(draft) Casual Inference and Discovery: Final Project Report

Pupipat Singkhorn

May 2025

Dataset: Road accidents list from Y51-58 festival period. `https://data.go.th/dataset/item_7d61f508-d2e1-4f0c-8408-dfde29f111f5`

1. Top Factors Affecting Survivability

Q: What factors matter the most to the survivability of the person in the accident?

Analysis: Trained a Random Forest on

> Age, sex, day_of_month, month, hour, road type, status (passenger/pedestrian), injured vehicle, counterpart vehicle, safety measure, and alcohol use

and extracted feature importances.

Top 5 Predictors:

1. **Age** (0.328)
2. **Hour of day** (0.232)
3. **Day of month** (0.100)
4. **Unknown alcohol use** (0.044)
5. **Sex** (0.021)

Interpretation: Older victims have markedly lower survival rates, followed by the time of day (possibly reflecting traffic or light conditions), then day-of-month effects (first vs. second half of January festival), with unknown alcohol status and gender playing smaller but non-negligible roles.

2. Helmet Effect in Motorcycle Accidents

Q: How much does helmet help survivability in motorcycle accident?

- **Sample:** Motorcyclists only, excluding "unknown" measures
- **Treatment:** helmet = 1 if "wore helmet"
- **Method:** DoWhy propensity-score matching

- **Estimate (ATE):** +0.0025 (i.e. +0.25 pp absolute increase in survival)
- **Naïve diff-in-means:** +0.0076 (0.76 pp)
- **Robustness checks:** Placebo treatment $p = 0.84$; random common cause $p = 1.0$

Interpretation: Wearing a helmet yields a small but positive survival benefit (∼0.25 percentage points) after adjusting for confounders, though the naïve estimate overstates it (0.76 pp) by failing to account for selection bias.

3. Seatbelt Effect in Car Accidents

Q: How much does seatbelt help survivability in car accident?

- **Sample:** Cars (sedan/taxi, pickup, van), excluding unknown measures
- **Treatment:** seatbelt = 1 if "wore seatbelt"
- **Method:** DoWhy propensity-score matching
- **Estimate (ATE):** +0.0196 (1.96 pp absolute increase)
- **Naïve diff-in-means:** +0.0199 (1.99 pp)
- **Robustness checks:** Placebo $p = 0.93$; random common cause $p = 1.0$

Interpretation: Seatbelt use increases survival by about 2 percentage points. Adjusted and naïve estimates are nearly identical, suggesting limited confounding in this subset.

4. Alcohol's Impact on Survivability

Q: Does alcohol factor into survivability given the dataset?

- **Sample:** All vehicles, excluding "unknown" alcohol status
- **Treatment:** alcohol = 1 if "drank"
- **Method:** DoWhy propensity-score matching
- **Estimate (ATE):** +0.0064 (0.64 pp)
- **Naïve diff-in-means:** +0.0042 (0.42 pp)
- **Robustness checks:** Placebo $p = 0.70$; random common cause $p = 1.0$

Interpretation: Surprisingly, after adjusting for confounders, alcohol consumption appears to increase survival by ∼0.6 pp—likely reflecting residual confounding or "drinker" correlations (e.g. drinking happens on safer roads/times). The small effect and non-significant refutations ($p > 0.05$) counsel caution.

5. Hospital Effect on Survivability

Q: Does the hospital affect the survivability?

- **Model:** Logistic regression predicting survival from demographics, road and crash factors
- **Metric:** For each hospital (by ID), compare **observed** vs. **expected** survival rate
- **Result:** Standard deviation of hospital-level survival difference = **0.0538**

Interpretation: Hospitals differ in their mortality outcomes by roughly ±5 pp (one standard-deviation spread). This suggests meaningful variability in care quality or patient mix across hospitals.

6. Hour-of-Day Survival Patterns (Exploratory)

Q: Any other interesting information you can gain from this dataset?

- **Plot:** Survival rate by hour (0–23) shows a trough around 4 am (∼96.2%) and a peak around 12 noon–1 pm (∼98.9%), then a gradual decline into the evening.

Insight: Very early-morning accidents (around dawn) have the lowest survival—perhaps reflecting emergency-response delays or impaired visibility—whereas midday incidents see the highest survival.

Overall Conclusions

- **Demographics** (age, sex) and **temporal factors** (hour/day) dominate survival risk.
- **Protective measures** (helmets, seatbelts) show modest but real increases in survival.
- **Alcohol** effects are confounded and require deeper investigation.
- **Hospital performance** varies substantially, indicating opportunities for targeted quality improvements.
- **Time-of-day** patterns point to resource-allocation or public-safety messaging opportunities during vulnerable periods (e.g. pre-dawn hours).

This causal analysis blends machine-learning–driven discovery (feature importances) with formal causal inference (DoWhy estimands and refutation tests), offering both **what** matters and **how much** it matters to accident survivability.

Appendix