Categorical Data Analysis

B.D. in Business Administration and Economics
Course in Quantitative Methods III

Rosario Barone
University of Rome "Tor Vergata"

rosario.barone@uniroma2.it

Contingency tables

The joint distribution between two categorical variables determines their relationship. This distribution also determines the marginal and conditional distributions.

Let X and Y denote two categorical response variables, X with I categories and Y with J categories. Classifications of subjects on both variables have IJ possible combinations. The responses (X, Y) of a subject chosen randomly from some population have a probability distribution.

We define as *contingency table* a rectangular tables with I rows and J columns, containing the frequencies of the outcome for each of the variables.

Notation tables

	1	**2**	**I**
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
J	n_{+1}	n_{+2}	n

	1	**2**	**I**
1	p_{11} $(p_{1\|1})$	p_{12} $(p_{2\|1})$	p_{1+} (1)
2	p_{21} $(p_{1\|2})$	p_{22} $(p_{2\|2})$	p_{2+} (1)
J	p_{+1}	p_{+2}	1

Contingency tables

Table cells at the intersections of rows and columns indicate frequencies of both events coinciding.

For example, the table below shows the preferred financial asset by a group of 223 investors .

	Bonds	**Equity**	**Row Tot**
Male	66	40	106
Female	30	87	117
Col Tot	96	127	223

Contingency tables

Contigency tables helps in calculating probabilities:

	Bonds	**Equity**	**Row Tot**
Male	66	40	106
Female	30	87	117
Col Tot	96	127	223

1. Joint Probability: $p_{ij} = \frac{n_{ij}}{\sum_i \sum_j n_{ij}}$:

$$P(\text{Female} \cap \text{Bond}) = 30/223 = 0.135$$

2. Marginal Probability: $p_i = \frac{\sum_j n_{ij}}{\sum_i \sum_j n_{ij}}$

$$P(\text{Bond}) = 96/223 = 0.431$$

Contingency tables

- Conditional Probability: $p_{i|j} = p_{ij}/p_j$, where p_{ij} is calculated as in (1) and p_j calculated as in (2).

$$P(\text{Female}|\text{Bond}) = \frac{P(\text{Female} \cap \text{Bond})}{P(\text{Bond})} = \frac{0.135}{0.431} = 0.313$$

However, via contingency tables:

	Bonds	**Equity**	**Row Tot**
Male	66	40	106
Female	**30**	87	117
Col Tot	**96**	127	223

$$P(\text{Female}|\text{Bond}) = 30/96 = 0.313$$

Chi-squared test

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the *expected frequencies* and the observed frequencies in one or more categories of a contingency table.

	Bonds	**Equity**	**Row Tot**
Male	45.632	$\mathbb{E}(M \cap E)$	**106**
Female	$\mathbb{E}(F \cap B)$	$\mathbb{E}(F \cap E)$	117
Col Tot	**96**	127	**223**

$$\mathbb{E}(M \cap B) = 96 \times 106/223 = 45.632$$

$$\ldots$$

Chi-squared test

- Table of the expected frequecies:

	Bonds	**Equity**
Male	45.632	60.368
Female	50.368	66.632

- The Chi-squared statistic is calculated as

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(E(n_{ij}) - n_{ij})^2}{E(n_{ij})}$$

	Bonds	**Equity**
Male	9.091	6.872
Female	8.236	6.226

Chi-squared test

- $\chi^2 = 9.091 + 6.872 + 8.236 + 6.226 = 30.425$
- $df = (nrow - 1) \times (ncol - 1) = 1$
- In general, the larger the difference between the observed and expected values, the greater is the value of the χ^2 statistic.
- `1-pchisq(30.425,df=1)` $= 0$
- A significant result of this test means that the cells of a contingency table should be interpreted.

Poisson Sampling

The probability distributions introduced in the yesterday class, extend to cell counts in contingency tables. For instance, a Poisson sampling model treats cell counts Y_{ij} as independent Poisson random variables with parameters λ_{ij}.

The joint probability mass function for potential outcomes n_{ij} is then the product of the Poisson probabilities $P(Y_{ij} = n_{ij})$ for the IJ cells, that is

$$\prod_i \prod_i \frac{\exp(-\mu_{ij})\mu_{ij}^{n_{ij}}}{n_{ij}!}$$

Multinomial Sampling

When the total sample size n is fixed but the row and column totals are not, a multinomial sampling model applies. The IJ cells are the possible outcomes. The probability mass function of the cell counts has the multinomial form

$$[n!/(n_{11}! \cdots n_{IJ}!)] \prod_i \prod_j p_{ij}^{n_{ij}}$$

Comparing two proportions

- Relative risk: simply the ratio of proportions p_1/p_2.
- Odds-ratio: the ratio of odds $\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$.
- Risk difference: difference of proportions.
- In `R`: `odds.ratio` in `library(questionr)`

Let suppose to observe two groups of individuals, respectively G_T (Treated group) and G_C (Control Group).

	Treatment (T)	**Control (C)**
Event (E)	TE	CE
Non-event (N)	TN	CN

Relative Risk

Relative risk:

$$RR = (TE/(TE + TN))/(CE/(CE + CN))$$

$$RR = p_T/p_C$$

- $RR = 1$ the treatment does not affect the outcome;
- $RR < 1$ the risk of the outcome is decreased by the treatment, which is a "protective factor";
- $RR > 1$ the risk of the outcome is increased by the treatment, which is a "risk factor".

Odds Ratio

Odds Ratio:

$$OR = (p_T/(1-p_T))/(p_C/(1-p_C))$$

- $OR = 1$ the treatment is not associated with the outcome;
- $OR < 1$ the treatment might be a "protective factor" against the outcome;
- $OR > 1$ the treatment might be a "risk factor" for the outcome.

Risk difference

Relative risk:

$$RR = (TE/(TE + TN)) - (CE/(CE + CN))$$

$$RR = p_T - p_C$$

- $RR = 0$ the treatment does not affect the outcome;
- $RR < 0$ the risk of the outcome is decreased by the treatment, which is a "protective factor";
- $RR > 0$ the risk of the outcome is increased by the treatment, which is a "risk factor".

RR and OR

- RR more interpretable
- OR can always be computed while RR and RD only when outcomes are not fixed ("prospective" studies)
- The OR asymptotically approaches the RR for small probabilities of outcomes.
- Precisely,

$$OR = RR\frac{1-p_2}{1-p_1}$$

Multi-way tables

It is rarely possible to claim a causal link from two-way tables. To see this, let us consider multi-way tables: these are two-way tables stratified by further variables.

TABLE 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty		Percent
		Yes	No	Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Source: M. L. Radelet and G. L. Pierce, *Florida Law Rev.* **43**: 1–34 (1991). Reprinted with permission from the *Florida Law Review.*

*More details: Alan Agresti, "Categorical Data Analysis", 2nd Chapter.

Simpson's paradox

- Marginal and conditional associations can have different directions, i.e. a trend appears in several groups of data but disappears or reverses when the groups are combined.
- Famous example: UC Berkeley gender bias data

	All	All	Men	Men	Women	Women
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	**44%**	4,321	**35%**

Simpson's paradox: UC Berkeley gender bias

"The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance".

"However, when examining the individual departments, it appeared that 6 out of 85 departments were significantly biased against men, while 4 were significantly biased against women. In total, the pooled and corrected data showed a **small but statistically significant bias in favor of women**".

P.J. Bickel, E.A. Hammel and J.W. O'Connell. *Sex Bias in Graduate Admissions: Data From Berkeley*,1975, *Science*.

Causal statistical reasoning

By stratifying (basing on the Gender) the sample we are conditioning: it is equivalent to including a Gender variable in a regression model (multivariate regression model in presence of more covariates).

So the question is: when should we include an additional variable (not of main interest), regardless of its p-value?

Simpson's paradox: UC Berkeley gender bias

Data from the six largest departments

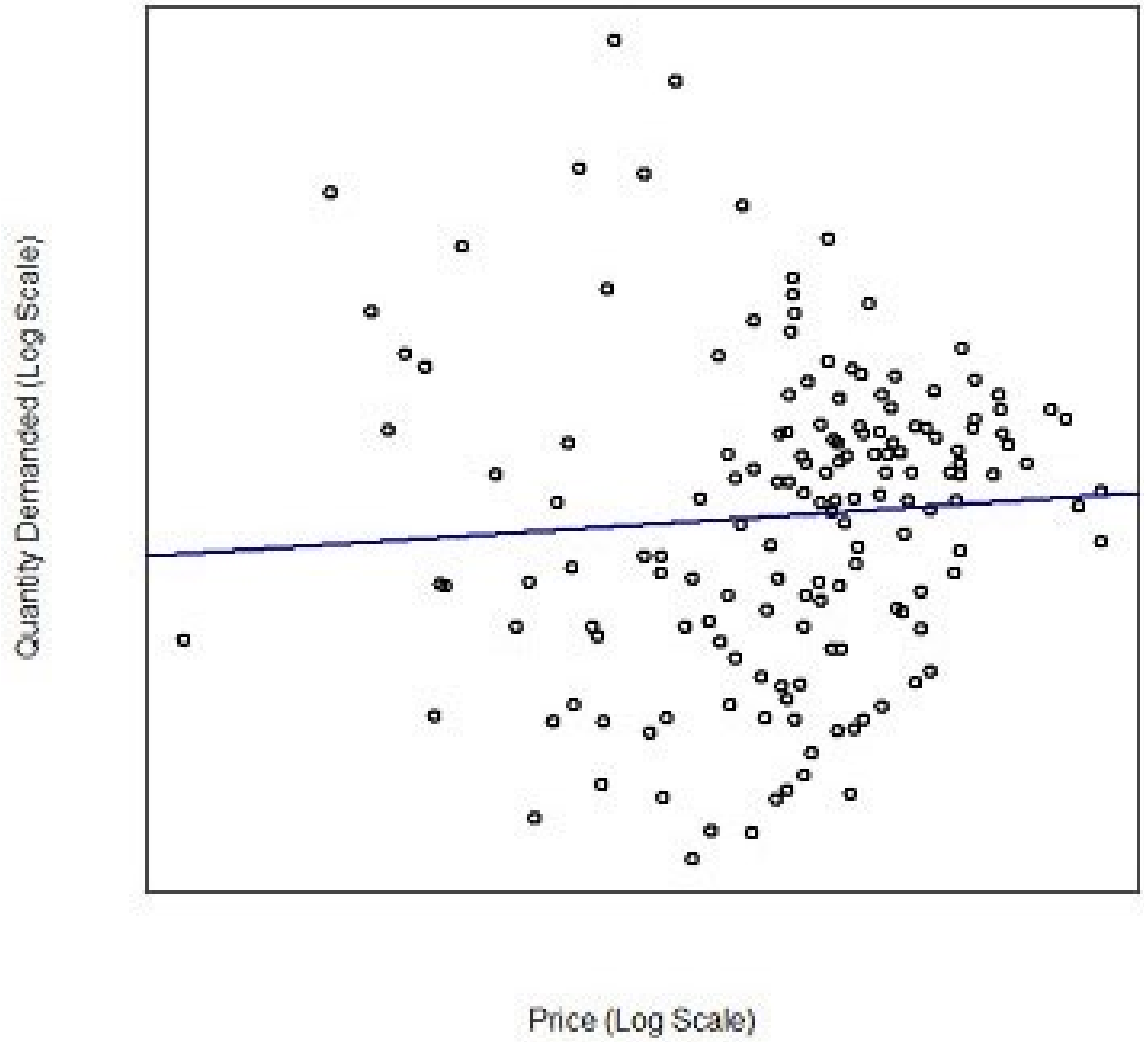
Department	All	All	Men	Men	Women	Women
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	825	62%	108	**82**%
B	585	63%	560	63%	25	**68**%
C	918	35%	325	**37**%	593	34%
D	792	34%	417	33%	375	**35**%
E	584	25%	191	**28**%	393	24%
F	714	6%	373	6%	341	**7**%
Total	4526	39%	2691	45%	1835	30%

Price sensitivity data

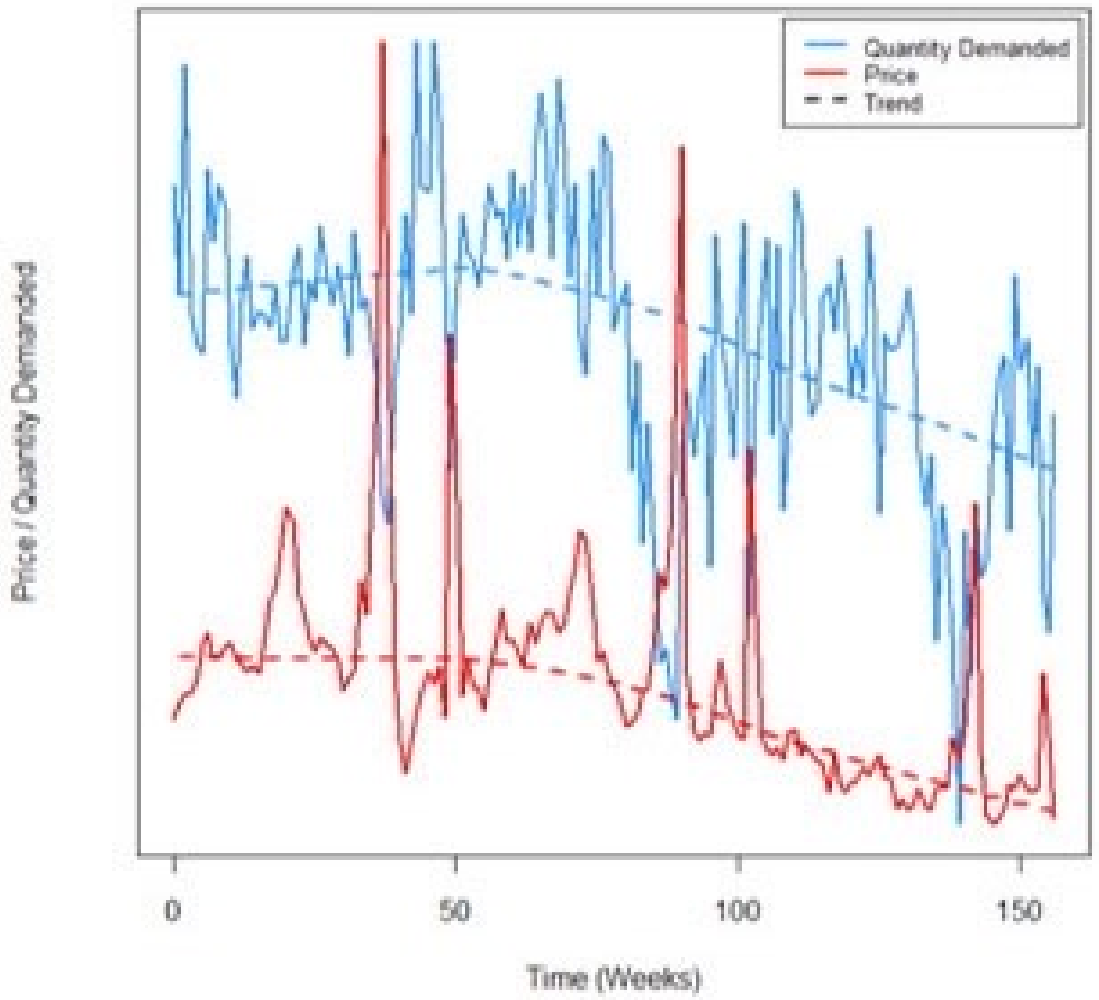
"The key to data-driven pricing strategies is the relationship between the price of a product and the amount that is sold at that price. Economists refer to this relationship as the price elasticity of demand (or simply price elasticity), but it is more commonly known as price sensitivity".

S. Berman, L. DalleMule, M. Greene, J. Lucker, *Simpson's Paradox: a cautionary tale in advanced analytics*, 2012, *Significance*.

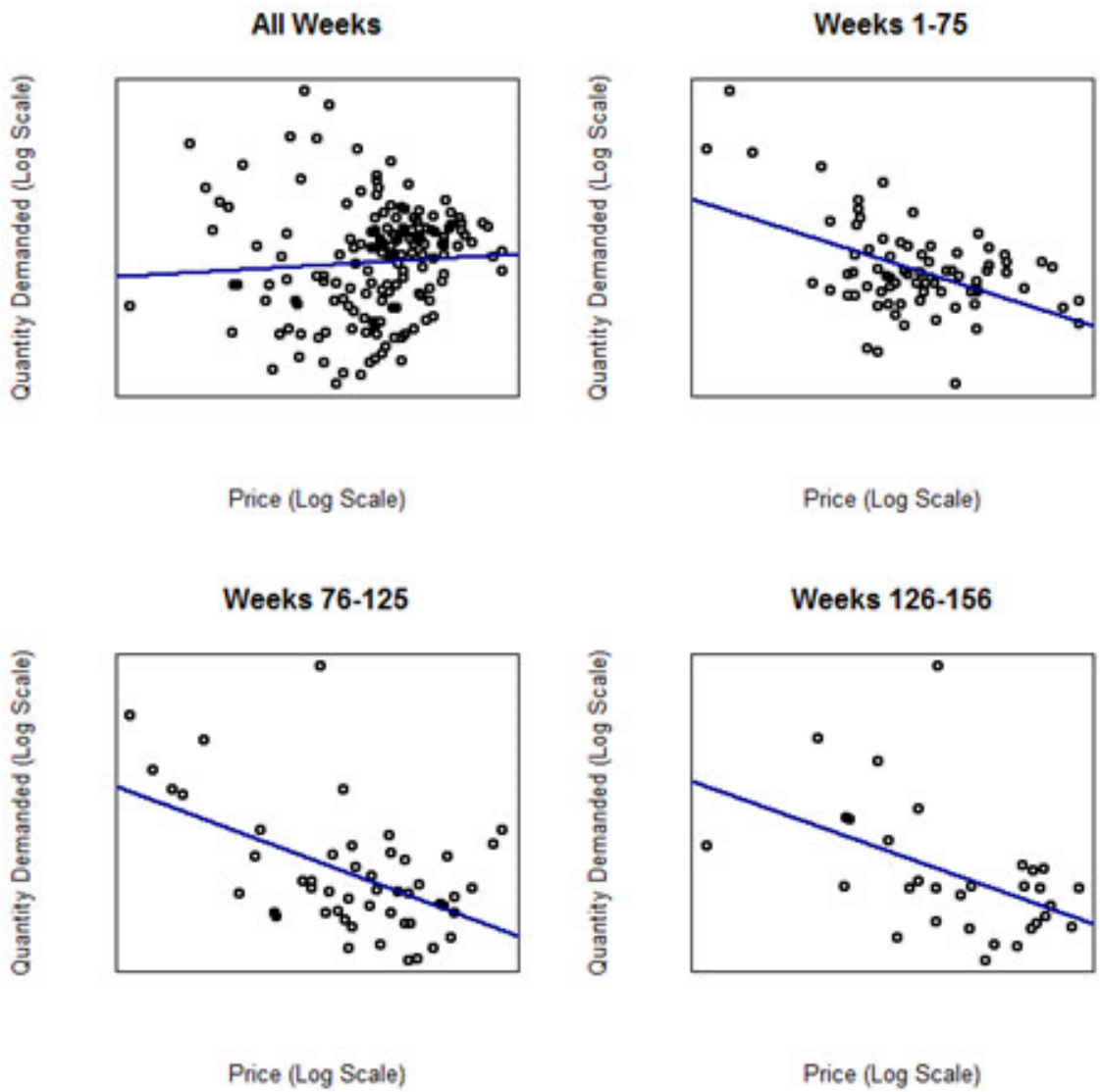
Price sensitivity data



Price sensitivity data



Price sensitivity data



Confounders, Colliders and Mediators

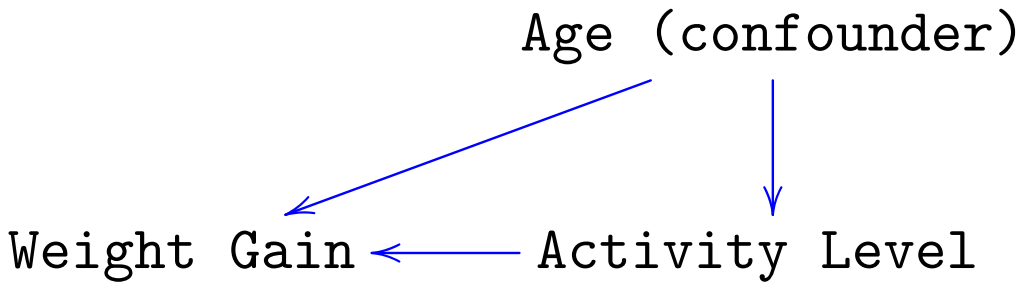
- Confounder: a variable that is having an effect both on the treatment and outcome
- Collider: a variable that is influenced both by treatment and outcome
- Mediator: a variable that is influenced by the treatment, and hence it influences the outcome

Confounders

Confounding factor are extraneous variables related to an independent and tne dependent variables considered in a study.

- It must be correlated with the independent variable (not necessairely a causal relationsih)
- It must be causally related to the dependent variable.

Example:



How to reduce the impact of confounders?

- **Restriction**: restrict the treatment group by only including subjects with the same values of potential confounding factors.
  - Relatively easy to implement
  - Restricts your sample a great deal
  - We might fail to consider other potential confounders

How to reduce the impact of confounders?

- **Matching**: each member of the comparison group should have a counterpart in the treatment group with the same values of potential confounders, but different independent variable values.
  - Allows you to include more subjects than restriction
  - Difficult to implement since you need pairs of subjects that match on every potential confounding variable
  - Other variables that can not be matched on might also be confounding variables

How to reduce the impact of confounders?

- **Statistical Control** If data has already been collected, we may include the possible confounders as control variables in the regression models; in this way, we will control for the impact of the confounding variable.

  By estimating the effect of the potential confounding variable on the dependent we can "separate" the impact of the independent variables.

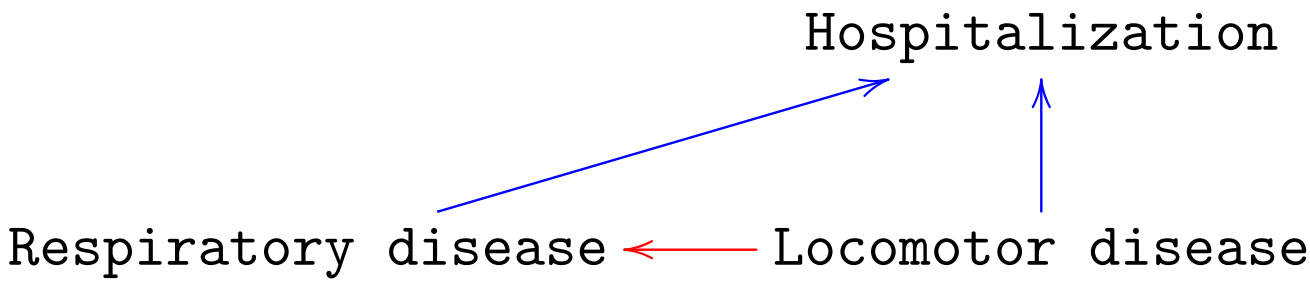
  - Easy to implement
  - Can be performed after data collection
  - Other confounding variables you have not accounted for might remain

How to reduce the impact of confounders?

- **Randomization**: randomize the values of your independent variable. You can randomly assign participants to each group. Randomization ensures that with a sufficiently large sample, all potential confounding variables (even those not directly observed) will have the same average value between different groups.
  - Considered the best method for minimizing the impact of confounding variables
  - Most difficult to carry out
  - Must be implemented prior to beginning data collection

Collider

When the dependent (outcome) and independent (exposure) variables independently affect a third variable, that variable is defined a "collider".

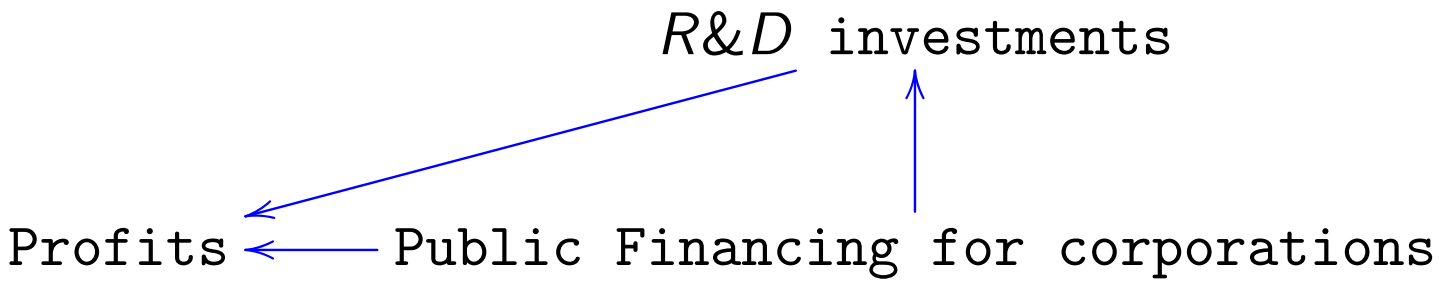


Inappropriately controlling for a collider variable results in collider bias.

D L Sackett. *Bias in analytic research*, 1979, *J. Chronic Dis*.

Mediator

When the independent (exposure) variable and a third variable affect the dependent variable, that third variable is defined a "mediator".



So, include or exclude?

- If we are dealing with confounding, we should include the confounder (if it is significant)
- If we are dealing with colliders, we should not include them (especially if they are significant!)
- Mediators can be included with some attention to the interpretation. Do not include them unless you know what you are doing.