Two Regression Theorems and Applications

Joe C

Wealthfront

2026

Roadmap

Roadmap

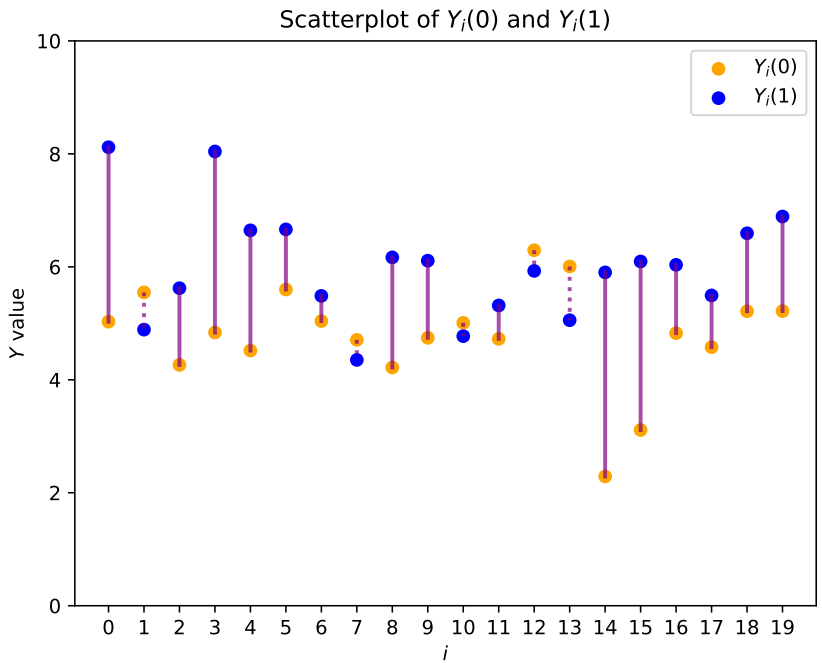
Defining our target parameter

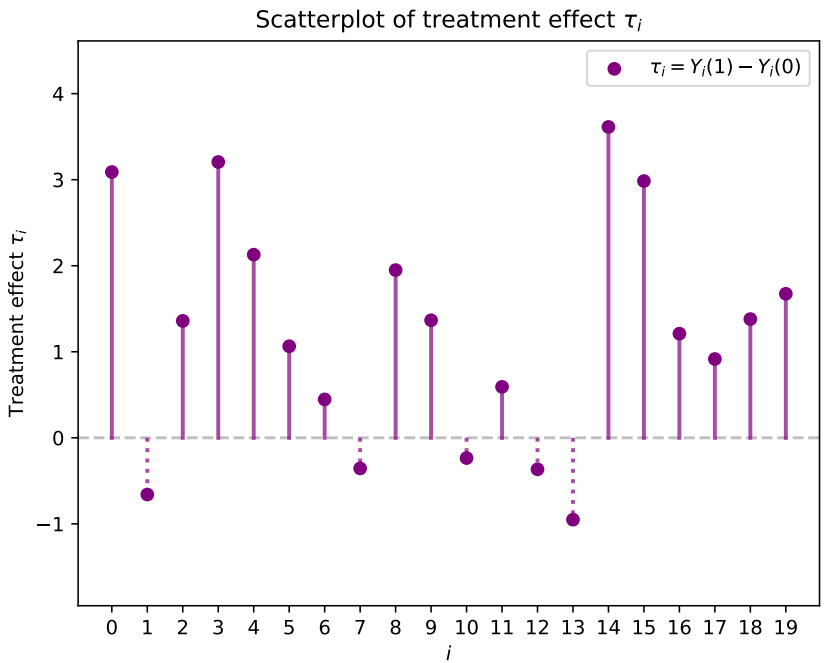
- Before we use an estimator we need to define our target parameter
- The target parameter throughout the rest of the presentation is the average treatment effect (ATE)
- To define the ATE I'll need to introduce the potential outcomes framework

Introducing the notation

- Let i index the units of interest
  - For example, clients
- Consider a binary treatment variable $D_i \in \{0, 1\}$
  - For example, two variants of an email
- Each unit i has two potential outcomes $Y_i(0)$ and $Y_i(1)$ corresponding to treatments $D_i = 0$ and $D_i = 1$ respectively
  - When $D_i = 0$ we say the unit is untreated, and when $D_i = 1$ we say the unit is treated
- For each unit i we observe $Y_i = (1 - D_i)\,Y_i(0) + D_i\,Y_i(1)$
  - This is the fundamental problem of causal inference: we observe $Y_i(0)$ or $Y_i(1)$ but never both

Potential Outcomes Visualization





Definition of the *ATE*

- Let $ATE = \mathbb{E}[Y_i(1) - Y_i(0)]$
  - This is the average treatment effect in the population
- In general, identifying the *ATE* is difficult because of selection bias
- However, identifying the *ATE* is easy when D_i is randomly assigned

Identification under random assignment.

Since $D_i \perp \{Y_i(0), Y_i(1)\}$ (independence)

$$\begin{aligned} ATE &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]}_{\text{Difference in Means}} \end{aligned}$$

Roadmap

ATE via Ordinary Least Squares

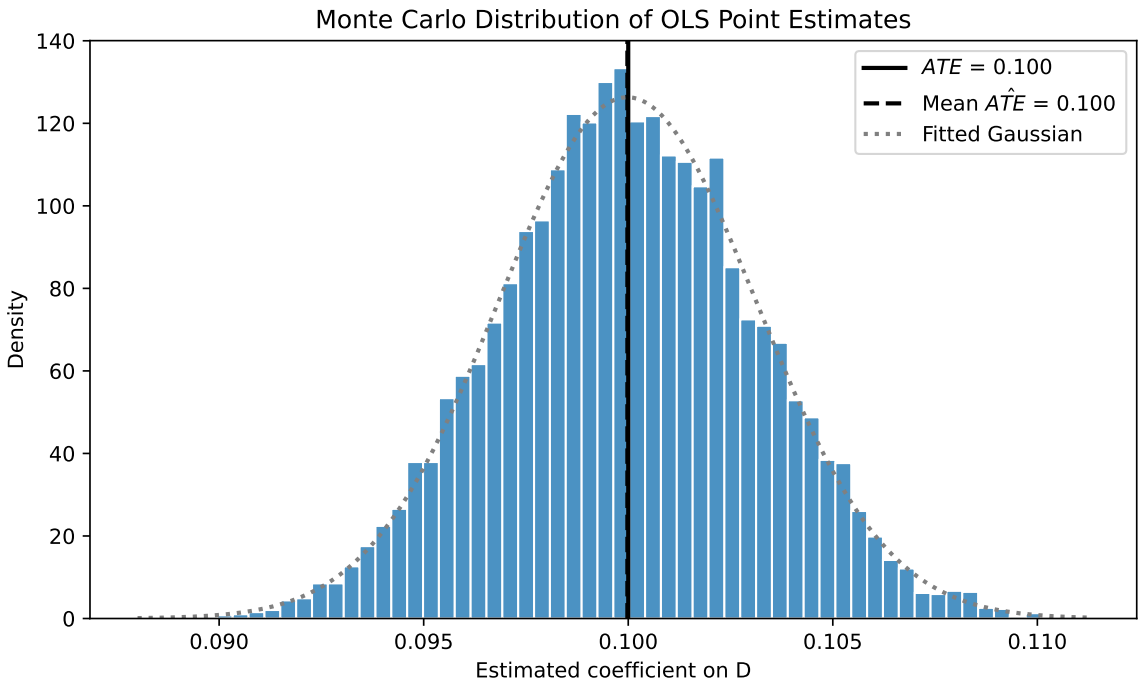
- Let $Y_i = \alpha + \beta D_i + u_i$ with D_i randomly assigned
  - Formally this means α, β solve $\min_{\alpha', \beta'} \mathbb{E}\left[(Y_i - \alpha' - \beta' D_i)^2\right]$
  - FOC of this objective function is that the residual is uncorrelated with each covariate
- $\alpha = \mathbb{E}[Y_i(0)]$ and $\beta = \mathbb{E}[Y_i(1) - Y_i(0)]$
  - Proof uses law of iterated expectation to split the objective function
- This means we can use a simple linear regression to identify the average treatment effect

Asymptotic Variance of Ordinary Least Squares Estimator

- In practice we observe $\hat{\beta}$ rather than β since we don't have infinite data
- As our sample size n approaches infinity we have $\sqrt{n}(\hat{\beta} - \beta) \rightarrow \mathcal{N}\left(0, \frac{\mathbb{V}[u_i]}{\mathbb{V}[D_i]}\right)$
  - I've assumed homoskedasticity here for simplicity
  - Motivates equal treatment split under random assignment since this maximizes $\mathbb{V}[D_i]$
- We want to minimize this asymptotic variance
  - It would be great if $\mathbb{V}[u_i]$ was smaller

Normality Doesn't Matter in Large Samples

```
df = pd.DataFrame({
    "i": [str(i) for i in range(N_OBS)],
    "D": rng.integers(low=0, high=2, size=N_OBS),
})
df["Y"] = rng.binomial(n=1, p=np.where(df["D"] == 1, .6, .5))
```



Roadmap

Omitted Variable Bias Formula

- Consider two regressions:
  - Short: $Y_i = \alpha^s + \beta^s D_i + u_i^s$
  - Long: $Y_i = \alpha^l + \beta^l D_i + \gamma W_i + u_i^l$

Theorem (Omitted Variable Bias Formula)

$\beta^s = \beta^l + \delta\gamma$ *where* $W_i = \eta + \delta D_i + v_i$

- Intuition: β^s captures the effect of D_i (β^l) as well as the effect of other variables that comove with it ($\delta\gamma$)

The Yule-Frisch-Waugh-Lovell Theorem

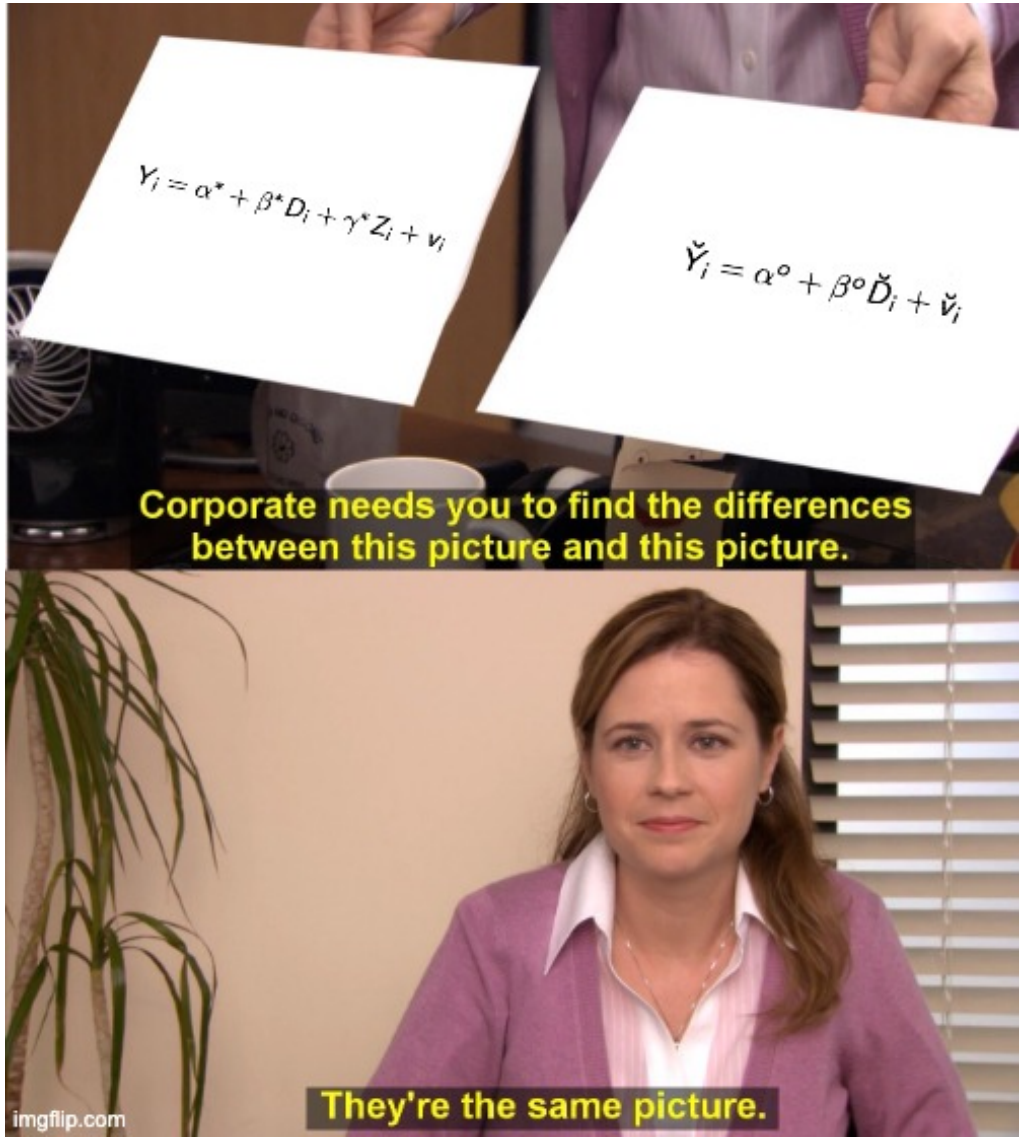
- Let $Y_i = \alpha^* + \beta^* D_i + \gamma^* Z_i + v_i$
- Let $\breve{Y}_i = \alpha^o + \beta^o \breve{D}_i + \breve{v}_i$ where
  - $Y_i = \delta Z_i + \breve{Y}_i$
  - $D_i = \eta Z_i + \breve{D}_i$
- Here $\breve{Y}_i$ and $\breve{D}_i$ contain only the variation that is orthogonal to variation in Z_i
  - They are the residuals that remain after regressing Z_i

Theorem (Yule-Frisch-Waugh-Lovell)

$\beta^* = \beta^o$ *and* $v_i = \breve{v}_i$

- Intuition: β^* is estimated using only the *independent* variation in D_i

Water break


$Y_i = \alpha^* + \beta^* D_i + \gamma^* Z_i + \nu_i$
$\check{Y}_i = \alpha^\circ + \beta^\circ \check{D}_i + \check{\nu}_i$
Corporate needs you to find the differences between this picture and this picture.
They're the same picture.
imgflip.com

Roadmap

Randomization eliminates OVB

- Consider two regressions:
  - $Y_i = \alpha^s + \beta^s D_i + u_i$
  - $Y_i = \alpha^l + \beta^l D_i + \gamma W_i + v_i$
- Assume, as before, that D_i is randomly assigned
  - This implies that $W_i = \eta + \delta D_i + \varepsilon_i$ with $\delta = 0$
- Then from the OVB formula we have $\beta^s = \beta^l + 0 \cdot \gamma = \beta^l$
  - Both regressions identify the same parameter!
- So why might we prefer one over the other?
  - Because we don't have infinite data

YFWL for asymptotic variance

- Consider two regressions:
  - $Y_i = \alpha^s + \beta^s D_i + u_i$
  - $Y_i = \alpha^l + \beta^l D_i + \gamma W_i + v_i$
- Focus on the long regression: $Y_i = \alpha^l + \beta^l D_i + \gamma W_i + v_i$
- From YFWL we know that the long regression is equivalent to $\check{Y}_i = \alpha^o + \beta^l \check{D}_i + v_i$
- Because D_i is randomly assigned we have $\check{D}_i = D_i - \mathbb{E}[D_i]$
- This is a simple linear regression so we can use the same variance expression from before, with v_i in place of u_i
  - $\sqrt{n}(\hat{\beta}^l - \beta^l) \to \mathcal{N}\left(0, \frac{\mathbb{V}[v_i]}{\mathbb{V}[D_i]}\right)$

Algebra

- Consider two regressions:
  - $Y_i = \alpha^s + \beta^s D_i + u_i$
  - $Y_i = \alpha^l + \beta^l D_i + \gamma W_i + v_i$
- Last step is to show that $\mathbb{V}[v_i] \leq \mathbb{V}[u_i]$.

$$
\begin{aligned}
u_i &= Y_i - \alpha^s - \beta^s D_i \\
&= Y_i - \alpha^s - \beta^l D_i \\
&= Y_i - \alpha^s - \beta^l D_i - \gamma W_i + \gamma W_i - \alpha^l + \alpha^l \\
&= (Y_i - \alpha^l - \beta^l D_i - \gamma W_i) + \gamma W_i + (\alpha^l - \alpha^s) \\
&= v_i + \gamma W_i + (\alpha^l - \alpha^s)
\end{aligned}
$$

- Last term is a constant and first two terms are uncorrelated by construction $\implies \mathbb{V}[v_i] = \mathbb{V}[u_i] - \mathbb{V}[\gamma W_i]$

Summary

- We can run the long and short regressions
  - $Y_i = \alpha^s + \beta^s D_i + u_i$
  - $Y_i = \alpha^l + \beta^l D_i + \gamma W_i + v_i$
- With infinite data, both regressions are equivalent since $\beta^s = \beta^l$
- With finite data we achieve increased precision with the long regression since $\mathbb{V}[v_i] \leq \mathbb{V}[u_i]$
- Intuition: Additional control variables W_i reduce the residual variance and thus tighten standard errors
  - The best control variables are those that are highly predictive of the outcome variable

Wealthfront Example

- C2I Boost Incentive aimed to increase C2I with a 0.5% APY boost for cash users who created an investing account
- Analysis compares the performance of two email variants for users with recent log ins
- Outcome variable is open rate and I consider two regression specifications
  - Short: constant, treatment indicator
  - Long: constant, treatment indicator, historical click rate, historical open rate
- Standard error of the treatment effect is 34% smaller with the long specification
  - Equivalent to an increase in sample size of 78%

Roadmap

Extensions

- With imbalanced assignment (i.e. not a 50-50 split) you should technically interact covariates with the treatment indicator
  - Winston Lin Paper
  - Winston Lin Blog Post 1
  - Winston Lin Blog Post 2
- This logic actually extends more broadly to nonlinear regression adjustment
  - Extension to ML methods