Crash Course in Causality Assignment

Sathvik Vadavatha

April 1, 2025

Chapter 1

Deep Dive into Causality

1.1 Introduction to Causality

Causality lies at the heart of scientific discovery and is a foundational principle in both Data Science and Machine Learning. While predictive models are often concerned with correlations, understanding *why* something happens — and how to intervene — requires a causal lens.

In Data Science, causal inference allows us to answer questions such as:

- What is the effect of changing tire strategy on the probability of winning an F1 race?
- If a safety car is deployed mid-race, how does it impact final driver positions?

The goal of this chapter is to guide you from foundational concepts to practical tools for drawing causal conclusions from data, particularly using real-world analogies from Formula 1 racing.

Why Causality Matters in Machine Learning

Traditional machine learning models, such as regression or classification algorithms, optimize for predictive accuracy. However, they often fall short in scenarios that require decision-making or policy interventions.

Example: Suppose a model predicts that rain increases the number of DNFs (Did Not Finish) in Formula 1 races. Does that mean rain *causes* DNFs, or are there other lurking variables (e.g., lower visibility, different tire compounds)?

Types of Questions Causal Inference Can Answer

- **Interventional**: What happens to Y if we do X?
- **Counterfactual**: What would have happened if we had not done X?
- **Attributional**: How much of Y was caused by X?

Understanding these distinctions enables a practitioner to go beyond prediction and into explanation, diagnosis, and optimization.

1.2 Correlation vs Causation

A common misconception in data analysis is equating correlation with causation. While correlated variables often appear related, they may not share a causal connection.

F1 Case Study: Pit Stops and Winning Probability

Let's say we analyze F1 race data and observe that drivers who pit early are more likely to finish in the top 5. This correlation may lead to the false conclusion that early pit stops *cause* better outcomes.

However, there may be a hidden variable: top teams with faster cars often pit earlier due to flexible strategy. So, it's not the pit stop timing that directly causes the better placement — it's team performance.

Equation Formulation

We often use the Pearson correlation coefficient r to measure linear association:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

However, a high r value does not imply a causal relationship.

Example Table: Driver Pit Stop Timing vs Placement

Table 1.1: Pit Stop Timing and Race Results (Simplified Example)

Driver	Pit Lap	Finish Position	Team
Hamilton	12	1	Mercedes
Verstappen	14	2	Red Bull
Alonso	10	5	Aston Martin
Zhou Guanyu	24	18	Alfa Romeo
Magnussen	25	16	Haas

This table may suggest a trend, but we must dig deeper to establish a causal explanation.

Key Takeaway

Correlation may help identify patterns, but only causal inference can guide decisions and interventions.

1.3 Causal Diagrams (Directed Acyclic Graphs)

Directed Acyclic Graphs (DAGs) are visual representations of causal relationships between variables. Each node represents a variable, and directed edges (arrows) represent causal influence.

Why DAGs Matter

DAGs help us:

- Identify confounding variables
- Determine valid adjustment sets
- Design better experiments and interventions

Formula 1 Example: Weather, Pit Strategy, and Finishing Position

Suppose we want to understand the impact of pit strategy on race outcome. The weather might influence both the strategy and the outcome.

Explanation:

- Weather → Pit Strategy (teams choose different tires in wet/dry conditions)
- Pit Strategy → Finishing Position (strategy can affect pace)
- Weather → Finishing Position (indirectly via visibility, grip)

In this case, to estimate the causal effect of pit strategy on the outcome, we must adjust for the confounder — Weather.

Backdoor Criterion

To identify a valid adjustment set, we use the **Backdoor Criterion**:

> A set of variables Z satisfies the backdoor criterion relative to an ordered pair of variables (X, Y) in a DAG if:
>
> - No node in Z is a descendant of X
> - Z blocks every path between X and Y that contains an arrow into X

This ensures we're isolating the true causal path from X to Y.

1.4 Counterfactual Reasoning

Counterfactuals ask "what if" questions and are crucial for personalized decision-making. They imagine an alternate world where a different action was taken.

F1 Example: What if Hamilton had taken Soft Tires instead of Mediums?

Imagine Hamilton finished second using medium tires. We want to know: *Would he have won the race had he switched to soft tires during the final pit stop?*

This is a counterfactual query:

$$Y_{X=soft}$$

where:

- Y: Final position
- X: Tire choice (actual = Medium, counterfactual = Soft)

Structural Causal Models (SCMs)

Counterfactuals are often evaluated using SCMs:

- Define a set of equations: $Y = f_X(X, U)$
- Replace X with desired counterfactual value
- Compute the outcome using the same noise term U

Example SCM

$$U_1 \sim \mathcal{N}(0, 1)$$
$$\text{Lap Time} = f_1(\text{Tire Type}, U_1)$$
$$\text{Final Position} = f_2(\text{Lap Time}, \text{Pit Strategy})$$

By changing Tire Type and keeping U_1 fixed, we simulate the counterfactual world.

Use Cases in Data Science

- Recommender Systems: Would the user have clicked a different item?
- Marketing: What if we offered a 20% discount instead of 10%?
- Medical ML: Would the patient have improved under an alternative treatment?

Counterfactuals power **individual-level** decisions — a crucial step toward actionable AI.

1.5 Causal Inference Techniques

Once we understand causal diagrams and counterfactual reasoning, the next step is to estimate causal effects using real-world data. This section introduces widely-used causal inference techniques.

1. Randomized Controlled Trials (RCTs)

RCTs are the gold standard for causal inference. Participants are randomly assigned to treatment and control groups, ensuring there are no confounding variables.

F1 Analogy: Imagine FIA randomly assigns different tire compounds to teams before a race. Since the assignment is random, any performance differences can be causally attributed to the tire type.

However, RCTs are often impractical or unethical, especially in observational domains like business or healthcare.

2. Matching Methods

Matching tries to simulate a randomized experiment by pairing individuals with similar characteristics but different treatments.

Example: To assess the effect of pit strategy, we can match drivers with similar qualifying positions, car performance, and weather conditions — then compare outcomes based on different strategies.

3. Propensity Score Matching (PSM)

Propensity scores represent the probability of receiving treatment given observed covariates. PSM reduces bias by comparing treated and untreated units with similar scores.

Formally:

$$e(X) = P(T = 1 \mid X)$$

Where T is treatment assignment and X are observed covariates.

4. Instrumental Variables (IV)

When unobserved confounding exists, IVs can be used. An instrument is a variable that affects the treatment but not the outcome (except through the treatment).

F1 Example: A rule change (e.g., mandatory pit window) that affects pit stop timing but not final performance directly could act as an instrument.

5. Difference-in-Differences (DiD)

Used in time-series data to estimate causal effects when before-and-after data is available for both treatment and control groups.

F1 Example: Compare team performance before and after a regulation change (e.g., cost cap) using teams unaffected by the rule as controls.

$$\text{DiD} = (\bar{Y}_{\text{treatment, after}} - \bar{Y}_{\text{treatment, before}}) - (\bar{Y}_{\text{control, after}} - \bar{Y}_{\text{control, before}})$$

Each method has assumptions — so understanding the data context is critical for applying the right one.

1.6 Applications of Causal ML in the Real World

Causal reasoning is reshaping how we approach problems across industries, shifting from *prediction* to *actionable decision-making*.

1. Healthcare

- **Question:** Does a new treatment reduce patient recovery time? - Causal ML helps simulate randomized trials from observational EHR data. - Techniques: Propensity score matching, uplift modeling.

2. Marketing and Business Strategy

- **Question:** What's the effect of a 10% discount on customer retention? - A/B testing + uplift modeling can optimize ad targeting and promotions.

3. Economics and Policy

- Used to estimate impact of laws, subsidies, or interventions. - Techniques like DiD and IVs are frequently applied.

4. Sports Analytics — Formula 1 Case Study

Formula 1 provides a rich testbed for causal analysis:

- **Tire Strategy:** Evaluate how soft vs hard compounds affect race pace.
- **Weather:** Estimate the effect of rain on DNF probabilities.
- **Regulations:** Measure performance impact of cost cap or ground-effect rule.
- **DRS Zones:** Quantify overtaking likelihood with and without DRS.

Example: Using historical data, we estimate:

$$P(\text{Win} \mid \text{Soft tires}) - P(\text{Win} \mid \text{Hard tires})$$

This gives a treatment effect of tire choice on race win probability, enabling strategic decisions.

5. Personalization and Recommendation

- Counterfactuals help answer: "What would the user have done if shown another recommendation?" - Applications in e-commerce, music, and news feed curation.

Summary

From improving race strategies in F1 to saving lives in healthcare, causal ML is critical for systems that don't just learn patterns but also help us act wisely in the real world.

1.7 Tools and Frameworks for Causal Inference

With the rising importance of causal analysis in data science, several powerful Python libraries have emerged to simplify causal modeling, estimation, and validation.

1. DoWhy

DoWhy is a Python library by Microsoft for causal inference that emphasizes transparency and testability.

Key Features:

- Graph-based causal modeling
- Estimation using matching, propensity scores, and more
- Refutation tests to validate causal claims

```
pip install dowhy
```

2. EconML

Developed by Microsoft, EconML focuses on heterogeneous treatment effects using machine learning.

Example Use: Estimating how the effect of pit strategies varies across circuits in F1.

```
pip install econml
```

3. PyWhy (CausalPy, Pyro Causal)

An evolving ecosystem for probabilistic causal modeling. It supports Bayesian inference and integration with Pyro.

4. CausalNex

Developed by QuantumBlack (McKinsey), CausalNex enables creation of DAGs from data, intervention simulations, and Bayesian modeling.

5. Other Tools

- **causalimpact** — Google's Bayesian structural time series for impact estimation.
- **TETRAD** — GUI-based tool from Carnegie Mellon for causal discovery.
- **Lingam** — For discovering non-Gaussian causal structures.

Sample Workflow with DoWhy

1. Define the causal graph
2. Identify valid estimators using backdoor/IV criteria
3. Estimate the treatment effect
4. Run refutation checks (e.g., placebo treatment)

These tools empower analysts to move beyond black-box modeling and reason about interventions in a principled way.

1.8 Challenges and Limitations

While causal inference is powerful, it comes with its own challenges that must be carefully considered.

1. Unobserved Confounding

If a confounder is not measured or available in the data, it can bias the causal estimate.

F1 Example: Car upgrades between races may influence both qualifying performance and final position — but if we don't record them, our estimates may be flawed.

2. Selection Bias

When the data collected is not representative of the population, the conclusions drawn may not generalize.

Example: Analyzing pit strategies only from top teams may miss how they affect mid-field drivers differently.

3. Model Misspecification

Incorrect assumptions in the causal model (e.g., wrong DAG structure or treatment relationships) can lead to invalid inferences.

4. Identifiability

In some cases, it's not possible to estimate the causal effect from available data — no adjustment set exists, or instrumental variables are weak.

5. Data Limitations

Temporal resolution, missing data, and poor variable definitions can all limit the success of causal methods.

Best Practice: Always perform robustness checks, sensitivity analyses, and clearly document assumptions.

1.9 Conclusion and Takeaways

Causality transforms the way we use data — from predicting outcomes to understanding mechanisms, making decisions, and optimizing interventions. In the realm of Data Science and Machine Learning, causal thinking adds depth and responsibility to our models.

Key Lessons from this Chapter

- Correlation is not causation — we need causal models and assumptions.
- DAGs help visualize and reason about causal paths and confounders.
- Techniques like matching, IV, DiD, and counterfactuals enable causal estimation from data.
- Tools like DoWhy and EconML bring these techniques to practice.
- Real-world applications in Formula 1 illustrate the power of causal reasoning — from strategy design to performance analysis.

Final Thoughts

As the world becomes more data-driven, the need for causal literacy is greater than ever. Whether you're optimizing race strategies, tailoring medical treatments, or evaluating policy interventions, causality equips you with the tools to ask better questions — and answer them with confidence.

"Prediction is good, but understanding is better."

Bibliography

[1] Judea Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[2] Guido W. Imbens and Donald B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[3] K. Sharma, A. Ghosh, A. Agarwal, J. Letham, and E. Horvitz, "DoWhy: An End-to-End Library for Causal Inference," Microsoft Research, `https://github.com/py-why/dowhy`, 2020.

[4] Microsoft Research, "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation," `https://github.com/microsoft/EconML`, 2019.

[5] Kay H. Brodersen et al., "Inferring causal impact using Bayesian structural time-series models," *Annals of Applied Statistics*, 2015.

[6] Ergast Developer API, "Formula 1 Race Data," `https://ergast.com/mrd/`, Accessed 2025.

[7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.