Regression, causality, statistical paradoxes and other fairy tales

Javier González

November 17, 2020

Microsoft Research Cambridge

"I checked it very thoroughly, said the computer, and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Douglas Adams, The Hitchhiker's Guide to the Galaxy (1979)

Old friends, new friends

Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Old friends, new friends

Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Bayesian linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \;\; \beta \sim \mathcal{N}(0, \Sigma_p)$$

Old friends, new friends

Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Bayesian linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \;\; \beta \sim \mathcal{N}(0, \Sigma_p)$$

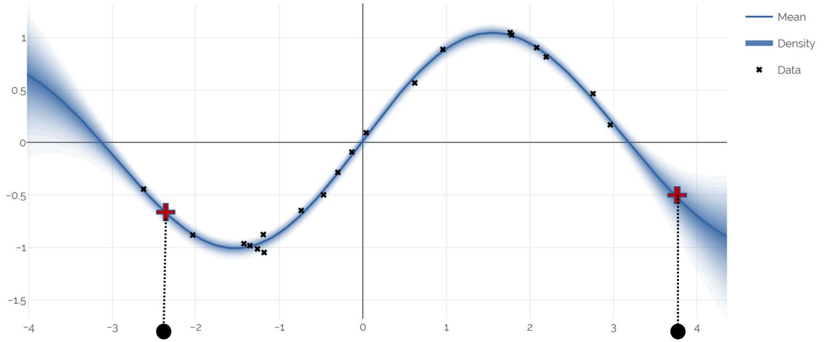
Bayesian non-linear regression (Gaussian process):

$$Y = f(X_1, \dots, X_p) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad f \sim \mathcal{GP}(0, K)$$

$$Y = \sum_k^{d_F} w_k \phi_k(X_1, \dots, X_p) + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \;\; w \sim \mathcal{N}(0, \Sigma_{d_F})$$

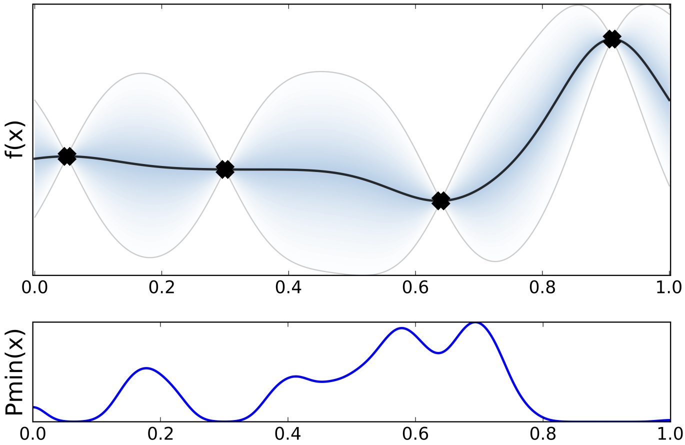
What can I do with a regression model?

1. I can make a predictions:



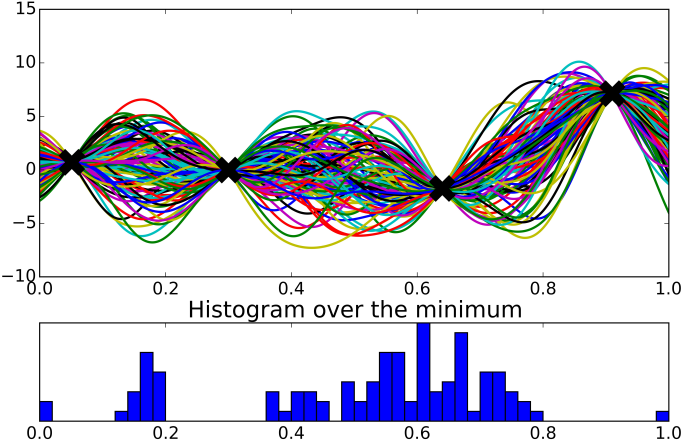
What can I do with a regression model?

2. I can learn about about a latent property of $f(x)$.



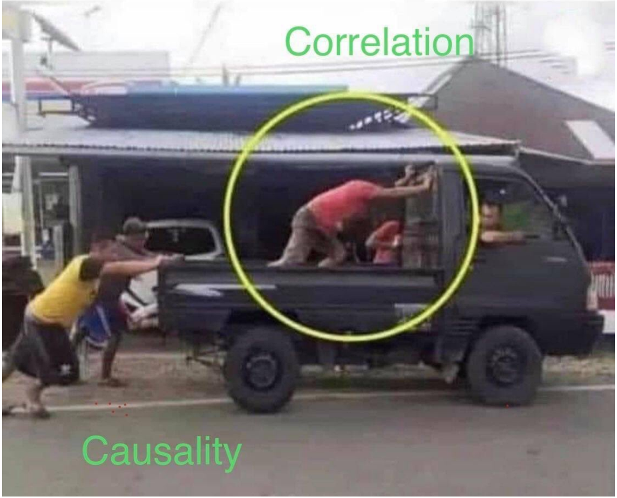
What can I do with a regression model?

2. I can learn about about a property of $f(x)$.



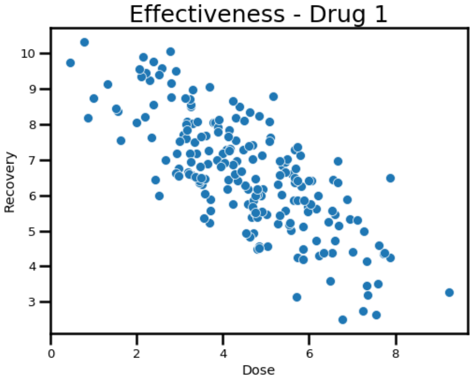
What can I do with a regression model?

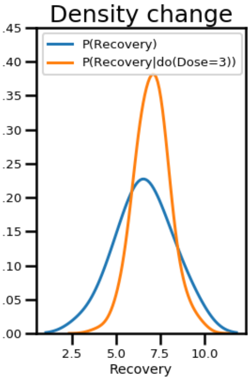
3. I can estimate a causal effect:



Ok, but what is exactly a causal effect?

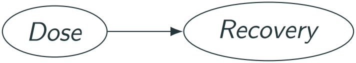
T causally affects Y if **intervening** on T changes the distribution of Y.





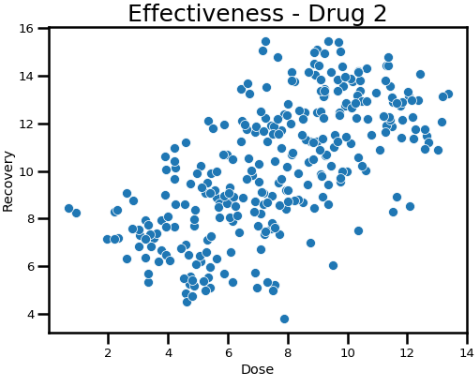
Ok, but what is exactly a causal effect?

$$\mathbb{P}(\mathit{recovery}) \neq \mathbb{P}(\mathit{Recovery}|\mathit{do}(\mathit{Dose} = 3))$$



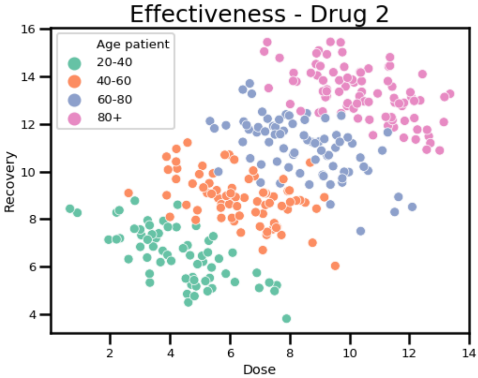
- A causal effect IS a ‘physical’ mechanisms.
- A causal effect IS NOT a property of the data.
- Intervening = experiment (change the laws of physics).
- *do* notation to represent an experiment.
- In general $\mathbb{P}(Y|\mathit{do}(T = t)) \neq \mathbb{P}(Y|T = t)$

Another example - drug 2



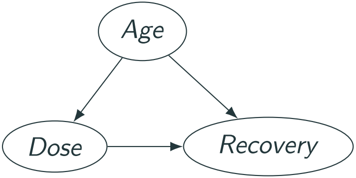
Increasing the dose in drug 2 seems to make patients to spend more time at the hospital (!!).

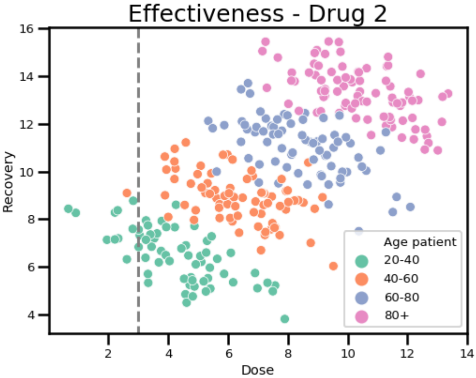
Days of recovery vs Dose - drug 2

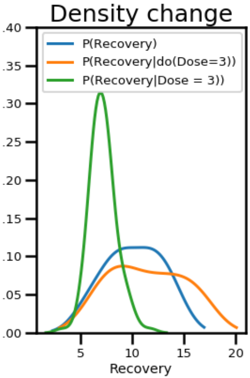


Age is a **confounder**. The drug is effective but older people suffer the disease more severely and require a larger dose.

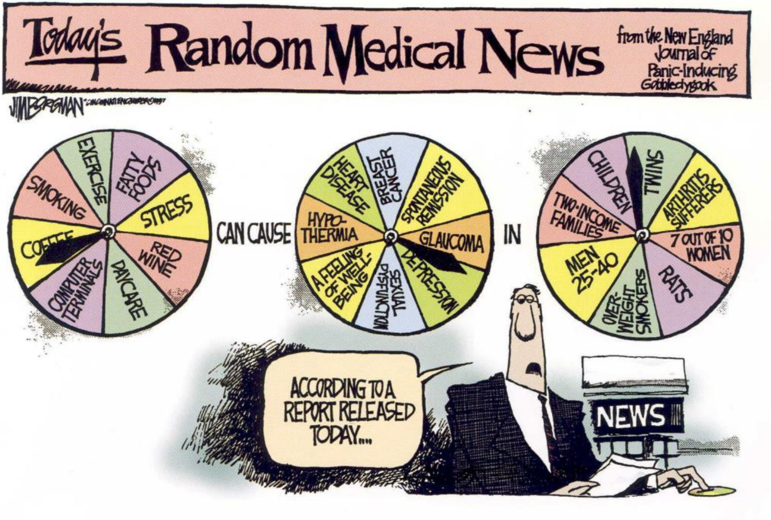
Days of recovery vs. Dose - drug 2







Correlation is not causation ... but is very easy to forget!



Source: Borgman, J (1997). The Cincinnati Enquirer. King Features Syndicate.

Simpson's paradox

'A trend that appears in several different groups of data may disappear or reverse when these groups are combined.'

Example: Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	Treatment B
78% (273/350)	**83% (289/350)**

Which treatment is better?

Example: Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Treatment B

Example: Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Treatment B

Ok, wait, are we sure? let's have a look to the data again....

Confounders

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**

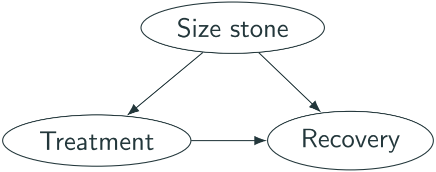
The size of the stone is a confounder.

Confounders

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**

The size of the stone is a confounder.



Solution

Weighting the effect of each treatment by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

Solution

Weighting the effect of each treatment by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

$$\begin{aligned}\mathbb{P}(Recover|do(T=A)) &= \mathbb{P}(small)\mathbb{P}(Recover|small, A) \\ &+ \mathbb{P}(big)\mathbb{P}(Recover|big, A) \\ &= \mathbf{0.8325}\end{aligned}$$

Solution

Weighting the effect of each treatment by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

$$\begin{aligned}
\mathbb{P}(Recover|do(T=A)) &= \mathbb{P}(small)\mathbb{P}(Recover|small, A) \\
&+ \mathbb{P}(big)\mathbb{P}(Recover|big, A) \\
&= \mathbf{0.8325} \\
\mathbb{P}(Recover|do(T=B)) &= \mathbb{P}(small)\mathbb{P}(Recover|small, B) \\
&+ \mathbb{P}(big)\mathbb{P}(Recover|big, B) \\
&= \mathbf{0.7788}
\end{aligned}$$

Treatment A is indeed better.

How to remove the effect of confounders

General adjustment formula

If Z is a **admissible adjustment set (confounders)** then:

$$\mathbb{P}(Y|do(T=t)) = \sum_{z} \mathbb{P}(Y|T=t, Z=z)\mathbb{P}(Z=z)$$

$$\mathbb{P}(Y|do(T=t)) = \int \mathbb{P}(Y|T=t, Z=z)\mathbb{P}(Z=z)dz$$

- Causal effects with observational data! No experiments!
- We only need to control by Z, nothing else.
- Knowing and observing all elements in Z is very hard.
- Adjusting by variables not in Z can be a terrible idea...

Berkson's paradox

'Two independent events A and B may become dependent when conditioning on a common effect (collider)'.

Berkson's paradox

We know that the is no causal effect between the two diseases:

$$\mathbb{P}(Bone|do(Respiratory = Yes)) = \mathbb{P}(Bone)$$

	General population		
	Bone disease		
Respiratory disease	Yes	No	%Yes
Yes	17	207	**8.4%**
No	184	2376	**7.7%**

Berkson's paradox

	General population			Hospitalizations last 6 months		
	Bone disease			Bone disease		
Respiratory disease	Yes	No	%Yes	Yes	No	%Yes
Yes	17	207	**7.6%**	5	15	**25%**
No	184	2376	**7.2%**	18	219	**7.6%**

Berkson's paradox

- The respiratory and bone diseases are independent.
- But they are conditionally dependent given hospitalization.

Adjusting by hospitalization is wrong!

$$\mathbb{P}(Bone|do(Re. = Yes)) = \mathbb{P}(Bone) \neq \int \mathbb{P}(Bone|Re. = Yes, Hos.)\mathbb{P}(Hos.)$$

Estimating an causal effect

Case 1: I can run experiments. EASY.

- Intervene in the world and check.

Case 2: I cannot run experiments. HARD.

- What is the causal relationship of interest?
- What experiment could capture the causal effect of interest?
- What is your identification strategy (confounders)?
- What is your mode of statistical inference (model)?

From regression to causation: average treatment effect

T: Treatment

Z: Confounders

Y: Response

Let's compute $ATE(t_1, t_2) := \mathbb{E}[Y|do(T = t_1)] - \mathbb{E}[Y|do(T = t_2)]$.

Step 1: Identification.

Find and observe all confounders Z or substitute confounders.

From regression to causation: average treatment effect

Step 2: Estimation.

Build a model that predicts the response Y using T, Z.

Linear regression: $\mathbb{E}[Y|T,Z] = w_0 + \tau T + wZ$

Gaussian process: $\mathbb{E}[Y|T,Z] = m(T,Z)$

where $m(\cdot)$ is the posterior mean of a Gaussian process.

From regression to causation: average treatment effect

Step 3: Marginalization

$$\text{Approximate } \mathbb{E}_Z[\mathbb{E}[Y|T = t_1, Z]] - \mathbb{E}_Z[\mathbb{E}[Y|T = t_2, Z]]$$

For a sample $\{t_i, z_i, t_i\}_{i=1}^n$ compute

$$A\hat{T}E(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} m(T = t_1, Z = z_i) - \frac{1}{n}\sum_{i=1}^{n} m(T = t_2, Z = z_i)$$

Fun fact

If you are using a linear regression model where

$$\mathbb{E}[Y|T,Z] = w_0 + \tau T + wZ$$

then:

- $\mathbb{E}[Y|do(T = t_1)] = \tau t_1$
- $\frac{\partial \mathbb{E}[Y|do(T=t)]}{\partial t} = \tau$

Linear models are pretty useful to compute causal effects!

Summary

Cool, isn't it? Now we can:

- Emulate experiments without experimentation.
- Learning how the world works, not just describing it.
- We can do all this with a Gaussian processes! ;-).

Ok, not it is time for some fairy tales...

Statistical fairy tale 1

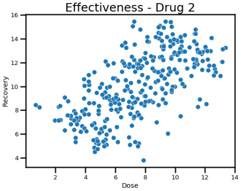
'To estimate an effect all I need is more data points'

Statistical fairy tale 1

'To estimate an effect all I need is more data points'

False!!

Identification and estimation are orthogonal steps.



Statistical fairy tale 2

'To estimate an effect it is fine if I just add all the observed variables to the model'

Statistical fairy tale 2

'To estimate an effect it is fine if I just add all the observed variables to the model'

False!!

Using colliders as confounders may introduce dependencies where they don't exist.

Statistical fairy tale 3

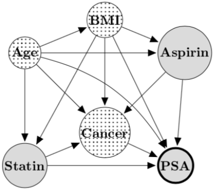
'I can do hypotesis-free causal inference'

Statistical fairy tale 3

'I can do hypotesis-free causal inference'

False!!

Causal inference ALWAYS involve making causal (and modelling) assumptions. These can be made explicit using causal graphs.



Statistical fairy tale 4

'All the validation I need to do, I can do it with my dataset.

False!!

It is usually VERY hard to know if there are unobserved confounders. In those cases, external validation is needed (an experiment).

Unknown unknowns

Questions?