly correct*, we can still do robust inference on the approximation. Classical regression inference was dishonest because of specification assumptions.
- Assumptions here are those that permit the WLLN and CLT: sample (Y_i, X_i) values at random from a population with finite or well behaved higher order moments, in which case (Y_i, X_i) can be treated as iid and well behaved.

The "agnostic" mindset

- Recall our goal: machinery for modeling that is honest about approximation.
- New mindset: "honest" approximation inference.
- Even if we don't get the CEF *exactly correct*, we can still do robust inference on the approximation. Classical regression inference was dishonest because of specification assumptions.
- Assumptions here are those that permit the WLLN and CLT: sample (Y_i, X_i) values at random from a population with finite or well behaved higher order moments, in which case (Y_i, X_i) can be treated as iid and well behaved.
- Next class we will bring this back to causal inference.