Chapter 15: Outcome Regression and Propensity Scores

Contents

This chapter explores **outcome regression** and **propensity scores** in greater depth, clarifying their roles in causal inference. We examine when simple regression adjustment is sufficient, when it fails, and how propensity scores can be used for confounding adjustment through matching, stratification, or weighting.

This chapter is based on Hernán and Robins (2020, chap. 15, pp. 207-226).

Key theme: Outcome regression and propensity scores are tools, not magical solutions. Their success depends on correctly specifying models and satisfying causal assumptions (especially conditional exchangeability and positivity).

1 15.1 Outcome Regression (pp. 207-210)

Outcome regression estimates causal effects by modeling the outcome as a function of treatment and confounders.

1.1 The Outcome Regression Approach

Definition 1.1 (Outcome Regression)**. Outcome regression** for causal inference:

1. Fit a model for $E[Y \mid A, L]$
2. Use the model to compute standardized means (g-formula)
3. Estimate causal effects as contrasts of standardized means

For simple cases, the treatment coefficient may approximate the causal effect, but this requires strong assumptions.

1.2 When Does the Treatment Coefficient Equal the Causal Effect?

Model: $E[Y \mid A, L] = \beta_0 + \beta_1 A + \beta_2^\top L$

Question: When does $\beta_1 = E[Y^{a=1}] - E[Y^{a=0}]$?

Answer: Only under restrictive conditions:

1. **No confounding**: $Y^a \perp\!\!\!\perp A$ (conditional exchangeability not needed)
2. **No effect modification**: The causal effect doesn't vary with L
3. **Correct model specification**: Linear model is correct

If effect modification exists, β_1 is a weighted average of conditional effects, not generally equal to the marginal causal effect.

Why this matters:

Many researchers fit $Y = \beta_0 + \beta_1 A + \beta_2^\top L$ and interpret β_1 as the causal effect. This is valid ONLY if:

- No unmeasured confounding (always required)
- No effect modification by measured confounders
- Correct linear model

When effect modification exists, use the g-formula to properly standardize, or include $A \times L$ interactions in the model.

1.3 Example: NHEFS Study

Simple model:

$$E[\text{Weight Change} \mid A, L] = \beta_0 + \beta_1 \text{Quit} + \beta_2 \text{Age} + \beta_3 \text{Sex} + ...$$

Issues:

- Assumes effect of quitting is the same for all individuals
- If the effect varies by age, sex, or other factors, β_1 doesn't equal the marginal causal effect
- Need to add interactions or use g-formula

Better approach:

$$E[Y \mid A, L] = \beta_0 + \beta_1 A + \beta_2^\top L + \beta_3^\top (A \times L)$$

Then use g-formula to compute marginal effect.

2 15.2 Propensity Scores (pp. 210-213)

The **propensity score** is the probability of receiving treatment given confounders. It plays a central role in observational studies.

Definition 2.1 (Propensity Score)**.** The **propensity score** is:

$$e(L) = \Pr[A = 1 \mid L]$$

For individual i with covariates L_i, the propensity score is $e(L_i) = \Pr[A = 1 \mid L = L_i]$.

2.1 Balancing Property

Key theorem: If $Y^a \perp\!\!\!\perp A \mid L$, then:

$$Y^a \perp\!\!\!\perp A \mid e(L)$$

Interpretation: Conditional on the propensity score, treatment assignment is independent of potential outcomes.

Implication: We can adjust for confounding by adjusting for the propensity score alone, rather than all components of L.

Why this is useful:

- With many confounders, stratification on L is difficult (curse of dimensionality)
- The propensity score reduces multidimensional L to a single dimension
- We can match, stratify, or weight on $e(L)$ instead of on the full vector L

Important caveat: This works only if:

1. The propensity score model is correctly specified
2. Positivity holds: $0 < e(L) < 1$ for all L with $\Pr[L] > 0$

2.2 Estimating Propensity Scores

Common approach: Logistic regression

$$\operatorname{logit} \Pr[A = 1 \mid L] = \alpha_0 + \alpha_1^\top L$$

Estimation:

1. Fit logistic regression with treatment A as outcome, confounders L as predictors
2. Predict $\hat{e}(L_i) = \widehat{\Pr}[A = 1 \mid L_i]$ for each individual
3. Use $\hat{e}(L_i)$ for matching, stratification, or weighting

Model selection:

- Include all confounders
- Consider interactions and nonlinear terms
- Assess balance after adjustment (see Section 15.3)

3 15.3 Propensity Stratification and Standardization (pp. 213-216)

Propensity scores can be used to stratify the population and then standardize.

3.1 Propensity Score Stratification

Procedure:

1. Estimate propensity score $\hat{e}(L_i)$ for all individuals
2. Create strata (e.g., quintiles) of the propensity score
3. Within each stratum, compute $\hat{E}[Y \mid A = a, \text{stratum } s]$
4. Standardize across strata:

$$\hat{E}[Y^a] = \sum_{s=1}^{S} \hat{E}[Y \mid A = a, \text{stratum } s] \times \Pr[\text{stratum } s]$$

3.2 Checking Balance

After stratification, check whether confounders are balanced within strata:

Balance: Within stratum s, the distribution of L should be similar for treated and untreated.

Diagnostics:

- Compare means/proportions of L across treatment groups within strata
- Standardized differences: $\frac{\bar{L}_{A=1,s} - \bar{L}_{A=0,s}}{SD_{\text{pooled}}}$
- Target: Standardized differences < 0.1 (rule of thumb)

If balance is poor, refine the propensity score model (add interactions, polynomials, etc.).

Why check balance:

The propensity score should create "pseudo-randomization" within strata. If confounders aren't balanced, either:

1. The propensity score model is misspecified
2. There are violations of positivity (some strata have only treated or only untreated)

Balance checking is a diagnostic tool, not a formal test. The goal is to achieve good balance so that confounding is minimized within strata.

3.3 Example: Quintile Stratification

Steps:

1. Fit logistic regression for $\Pr[A = 1 \mid L]$
2. Divide individuals into 5 groups (quintiles) based on $\hat{e}(L)$
3. Within each quintile, compare treated vs untreated outcomes
4. Standardize across quintiles using quintile proportions as weights

Common finding: Most of the confounding is removed by stratifying on propensity score quintiles, though finer stratification may improve balance.

4 15.4 Propensity Matching (pp. 216-219)

Propensity score matching creates pairs (or sets) of treated and untreated individuals with similar propensity scores.

4.1 Matching Algorithms

1-to-1 nearest neighbor matching:

1. For each treated individual, find the untreated individual with the closest propensity score
2. Form matched pairs
3. Compute the effect as the average within-pair difference

Matching with replacement:

- Each untreated individual can be matched to multiple treated individuals
- Reduces bias but complicates variance estimation

Caliper matching:

- Only match if propensity scores are within a specified distance (caliper)
- Individuals without a close match are excluded
- Improves balance but may reduce sample size

Matching vs Stratification:

- **Matching**: Creates a matched dataset; analyze as if randomized within pairs
- **Stratification**: Creates strata; standardize across strata

Both use the propensity score to reduce confounding. Matching may be more intuitive and allows visual inspection of matched pairs.

Limitations:

- Discards unmatched individuals (efficiency loss)
- Complex variance estimation (pairs are dependent)
- Exact matching on propensity score is impossible (continuous variable)

4.2 Assessing Match Quality

After matching, assess balance:

1. **Standardized differences**: Compare means of L in matched treated vs untreated
2. **Love plots**: Graphical display of standardized differences before and after matching
3. **Distribution plots**: Compare distributions of confounders in matched samples

Target: Standardized differences < 0.1 for all confounders

4.3 Example: NHEFS Matching

Procedure:

1. Estimate propensity score for quitting smoking
2. Match each quitter to a non-quitter with similar propensity score
3. In matched sample, compare weight change between quitters and non-quitters
4. Estimate causal effect as mean difference in matched pairs

Advantages: Intuitive, allows checking balance on all confounders

Disadvantages: Discards some individuals, may not achieve perfect balance

5 15.5 Propensity Models, Treatment Models, and Marginal Structural Models (pp. 219-222)

Clarifying terminology: propensity scores, treatment models, and MSMs.

5.1 Definitions

Propensity score: $e(L) = \Pr[A = 1 \mid L]$

Treatment model: Any model for $\Pr[A \mid L]$ (or $f(A \mid L)$ for non-binary A)

Marginal structural model (MSM): Model for $E[Y^a]$ or $E[Y^a \mid V]$

5.2 Relationship

IP weighting uses the treatment model to create weights:

$$W^A = \frac{1}{\Pr[A \mid L]}$$

For binary A, this uses the propensity score:

$$W^A = \frac{1}{e(L)} \text{ if } A = 1, \quad W^A = \frac{1}{1 - e(L)} \text{ if } A = 0$$

MSM is then fit using IP weights:

$$E[Y^a] = \beta_0 + \beta_1 a$$

Distinction:

- **Propensity score**: A specific quantity, $\Pr[A = 1 \mid L]$
- **Treatment model**: A statistical model for $\Pr[A \mid L]$, which can be used to estimate propensity scores
- **MSM**: A model for potential outcomes, estimated using IP weights derived from the treatment model

These are related but distinct concepts. The propensity score can be used for:

1. IP weighting (via treatment model)
2. Stratification
3. Matching

All three methods aim to adjust for confounding.

6 15.6 Propensity Scores and Outcome Regression (pp. 222-224)

Can we combine propensity scores with outcome regression?

6.1 Doubly Robust Estimation

Idea: Use both a treatment model and an outcome model.

Estimator: Fit outcome model within propensity score strata (or matched sets), then standardize.

Double robustness: The estimator is consistent if EITHER:

1. The propensity score model is correct, OR
2. The outcome model is correct

(But not necessarily both)

6.2 Augmented IP Weighting (AIPW)

Advanced approach: Combine IP weighting with outcome modeling:

$$\hat{E}[Y^a] = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{I(A_i = a) Y_i}{f(a \mid L_i)} - \frac{I(A_i = a) - f(a \mid L_i)}{f(a \mid L_i)} m(a, L_i) \right]$$

where $m(a, L) = \hat{E}[Y \mid A = a, L]$ is the outcome model.

Properties:

- Doubly robust: Consistent if either model is correct
- More efficient than IP weighting alone when outcome model is correct
- Locally efficient (optimal variance) when both models are correct

Practical advice:

Doubly robust methods provide insurance against misspecification of one model. However:

1. If both models are wrong, the estimator is generally inconsistent
2. Model checking is still important for both models
3. These methods are more complex to implement and require specialized software

Many researchers use simpler approaches (IP weighting OR outcome regression) and conduct sensitivity analyses by trying both methods.

7 15.7 Propensity Scores for Continuous Treatments (pp. 224-226)

Propensity scores extend to **continuous treatments**, though with additional complexity.

7.1 Generalized Propensity Score

For continuous treatment A, the **generalized propensity score** is the conditional density:

$$f(A \mid L)$$

Balancing property: Under conditional exchangeability,

$$Y^a \perp\!\!\!\perp A \mid f(A \mid L)$$

7.2 Estimation

Common approach: Model the conditional distribution of A given L.

Example: Normal model

$$A \mid L \sim \text{Normal}(\mu(L), \sigma^2)$$

where $\mu(L) = \alpha_0 + \alpha_1^\top L$

GPS: $f(A_i \mid L_i) = \phi\left(\frac{A_i - \mu(L_i)}{\sigma}\right)$ where ϕ is the standard normal density.

7.3 Using the GPS

IP weighting: Create weights

$$W_i = \frac{f(A_i)}{f(A_i \mid L_i)}$$

where $f(A_i)$ is the marginal density of A (unconditional).

Stratification: Stratify on the GPS and standardize within strata.

Challenges with continuous treatments:

1. **Density estimation**: Requires choosing a parametric model for $f(A \mid L)$
2. **Positivity**: Need $f(A \mid L) > 0$ for all observed (A, L) combinations
3. **Practical positivity**: Even if positive, very small densities lead to extreme weights

4. **Model diagnostics**: Harder to check balance with continuous treatment

Despite these challenges, GPS methods can be useful for continuous exposures (dose, duration, intensity).

8 Summary

Key concepts:

1. **Outcome regression**: Models $E[Y \mid A, L]$ to estimate causal effects via g-formula
2. **Propensity score**: $e(L) = \Pr[A = 1 \mid L]$, reduces confounding adjustment to a single dimension
3. **Balancing property**: Conditioning on propensity score achieves conditional exchangeability
4. **Propensity stratification**: Create strata by propensity score and standardize
5. **Propensity matching**: Match treated and untreated with similar propensity scores
6. **Double robustness**: Combining treatment and outcome models for robustness
7. **Continuous treatments**: Generalized propensity score extends to non-binary treatments

Methods comparison:

Method	Uses	Advantages	Disadvantages
Outcome regression	$E[Y \mid A, L]$	Natural, efficient when correct	Requires correct outcome model
IP weighting	$\Pr[A \mid L]$	Natural for MSMs, handles time-varying	Can be unstable, needs correct treatment model
Propensity matching	$e(L)$	Intuitive, easy to check balance	Discards data, complex inference
Propensity stratification	$e(L)$	Reduces dimensionality	Requires choosing # of strata
Doubly robust	Both models	Robust to one misspecification	More complex, needs both models

Practical recommendations:

1. **Always check balance**: After propensity score adjustment, assess whether confounders are balanced
2. **Model carefully**: Propensity scores are only as good as the treatment model
3. **Check positivity**: Extreme propensity scores (near 0 or 1) indicate violations
4. **Use multiple methods**: Try outcome regression, IP weighting, and propensity methods as sensitivity analyses
5. **Consider double robustness**: When feasible, doubly robust methods provide insurance against misspecification

Historical note:

Propensity scores were popularized by Rosenbaum and Rubin (1983) and have become extremely popular in observational studies. However, they are not a panacea:

- They don't solve the problem of unmeasured confounding
- They still require correct model specification
- They work best when there's good overlap in propensity scores between treated and untreated

Looking ahead: Chapter 16 introduces instrumental variables, which can help with unmeasured confounding under different assumptions. Part III addresses time-varying treatments where IP weighting and g-formula shine.

Hernán, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC. https://miguelhernan.org/whatifbook.