

Stata Conference 2024

Portland, Oregon
(1–2 August 2024)

Optimal policy learning with observational data in multi-action scenarios: Stata implementation

Giovanni Cerulli & Antonio Zinilli
CNR-IRCRES

Research Institute on Sustainable Economic Growth,
National Research Council of Italy

Outline

Introduction

- Presentation of a new Stata command for **optimal policy learning (OPL)** with observational data
- Focus on data-driven optimal decision-making in **multi-action (or multi-arm)** settings

Main components of the commands

- **Estimation:** Techniques for deriving optimal policies
- **Uncertainty:** Accounting for different attitudes towards risk in decision-making
- **Regret Estimation:** Measuring the potential regret from decisions using three methods:
 - *Regression Adjustment*
 - *Inverse Probability Weighting*
 - *Doubly Robust Estimators*

Stata implementation

- Introduction to the syntax and usage of the new Stata commands: **opl_ma_fb** and **opl_ma**

Application Example

- Focus on an application related to labour and agricultural policies

Introduction to OPL

Optimal Policy learning (OPL) is a technique used for **automizing ex-ante decision** processes operated by agents who have the purpose of carrying out a specific intelligent task

For this purpose, **OPL** makes use of experience/information from **past decisions** (i.e., accumulated data) and exploits them for “optimal” decision making

Definition of OPL

OPL is a kind of **data-driven decision making** using **machine learning (ML)** and **causal inference (CI)** for suggesting optimal ex-ante decisions based on a past accumulated experience about decisions undertaken to carry out specific tasks

$$\text{OPL} = \text{DATA} + \text{ML} + \text{CI}$$

OPL: main objects

Agent = The person/institution in charge of the choice

Task = Objective for taking a certain decision

Environment = external-to-agent conditions

Action = available choice alternatives

Reward = Positive achievement measured after decision is made

Real examples of OPL

Example 1. Commuting to work

Agent = one single person

Task = commuting to work

Environment = temperature, humidity, traffic jam

Action = Commuting modes: car (A), walk (B), public transportation (C)

Reward = Positive feeling in a scale (0-10) measured at the end of the day

Example 2. Business advertising decision

Agent = a business

Task = raising weekly profits

Environment = competitor strength, operative costs, productivity

Action = Advertising modes: internet (A), newspapers (B), television(C)

Reward = Weekly net-profits

Example 3. Medical treatment

Agent = a doctor

Task = providing a treatment to a patient

Environment = symptoms, blood tests, reporting

Action = treatment modes: T1 (A), T2 (B), T3 (C)

Reward = Positive recovery probability

Example 4. Enrollment in a policy program

Agent = a policy-maker

Task = selecting policy beneficiaries

Environment = individual characteristics

Action = selecting modes: Yes (A), No (B)

Reward = Positive effect of the policy on the individual

Environment - Action - Reward

Y = return (outcome, payoff, reward)

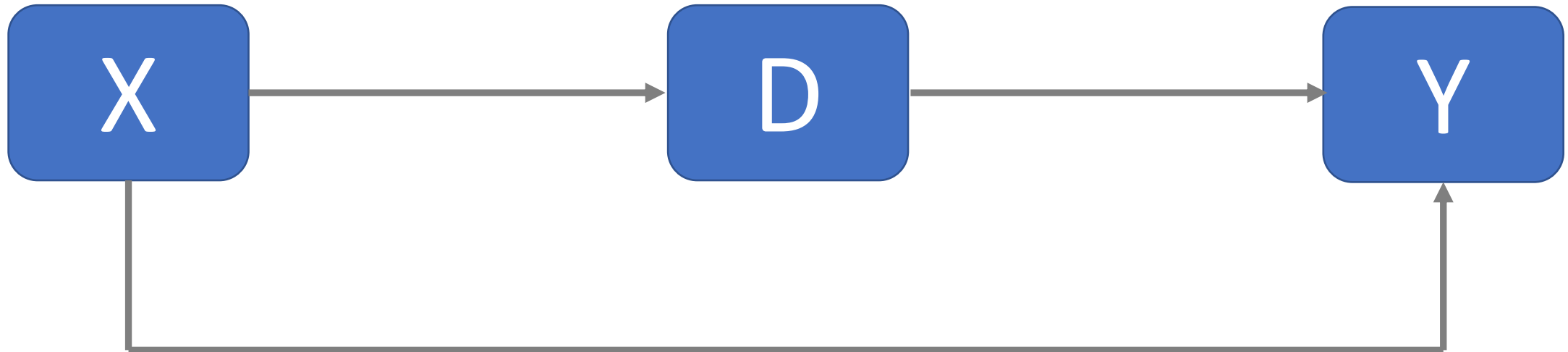
X = environment (state of the world, context)

D = choice (action, policy, decision)

Environment

Action

Reward



The decision-making setting

Consider an agent i having to choose at time t between a set of $J + 1$ different actions $D_{it} = \{0, 1, 2, \dots, j, \dots, J\}$ with corresponding set of rewards $\{Y_{it}(0), Y_{it}(1), \dots, Y_{it}(J)\}$ with distributions $\{\mathcal{F}_{it}(0), \mathcal{F}_{it}(1), \dots, \mathcal{F}_{it}(J)\}$.

Define the agent's action indicator function $d_{it}(j) = \mathbf{1}[D_{it} = j]$, with $j = 0, \dots, J$, taking value 1 when the agent selects action j and 0 otherwise.

Define as \mathbf{x}_{it} a set of p features that represent the signal from the environment (the agent i belongs to) at time t .

Potential outcomes

These are **potential outcomes**: we can observe *only one* of them, never all at the same time:

$$D_{it} = \{0, 1, 2, \dots, j, \dots, J\}$$



$$\{Y_{it}(0), Y_{it}(1), \dots, Y_{it}(J)\}$$

Non-identification of the optimal choice

Set of actions: $\{0, 1, 2\}$

Time	Y_t	D_t	$d_t(0)$	$d_t(1)$	$d_t(2)$	$Y_t(0)$	$Y_t(1)$	$Y_t(2)$	X_t
1	Y_1	0	1	0	0	Y_1	.	.	X_1
2	Y_2	0	1	0	0	Y_2	.	.	X_2
3	Y_3	0	1	0	0	Y_3	.	.	X_3
4	Y_4	1	0	1	0	.	Y_4	.	X_4
5	Y_5	1	0	1	0	.	Y_5	.	X_5
6	Y_6	1	0	1	0	.	Y_6	.	X_6
7	Y_7	1	0	1	0	.	Y_7	.	X_7
8	Y_8	1	0	1	0	.	Y_8	.	X_8
9	Y_9	2	0	0	1	.	.	Y_9	X_9
10	Y_{10}	2	0	0	1	.	.	Y_{10}	X_{10}

Table 1: Example of non-identification of counterfactual rewards. Observe that X_t is the signal from the environment.

Potential outcome

At each time t , agent i can choose only one out of the $J + 1$ possible alternatives. His observed reward, Y_{it} , is thus equal to:

$$Y_{it} = d_i(0)Y_{it}(0) + \dots + d_i(j)Y_t(ij) + \dots + d_i(J)Y_{it}(J)$$

as only one of the $J + 1$ potential rewards can be observed for individual i at time t .

The decision-making setting

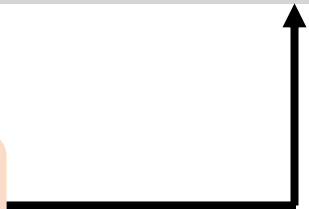
Given a certain configuration of the environment (that is, for a given \mathbf{x}_{it}), we define the *conditional expected reward* if agent i when he chooses action j as:

$$\mu_{it}(j, \mathbf{x}_{it}) = \mathbb{E}(Y_{it}(j) | \mathbf{x}_{it})$$

which implies that the optimal action j^* to select is:

$$j^* = \{j : \max\{\mu_{it}(j, \mathbf{x}_{it})\}, j = 1, 0, \dots, J\}$$

**Optimal Decision Rule
(ODR)**



Unfortunately, this optimal choice cannot be identified by data as it entails the knowledge of counter-factual quantities. Indeed, for each j , $\mu_{it}(j, \mathbf{x}_{it})$ is inherently unknown.

Identification

Following Cattaneo (2010) and Cattaneo and Drukker (2013), however, two assumptions allow to identify the conditional expected reward from data:

A1. Selection-on-observables. For all $j = 0, 1, \dots, J$:

$$Y(j) \perp d(j) | \mathbf{x}$$

Given \mathbf{x} the potential outcome and the choice dummy are independent

A2. Overlapping. For all $j = 0, 1, \dots, J$:

$$0 < p_{min} < p_j(\mathbf{x}) \text{ with } p_j(\mathbf{x}) = P(D = j | \mathbf{x}).$$

The *propensity score* is never equal to 0

Under assumptions A1 and A2, we can prove that:

$$\mu_{it}(j, \mathbf{x}_{it}) = E(Y_{it} | D_{it} = j, \mathbf{x}_{it})$$

The counterfactual becomes function of observable elements

Estimable optimal decision rule (ODR)

Procedure 1. *Optimal action selection under assumptions A1 and A2*

- Generate the mapping between Y_{it} and \mathbf{x}_t for each $D_{it} = 0, 1, \dots, J$ using a specific learner, and obtain the following set of J predictors:

$$\mathcal{M}_{it} = \{\hat{\mu}_{it}(0, \mathbf{x}_{it}), \{\hat{\mu}_{it}(1, \mathbf{x}_{it}), \dots, \hat{\mu}_{it}(j, \mathbf{x}_{it}), \dots, \hat{\mu}_{it}(J, \mathbf{x}_{it})\}\}$$

- Given a new environment signal \mathbf{x}_{t-1} , evaluates the previous set of predictions at $t + 1$, thus getting:

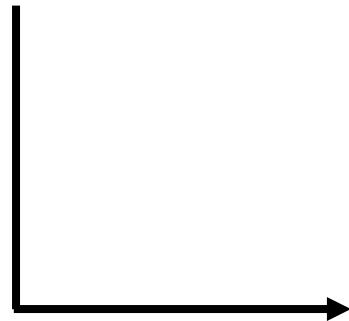
$$\mathcal{M}_{i,t+1} = \{\hat{\mu}_{i,t+1}(0, \mathbf{x}_{i,t+1}), \{\hat{\mu}_{i,t+1}(1, \mathbf{x}_{i,t+1}), \dots, \hat{\mu}_{i,t+1}(j, \mathbf{x}_{i,t+1}), \dots, \hat{\mu}_{i,t+1}(J, \mathbf{x}_{i,t+1})\}\}$$

- Select the best action to undertake at $t + 1$ according to this rule:

$$j_{t+1}^* = \{j : \max\{\mathcal{M}_{i,t+1}\}, j = 1, 0, \dots, J\}$$

Estimate of the mapping between Y_D and $[X | D]$

- The **mapping** between Y_D and $[X | D]$ is uncertain (or stochastic)
- We have to consider the “**expected return**”



$$E(Y_D | X, D)$$



Expected **potential** return,
given environment X and choice D

An example with 3 actions: A, B, C

- Define the conditional expectation of Y_D given X in the three states of the world $\{A, B, C\}$
- Make the decision having the largest conditional expectation

$$\mu_{iA} = E(Y_A \mid X_i, D_i = A)$$

$$\mu_{iB} = E(Y_B \mid X_i, D_i = B)$$

$$\mu_{iC} = E(Y_C \mid X_i, D_i = C)$$

These are three **potential returns** arising when action is either A, B, or C

In general it is not identified by observation

Optimal decision rule

Optimal decision

$$D_i^* = \{D: \max[\mu_{iD}], D = (A, B, C)\}$$

ML-based Regression Adjustment estimation of μ_{iD}

Under assumption (i) and (ii), suppose to have the following i.i.d. dataset $\{Y_i, X_i, D_i\}$ with $i = 1, \dots, N$, with D made of three actions $D = \{0, 1, 2\}$, then an estimation of μ_{iD} is:

$$\mu_{iD} = \mu_D(X_i) = E(Y_i | X_i, D_i)$$

can be obtained using a prediction of Y_i obtained from an **ML regression** of Y on X in the sub-group of units having D_i .

In this way, we have an estimate of all the counterfactuals for each unit i .

PROCEDURE:

1. Generate the mapping between Y_i and X_i under $D=0, 1, 2$ by an ML-RA
2. Given X_i , compute the predictions $\hat{\mu}_{i0}, \hat{\mu}_{i1}, \hat{\mu}_{i2}$ using the previous ML-mappings
3. Apply the ODR to select the action to undertake

Optimal action selection

Identifier	Y_t	D_t	X_t	$\hat{\mu}_t(0, X_t)$	$\hat{\mu}_t(1, X_t)$	$\hat{\mu}_t(2, X_t)$
1	Y_1	0	X_1	$\hat{Y}_{1,0}$	$\hat{Y}_{1,1}$	$\hat{Y}_{1,2}$
2	Y_2	0	X_2	$\hat{Y}_{2,0}$	$\hat{Y}_{2,1}$	$\hat{Y}_{2,2}$
3	Y_3	0	X_3	$\hat{Y}_{3,0}$	$\hat{Y}_{3,1}$	$\hat{Y}_{3,2}$
4	Y_4	1	X_4	$\hat{Y}_{4,0}$	$\hat{Y}_{4,1}$	$\hat{Y}_{4,2}$
5	Y_5	1	X_5	$\hat{Y}_{5,0}$	$\hat{Y}_{5,1}$	$\hat{Y}_{5,2}$
6	Y_6	1	X_6	$\hat{Y}_{6,0}$	$\hat{Y}_{6,1}$	$\hat{Y}_{6,2}$
7	Y_7	1	X_7	$\hat{Y}_{7,0}$	$\hat{Y}_{7,1}$	$\hat{Y}_{7,2}$
8	Y_8	2	X_8	$\hat{Y}_{8,0}$	$\hat{Y}_{8,1}$	$\hat{Y}_{8,2}$
9	Y_9	2	X_9	$\hat{Y}_{9,0}$	$\hat{Y}_{9,1}$	$\hat{Y}_{9,2}$
10	Y_{10}	2	X_{10}	$\hat{Y}_{10,0}$	$\hat{Y}_{10,1}$	$\hat{Y}_{10,2}$
11	100	0	X_{11}	$\hat{\mu}_t(0, X_{11}) = 100$	$\hat{\mu}_t(1, X_{11}) = 50$	$\hat{\mu}_t(2, X_{11}) = 30$

Training data

New decision to make

Prediction

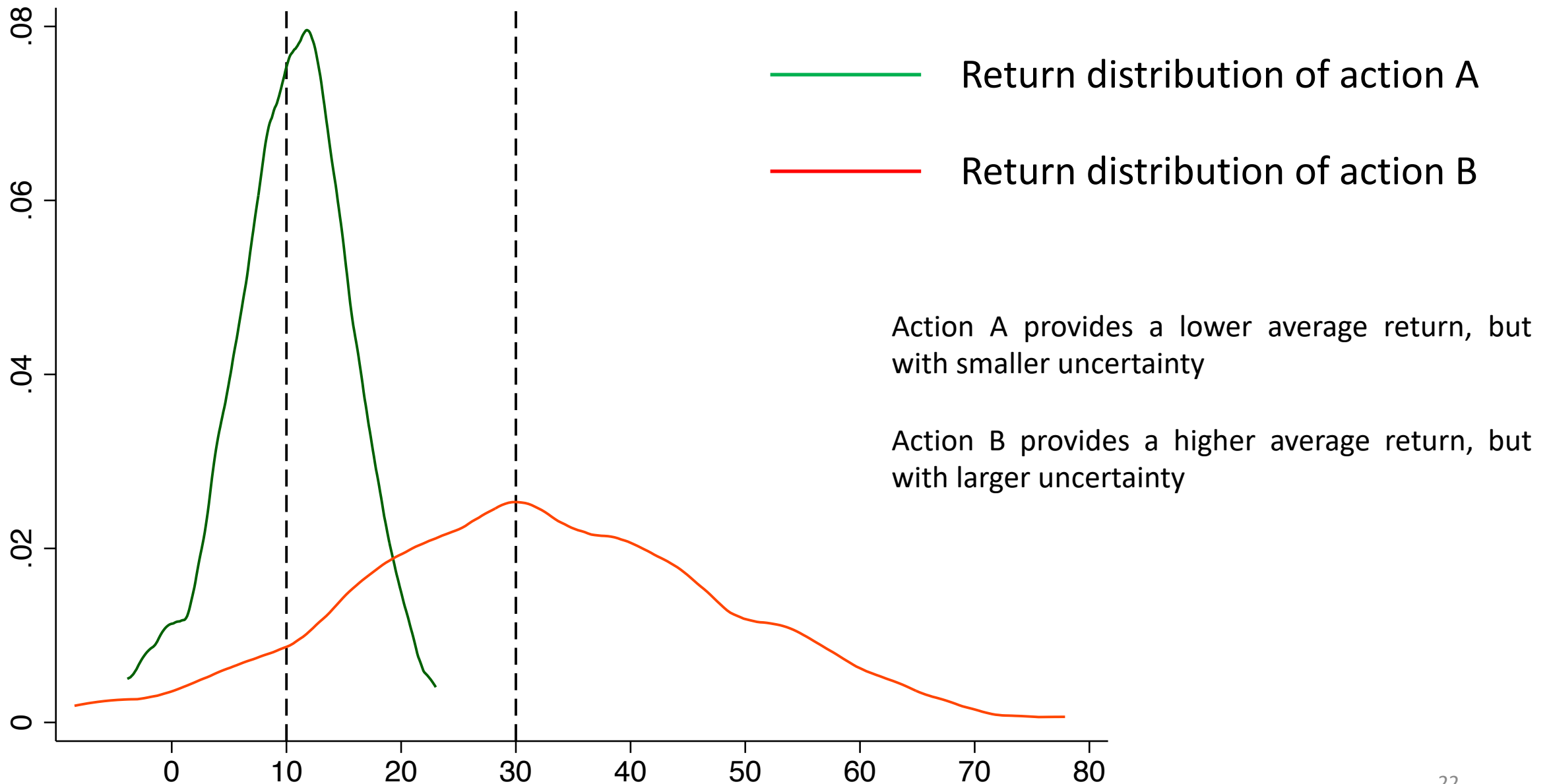
Best action is **0**

Uncertainty

Return and uncertainty - 1

- High returns can be associated to higher **uncertainty**
- Choosing action A instead of B, depends not only on the average returns of each option, but also on the uncertainty in getting such returns
- Decision making must ponder **return** and **uncertainty**

Return and uncertainty - 2



Measuring uncertainty: **variance**

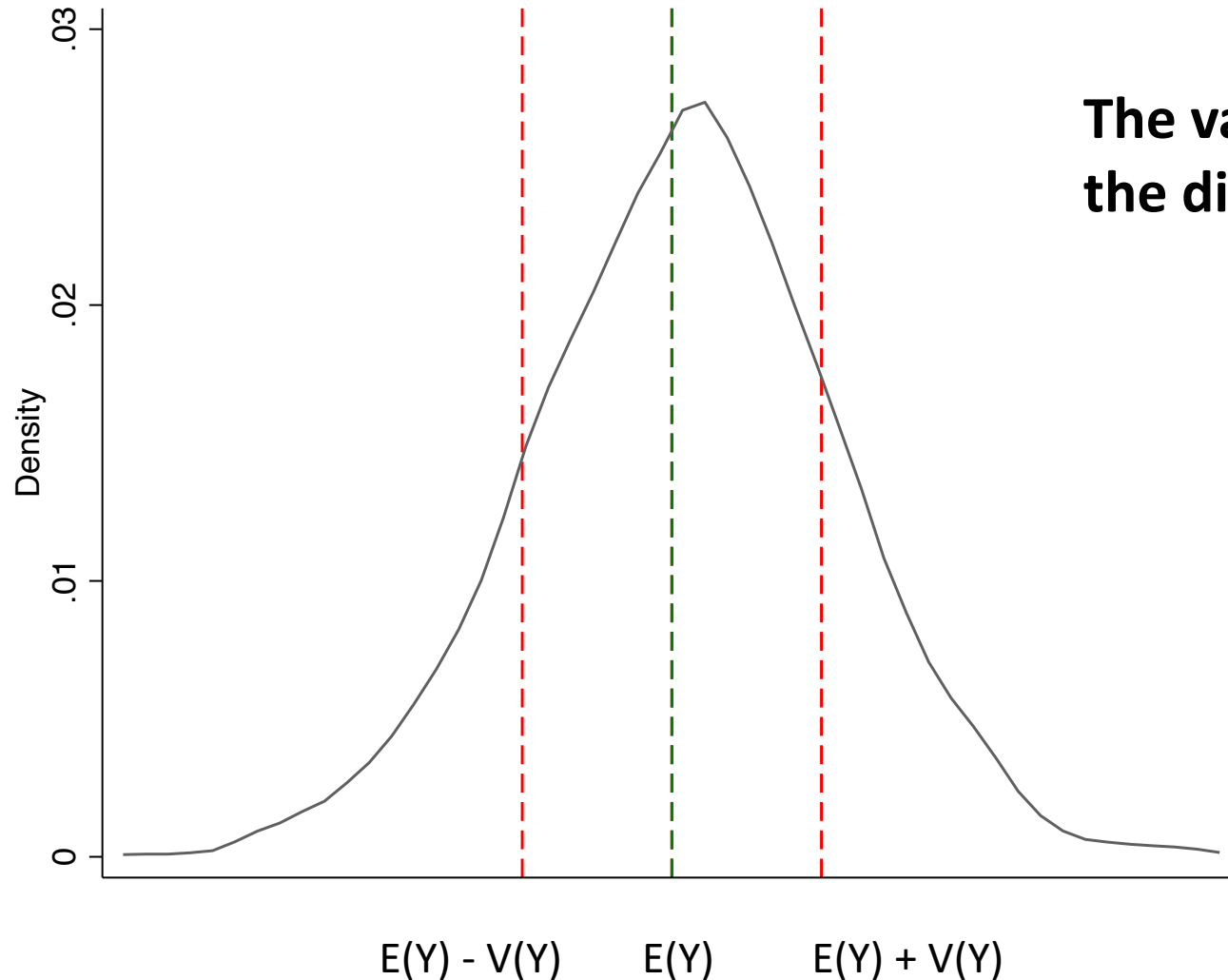
In statistics, **uncertainty** is generally measured via the **variance** of the distribution of Y . Thus, while the expected return of Y is $E(Y)$, its variance is $V(Y)$:

$$V(Y) = E[Y - E(Y)]^2 = E(Y^2) - \{E(Y)\}^2$$



This equality is precious as it simplifies variance's computation in many contexts

Variance as a measure of distribution **spread**



The variance is a measure of how **spread** is the distribution around its **central value**

Conditional variance - definition

Conditional **uncertainty** can be measured via the **conditional variance** of the distribution of $Y|X$. Thus, while the conditional expected return of $Y|X$ is $E(Y|X)$, its conditional variance is $V(Y|X)$:

$$V(Y|X) = E[Y - E(Y|X)]^2 = E(Y^2|X) - \{E(Y|X)\}^2$$

Conditional variance - estimation

The **estimation** of the **conditional variance** is rather simple:

$$V(Y | X) = E[Y - E(Y) | X]^2 = E(Y^2 | X) - \{E(Y | X)\}^2$$

Regression of Y^2 on X Regression of Y on X

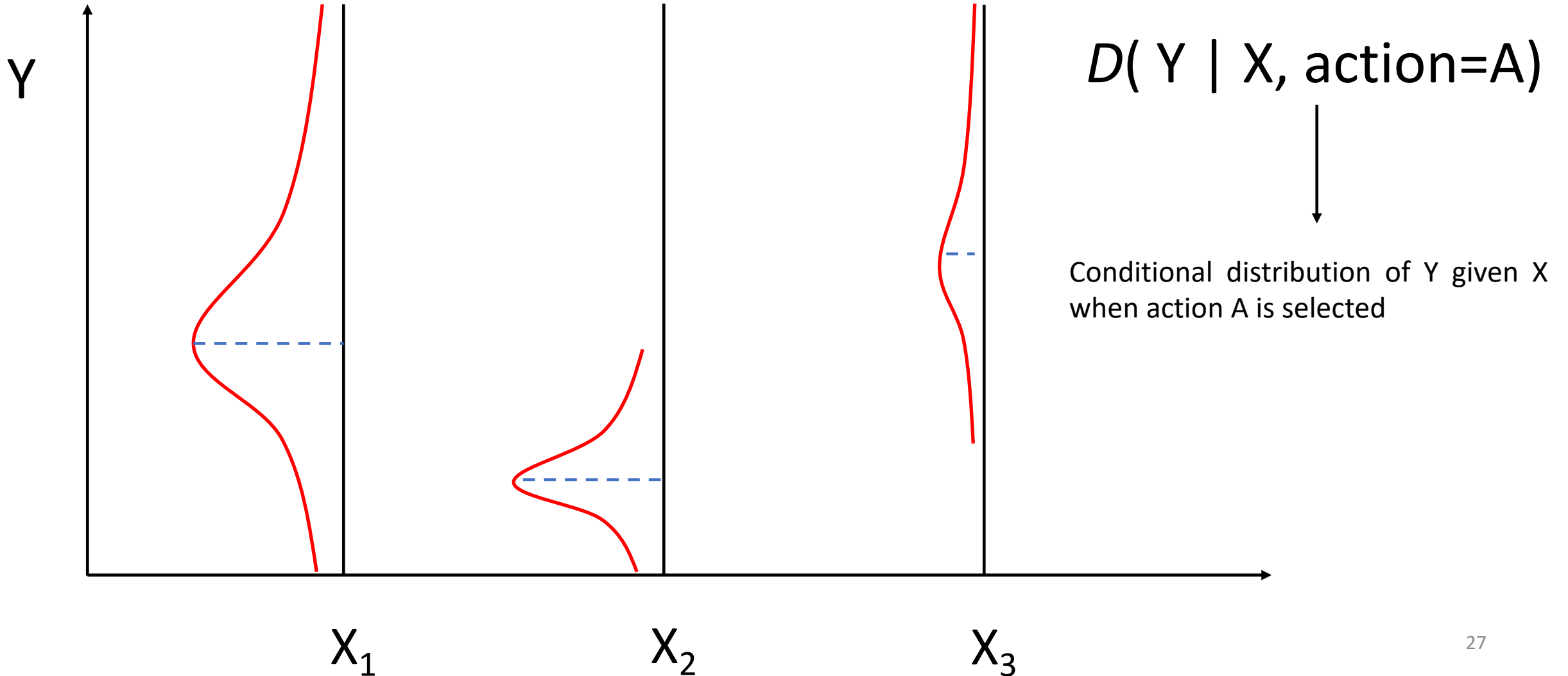
NOTE: whatever learner can be used to estimate these two regressions:
boosting, neural nets, random forests, etc.

Conditional return and conditional uncertainty

Medium return
Medium uncertainty

Low return
Low uncertainty

High return
High uncertainty



Conditional variance **by action**

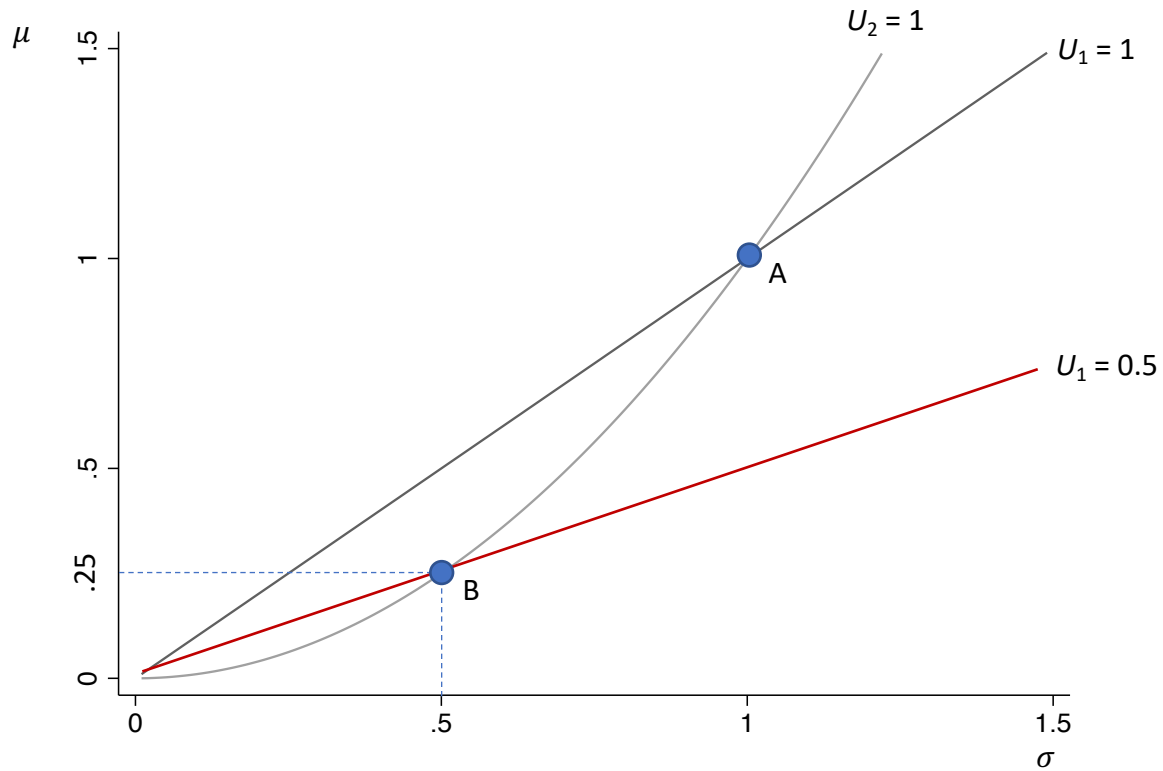
We define the conditional variance by action as:

$$\sigma^2_{iA} = v_A(X_i) = \text{Var}(Y \mid X_i, \text{if action} = A)$$

$$\sigma^2_{iB} = v_B(X_i) = \text{Var}(Y \mid X_i, \text{if action} = B)$$

$$\sigma^2_{iC} = v_C(X_i) = \text{Var}(Y \mid X_i, \text{if action} = C)$$

Different preferences over μ and σ



Utility functions:

$$U_1 = \frac{\mu}{\sigma}$$

$$U_2 = \frac{\mu}{\sigma^2}$$

Utility indifferent functions:

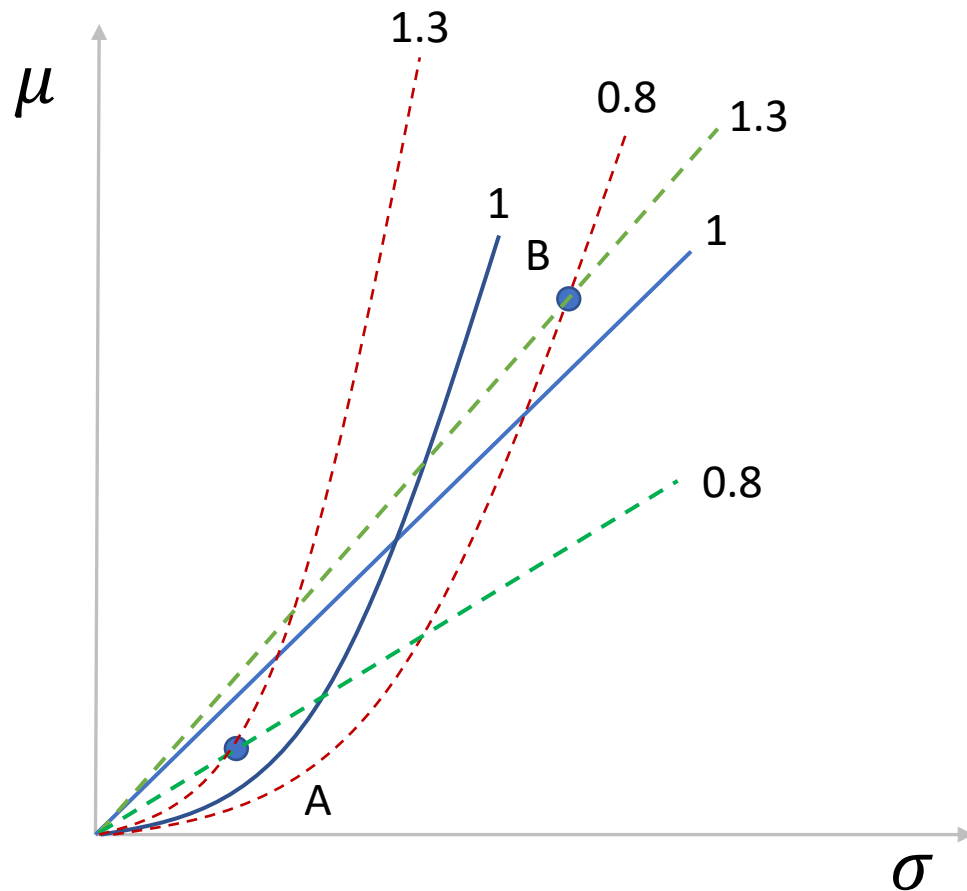
$$U_1 = \frac{\mu}{\sigma} = k \longrightarrow \mu = k \cdot \sigma$$

$$U_2 = \frac{\mu}{\sigma^2} = k \longrightarrow \mu = k \cdot \sigma^2$$

Action A \succ B under a linear indifferent curve (U_1)

Action A \sim B under a quadratic indifferent curve (i.e., U_2)

Preference inversion over μ and σ



According to $U_1 = \frac{\mu}{\sigma} \Rightarrow B \succ A$

According to $U_2 = \frac{\mu}{\sigma^2} \Rightarrow A \succ B$



This implies that we can have **preference inversion** depending on the attitude toward risk

Regret

Definition of REGRET

In policy learning, "**regret**" refers to the difference in performance between a **generic policy** and the **optimal policy** learnt from data.

$$R = \mathbb{E}[W(\pi^*)] - \mathbb{E}[W(\hat{\pi})]$$

Expected welfare of
the optimal policy



Expected welfare of
a generic policy



Regret estimation: RA

Regression adjustment (RA). This approach estimates the value function using regression estimates of the counterfactual (potential) outcomes. As such, it is also known as the *direct method*. The regression adjustment formula is:

$$\hat{V}_{RA}(\pi) = \frac{1}{N} \sum_{i=1}^N \hat{\mu}_i(\pi(\mathbf{x}_i), \mathbf{x}_i)$$

where $\hat{\mu}_i(\pi(\mathbf{x}_i), \mathbf{x}_i) = \sum_{j=0}^J \hat{\mu}_i(j, \mathbf{x}_i) \cdot \pi_{ij}$ with $\pi_{ij} = 1[\pi_i = j]$. The RA approach provides a consistent estimation of the value function provided that the functional form of the regression model is correct. If this is not the case, this approach can be highly biased.

Regret estimation: IPW

Inverse probability weighting (IPW). The formula of this estimator of the value-function is:

$$\hat{V}_{IPW}(\pi) = \frac{1}{N} \sum_{i=1}^N \frac{1[D_i = \pi(\mathbf{x}_i)]Y_i}{\hat{p}_{D_i}(\mathbf{x}_i)}$$

where $\hat{p}_{D_i}(\mathbf{x}_i)$ is an estimate of the propensity score. The *IPW* approach does not require an estimation of the mean potential outcomes; rather, it uses directly the values of the observed outcome variable Y . Unfortunately, this estimation method is biased when the propensity score functional form is misspecified.

Regret estimation: DR

Doubly-robust (DR). This estimator of the value-function, derived from the optimal influence function, takes on this formula:

$$\hat{V}_{DR}(\pi) = \frac{1}{N} \sum_{i=1}^N \left[\frac{[Y_i - \hat{\mu}_i(D_i, \mathbf{x}_i)] \cdot 1[D_i = \pi(\mathbf{x}_i)]}{\hat{p}_{D_i}(\mathbf{x}_i)} + \hat{\mu}_i(\pi(\mathbf{x}_i), \mathbf{x}_i) \right]$$

Unlike the *RA* and *IPW* approaches, the *DR* does not require for its consistency that both the propensity score and the conditional mean are simultaneously correctly specified. Only one out of the two must be correctly specified, with the other being potentially also misspecified.

Implementation

Two modes to learn the **optimal decision**:

- **Offline learning**

we learn the new optimal action by re-fit the model over the entire dataset

- **Online learning**

we learn the new optimal action incrementally, as new information gets in

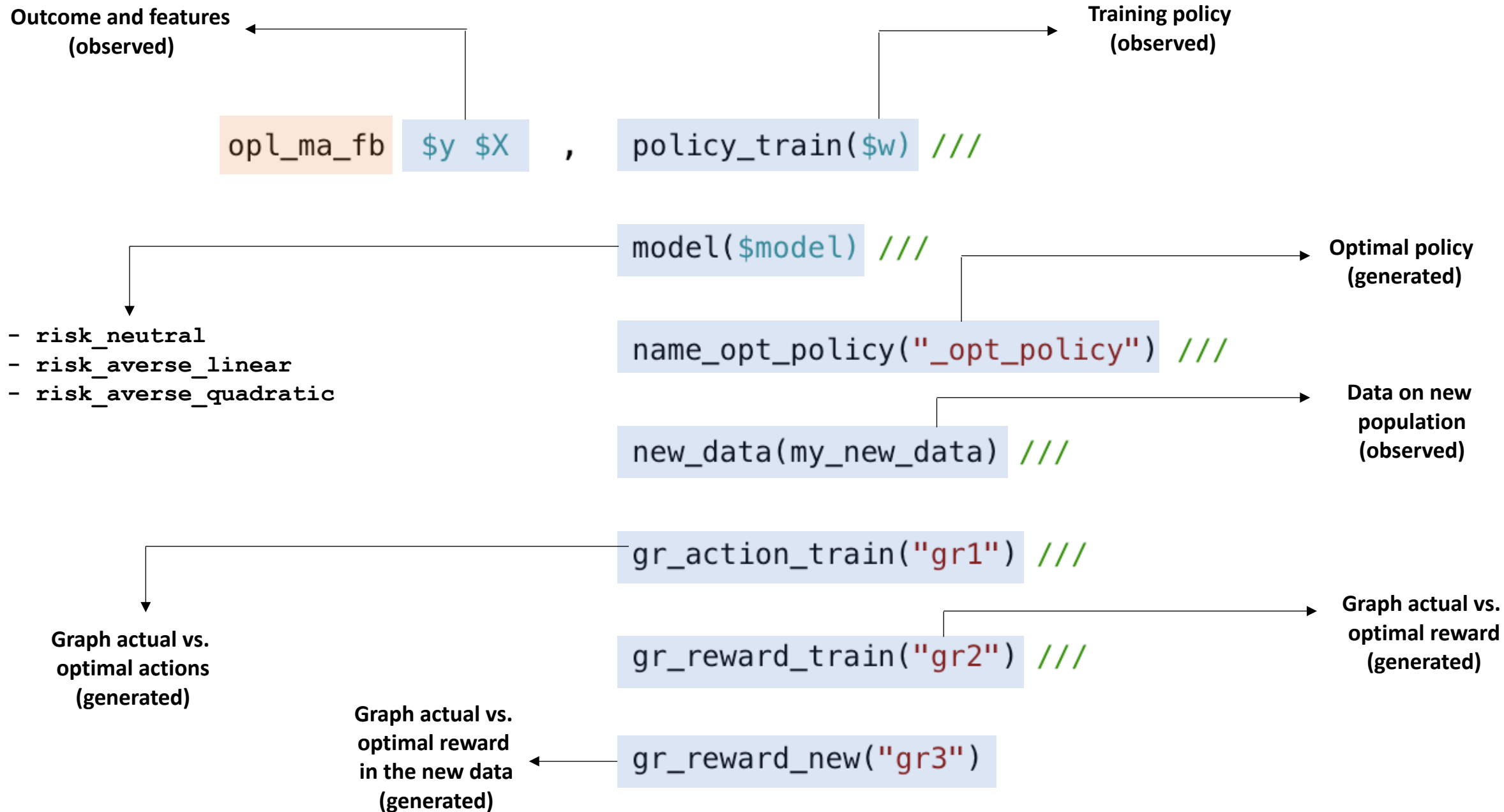
Software: Stata commands

Stata commands:

- **opl_ma_fb** : for estimating the *optimal policy*
- **opl_ma**: for estimating the *regret*

soon available on SSC.

Stata command: `opl_ma_fb`



Stata command: `opl_ma`

```
*****>
* REGRET ESTIMATION USING "opt_ma.ado"
*****>
* Value-function "first-best policy"
```

```
opl_ma $y $X , policy_train($w) policy_new(_opt_policy)
```

Training policy
(observed)

Optimal policy from `opl_ma_fb`
(generated)

```
gl EV_RA_opt=e(RA) // regression adjustment
gl EV_IPW_opt=e(IPW) // inverse probability weighting
gl EV_DR_opt=e(DR) // double robust
```

```
* Value-function "training policy"
cap drop _D* _pi*
opl_ma $y $X , policy_train($w) policy_new($w)
gl EV_RA_curr=e(RA)
gl EV_IPW_curr=e(IPW)
gl EV_DR_curr=e(DR)
```

```
* Regret estimation
global regret_RA=$EV_RA_opt-$EV_RA_curr
di in red "Regret RA = "$regret_RA
global regret_IPW=$EV_IPW_opt-$EV_IPW_curr
di in red "Regret IPW = "$regret_IPW
global regret_DR=$EV_DR_opt-$EV_DR_curr
di in red "Regret DR = "$regret_DR
*****>
* End
*****>
```


Application 1. Labour Policy

As an illustrative example, I utilize the well-known LaLonde (1986) dataset `jtrain2.dta`, which was employed by Dehejia and Wahba (1999) to assess various propensity-score matching methods in an ex-post policy evaluation. In their investigation, the authors aimed to estimate the impact of participating in a job training program administered in 1976 (indicated by the binary variable `train`, taking the value 1 for treated individuals and 0 for untreated) on real earnings in 1978 (variable `re78`) for a group of individuals in the United States. The dataset comprises a total of 445 observations, with 185 individuals treated and 260 untreated.

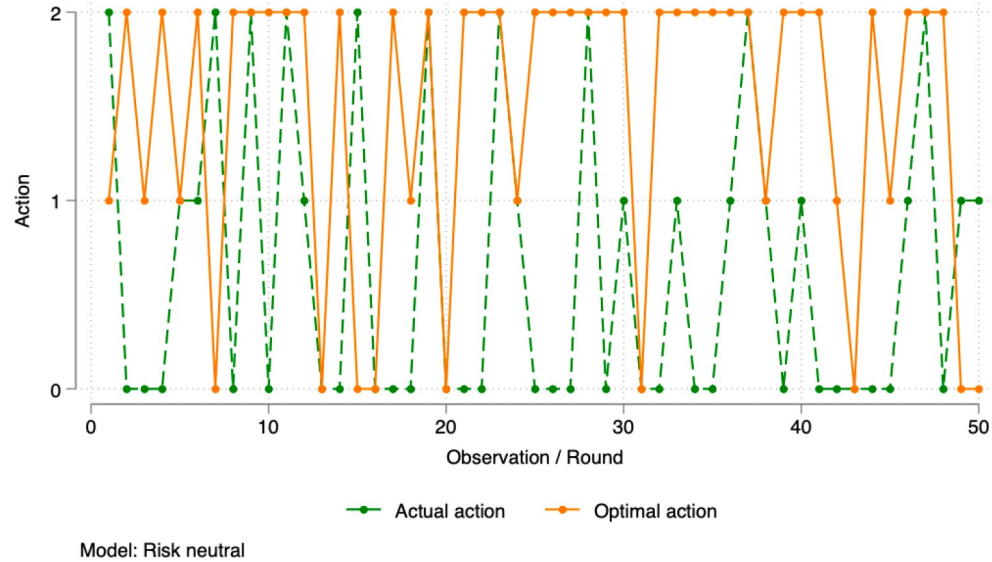
Setting

In our study, we designate the number of months of training (variable `mostrn`) as the treatment variable D , ranging from 0 to 24 months. The median for treated individuals is 21 months. Consequently, I construct a 3-arm set of actions:

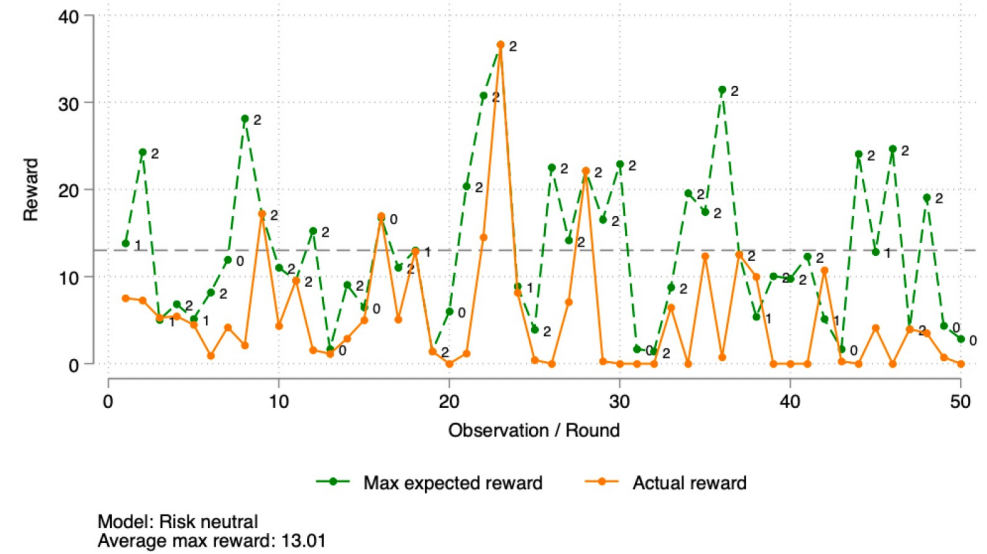
- Action 1: no training, $D = 0$, $N_0 = 260$;
- Action 2: training between 1 month and 21 month, $D = 1$, $N_1 = 107$;
- Action 3: training lasting from 22 to 24 months, $D = 2$, $N_2 = 78$;

Case 1. Risk-neutral setting

Actual vs. optimal action allocation
(Training dataset)



Actual vs. maximal expected reward
(Training dataset)

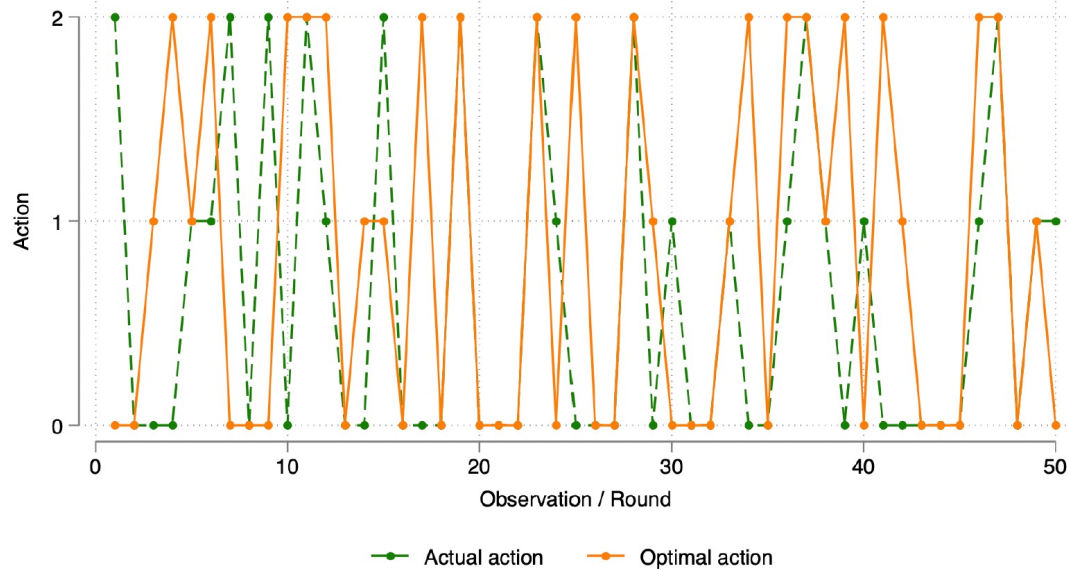


Variable	Obs	Mean	Std. dev.
_match	50	.3	.46291

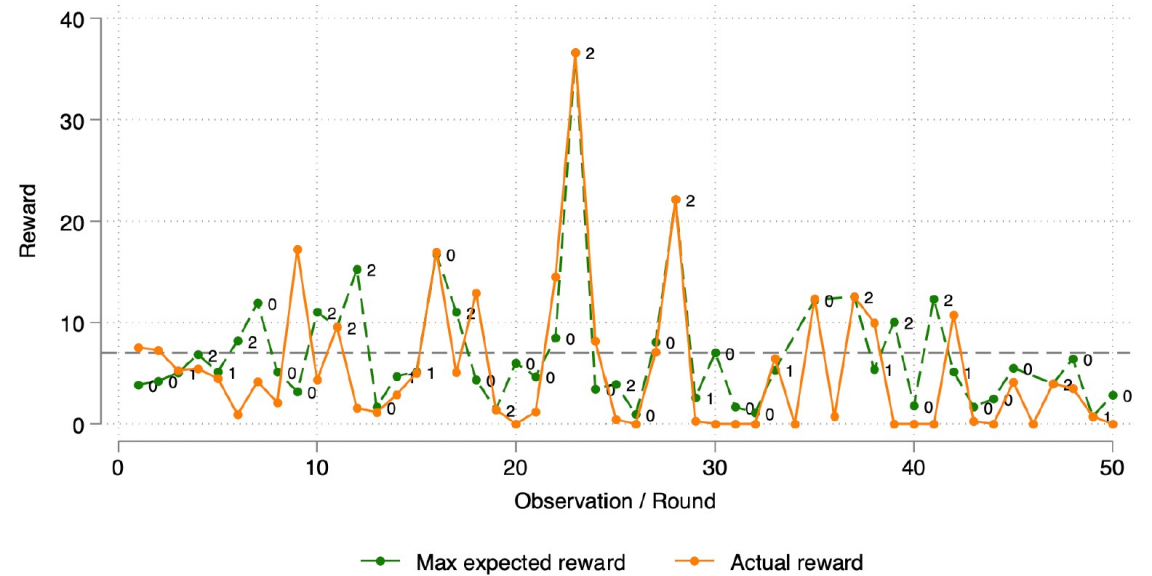
 Regret RA = 8.891423
 Regret IPW = 3.7557106
 Regret DR = 7.3346037

Case 2. Risk-adverse linear setting

Actual vs. optimal action allocation
(Training dataset)



Actual vs. maximal expected reward
(Training dataset)



Variable	Obs	Mean	Std. dev.
_match	50	.54	.5034574

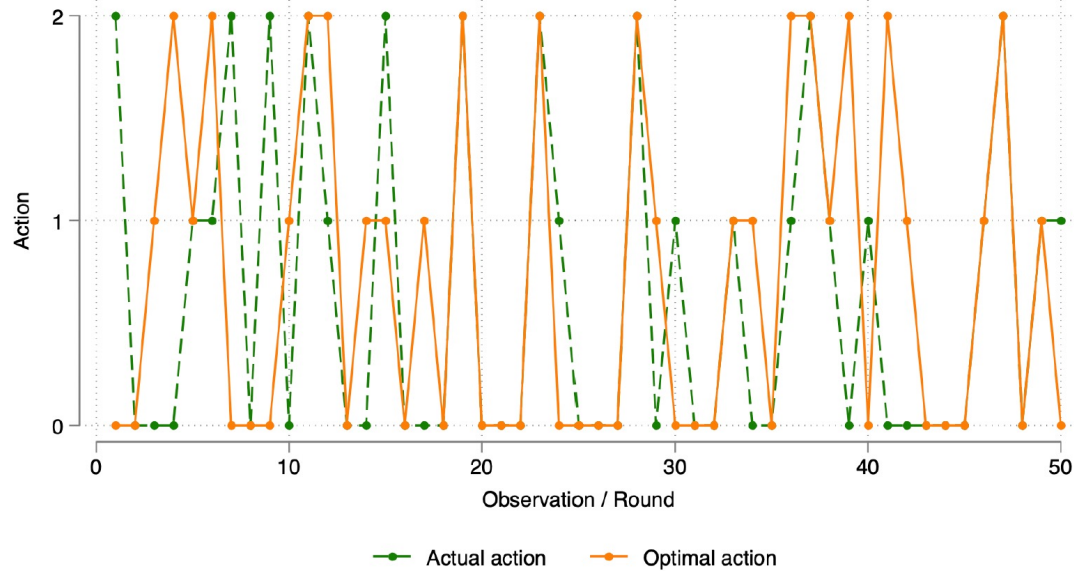
Regret RA = 3.4163201

Regret IPW = .55887842

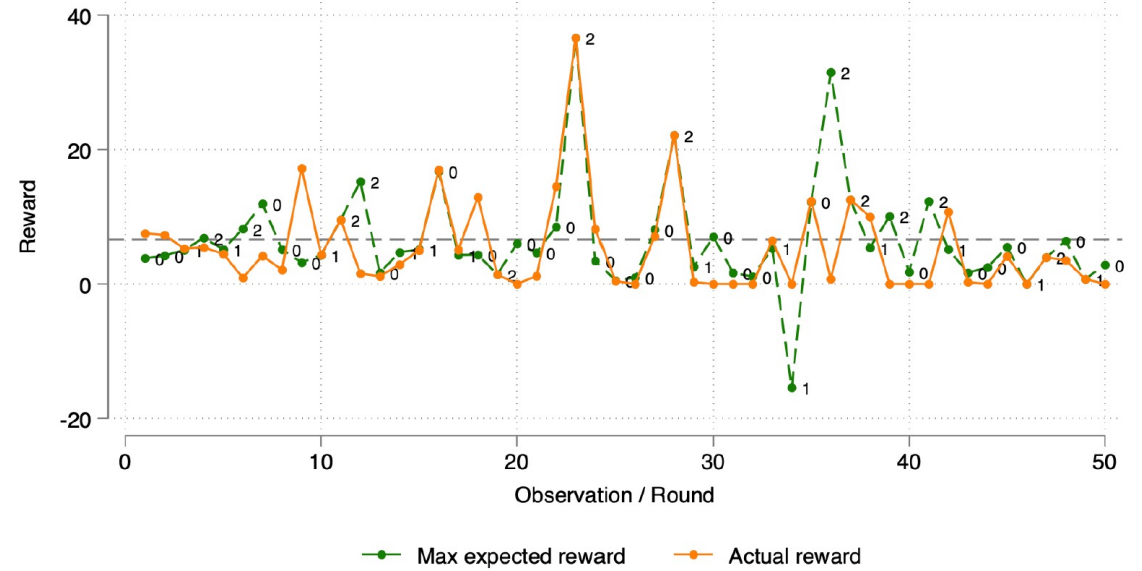
Regret DR = 2.5841078

Case 3. Risk-adverse quadratic setting

Actual vs. optimal action allocation
(Training dataset)



Actual vs. maximal expected reward
(Training dataset)



Variable	Obs	Mean	Std. dev.
_ match	45	.2	.4045199

Regret RA = -5.0857218

Regret IPW = .03672314

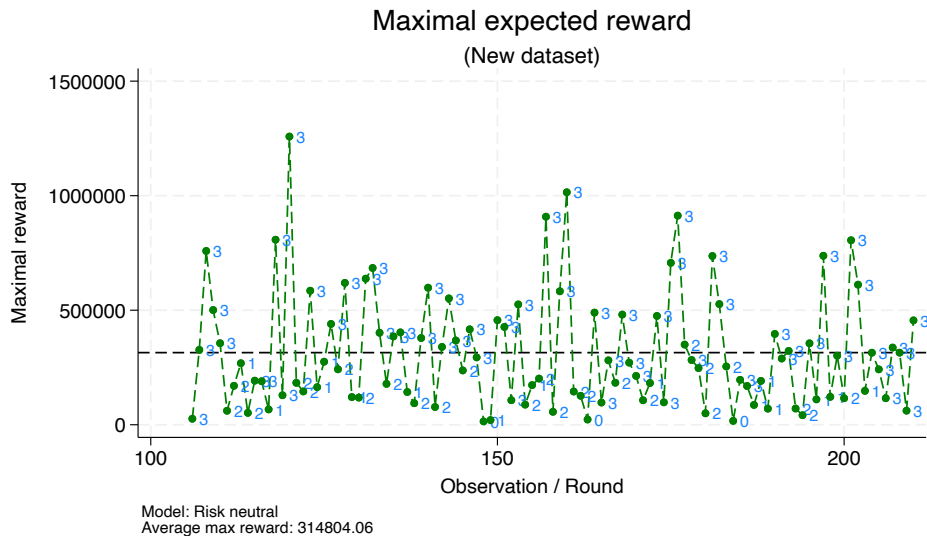
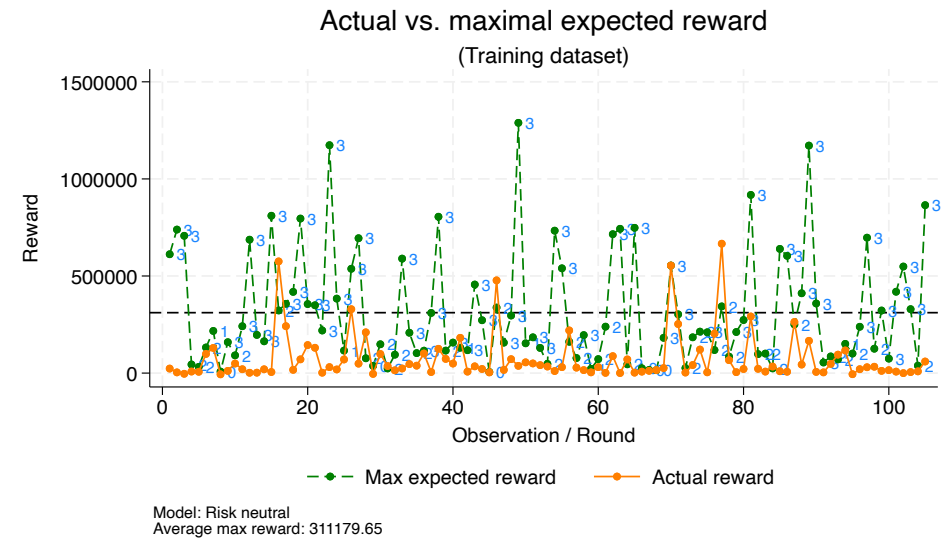
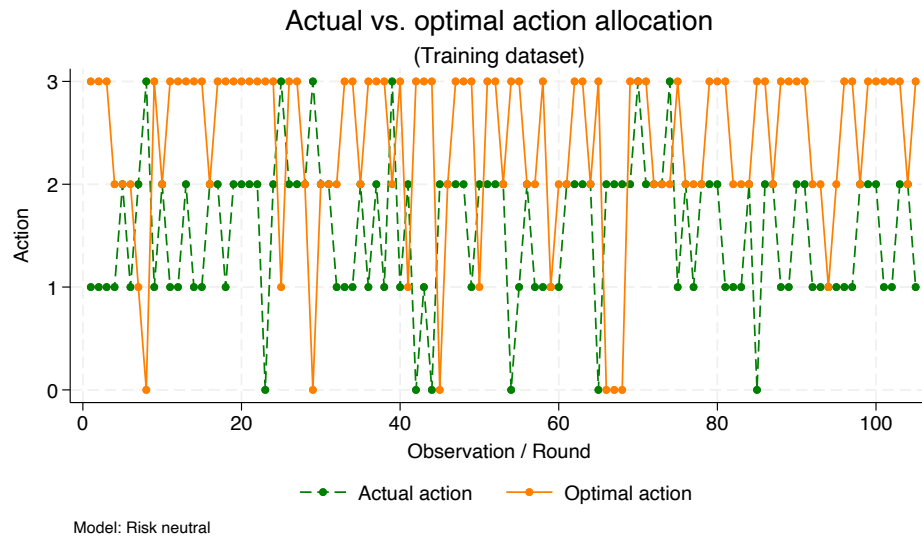
Regret DR = 1.0449446

Application 2. Agricultural Policy

Dataset: **Farm Accountancy Data Network (FADN)**. Collection of 140,788 observations from 31,813 unique agricultural holdings over a 12-year period from 2010 to 2022. We consider only year **2022**:

- **Data**: Subset of around 200 firms
- **Outcome**: Farm net-income
- **Treatment**: 0, 1, 2, 3 (i.e., 0 = no-treatment, 1 = direct payments, 2 = market enhancing measures, 3 = national and regional subsidies)
- **Features**: Farm characteristics, like: size, location, type of crop production, etc.

Results (risk neutral)



```
. sum _match if _index==0
```

Variable	Obs	Mean	Std. dev.	Min	Max
_match	105	.2190476	.4155847	0	1

Regret RA = 239577
Regret IPW = 62097
Regret DR = 219192

Percentage of farms with optimally allocated treatment

Conclusion

- OPL is based on *imitation learning*
- OPL implemented in **Stata** (soon in **Python** as well)
- OPL relies on assumptions A1 and A2
- Weak overlap (failure of A2) poses severe limitations
- Risk preferences are key for learning the optimal policy

Publication



Cornell University

We gratefully acknow
[mem](#)

arXiv > stat > arXiv:2403.20250

Search..

Help | Ad

Statistics > Machine Learning

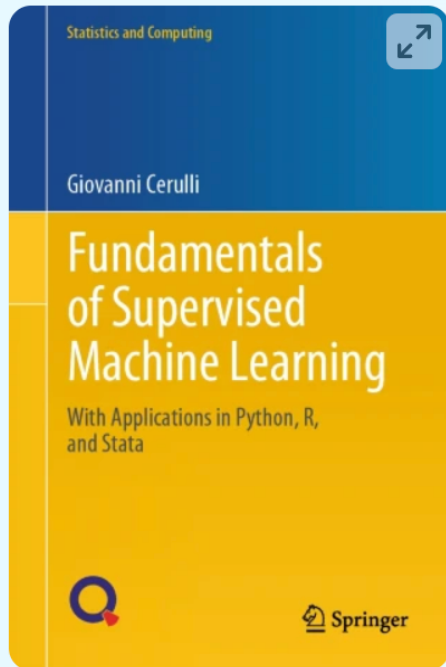
[Submitted on 29 Mar 2024]

Optimal Policy Learning with Observational Data in Multi-Action Scenarios: Estimation, Risk Preference, and Potential Failures

[Giovanni Cerulli](#)

This paper deals with optimal policy learning (OPL) with observational data, i.e. data-driven optimal decision-making, in multi-action (or multi-arm) settings, where a finite set of decision options is available. It is organized in three parts, where I discuss respectively: estimation, risk preference, and potential failures. The first part provides a brief review of the key approaches to estimating the reward (or value) function and optimal policy within this context of analysis. Here, I delineate the identification assumptions and statistical properties related to offline optimal policy learning estimators. In the second part, I delve into the analysis of decision risk. This analysis reveals that the optimal choice can be influenced by the decision maker's attitude towards risks, specifically in terms of the trade-off between reward conditional mean and conditional variance. Here, I present an application of the proposed model to real data, illustrating that the average regret of a policy with multi-valued treatment is contingent on the decision-maker's attitude towards risk. The third part of the paper discusses the limitations of optimal data-driven decision-making by highlighting conditions under which decision-making can falter. This aspect is linked to the failure of the two fundamental assumptions essential for identifying the optimal choice: (i) overlapping, and (ii) unconfoundedness. Some conclusions end the paper.

[Home](#) > [Textbook](#)



Fundamentals of Supervised Machine Learning

With Applications in Python, R, and Stata

Textbook | © 2023

✔ Access provided by Consiglio Nazionale delle Ricerche Direzione Centrale Servizi

[Download book PDF](#) ↓

[Download book EPUB](#) ↓