From the stars to "Poor Law Statistics"

- Almost a century after Gauss
- Scientists correlating/regressing anything
- Problem: what does it mean?

e.g. Francis Galton correlated numeric traits between generations of organisms...

But *why*? "Nature versus nurture" debate (still unresolved?)

e.g. Udny Yule and others correlated poverty ("pauperism") with welfare ("out-relief")...

But *why*? "Welfare trap" debate (still unresolved?)

Origin of multiple regression

- Udny Yule (1871-1951)
- Studied this poverty question
- First paper using multiple regression in 1897
- Association between poverty and welfare while "controlling for" age

Yule, in 1897:

> Instead of speaking of "causal relation," ... we will use the terms "correlation," ...

- Variables, roughly:
  - $Y =$ prevalence of poverty
  - $X_1 =$ generosity of welfare policy
  - $X_2 =$ age
- Positive correlations:
  - $\text{cor}(Y, X_1) > 0$
  - $\text{cor}(X_2, X_1) > 0$

Do more people enter/stay in poverty if welfare is more generous?

Or is this association "due to" age?

Yule, in 1897:

> The investigation of **causal relations** between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions.
>
> Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and *discuss as best he can the relations of a whole group of changes*; he **cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics**.

[We] cannot [...] narrow down the issue to the effect of one variation at a time

but... isn't this how *almost everyone* interprets regression coefficients?...

(yes! and they are wrong!!!!)

the next slide is about some common mistakes people make when interpreting regression coefficients

(don't try to memorize the formulas)

Interpreting regression coefficients

People *want* these things to be true:

- "The linear **model** and our **estimates** are both good"

$$\frac{\partial}{\partial x_j}\mathbb{E}[\mathbf{y}|\mathbf{X}] = \beta_j \approx \hat{\beta}_j$$

- "We can interpret β_j as a causal parameter," i.e. **intervening** to increase x_j by 1 unit would result in conditional average of y changing by β_j units

$$\text{If } (x_j \mapsto x_j + 1) \text{ then } (\mathbb{E}[y] \mapsto \mathbb{E}[\mathbf{y}|\mathbf{X}] + \hat{\beta}_j)$$

But this *almost never* works!

Many textbooks tell us something like:

> "The coefficient $\hat{\beta}_j$ estimates the relationship between the (conditional mean of the) outcome variable and x_j *while holding all other predictors constant*"

i.e. "**ceteris paribus**" or "other things equal" (unchanged)

Fundamental problem of interpreting regression coefficients:

"holding all other predictors constant" is (almost) never applicable in the real world, i.e. ceteris is (almost) never paribus

Reasons we'll highlight today: **causality** and **nonlinearity**

Interpreting causality

Back to Yule. What does $\hat{\beta}_{\text{welfare}}$ mean?

```
lm(poverty ~ welfare + age) |> broom::tidy() |> knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.009	0.046	-0.19	0.849
welfare	0.491	0.015	31.97	0.000
age	0.267	0.083	3.21	0.001

```
lm(welfare ~ poverty + age) |> broom::tidy() |> knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.017	0.067	-0.262	0.794
poverty	1.032	0.032	31.973	0.000
age	0.484	0.120	4.027	0.000

Are these associations "causal"?

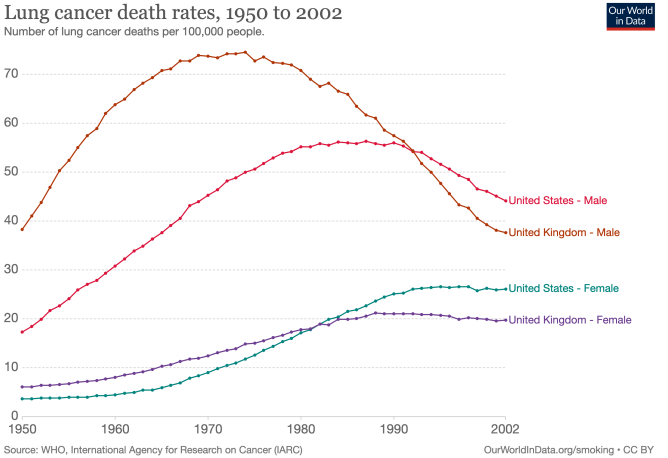
Yule found a positive association between `welfare` and `poverty` after "controlling for" `age`

Which is the cause and which is the effect?

Both? Neither?

Another important historic example

Smoking and lung cancer

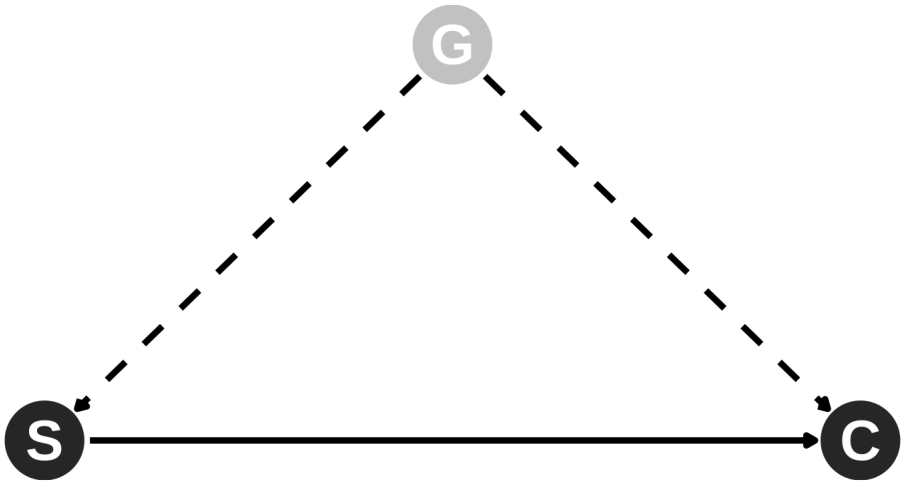


(don't smoke)

R. A. Fisher on smoking and lung cancer (in 1957)

> ... the B.B.C. gave me the opportunity of putting forward examples of the two classes of alternative theories which **any statistical association, observed without the predictions of a definite experiment**, allows--namely, (1) that the supposed effect **is really the cause**, or in this case that incipient cancer, or a pre-cancerous condition with chronic inflammation, is a factor in inducing the smoking of cigarettes, or (2) that cigarette smoking and lung cancer, though not mutually causative, are **both influenced by a common cause**, in this case the individual genotype ...

Graphical notation for causality

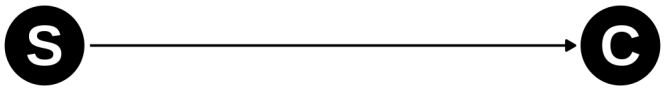


Variables: vertices (or nodes)

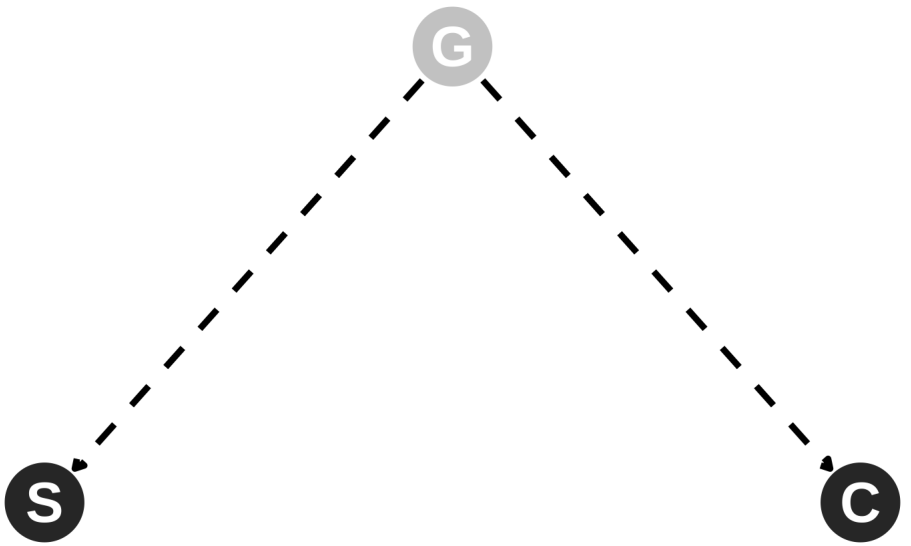
Relationships: directed edges (arrows)

Shaded node / dashed edges: unobserved variable

Smoking causes cancer?



Genotype is a common cause?



Fisher: association is not causation

(He did not use graphical notation like this)

Idea: adjusting for confounders

Confounders: other variables that obscure the (causal) relationship from X to Y, e.g.

- Y: health outcome
- X: treatment dose
- Z: disease severity

Without considering Z, it might seem like larger doses of X correlate with worse health outcomes

Solution: add more variables to the model

Including (measured) confounders in the regression model may give us a more accurate estimate

(My conjecture: Fisher used genes as his example confounder because, in his day, they could not be measured, so his theory would be harder to disprove)

Confounder adjustment is why some people think **multiple** regression is One Weird Trick that lets us make causal conclusions

(Statisticians Don't Want You To Know!)

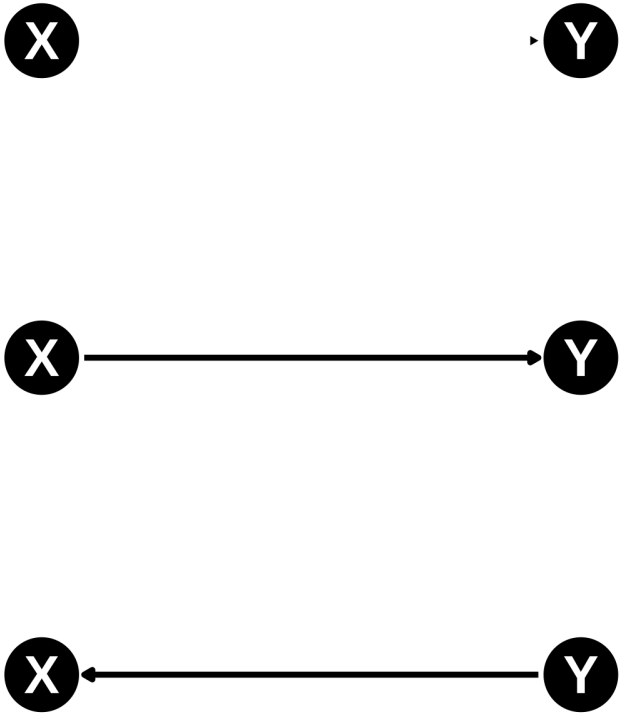
It's not that simple, and DAGs can help us understand why!

Simple models for causality

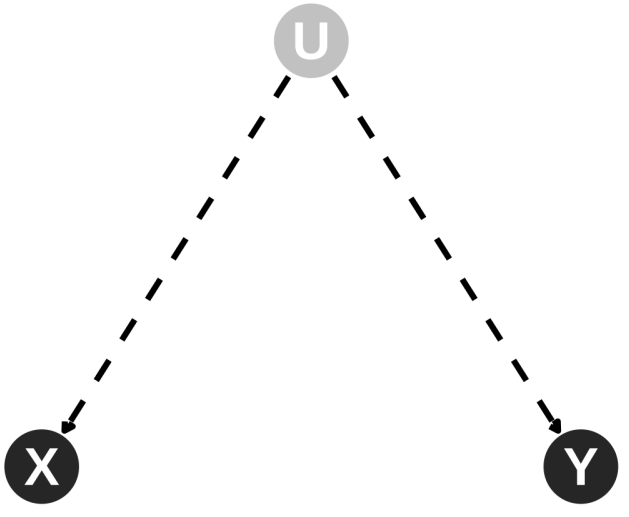
Think about **interventions** that change some target variable T

- Forget about the arrows pointing into T (intervention makes them irrelevant)
- Change T, e.g. setting it to some arbitrary new value $T = t$
- This change propagates along directed paths out of T to all descendant variables of T in the graph, causing their values to change

(All of these changes could be deterministic, but most likely in our usage they are probabilistic)



Exercise: in each of these cases, if we intervene on X which other variable(s) are changed as a result?



Explaining an observed correlation

We find a statistically significant correlation between X and Y

What does it mean?

1. False positive (spurious correlation)
2. X causes Y
3. Y causes X
4. Both have common cause U [possibly unobserved]

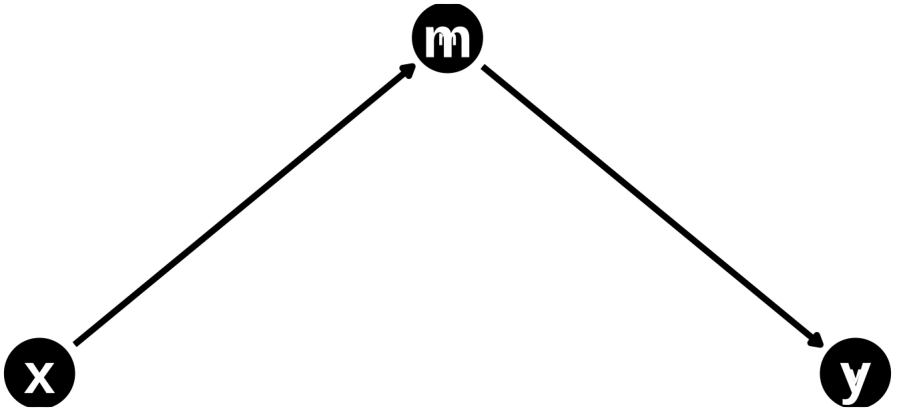
Statistically indistinguishable cases (without "experimental" data)

Importantly different consequences!

Computing counterfactuals

If we know/estimate *functions* represented by edges, we can simulate/compute the consequences of an intervention

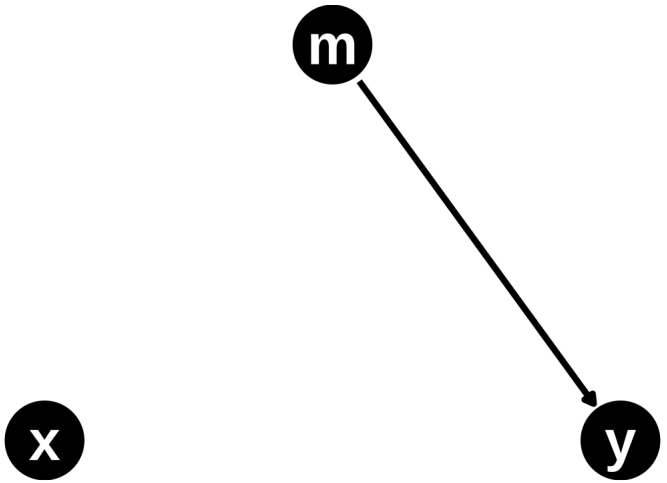
$$x = \text{exogeneous}, \quad m = f(x) + \varepsilon_m, \quad y = g(m) + \varepsilon_y$$



$$x \leftarrow x', \quad m \leftarrow f(x') + \varepsilon_m, \quad y \leftarrow g(f(x') + \varepsilon_m) + \varepsilon_y$$

If we intervene on m instead:

$$x = x, \quad m \leftarrow m', \quad y \leftarrow g(m') + \varepsilon_y$$



We can ask different causal questions about the same model, and communicate clearly/visually

Strategy: two staged regression

You might have learned "two-stage least squares" (**2SLS**)

Suppose we want to learn the causal relationship of D on Y, but (**Exercise**: draw the DAG for this)

$$Y = D\theta + X\beta + \varepsilon_Y$$

$$D = X\alpha + \varepsilon_D$$

In words: X is confounding the relationship

- First stage: regress out X
- Second stage: using residuals from first stage,

$$\text{regress } Y - X\hat{\beta} \text{ on } D - X\hat{\alpha}$$

Strategy: double machine learning (DML)

For various reasons (e.g. nonlinearity) we might replace linear regression in 2SLS with more complex, machine learning predictive models

- First stage: regress out X using ML models
- Second stage: using residuals from first stage,

$$\text{regress } Y - \hat{Y} \text{ on } D - \hat{D}$$

(This is an exciting and active field of research now!)

This is pretty cool

To see why, let's remember the other of the two common reasons regression coefficients are often misinterpreted: **nonlinearity**

Non-linear example

Suppose there is one predictor x, and a (global) non-linear model fits the CEF:

$$\mathbb{E}[\mathbf{y}|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

We don't know the β's but we have some data, and we use multiple linear regression to fit the coefficients

```
x2 <- x^2
lm(y ~ x + x2)
```

The model fits well, but there's an **interpretation problem**:

$$\frac{\partial}{\partial x}\mathbb{E}[\mathbf{y}|x] = \beta_1 + 2\beta_2 x \neq \beta_1 \approx \hat{\beta}_1$$

What went wrong?

In this simple example we know the problem is that x_2 is actually a function of x. **Solution**: interpret $\frac{\partial}{\partial x}$ locally as a function of x, not as a global constant

Sometimes simplifying assumptions are *importantly wrong*. Sometimes we must reject simple interpretations and use more complex models (ML)

Problem: ML models may be more difficult to interpret, e.g. not having coefficients like regression models

Preview: later in the course we will learn new methods for interpreting some ML models

Conclusions

Wisdom from one of the great early statistical explorers

Udny Yule:

> **Measurement does not necessarily mean progress**. Failing the possibility of measuring that which you desire, the lust for measurement may, for example, merely result in your measuring something else - and perhaps forgetting the difference - or in your ignoring some things because they cannot be measured.

Remember: regression coefficients do not necessarily mean causal relationships

Experiments

Actually do interventions while collecting data

Observational studies

Try to infer causal relationships without interventions, by using ~~dark arts~~ more/specialized assumptions/methods that require careful interpretation

(increasingly common due to superabundance of data)

Scientific progress: be wrong in more interesting/specific ways

Causal inference isn't easy!

Predictive machine learning is about

$$p_{Y|X}(y|x)$$

and regression--conditional expectation, conditional quantile, etc. If we passively observe some value of x, what would we observe about y?

Causal inference is about (various notations)

$$p(y|\mathrm{do}[X = x]), \quad \text{i.e.} \quad p(y|X \leftarrow x)$$

i.e. what happens to Y when we actually **intervene** on X

Causal inference

An exciting interdisciplinary field

Practically important, connections to ML

> "Data scientists have hitherto only predicted the world in various ways; the point is to change it" - Joshua Loftus