PODS Lab 7:
Experimental and statistical control

Hamza Alshamy

Center for Data Science, NYU
ha2486@nyu.edu

03/07/25

How to contextualize these topics?

- **Causality**: Does x cause y?
- **Confound**: A variable (z) that affects both the independent variable (x) and the dependent variable (y).

Method	Causality?	Controls Confounds?
Experiment (RCTs)	Yes (if well-designed)	Yes (via randomization)
Natural Experiment	Yes (under assumptions)	Yes
Partial Correlation	No	Yes (removes one known confound)
Multivariate Regression	No	Yes (for observed confounders)

Experiments

- **Experiment:** A study in which treatment assignment is directly under the control of the researcher.
- **Ideally,** random assignment of treatment (IV)
  - **E.g.** clinical trial where participants are randomly assigned to receive either a new drug or a placebo to measure its effectiveness.

- **Why does randomization *and* a large sample size ensure causality and control?**
  - **Randomization:** Ensures that **treatment** and **control** groups are statistically equivalent on all confounding variables, both observed and unobserved.
  - **Large Sample Size:** Reduces random fluctuations and ensures that any **differences observed are due to the treatment** rather than chance.

Natural Experiment

- **Natural Experiment:** Assignment to treatment is outside the control of the researcher but is **as if random**.

Natural Experiment

- **Natural Experiment:** Assignment to treatment is outside the control of the researcher but is **as if random**.
- **Example:** Effect of Institutions (North and South Korea)
  - **After WW2**, Korea split into:
  - **North Korea** with institutions based on authoritarian communism.
  - **South Korea** with democratic capitalism.
  - The assignment of **treatment** (institution style) was outside of control of researchers.

Natural Experiment: South Korea and North Korea WW2 Split

- **In 2023,** South Korea's GDP per capita is about **22 times higher** than that of North Korea's.

South and North Korea at night as seen from a satellite. The stark contrast in illumination reflects economic disparity.

Summary of Experiments

- **Experiments** allow us to infer causality.
- **Causality:** The relationship where a change in one variable directly influences a change in another.
- Experiments do that through **randomized treatment assignment** and **large sample size**.

Method	Causality?	Controls Confounds?
Experiment (RCTs)	Yes (if well-designed)	Yes (via randomization)
Natural Experiment	Yes (under assumptions)	Yes

Partial Correlation: What if you don't have resources for a study and randomization?

Two Assumptions:

1. **You cannot conduct an experiment.**
   - Experiments can be expensive, time-consuming, or raise ethical concerns.

Partial Correlation: What if you don't have resources for a study and randomization?

Two Assumptions:

1. **You cannot conduct an experiment.**
   - Experiments can be expensive, time-consuming, or raise ethical concerns.
2. **You are aware of one confounder z in your study.**

- **Partial Correlation:** Use when there is **one known confounder** and conducting an experiment is not possible.
  - **Goal:** Correlation between two variables, controlling for the effect of a third variable (confound z).

Okay, but what is partial correlation?

Remember: Simple Linear Regression

$$Y = \underbrace{\beta_0 + \beta_1 X_1}_{\hat{y}} + \underbrace{\varepsilon}_{\text{residual}}$$

where

- $\hat{Y} = \beta_0 + \beta_1 X_1$ is the **predicted value** of y.
- ε is the **residual**, the difference between the actual and predicted values.
- **Key Idea**: The residual represents the part of Y that is not explained by X_1.

Okay, but what is partial correlation?

- **Idea of Partial Correlation:**
  - It is the correlation between the **residual of** X after regressing on Z and the **residual of** Y after regressing on Z.

Algorithm:

1. Perform simple linear regression predicting X (IV) from Z (Confounder)
   - $X = \beta_0 + \beta_1 Z + \varepsilon_X$
2. Perform simple linear regression predicting Y (DV) from Z (Confounder)
   - $Y = \beta_0 + \beta_1 Z + \varepsilon_Y$
3. Compute the correlation between the **residuals** of the two regressions
   - $r_{XY \cdot Z} = r_{\varepsilon_X \varepsilon_Y}$

How interpret Partial Correlation?

- **How does the correlation between the residuals control for the confounder?**
  - By regressing X and Y on Z, we remove the variance in X and Y that is **explained by** Z.
  - The residuals ε_X and ε_Y now represent the variation in X and Y that is **independent of** Z.
  - The correlation between ε_X and ε_Y measures the **direct association** between X and Y after removing the influence of Z.

Multivariate (multiple) Linear Regression

- **Question:** What if there are multiple possible predictors and/or confounders?

Multivariate (multiple) Linear Regression

- **Question:** What if there are multiple possible predictors and/or confounders?
- **Do multiple linear regression!**

Multivariate Linear Regression Equation

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

where

1. Y: **outcome**.
2. β_0: Offset (in a simple linear regression, it is the y-intercept).
3. $\beta_1, \beta_2, \ldots \beta_n$: The **weights** of each predictor X (how much each predictor matters).
4. $X_1, X_2, \ldots, X_n$: **Predictors** and/or possible confounders.
5. ε: **Residual/error**, capturing variation in Y not explained by the predictors.

Multivariate Linear Regression: Interpretation

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- **"Ceteris Paribus" = "All else equal"**
  - Assumes that everything else is held constant.
- In multivariate regression, we **invoke Ceteris Paribus**
  - This means isolating the effect of one variable on another by assuming that all other variables remain unchanged.
- $\beta_1, \beta_2, \ldots, \beta_n$ represent the effect of each predictor.
  - **Interpretation of β_4:**
  - β_4 measures the expected change in the outcome Y for a one-unit increase in predicotr X_4, **holding all other predictors constant**.

Regression Evaluation Metrics:

- **Questions:**
  - How do we compare different models?
  - How do we compare a model with **10 predictors that capture 90%** of the variation vs. a model with **5 predictors that capture 80%** of the variation?
- **Regression Evaluation Metrics:**
  1. Multiple Correlation (R)
  2. Coefficient of Determination (COD or R^2)
  3. Root Mean Squared Error (RMSE)

Regression Evaluation Metrics: Multiple Correlation and R^2

- **(1) Multiple Correlation (R):** Measures the correlation between predicted values $\hat{Y}$ and actual values Y.
  - **Equation:**

$$R_{Y,\hat{Y}} = \frac{\text{Cov}(Y, \hat{Y})}{\sigma_Y \cdot \sigma_{\hat{Y}}}$$

  - **Range:** $R \in [-1, 1]$

Regression Evaluation Metrics: Multiple Correlation and R^2

- **(1) Multiple Correlation (R):** Measures the correlation between predicted values $\hat{Y}$ and actual values Y.
  - **Equation:**
  $$R_{Y,\hat{Y}} = \frac{\text{Cov}(Y,\hat{Y})}{\sigma_Y \cdot \sigma_{\hat{Y}}}$$
  - **Range:** $R \in [-1,1]$
- **(2) Coefficient of Determination (R^2):** Proportion of variance in Y that is explained by the model.
  - **Equation:**
  $$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$
  - **Range:** $R^2 \in (-\infty, 1] \rightarrow$ (**Note:** R^2 can be negative when the model fits worse than a horizontal line at $\bar{Y}$, especially if there's no intercept.)
  - **Interpretation (e.g., $R^2 = 0.6$):** The model explains **60% of the variance** in the outcome variable.

Regression Evaluation: Root Mean Squared Error (RMSE)

(3) RMSE: Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- **Definition:** Measures the **average distance of prediction errors**.
- **Interpretation:** Lower RMSE indicates better model performance, meaning predictions are closer to actual values.
- **Units:** RMSE has the same units as Y, making it **directly interpretable** in the context of the outcome variable.
- **Range:** $RMSE \in [0, \infty]$

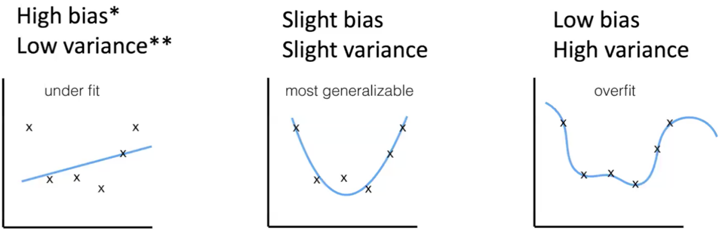
Why Not Just Add More Predictors?

- **Question:** If adding more predictors means controlling for more variables, **why not include as many predictors as possible?**
- **Two Considerations:**
  1. **Multi-collinearity:** High correlation among predictors can distort coefficient estimates and make the model unstable.
  2. **Overfitting:** The model starts fitting noise in the data instead of capturing the true underlying relationship.
     - The model loses generalizability to new unseen data.

Ideal model should be balanced

- **Ideal model should be balanced:**
  1. Accounts for variance
  2. Be as simple as possible
     - **Occam's Razor:** If two theories have the same explanatory power, we should prefer the simpler one
  3. No multi-collinearity
     - When the predictors themselves are correlated
  4. No overfitting (to noise)

Bias/Variance Tradeoff



1. **Bias (Underfitting):** When the model is **too simple** to capture the true relationship between predictor and outcome
2. **Variance (Overfitting):** When the model is **too flexible** in order to fit to the data

- **Takeaway:** Balance bias (simplicity) and variance (complexity)

Regularized Regression

- **OLS Regression** minimizes squared residual but can lead to **overfitting**, especially with many predictors.
- To improve generalization, we can **intentionally introduce bias** to reduce variance
- **Regularization** adds a penalty to the model's complexity, preventing it from fitting noise in the data.
    - Penalizes large β coefficients
- **Two Common Types of Regularized Regression:**
    1. **Ridge Regression (L2 Regularization)**: Penalizes large coefficients by adding a squared penalty term.
    2. **LASSO Regression (L1 Regularization)**: Encourages sparsity by shrinking some coefficients to zero.

(1) Ridge Regression – L2 Regularization

Ridge Regression (L2):

$$\arg\min_{\beta} \sum_{i=1}^{n} \underbrace{(Y_i - \hat{Y}_i)^2}_{\text{Residuals}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{Penalty Term}}$$

- **Penalty term:** Introduces bias to reduce variance.
- Shrinks regression coefficients β_j **towards** zero, but **never exactly to zero**.
- Penalizes large β_j values, preventing overfitting.
- As λ increases, regularization strengthens, shrinking coefficients more aggressively.

- **How to find the optimal λ?**
  - Through **cross-validation**: Iteratively testing different λ values to minimize validation error.

(2) Least Absolute Shrinkage and Selection Operator (LASSO) Regression – L1 Regularization

LASSO Regression (L1):

$$\arg\min_{\beta} \sum_{i=1}^{n} \underbrace{(Y_i - \hat{Y}_i)^2}_{\text{Residuals}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{Penalty Term}}$$

- **Feature Selection:** LASSO can **set some β_j coefficients exactly to 0**.
  - As λ increases, more β_j coefficients shrink to 0, removing less important predictors.
  - Leads to a **simpler model** with fewer predictors.
- **Addresses Overfitting & Multicollinearity:**
  - Reduces complexity by eliminating non-informative predictors.
  - Helps in high-dimensional datasets where many predictors are correlated.