STAT 333: Multiple Linear Regression and Confounding (Jan. 30 Lecture)

Lecture Notes

February 18, 2025

Overview

These notes continue from our discussion of simple linear regression and the two-sample t-test as a special case. We now move on to:

- The motivation for **multiple** linear regression.
- The fundamental idea of **confounding**.
- The role of **randomized controlled trials (RCTs)** as the "gold standard" for causal inference, and why we often use regression (especially when randomization is not feasible).

---

1 Multiple Linear Regression Model

1.1 General Setup

In *simple* linear regression, we modeled

$$Y_i \;=\; \beta_0 \;+\; \beta_1\, x_i \;+\; \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where x_i was a single predictor (possibly a 0–1 treatment indicator). To allow for multiple predictors, we extend this to:

$$Y_i \;=\; \beta_0 \;+\; \beta_1\, x_{i1} \;+\; \beta_2\, x_{i2} \;+\; \cdots + \;\beta_p\, x_{ip} \;+\; \varepsilon_i.$$

Here:

- Y_i is again the outcome (e.g. *lifespan*, *GPA*, *miles per gallon*, etc.).
- $x_{i1}, x_{i2}, \ldots, x_{ip}$ are p different predictors (features).
- $\beta_0, \beta_1, \ldots, \beta_p$ are unknown regression coefficients.
- ε_i is an error/noise term, often assumed i.i.d. with mean 0 (and sometimes normally distributed with variance σ^2).

1.2 Example: Vitamins and Lifespan

- Let x_{i1} be an indicator for "Takes vitamins" (1) vs. "Does not take vitamins" (0).
- Let $x_{i2}, x_{i3}, \ldots, x_{ip}$ be other *confounders* or relevant covariates such as
  - Exercise frequency (hours/week).
  - Cigarette use (packs/day) or a 0–1 indicator for any smoking.
  - Daily diet metrics (e.g., fruit/vegetable intake).
  - Sleep duration (hrs/night).
  - Socioeconomic status (income, etc.).
- The multiple regression model would then be

$$Y_i \;=\; \beta_0 \;+\; \beta_1(\text{vitamin usage})_i \;+\; \sum_{k=2}^{p} \beta_k \, x_{ik} \;+\; \varepsilon_i,$$

  where Y_i might be *lifespan* in years. If all relevant confounders are measured and included, then (under linearity assumptions) β_1 can be interpreted as the effect of taking vitamins "holding the other predictors constant."

---

2 Confounding and Randomized Controlled Trials (RCTs)

2.1 Definition of Confounding

Confounding arises when some third variable (the *confounder*) *causes* both the treatment X and the outcome Y, thereby inducing a correlation between X and Y even if there is no direct causal path from X to Y.

$$\begin{array}{c} \text{Confounder} \\ X \swarrow\!\searrow Y \end{array}$$

In the vitamins–lifespan scenario, "*healthy habits*" could be a broad confounder:

$$\text{Healthy habits} \;\longrightarrow\; \text{(Takes vitamins?)} \;\text{ and }\; \longrightarrow\; \text{(Longer lifespan)}.$$

Merely observing who takes vitamins vs. not may thus *confound* the apparent relationship unless we account for those habits.

2.2 Gold Standard: Randomized Controlled Trials

- In an **RCT**, the researcher *assigns* treatment vs. control at random, breaking the arrow from confounder $\rightarrow$ treatment. Hence, there is no correlation *induced by* the confounder.
- If X (treatment) truly is randomly assigned, a simple t-test (or simple linear regression with a 0–1 dummy for treatment) is valid for causal inference.
- However, **many treatments** (e.g. smoking, certain demographics, or ethically problematic exposures) **cannot** be randomized. Also, RCTs can be expensive, time-consuming, or infeasible in many scenarios.

2.3 What if We Cannot Randomize?

- Then we try **multiple linear regression** (or other observational-study methods).
- If all important confounders are *measured* and included as covariates $x_{i2}, \ldots, x_{ip}$, then (under modeling assumptions) β_1 reflects a *causal* effect of x_{i1} on Y_i (e.g. the vitamins-lifespan question).
- In practice, we always worry about *unmeasured confounding*. If some key confounder is not included, the regression can yield biased inferences.

2.4 Downstream Variables (What Not to Include)

- If a variable is *caused by* Y (the outcome), or is a *downstream effect* of Y, including it in the regression can create new biases.
- Generally, we only want to control for (i.e. include) *predictors that cause* Y, not those that occur *after* Y. (Such variables are often called *mediators* or *descendants* of Y.)

---

3 Interpretation of Regression Coefficients in the Multiple Setting

When fitting the model

$$Y_i \;=\; \beta_0 \;+\; \beta_1\, x_{i1} \;+\; \cdots \;+\; \beta_p\, x_{ip} \;+\; \varepsilon_i,$$

each slope β_j can be thought of as a *partial effect* of x_{ij} on Y_i, *holding all other predictors fixed*. Symbolically, some texts write:

$$\beta_j \;=\; \frac{\partial \mathbb{E}[Y \mid X]}{\partial x_j},$$

assuming a linear form. Practically:

- β_1 is often the *treatment effect* of interest if x_{i1} indicates a treatment/exposure.
- The other β_j's adjust for any confounders x_{ij}. This is sometimes called *controlling for* x_{ij} or *conditioning on* those covariates.
- β_0 is the intercept (often not of primary interest); it is the mean of Y when *all* $x_{ij} = 0$.

---

4 Key Takeaways & Next Steps

1. **Multiple linear regression** allows us to include any number of predictors $x_{i1}, \ldots, x_{ip}$, potentially reducing the bias from confounding when we cannot do a randomized experiment.
2. We rely on the assumptions that (i) all relevant confounders are measured and included, (ii) the linear form is appropriate (or sufficiently flexible), and (iii) no unmeasured confounders are driving the treatment–outcome relationship.

3. **RCTs** remain the *gold standard* for establishing causality precisely because they *break* the link from hidden confounders to X. But in many real-world studies, randomization is infeasible or unethical.

4. In upcoming lectures, we will see how to *estimate* β_j's from data and how to interpret p-values, confidence intervals, and model fit.

End of Jan. 30 Lecture Notes.