Gender Wage Gap in Spain

A case study

Contents

Case Study

INE - Encuesta Cuatrienal de Estructura Salarial

- Employed workers working in industry or in the construction or service sector, excluding domestic service.

- Analyzes hourly/monthly/annual salary w.r.t. demographic and employer data.
  - Gender, age, region, spanish nationality, education level.
  - Job field, public or private, market, unit size.
  - Occupation, weekly hours, seniority, whether the employee has people in their charge.

- Stratified sampling with weighted samples.
  - 216,726 sampled individuals.

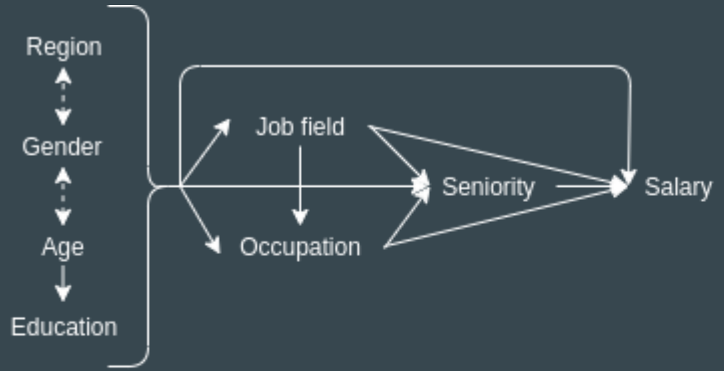
INE - Encuesta Cuatrienal de Estructura Salarial

- INE provides a report on this dataset analyzing the actual gender wage gap and providing possible explanations for this gap.
  - Job field segregation.
  - Full-time vs. half-time to balance family and work life.

- Some numbers (2018):
  - Annual wage gender gap (1 - female / male salaries) of 21.4%.
  - Hourly wage gender gap: full-time 6.7% vs. part-time 12.6%.
  - Hourly wage gender gap grouped by age range.

Causal Graph

- To analyze data in a causal manner, we need to define a causal graph.
  - Determine which variables affect others (seniority → salary).



- Bidirected dashed arrows represent latent confounding between these variables.
  - An unmeasurable variable that affects both of them, a latent confounder.

Do-calculus

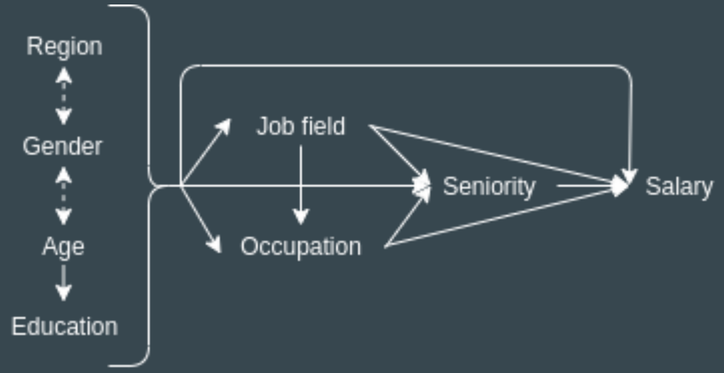
- Observational vs. Interventional data:
  - Observational = data obtained passively, without a randomized experiment.
  - Interventional = data obtained actively, through a randomized experiment.
    - Would our salary increase were we to get an undergraduate degree?
    - No matter what our “natural” education level would be, we *impose* (intervene) a certain value for that variable, and derive which other variables might be affected by this change.

Do-calculus

- Given this graph, we can compute causal queries using the rules of do-calculus.
- For example, what is the effect of gender (g) on salary (s)?
  - In this case, region (r) and age (a) constitute a *back-door admissible set*, which allow us to use the back-door adjustment formula:

$$p(s \mid do(g)) = \sum_{r,a} p(s \mid g, r, a) \cdot p(r, a)$$

- These terms can be derived with ML.
- If we can't find an estimand for a causal query, we say that the expression is **non-identifiable**.

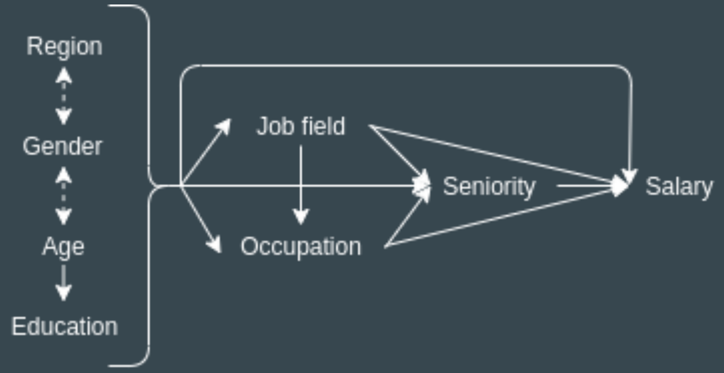


Do-calculus vs. sampling strategies

- Instead of using do-calculus, we can employ sampling-based strategies.
- Structural Causal Models (SCM):
  - This model defines each node in the graph as a function of its parents (+ a noise signal that provides stochasticity to the function).

$$education = f(age, \varepsilon_{education})$$

  - Computing causal queries in this graph amounts to sampling new values for all variables, except those that are intervened (their value is fixed).
    - Only if the causal query is identifiable!!! Were it not, two SCMs could give different results for the same causal query.



Do-calculus vs. sampling strategies

- Do-calculus pros:
  - Some non-identifiable queries can be answered by adding assumptions to the generative process, or by including data from randomized experiments.
- Do-calculus cons:
  - Some estimands derived from do-calculus might be intractable and require ad-hoc models.
  - Answering a specific causal query results in training a specific model.
    For any other causal queries, we would need to train a different model.
- SCMs avoid these two issues: we don't use the estimand, but simple sampling, and an SCM, once trained, can answer any (identifiable) causal query with it.

Causal Queries

- Let's answer some causal queries from this dataset.
  - What's the effect of gender on salary?
  - What's the effect of age, education and seniority on salary?
  - On average, were we to change the gender of a person, how much would their salary change?
  - What's the *direct* effect of gender on salary?

Causal Queries

- What's the effect of gender on (annual) salary? 5,627€ ± 153

$$\mathbb{E}[s \mid do(g = m)] - \mathbb{E}[s \mid do(g = f)]$$

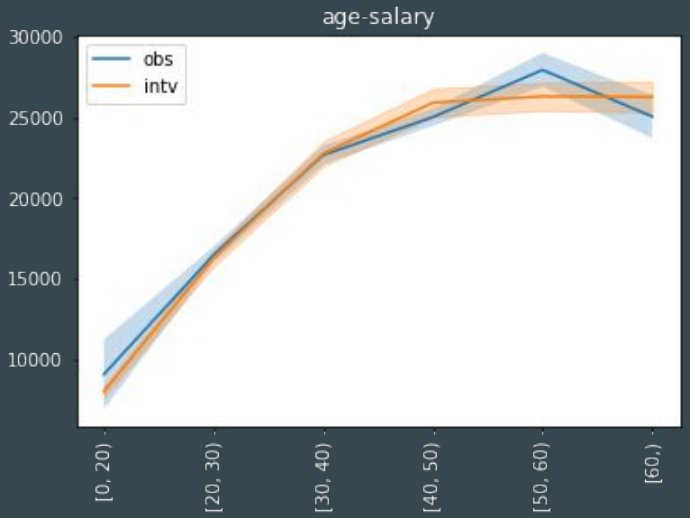
  - Note that this is not so different from the observational term: 5,771€.
  - This is because the effect of confounders is not so strong so as to bias the results of the observational term significantly.

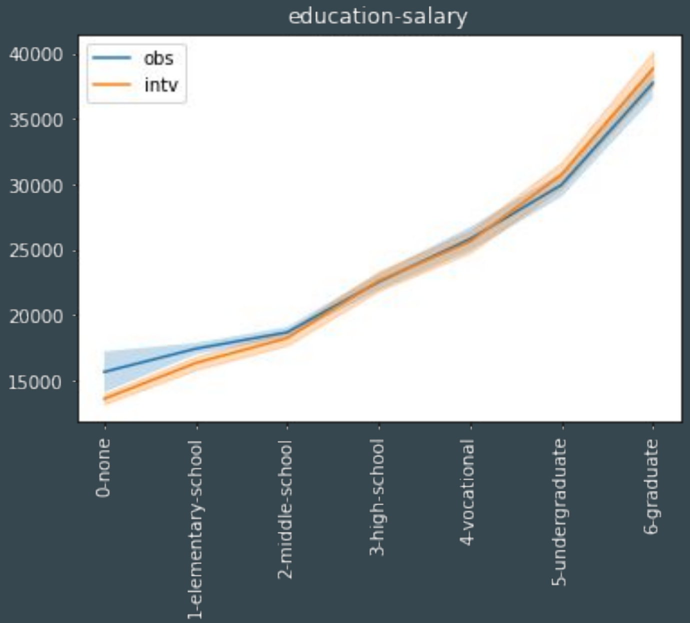
$$p(s \mid do(g)) = \sum_{r,a} p(s \mid g, r, a) \cdot p(r, a)$$

  - In other words, for this query (and for some others we will see next) there won't be significant differences between the interventional queries and their observational counterparts.

Causal Queries

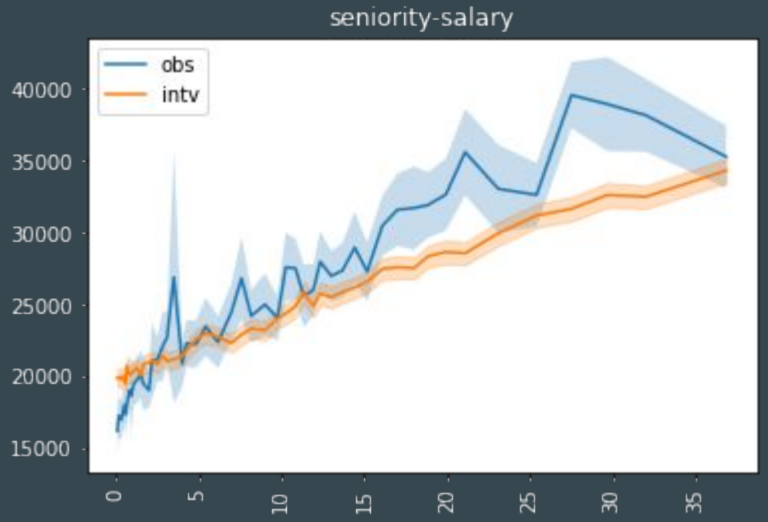
- What's the effect of age/education on salary?





Causal Queries

- What's the effect of seniority on salary?



  - In this case the difference between the observational curve and the intervention is significant.

Causal Queries

- On average, were we to change the gender of a person,
  how much would their salary change?
  - This is a different question, since we're interested, on an individual level, on how would an intervention on gender affect the outcome of that particular individual.
  - This is a **counterfactual** question, which can be answered by a 3-step process: abduction, intervention, prediction.
  - This is the salary of that particular person in a parallel world where their gender was swapped.
  - This can be used to **answer questions about particular individuals.**
  - If we average this expression for the whole population,
    we should get the effect of gender on salary, as before.
- The result is 5,536€ ± 132 (close to the interventional 5,627€ ± 153).

Causal Queries

- What's the *direct* effect of gender on salary?
  - The **direct effect** is the effect of the intervened variable when we remove any influence from the rest of the variables.
- In this case, the result is 4086€ ± 590.
- **Disclaimer:** when defining the causal graph, if we omitted any variables also affecting salary, were we to include them now, the direct effect could diminish as a result.
  - The direct effect should only be assessed when we are sure that we have included all of salary's ancestors in our graph.

Deep Causal Graphs

Deep Causal Unit

- How did we obtain the previous results? By using a sampling strategy called Deep Causal Graphs (DCGs).
- A DCG mimics a Structural Causal Model by learning each node's function with a Deep Neural Network.

  $$education = f(age, \varepsilon_{education})$$

- Each of these functions is defined by a Deep Causal Unit (DCU), which can be of any form, but must provide an implementation for three operation:
  - **Sample**: generate a sample from this node's distribution, given its parents' values.
  - **Likelihood**: compute the log-likelihood of a sample, given its parents' values.
    This is the model's training objective, and must be differentiable w.r.t. the model's parameters.
  - **Abduct**: derive the value for the noise signal (ε), given its value and its parents' values.

Deep Causal Unit

- With a DCU, we can:
    - **Sample**: use each DCU function in topological order to generate samples from the graph.
        - Intervened samples just use the intervened value for the intervened node, instead of the trained function. Then this value can propagate to its descendants.
    - **Likelihood**: compute the likelihood of a whole sample by computing each node's likelihoods.
    - **Counterfactual estimation**: we can use *abduct* to obtain the corresponding ε for each node, and then use these values to generate counterfactual samples using the 3-step process mentioned before.
- Most causal queries can be estimated using these three techniques.

Distributional Causal Nodes

- A possible DCU is the Distributional Causal Node (DCN).
- For this, we need to define which probability distribution this node follows, conditional on its parents.
- For example, a Gaussian DCN:
    - Sample: $f(pa_X, \epsilon_x) = \sigma(pa_X) \cdot \epsilon_x + \mu(pa_X),\ \epsilon_x \sim \mathcal{N}(0, 1)$
    - Likelihood: use the Gaussian's density, given its parameters μ and σ.
    - Abduct:

$$\epsilon_x = \frac{x - \mu(pa_X)}{\sigma(pa_X)}$$

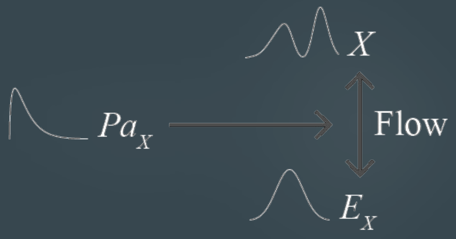
- μ and σ are computed with a NN that takes as input the node's parent values.

Distributional Causal Nodes

- Distributional Causal Nodes are too restrictive and cumbersome to use.
    - We're restricted to several known distributions that accomodate to the DCNs schema.
    - We need to explore every possible distribution for every continuous node until we find the one that fits.
- However, some distributions (Bernoulli, Categorical) work perfectly with DCNs.

Normalizing Causal Flows

- An alternative for DCUs are Normalizing Causal Flows (NCFs):
  - A Normalizing Flow transforms a random variable X, our node's distribution, into a signal E_X from a base distribution (a Gaussian, for example).
  - Using a **conditional** flow, we can model the distribution **conditional on its parents' values.**
- Note that with this technique we can perform the three DCU operations.
  - **Sample**: sample from the base distribution E_X and transform from E_X to X.
  - **Abduct**: transform from X to E_X.
  - **Likelihood**: use the flow's likelihood feature as usual.
- Any type of Conditional Normalizing Flow works as an NCF and can be used to model arbitrary continuous distributions.



Tips & Tricks

- Our case study was performed using DCGs:
  - Bernoulli and Categorical DCNs for discrete variables.
  - NCFs for continuous variables.
- Some tricks used to train this model:
  - Since the dataset includes a weight for each sample, we need to weight each sample's log-likelihood when training the model. This is done by simply computing the weighted average of the log-likelihoods.
  - All continuous distributions in this dataset are strictly non-negative. To improve results and satisfy this restriction, we can use a Softplus or an Exponential as the last layer of the flow, since they are diffeomorphisms compatible with the Flow structure.

Tips & Tricks

- Some tricks used to train this model:
  - When training small datasets, saving some training data for validation (for example, for early stopping) is quite wasteful. As an alternative, we can use **Cross Validation** to train K different models, and then use an **ensemble** of them for every node as the final graph.
    Note that a DCU can be assembled with several components by adjusting the sample/likelihood/abduct operations.
    - **Sample**: use a Categorical random sample to choose a component and sample with it.
    - **Likelihood**: get the likelihood for each component and compute the (weighted) average.
    - **Abduct**: sample from a Categorical to choose a component, and then use Importance Sampling to weight the resulting abducted sample.

Thank you!