

Mestrado Profissional em Avaliação e
Monitoramento de Políticas Públicas

Métodos Quantitativos I

Aula 4: Relações entre variáveis

Professores: Daniel Grimaldi e Arthur Bragança

3º Trimestre - 2025

Relação entre variáveis

Contexto

- Até agora, foco das aulas esteve em descrever variáveis e testar hipóteses, mas sempre partindo de distribuições independentes (que consideram apenas 1 variável).
- Em pesquisa social normalmente queremos investigar a relação entre múltiplas variáveis. Essas relações podem ser:
 - positivas, negativas ou inexistentes;
 - lineares ou não;
 - causais ou não.
- Para investigar essas relações, devemos entender como a distribuição de uma variável condicional (“afeta”) a distribuição das outras.

Variância vs Covariância

- ❖ Variância mede o grau de dispersão de uma variável - ou o quão provável é observarmos um valor distante da média populacional.

$$Var(X) \equiv \sigma_x^2 = E[(X - \mu_x)^2] = E[(X - \mu_x)(X - \mu_x)]$$

- ❖ Covariância mede a relação linear entre duas variáveis - ou o quão provável é observarmos um valor distante da média populacional para as duas variáveis conjuntamente.

$$Cov(X, Y) \equiv \sigma_{xy}^2 = E[(X - \mu_x)(Y - \mu_Y)]$$

Propriedades da Covariância

- ✚ Sendo a, b, c e d constantes e X e Y Variáveis Aleatórias, vale que:
 - i. $Cov(X, X) = \sigma_x^2$;
 - ii. $Cov(X, Y) = Cov(Y, X)$;
 - iii. $Cov(X, Y) = E(XY) - E(X)E(Y)$;
 - iv. $Cov(aX + b, cY + d) = ac Cov(X, Y)$
- ✚ Independência implica covariância zero (isso fica claro por *iii*), mas a recíproca não é verdadeira.

Correlação

- ❖ O Coeficiente de Correlação de Pearson (o mais comum) transforma a medida de covariância para termos um número diretamente interpretável - é a covariância padronizada.

$$Cor(X, Y) = \frac{E[(X - \mu_x)(Y - \mu_Y)]}{\sqrt{Var(X)Var(Y)}} \quad (1)$$

- ❖ Quanto mais perto de 1 (-1), mais positiva (negativa) é a relação linear entre as variáveis.

OLS e Correlação

- ❖ Considere um modelo de regressão linear simples entre duas variáveis aleatórias Y e X , definido por:
 $Y = \alpha + \beta X + \epsilon$. O Coeficiente estimado por OLS para β equivale a:

$$\hat{\beta}_{ols} = \frac{\sum(Y - \bar{Y})(X - \bar{x})}{\sigma_x} \quad (2)$$

- ❖ Vale notar que ele é fundamentalmente o coeficiente de correlação (dado pela Equação 1), mas sem a mesma padronização.
 - ❖ Veremos o modelo OLS em detalhes nas próximas aulas

Causalidade em ciências sociais

Carecas de ouvir...

✚ “In god we trust, all others must bring data”

W. Edwards Deming

✚ “Let the data speak for themselves”

Ronald Fisher

✚ “Correlação não é causalidade!”

Qualquer professor de STAT-001

... Então, o que os dados podem me dizer sobre causalidade?

Causalidade

Em ciências sociais, a defesa de uma relação de **causalidade** entre A e B exige que todo pesquisador atenda a três condições:

1. Demonstrar associação empírica entre A e B ;
2. Demonstrar relação temporal apropriada A e B ;
3. Convencer a todos que as duas condições anteriores não se devem a **fatores espúrios**.

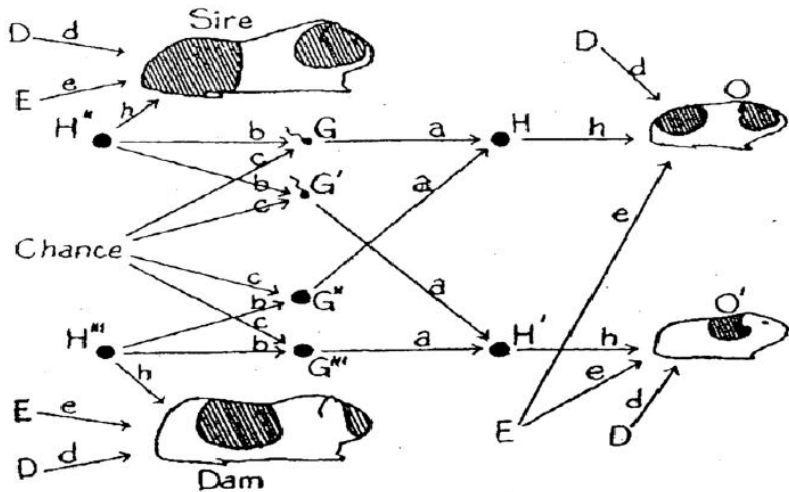
É virtualmente impossível atender à terceira condição só com dados...

Karl Pearson

- ✚ Pearson foi aluno de Francis Galton e o ajudou a desenvolver o conceito de regressão linear;
- ✚ Fundador da Biometrika, até hoje uma das importantes revistas de estatística do mundo.

“That a certain sequence has occurred and reoccurred in the past is a matter of experience to which we give expression in the concept of causation (...). Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated.”

Sewall Wright



Sewall vs. Niles

- Wright (1921) defende o uso de **path analysis** para separar correlação de causalidade em outros contextos fora da genética.
- Niles (1922) critica a proposta de Wright (1921)

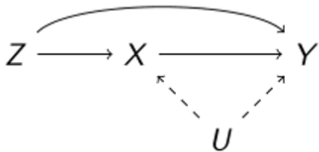
“To contrast “causation” and “correlation” is unwarranted because causation is simply perfect correlation. Incomplete correlation denotes partial causation, the effect here being brought about by more than one important cause. Many things show either high or perfect correlation that, on common-sense grounds, can not possibly be cause and effect. But we can not tell a priori what things are cause and effect (...).”

Judea Pearl

- ❖ Judea Pearl é professor de ciência da computação na UCLA e ganhador do prêmio Turing (2011);
- ❖ Desenvolveu o arcabouço de Directed Acyclical Graphs (DAGs) com o objetivo de criar uma “linguagem” formal para inferência causal (Pearl and Mackenzie 2018; Pearl, Glymour, and Jewell 2016; Chen and Pearl 2013).
- ❖ Ferramenta tem se tornado mais relevante em ciências sociais (Imbens 2020).

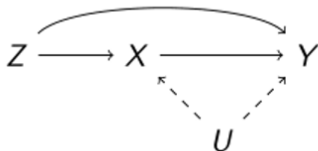
Show me the DAGs!

O que é um DAG?

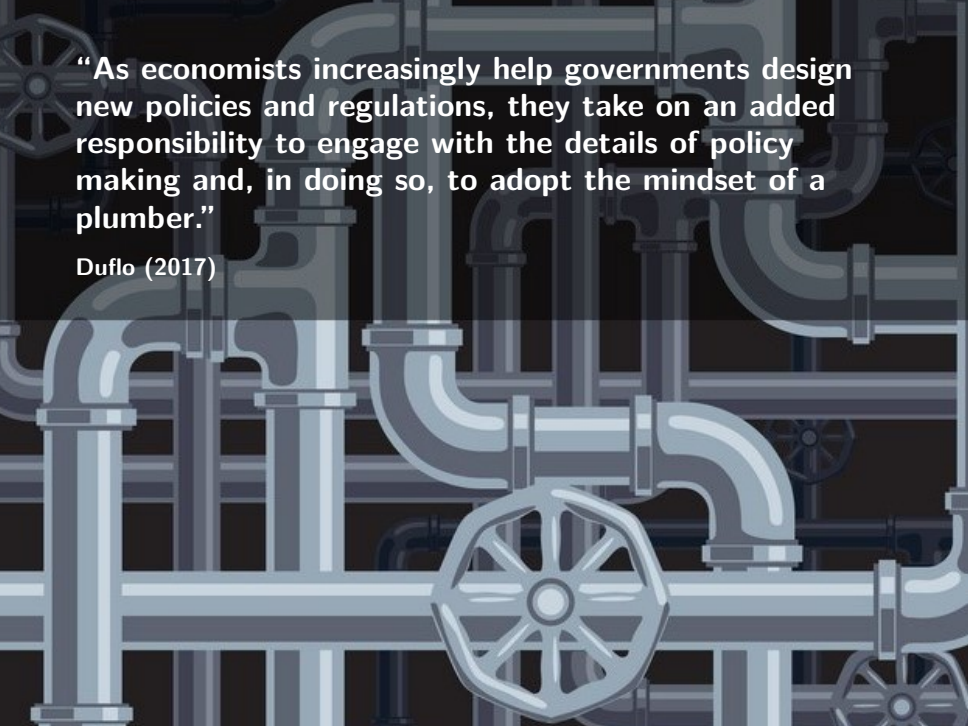


- ❖ **Causal Directed Acyclical Graphs (DAGs)** são uma representação gráfica (não-paramétrica) de uma relação causal teórica.
- ❖ Os nódulos representam variáveis
- ❖ As setas representam [potenciais] relações causais **diretas**
 - ❖ setas tracejadas indicam causalidade associada a variáveis não observáveis.

Terminologia



- ❖ **Children (Parents)** de um nóculo: todos os nóculos que diretamente são afetados (afetam)
- ❖ **Descendants (Ancestors)** de um nóculo: todos os nóculos que direta ou indiretamente são afetados (afetam)
- ❖ **Path**: uma associação (\neq relação causal) ligando dois ou mais nóculos.



“As economists increasingly help governments design new policies and regulations, they take on an added responsibility to engage with the details of policy making and, in doing so, to adopt the mindset of a plumber.”

Duflo (2017)

Quando o cano está aberto?

Um **mediator**, como X na Figura A, permite que a associação flua entre Z e Y, mas controlar/condicionar por um **mediator** (ou seus descendentes) fecha o fluxo.

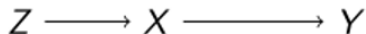


Figure A

Um **collider**, como X na Figura B, bloqueia associação entre Z e Y, mas controlar/condicionar por um **collider** (ou seus descendentes) abre o fluxo.

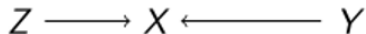
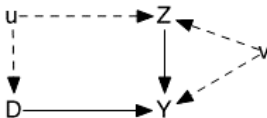


Figure B

Fechando os canos certos

- ❖ Em todo DAG existem os *paths* que desejamos avaliar e *paths* alternativos (***backdoor paths***).
- ❖ ***Backdoor criterion***: Devemos garantir que todos os *backdoor paths* estão bloqueados para medirmos apenas a relação de interesse.
 - ❖ É equivalente à hipótese de independência condicional (CIA) ou *unconfoundedness*.



DAGs na prática

Quando usamos DAGs?

- ❖ DAGs são úteis para orientar a reflexão e a apresentação da sua teoria (hipótese causal)
 - ❖ É claro que existem outras formas de se fazer isso, mas DAGs criam uma estrutura comum.
 - ❖ Sem teoria, um número com três asteriscos ao lado é... só um número com três asteriscos ao lado.
- ❖ Uma teoria consolidada orienta decisões sobre como usar os dados
 - ❖ Uma abordagem data-driven pura te torna refém dos dados
 - ❖ DAGs ajudam a tornar mais claros erros de especificação (*omitted variable bias*, *collider bias* e *bad controls*)

Hands On!

Características do programa

- ❖ Imaginem **um programa de recolocação profissional** (como em *Dehejia and Wahba, 2002*):
 - ❖ apenas funcionários que foram demitidos em t_0 são elegíveis para o programa;
 - ❖ renda depende positivamente de habilidades não observáveis (S) de cada indivíduo;
 - ❖ probabilidade de demissão e de participação no programa depende negativamente de S ;
 - ❖ **impacto do programa é *a priori* desconhecido**;

$$S \sim N(50, 10) ; \epsilon \sim N(0, 30) \quad (3)$$

$$P(D_i = 1|S = s_i) = \frac{1}{1 + e^{\left(\frac{s_i - \mu_S}{\sigma_S}\right)}} \quad (4)$$

$$P(P_i = 1|D = d_i, S = s_i) = d_i \cdot 1 - \frac{1}{2 + e^{\frac{s_i - \mu_S}{\sigma_S}}} \quad (5)$$

$$Y_{t_0} = 800 S_i - 400 D_i S_i + \epsilon_{t_0,i} \quad (6)$$

$$Y_{t_1} = 1,000 S_i - 500 D_i S_i - \delta P_i + \epsilon_{t_1,i} \quad (7)$$

Implementando PGD

```
pgd <- function(sample.size){  
  set.seed(13)  
  data <- data.frame(  
    S = rnorm(n=sample.size, mean=50, sd=30)) %>%  
    mutate(p_D = 1/(1+exp(1+scale(S))),  
           D = sapply(p_D, function(i) rbinom(1,1, prob=i)),  
           e_0 = rnorm(n=sample.size, sd=30),  
           Y_0 = 800*S - 400*D + e_0,  
           p_P = D*(1-1/(1+exp(2+scale(S)))),  
           P = sapply(p_P, function(i) rbinom(1,1,prob=i)),  
           e_1 = rnorm(n=sample.size, sd=30),  
           Y_1 = 1000*S - 500*D*D + delta*P + e_1)  
  data  
}
```

Simulação

```
data <- pgd(10000)
head(data)
```

##	S	p_D	D	e_0	Y_0	p_P	P	e_1	Y_1
## 1	66.62981	0.17335260	0	10.057444	53313.90	0.000000	0	40.1145951	66669.92
## 2	41.59184	0.32593095	1	5.943428	32879.42	0.848983	1	-0.7881657	40091.05
## 3	103.25490	0.05819142	0	-36.268493	82567.65	0.000000	0	7.9899497	103262.89
## 4	55.61960	0.23242120	0	-54.767298	44440.92	0.000000	0	4.4553285	55624.06
## 5	84.27578	0.10425491	0	6.837846	67427.47	0.000000	0	-20.0593818	84255.73
## 6	62.46578	0.19417356	0	28.298220	50000.93	0.000000	0	-12.0116126	62453.77

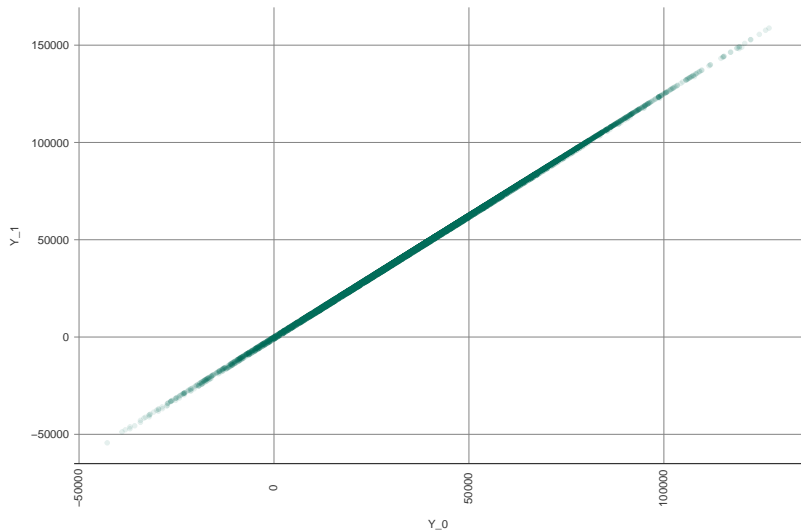
Caso 1: Relação entre programa e renda futura

- ✚ Temos acesso às seguintes informações:
 - ✚ status de participação no programa (P_i);
 - ✚ status sobre demissões (D_i);
 - ✚ renda dos trabalhadores (Y_{i,t_0} e Y_{i,t_1})
- ✚ **Queremos investigar a relação entre a participação no programa e renda futura dos trabalhadores.**
- ✚ É útil começarmos com uma inspeção visual...

Plot entre Y_1 e Y_0

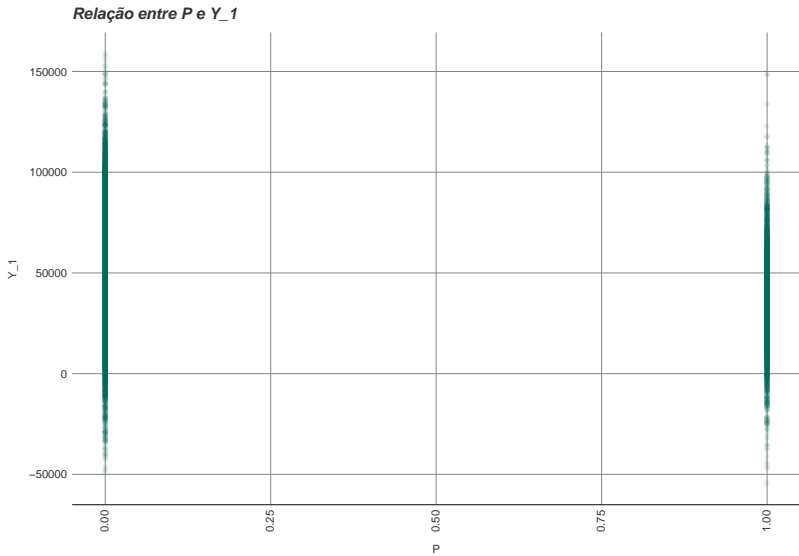
```
fig <- ggplot(data, aes(x=Y_0, y=Y_1)) +  
  geom_point(color=cores$verde_escuro,  
             fill=cores$verde_claro,  
             alpha=0.1) +  
  labs(title="Relação entre Y_0 e Y_1",  
        y="Y_1",  
        x="Y_0") +  
  tema_base_fundobranco()
```

Relação entre Y_0 e Y_1



geom_point

```
fig <- ggplot(data, aes(x=P, y=Y_1)) +  
  geom_point(color=cores$verde_escuro,  
             fill=cores$verde_claro,  
             alpha=0.1) +  
  labs(title="Relação entre P e Y_1",  
        y="Y_1",  
        x="P") +  
  tema_base_fundobranco()
```

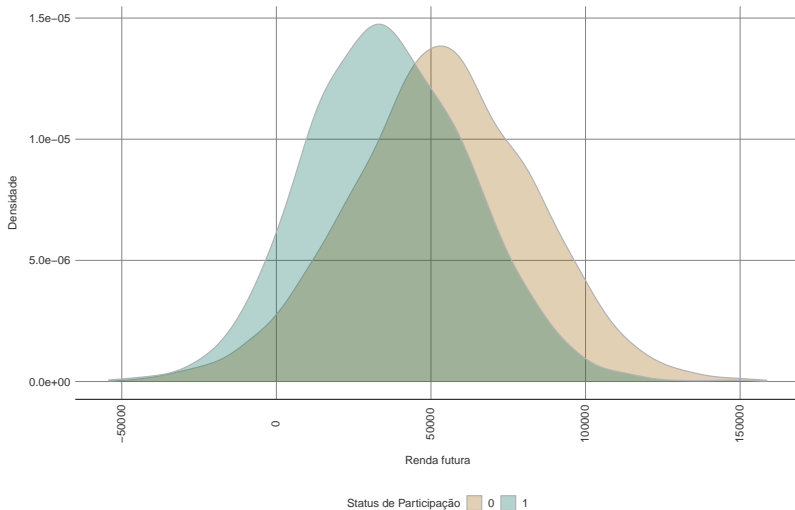


Histograma condicional: P

```
fig <- ggplot(data) +  
  geom_density(aes(Y_1, fill=as.factor(P)),  
               alpha=0.3,  
               color=cores$cinza_claro,  
               adjust=1.2) +  
  scale_fill_manual(values=c(cores$amarelo_fechado, cores$verde_escuro)) +  
  labs(title="Distribuição Renda Futura",  
        subtitle="Amostra completa",  
        y="Densidade",  
        x="Renda futura",  
        fill="Status de Participação") +  
  tema_base_fundobranco()
```

Distribuição Renda Futura

Amostra completa

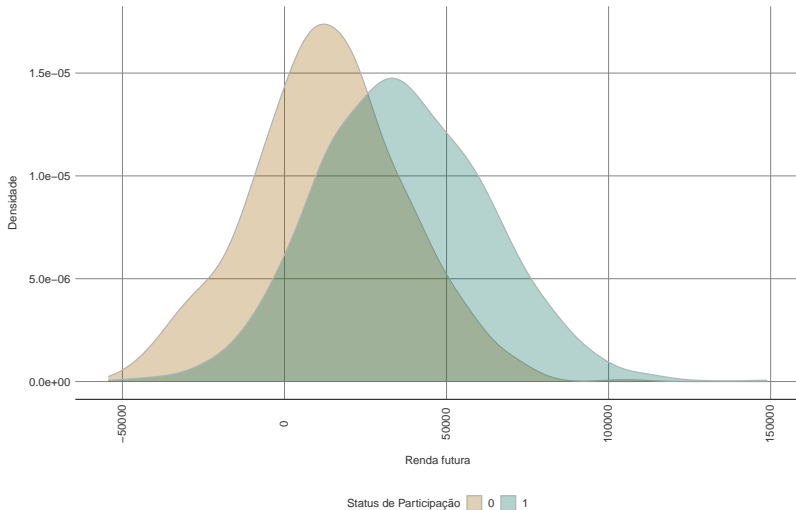


Histograma condicional: $P \mid D=1$

```
fig <- ggplot(filter(data, D==1)) +  
  geom_density(aes(Y_1, fill=as.factor(P)),  
               alpha=0.3,  
               color=cores$cinza_claro,  
               adjust=1.2) +  
  scale_fill_manual(values=c(cores$amarelo_fechado, cores$verde_escuro)) +  
  labs(title="Distribuição Renda Futura",  
       subtitle="Apenas entre demitidos",  
       y="Densidade",  
       x="Renda futura",  
       fill="Status de Participação") +  
  tema_base_fundobranco()
```

Distribuição Renda Futura

Apenas entre demitidos



Estimando correlações

```
cor_total <- cor(data$Y_1, data$P)  
cor_demitidos <- cor(data$Y_1[data$D==1], data$P[data$D==1])  
cor_total
```

```
[1] -0.2376758
```

```
cor_demitidos
```

```
[1] 0.3518679
```

Estimando OLS

```
reg1 <- lm(Y_1 ~ P, data=data)
summary(reg1)
```

```
##
## Call:
## lm(formula = Y_1 ~ P, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102106  -19300    -219    19638   112520
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53358.6      334.7   159.44  <2e-16 ***
## P           -16934.2      692.1   -24.47  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29290 on 9998 degrees of freedom
## Multiple R-squared:  0.05649,    Adjusted R-squared:  0.0564
## F-statistic: 598.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Construindo tabela de regressão

```
reg2 <- lm(Y_1 ~ P + D, data=data)
reg3 <- lm(Y_1 ~ P + Y_0, data=data)
stargazer(reg1, reg2, reg3,
           header = FALSE, keep.stat = c("n"),
           font.size = "tiny", no.space = TRUE,
           dep.var.labels.include = FALSE,
           dep.var.caption = "Renda após P",
           column.sep.width="1pt", digits=2,
           digits.extra=0)
```

Coeficientes de modelos OLS

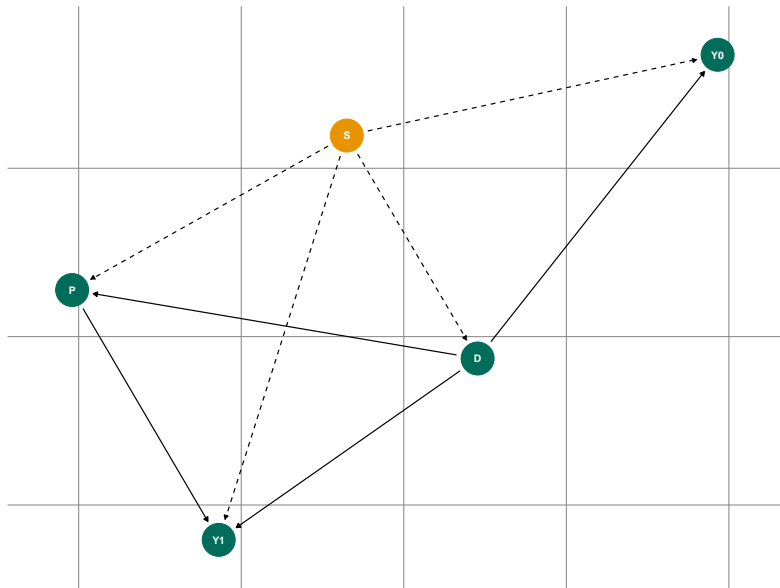
	Renda após P		
	(1)	(2)	(3)
P	-16,934.21*** (692.14)	23,216.16*** (1,177.80)	-999.69*** (1.17)
D		-44,104.97*** (1,085.25)	
Y_0			1.25*** (0.00)
Constant	53,358.60*** (334.67)	57,313.19*** (324.97)	-1.32 (1.04)
Observations	10,000	10,000	10,000
Note:		*p<0.1; **p<0.05; ***p<0.01	

DAG para a teoria de mudança

```
set.seed(13)
ex1_dag <- dagify(D ~ S,
                  Y0 ~ S + D,
                  P ~ S + D,
                  Y1 ~ S + D + P) %>%
  tidy_dagitty() %>%
  mutate(color_node=ifelse(name=="S", "Unobserved", "Observed"),
         edge_line=ifelse(name=="S", 2, 1))
```

Gerando Gráfico DAG

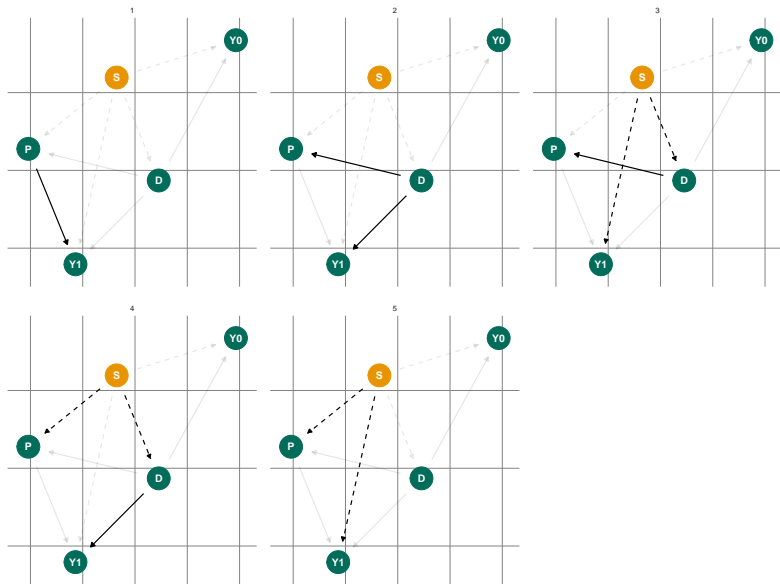
```
graf1 <- ggplot(data= ex1_dag, aes(
  x = x,
  y = y,
  xend = xend,
  yend = yend
)) +
  geom_dag_point(aes(colour=color_node, fill=color_node), show.legend = FALSE) +
  scale_color_manual(values = c(cores$verde_escuro, cores$amarelo_escuro)) +
  geom_dag_edges(aes(edge_linetype=edge_line))+
  geom_dag_text() +
  labs(title = "Teoria causal do Programa",
        y="",
        x="") +
  tema_base_fundobranco() +
  theme(axis.line.x = element_line(colour=NA),
        axis.text.x = element_text(size=0),
        axis.text.y = element_text(size=0))
```



Quais canos estão abertos?

```
graf2.data <- dag_paths(ex1_dag, from = "P", to="Y1") %>%  
  distinct(set, name, to, .keep_all=TRUE) %>%  
  mutate(show_node=ifelse(is.na(path), 0, 1),  
         show_line=case_when(is.na(path) ~ 0.1,  
                              TRUE ~ 1))
```

Paths entre Programa de Recolocação e Renda



Qual o modelo correto?

- Estimação (1) deixa todos os *paths* abertos, contaminando β_{ols} com correlações espúrias (*paths* 2 a 5).
- Estimação (2) fecha 3 dos 4 caminhos não-causais, mas ainda resta o *path* 5, que também contamina $\hat{\beta}_{ols}$
- Estimação (3) fecha todos os *paths* não-causais, porque controla por um descendente de S e de D .
 - $\delta = -1000$
- Também é possível recuperar o efeito causal se controlarmos simultaneamente por S e por D

Se tivermos informação sobre S

	Renda após P			
	(1)	(2)	(3)	(4)
P	-16,934.21*** (692.14)	23,216.16*** (1,177.80)	-999.69*** (1.17)	-1,000.70*** (1.41)
S				1,000.00*** (0.01)
D		-44,104.97*** (1,085.25)		-500.30*** (1.45)
Y_0			1.25*** (0.00)	
S:D				0.04 (0.02)
Constant	53,358.60*** (334.67)	57,313.19*** (324.97)	-1.32 (1.04)	-0.24 (0.83)
Observations	10,000	10,000	10,000	10,000

Note: *p<0.1; **p<0.05; ***p<0.01

Caso 2: Relação entre Habilidade e Demissão

- ✚ Por meio de uma pesquisa de campo, conseguimos uma medida confiável da habilidade dos trabalhadores
- ✚ **Queremos investigar a relação entre essa habilidade e a probabilidade de demissão.**
- ✚ Por simplificação, pulemos agora a inspeção visual...

Estimando coeficientes OLS

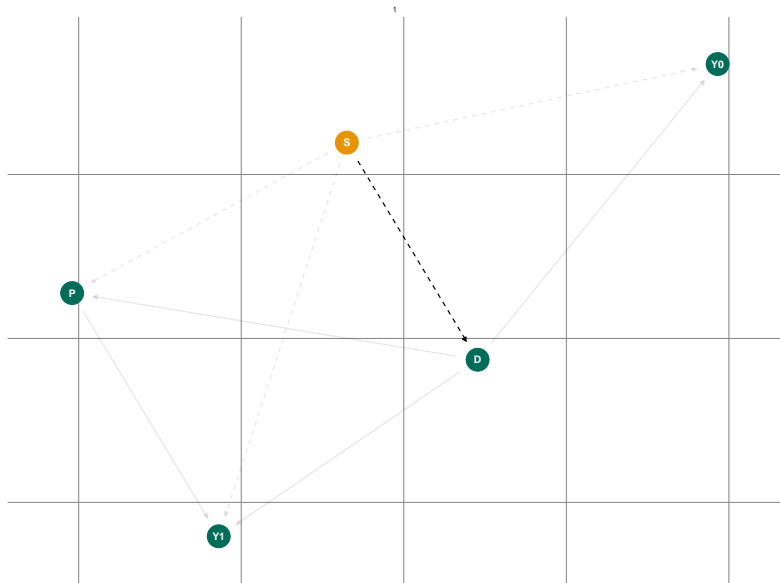
```
reg1 <- lm(D ~ S, data=data)
reg2 <- glm(D ~ S, family = binomial(link="logit"), data=data)
reg3 <- lm(D ~ S + Y_0, data=data)
reg4 <- glm(D ~ S + Y_0, family = binomial(link="logit"), data=data)
```

Estimando coeficientes OLS

	P(D)			
	<i>OLS</i> (1)	<i>logistic</i> (2)	<i>OLS</i> (3)	<i>logistic</i> (4)
S	-0.01*** (0.00)	-0.03*** (0.00)	1.94*** (0.00)	151.25 (17,298.69)
Y_0			-0.00*** (0.00)	-0.19 (21.60)
Constant	0.59*** (0.01)	0.64*** (0.04)	0.02*** (0.00)	-38.49 (6,404.72)
Observations	10,000	10,000	10,000	10,000

Note: * p<0.1; ** p<0.05; *** p<0.01

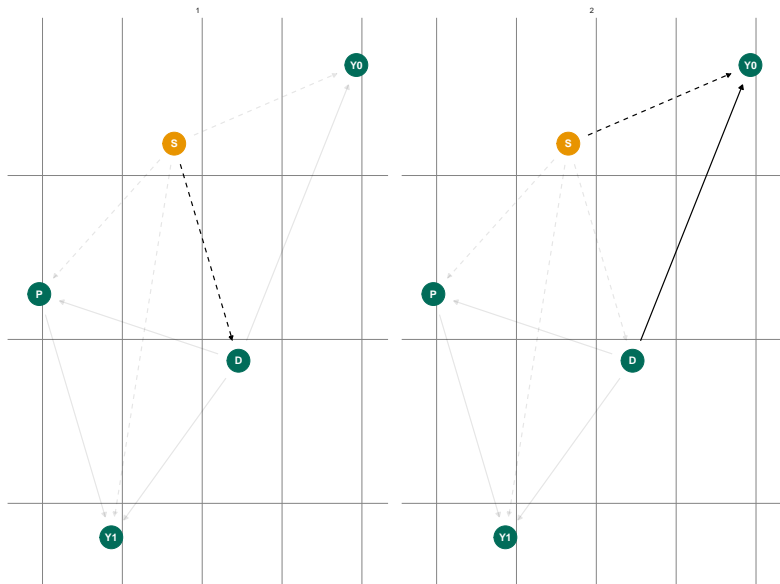
Paths entre Programa de Recolocação e Renda



Bad controls

```
fig.data <- dag_paths(ex1_dag,  
                      from = "S",  
                      to="D",  
                      adjust_for = "Y0") %>%  
distinct(set, name, to, .keep_all=TRUE) %>%  
mutate(show_node=ifelse(is.na(path), 0, 1),  
       show_line=case_when(is.na(path) ~ 0.1,  
                           TRUE ~ 1))
```

Paths entre Programa de Recolocação e Renda



Bad controls

- ❖ A inclusão de Y_{t0} na regressão equivale a controlar por um *collider*
 - ❖ Logo, ele abre um caminho não-causal (*path 2*) entre S e D , que acaba contaminando o coeficiente estimado
- ❖ Mais controles não implica, necessariamente, um modelo mais robusto
- ❖ É preciso ter uma teoria por trás a respeito de como deve operar a realidade, para que se possa julgar o modelo mais adequado
 - ❖ DAGs ajudam nisso

Mais sobre DAGs

- Chen, Bryant, and Judea Pearl. 2013. "Regression and Causation: A Critical Examination of Six Econometrics Textbooks." {SSRN} {Scholarly} {Paper} ID 2338705. Rochester, NY: Social Science Research Network. <https://papers.ssrn.com/abstract=2338705>.
- Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79. <https://doi.org/10.1257/jel.20191597>.
- Niles, Henry E. 1922. "Correlation, Causation and Wright's Theory of "Path Coefficients"." *Genetics* 7 (3): 258–73. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200533/>.
- Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.
- Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.
- Wright, Sewall. 1921. "Correlation and Causation." *J. Agric. Res.* 20: 557–80.

Obrigado!