UC San Diego

SPPS 255: Principles of Pharmacoeconomics

Bias and Confounding/Methods to address confounding

Monday, 28 April 2025

OBJECTIVES

Define and sketch a causal diagram using directed acyclic graphs (DAG)

Explain the properties of a confounder

Identify a confounder along the DAG

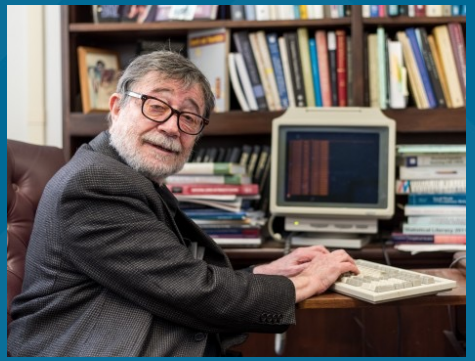
Explain the methods for controlling a confounder

DIRECTED ACYCLIC GRAPH (DAG)

A directed acyclic graph (DAG) is a visual representation of the causal relations that make the underlying relationships between factors explicit

A cause is a factor that produces an effect on another factor

A directed path is a sequence of arrows in which every arrow points in the same direction (no closed loop)



Judea Pearl

DIRECTED ACYCLIC GRAPH (DAG)—EXAMPLE

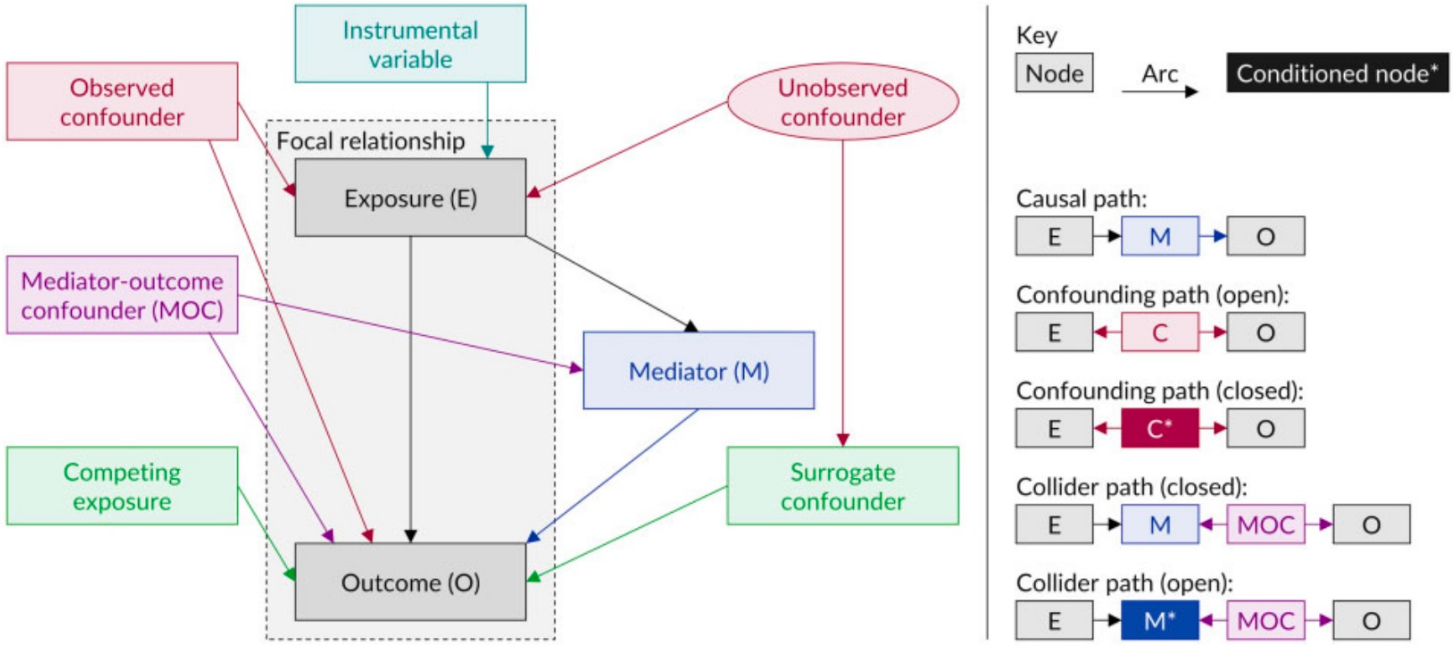


Figure 1 Illustration of the main components of a DAG, the most common types of contextual variables and the most common types of paths. The DAG has been visually arranged so that all constituent arcs flow from top-to-bottom.

Tennant, et al. International Journal of Epidemiology. 2021:50(2):620-632.

DAGITTY (DRAW YOUR OWN DAG DIAGRAMS)

DAGitty — draw and analyze causal diagrams

DAGitty is a browser-based environment for creating, editing, and analyzing causal diagrams (also known as directed acyclic graphs or causal Bayesian networks). The focus is on the use of causal diagrams for minimizing bias in empirical studies in epidemiology and other disciplines. For background information, see the "learn" page.

Launch	Download	Learn	Code
 Launch DAGitty online in your browser.	 Download DAGitty's source for offline use.	 Learn more about DAGs and DAGitty.	 The R package "dagitty" is available on CRAN or github.

DAGitty is developed and maintained by Johannes Textor (Tumor Immmunology Lab and Institute for Computing and Information Sciences, Radboud University Nijmegen).

Tweet #DAGitty | Tweet to @JohannesTextor

Versions

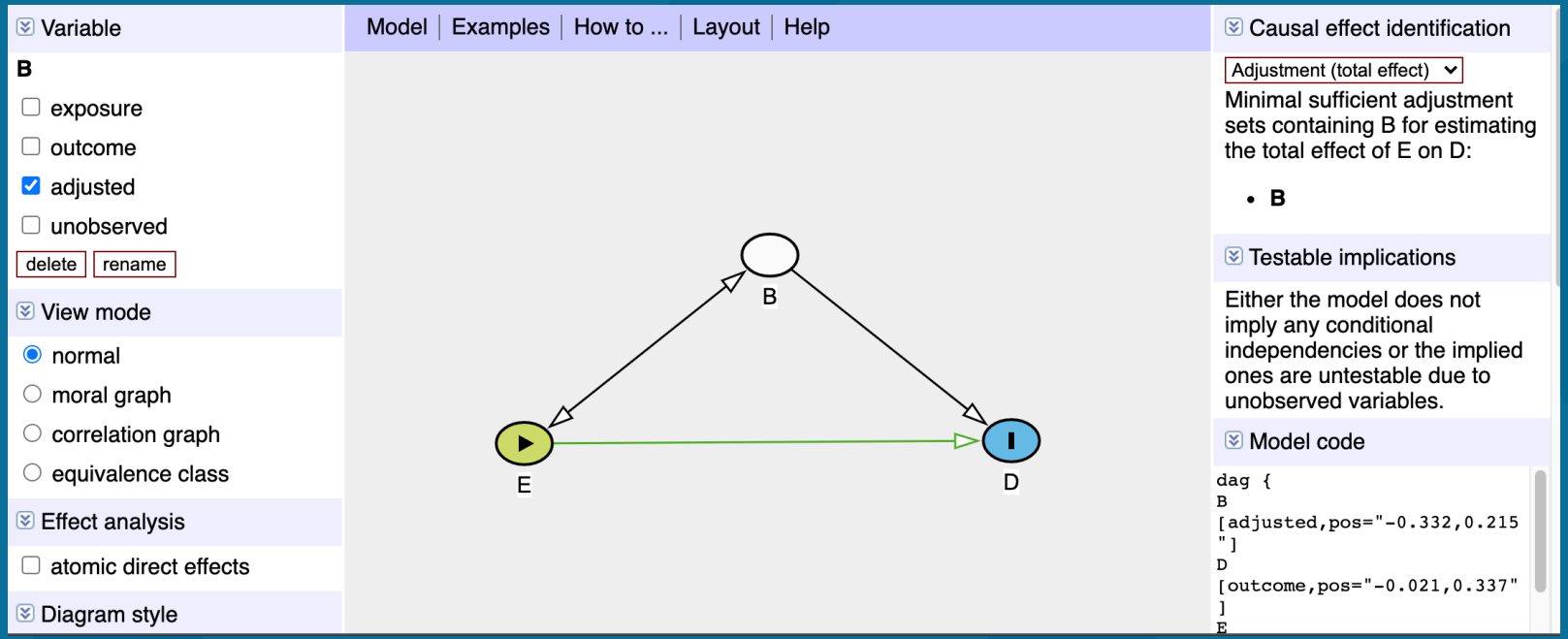
The following versions of DAGitty are available:

- Development version
  Recent development snapshot. May contain new features, but could also contain new bugs.
- Experimental version
  Most recent development snapshot. May not even work.
- 3.0: Released 2019-01-09
- 2.3: Released 2015-08-19
- 2.2: Released 2014-10-30
- 2.1: Released 2014-02-06
- 2.0: Released 2013-02-12
- 1.1: Released 2011-11-29
- 1.0: Released 2011-03-24
- 0.9b: Released 2010-11-24
- 0.9a: Released 2010-09-01

Changelog

http://www.dagitty.net/

DAGITTY (DRAW YOUR OWN DAG DIAGRAMS)—CONFOUNDER EX.



http://www.dagitty.net/

TYPES OF BIASES (1)

Bias is any systematic error in an epidemiologic study that results in an incorrect estimate of the association between exposure and the health outcome.

- Selection bias
  - Non-random sampling bias
  - Responder bias (Non-responder bias)
  - Berkson's bias
- Information bias
  - Measurement bias
  - Misclassification bias
  - Hawthorne effect

TYPES OF BIASES (2)

Table 1 Alphabetical list of biases, indicating their type and the design where they can occur

Specific name of bias	Group of bias	Subgroup of bias (next level to specific name)	Type of design affected
Allocation of intervention bias	Execution of an intervention		Trial
Apprehension bias	Information bias	Observer bias	All studies
Ascertainment bias	Selection bias	Inappropriate definition of the eligible population	Observational study
Berkson's bias	Selection bias	Inappropriate definition of the eligible population	Hospital based case-control study
Centripetal bias	Selection bias	Healthcare access bias	Observational study
Citation bias	Selection bias	Lack of accuracy of sampling frame	Systematic review/meta-analysis
Competing risks	Selection bias	Ascertainment bias	All studies
Compliance bias	Execution of an intervention		Trial
Confounding by group	Confounding		Ecological study
Confounding by indication	Confounding		Case-control study, cohort study
Contamination bias	Execution of an intervention		Trial, mainly community trials
Detection bias	Selection bias	Uneven diagnostic procedures in the target population	Case-control study
Detection bias	Information bias	Misclassification bias	Cohort study
Diagnostic/treatment access bias	Selection bias	Healthcare access bias	Observational study
Diagnostic suspicion bias	Selection bias	Detection bias	Case-control study
Diagnostic suspicion bias	Information bias	Detection bias	Cohort study
Differential maturing			Trial
Differential misclassification bias	Information bias	Misclassification bias	All studies
Dissemination bias	Selection bias	Lack of accuracy of sampling frame	Systematic review/meta-analysis
Ecological fallacy	Information bias		Ecological study
Exclusion bias	Selection bias	Inappropriate definition of the eligible population	Case-control study
Exposure suspicion bias	Information bias	Recall bias	Case-control study
Family aggregation bias	Information bias	Reporting bias	Observational study
Friend control bias	Selection bias	Inappropriate definition of the eligible population	Case-control study
Hawthorne effect	Information bias		Trial
Healthcare access bias	Selection bias	Ascertainment bias	Observational study
Healthy volunteer bias	Selection bias	Non-response bias	Observational study
Healthy worker effect	Selection bias	Inappropriate definition of the eligible population	Cohort study (mainly retrospective)

Journal of Epidemiology & Community Health 2004;**58**:635-641

THREE TYPES OF BIAS, THUS FAR…

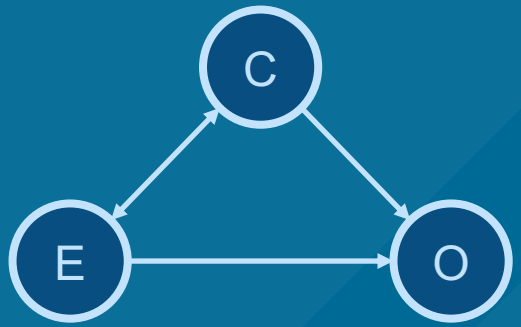
- Selection bias
  - Non-random sampling bias
  - Responder bias (Non-responder bias)
  - Berkson's bias
- Information bias
  - Measurement bias
  - Misclassification bias
  - Hawthorne effect
- Confounding

CONFOUNDER

A confounder is a distortion of the relationship between the exposure and outcome; it is a factor that is associated with the exposure and the outcome but not in the causal path between the exposure and outcome

A confounder has three criteria:

- Have an association with the outcome (must be a risk factor for the outcome)
- Have an association with the exposure
- Must not be in the causal path from the exposure to the outcome



STUDY DESIGN METHODS TO CONTROL FOR CONFOUNDING

Randomization: Eliminates links between the exposure and confounders (observed an unobserved)

Matching on the confounder: Balance the confounder between the groups

Restrict the sample: Focus on subjects with the same covariates (e.g., age, sex)

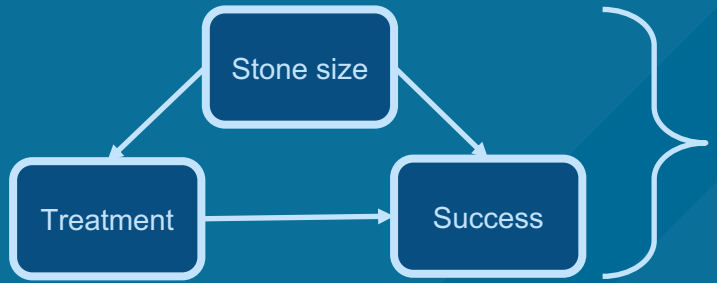
Statistical analyses: Mantel-Haenszel (MH) adjustment (stratification) and multivariable regression models

MOTIVATING EXAMPLE: SIMPSON'S PARADOX

Simpson's paradox is a statistical phenomena where the association of two factors may be independent or reversed when stratified into subgroups

Example from Charig (1986), which treatment is better at improving survival?

	Large stones	Small stones	Total
Old treatment	192 / 263 = 73%	81 / 87 = 93%	273 / 350 = 78%
New treatment	55 / 80 = 62%	234 / 270 = 83%	289 / 350 = 83%

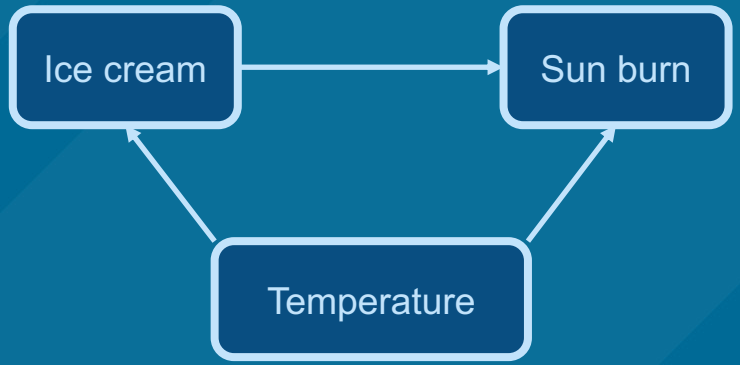


Surgeons were biased in their treatment selection; they preferred to treat patients with smaller stones with the newer treatment

MOTIVATING EXAMPLE: ICE CREAM AND SUNBURN

Let’s look at the proposed causal relationship

Does ice cream cause sunburn?



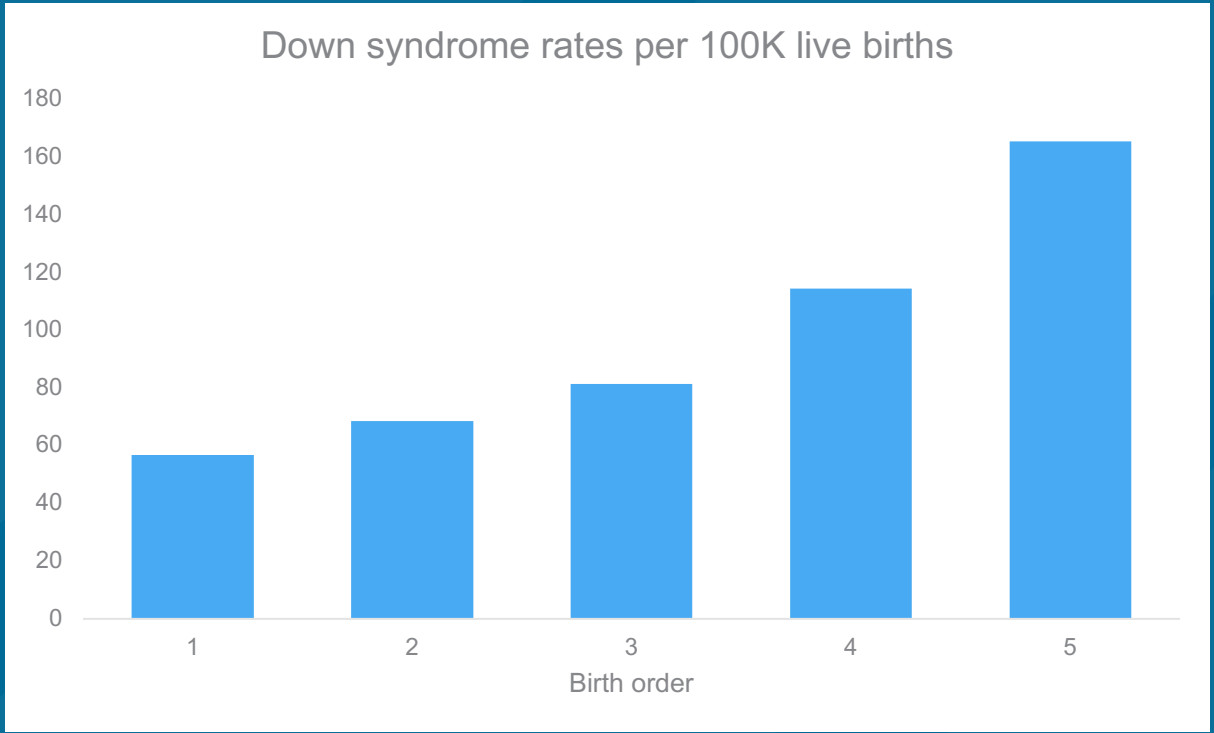
MOTIVATING EXAMPLE: DOWN SYNDROME AND BIRTH ORDER (1)

Let's look at the proposed causal relationship

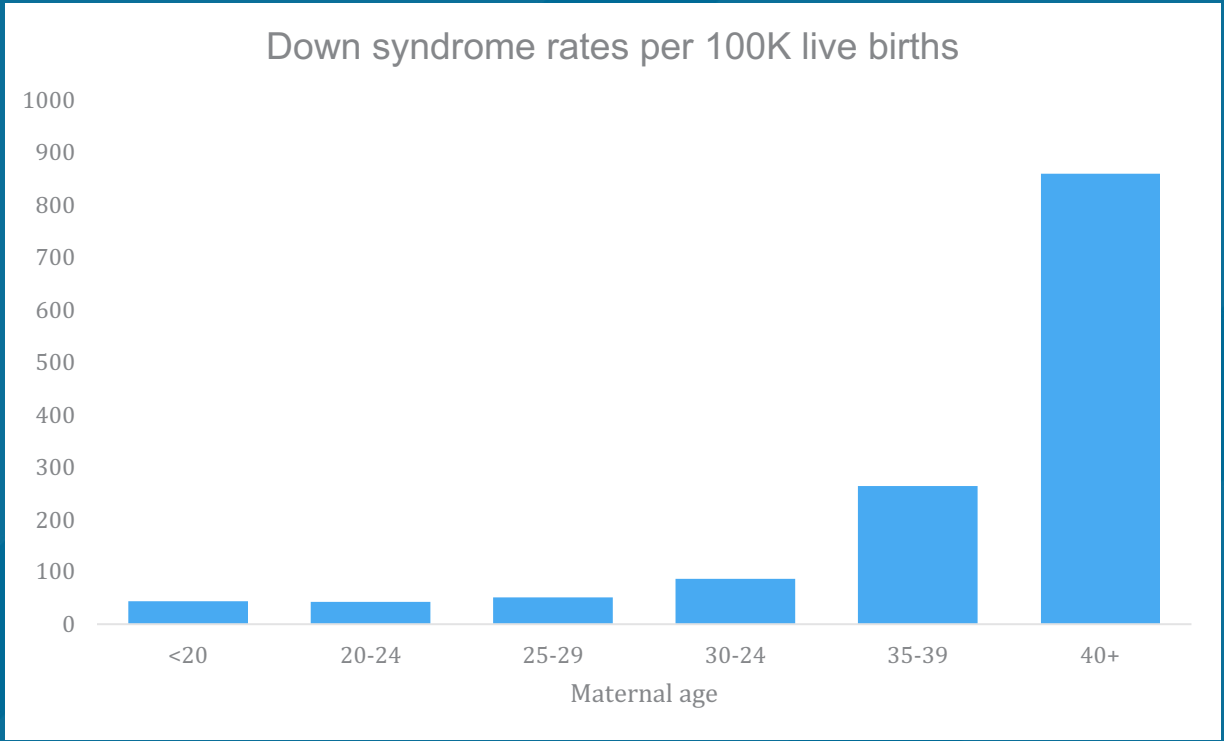
Does birth order cause Down syndrome?

Stark and Mantel (1996)

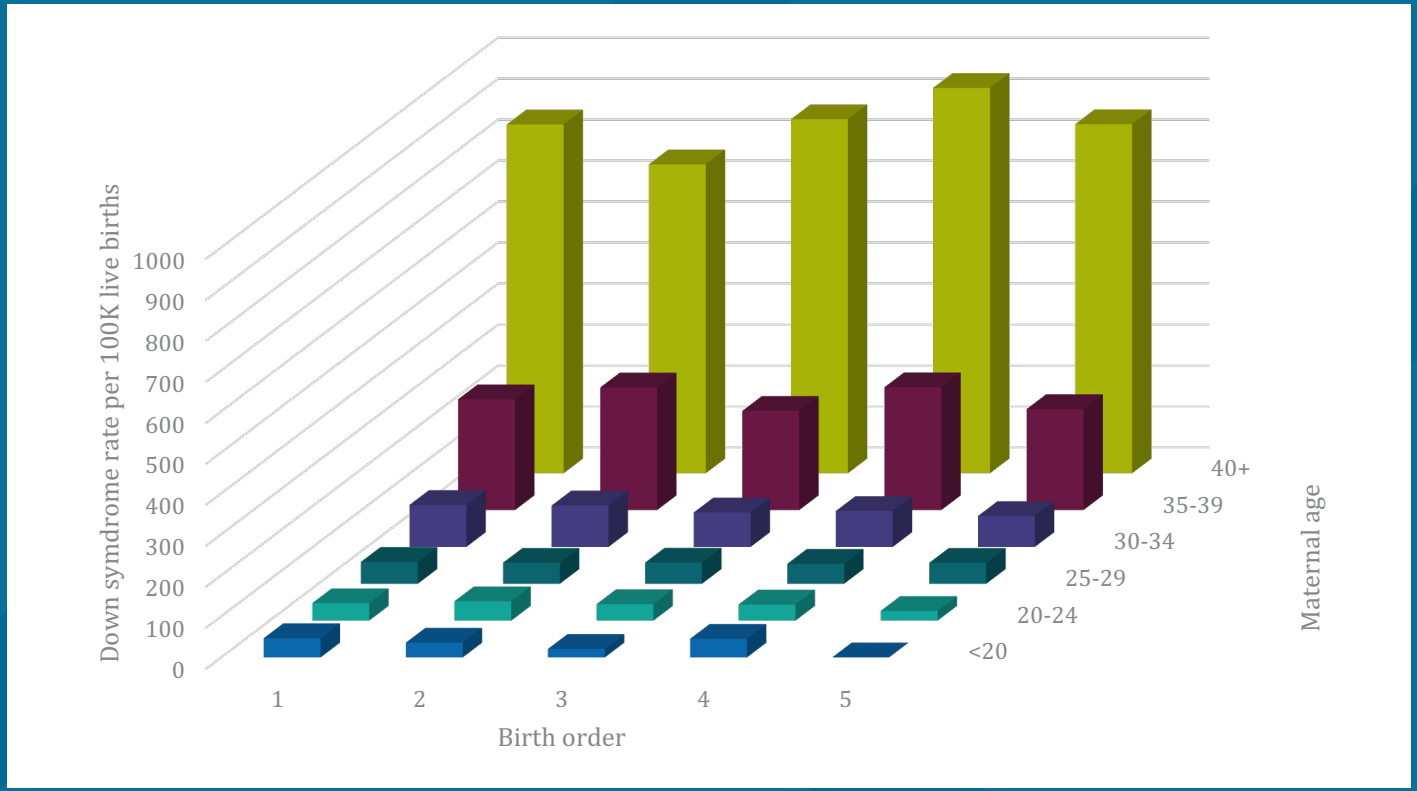
MOTIVATING EXAMPLE: DOWN SYNDROME AND BIRTH ORDER (2)



MOTIVATING EXAMPLE: DOWN SYNDROME AND BIRTH ORDER (3)



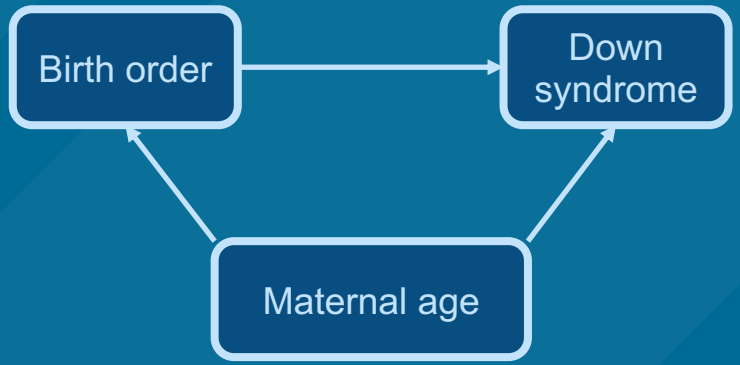
MOTIVATING EXAMPLE: DOWN SYNDROME AND BIRTH ORDER (4)



MOTIVATING EXAMPLE: DOWN SYNDROME AND BIRTH ORDER (5)

Let's look at the proposed causal relationship

Does birth order cause Down syndrome?



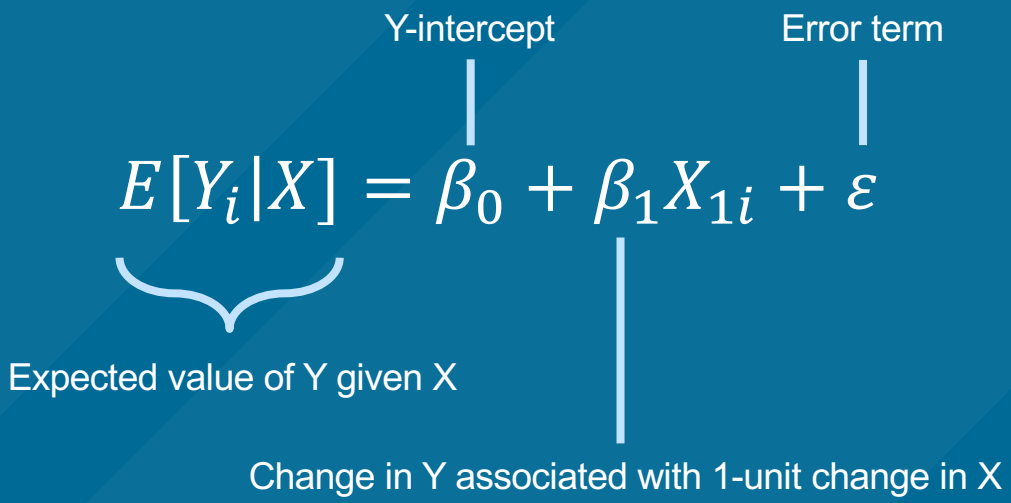
ADDRESSING CONFOUNDING WITH STATISTICS

Multivariable regression models

- Linear regression
- Logistic regression
- Poisson / Negative binomial regression
- Cox proportional hazard regression

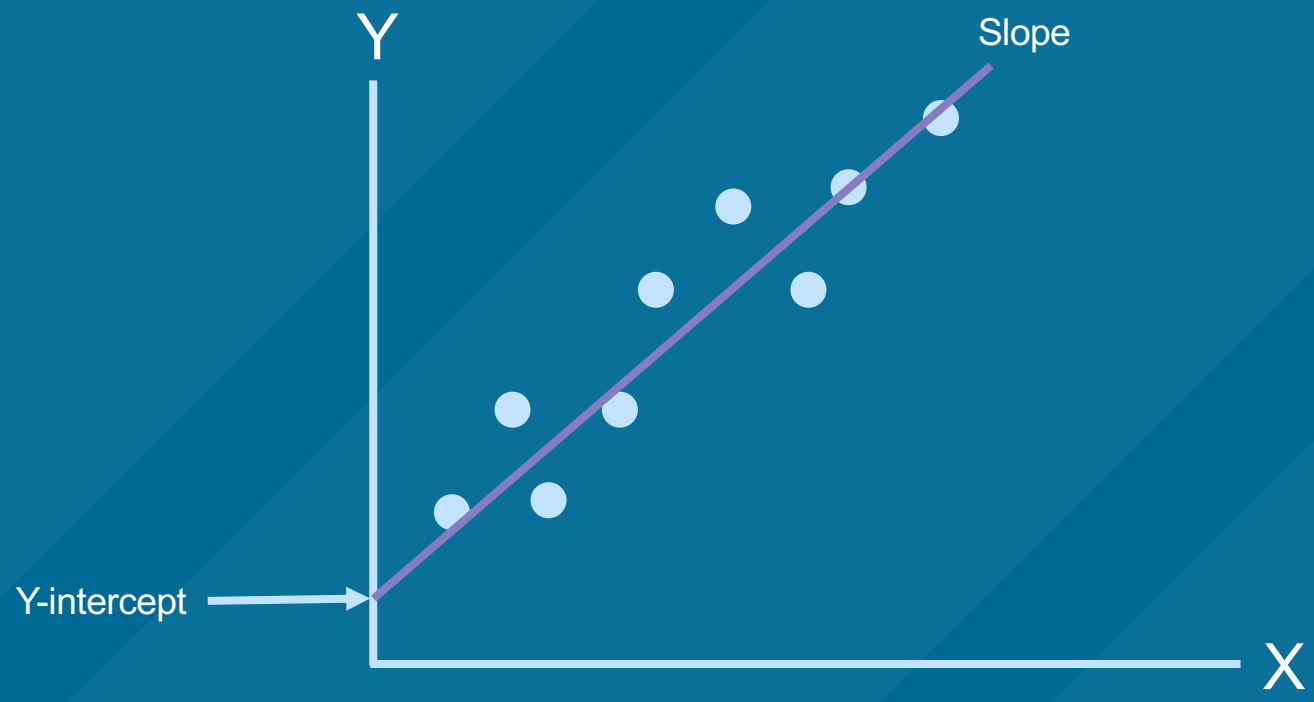
Propensity score matching

LINEAR REGRESSION—STRUCTURAL FORM



Y_i denotes the outcome (or dependent) variable for subject *i*
X_{1i} denotes the predictor of interest or the independent variable (X_1) for subject *i*
β_0 denotes the Y-intercept when X_1 is zero
β_1 denotes the slope or the change in Y with a 1-unit change in X_1
ε denotes the errors

LINEAR REGRESSION—VISUALIZATION



$$E[Y\ |X] = \beta_0 + \beta_1 X + \varepsilon$$

LINEAR REGRESSION—GLUCOSE ~ AGE (1)

Is Age associated with Glucose level?

Does a 4-unit change in Age significantly associated with the Glucose level?

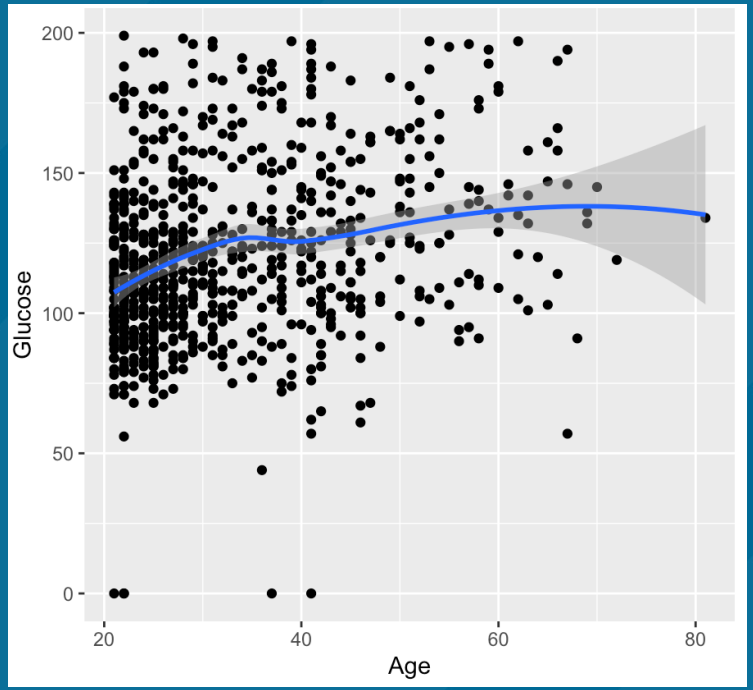
The null hypothesis is, there is no statistically significant association between Age and Glucose level

$$E[Y|X] = \beta_0 + \beta_1 X + \varepsilon$$

$$E[Glucose|Age] = \beta_0 + \beta_1(Age) + \varepsilon$$

LINEAR REGRESSION—GLUCOSE ~ AGE (2)

```
### Plot the association between the subject's age and glucose level
ggplot(diabetes.data, aes(x = Age, y = Glucose)) +
  geom_point() +
  stat_smooth()
```



LINEAR REGRESSION—GLUCOSE ~ AGE (3)

Is Age associated with Glucose level?

Does a 4-unit change in Age significantly associated with the Glucose level?

The null hypothesis is, there is no statistically significant association between Age and Glucose level

$$Glucose = \beta_0 + \beta_1(19) + \varepsilon$$

$$Glucose = \beta_0 + \beta_1(15) + \varepsilon$$

$$\Delta Glucose_{19-15} = \beta_1(19 - 15)$$

LINEAR REGRESSION—GLUCOSE ~ AGE (4)

```
> ### Linear regression model (Y = Age, X = Glucose)
> linear.model1 <- lm(Glucose ~ Age, data = diabetes.data)
> summary(linear.model1)

Call:
lm(formula = Glucose ~ Age, data = diabetes.data)

Residuals:
     Min       1Q   Median       3Q      Max
-126.453  -20.849   -3.058   18.304   86.159

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 97.08016    3.34095   29.06  < 2e-16 ***
Age          0.71642    0.09476    7.56 1.15e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.86 on 766 degrees of freedom
Multiple R-squared:  0.06944,   Adjusted R-squared:  0.06822
F-statistic: 57.16 on 1 and 766 DF,  p-value: 1.15e-13

> confint(linear.model1)
                2.5 %      97.5 %
(Intercept) 90.5216601 103.6386585
Age          0.5304001   0.9024361
```

(Link to R tutorial)

LINEAR REGRESSION—GLUCOSE ~ AGE (5)

$$Glucose = \beta_0 + \beta_1(19) + \varepsilon$$
$$Glucose = \beta_0 + \beta_1(15) + \varepsilon$$

$$\Delta Glucose_{19-15} = \beta_1(19 - 15)$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 97.08016    3.34095   29.06  < 2e-16 ***
Age          0.71642    0.09476    7.56 1.15e-13 ***
```

$$\beta_0 = 97.08$$
$$\beta_1 = 0.72$$

Glucose level for 19-year old

$$Glucose = 97.08 + 0.72(19)$$
$$110.76 = 97.08 + 0.72(19)$$

Glucose level for 15-year old

$$Glucose = 97.08 + 0.72(15)$$
$$107.88 = 97.08 + 0.72(15)$$

$$110.76 - 107.88 = 2.88$$

$$\Delta Glucose_{19-15} = 0.72 * 4 = 2.88$$

LINEAR REGRESSION—GLUCOSE ~ PREGNANCY (1)

Is Pregnancy associated with Glucose level?

Does a 1-unit change in Pregnancy (0 → 1) significantly associated with the Glucose level?

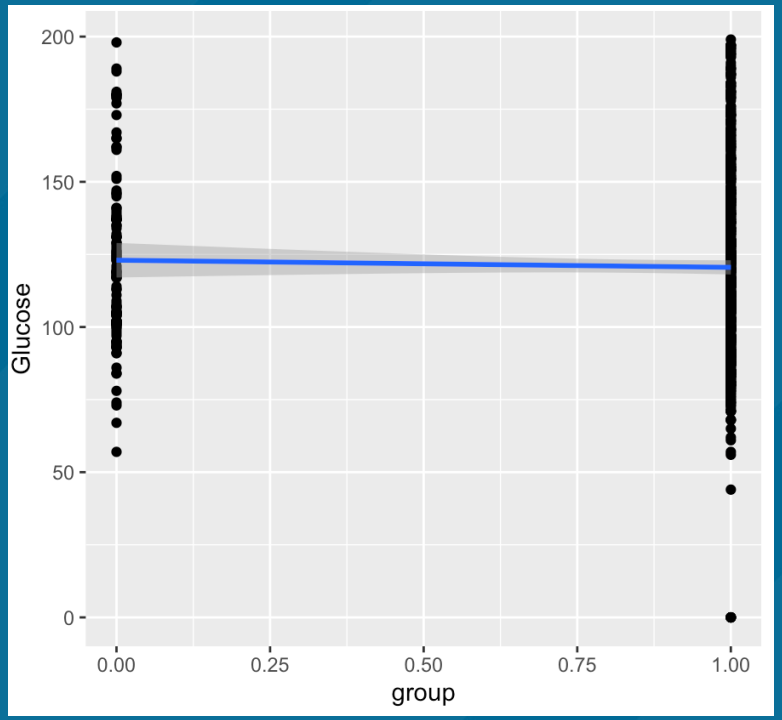
The null hypothesis is, there is no statistically significant association between Pregnancy and Glucose level

$$E[Y\ |X] = \beta_0 + \beta_1 X + \varepsilon$$

$$E[Glucose|Pregnancy] = \ \beta_0 + \beta_1(Pregnancy) + \varepsilon$$

LINEAR REGRESSION—GLUCOSE ~ PREGNANCY (2)

```
### Plot the association between the subject's pregnancy history and glucose level
ggplot(diabetes.data, aes(x = group, y = Glucose)) +
  geom_point() +
  stat_smooth(method = "lm")
```



LINEAR REGRESSION—GLUCOSE ~ PREGNANCY (3)

Is Pregnancy associated with Glucose level?

Does a 1-unit change in Pregnancy (0 → 1) significantly associated with the Glucose level?

The null hypothesis is, there is no statistically significant association between Pregnancy and Glucose level

$$Glucose = \beta_0 + \beta_1(1) + \varepsilon$$

$$Glucose = \beta_0 + \beta_1(0) + \varepsilon$$

$$\Delta Glucose_{1-0} = \beta_1(1 - 0)$$

$$\Delta Glucose_1 = \beta_1(1)$$

LINEAR REGRESSION—GLUCOSE ~ PREGNANCY (4)

Group represents Pregnancy and is a dichotomous variable
(0 = Not pregnant and 1 = Pregnant)

```
> ### Linear model (Glucose = B0 + B1(Pregnancy))
> linear.model2 <- lm(Glucose ~ group, data = diabetes.data)
> summary(linear.model2)

Call:
lm(formula = Glucose ~ group, data = diabetes.data)

Residuals:
     Min       1Q   Median       3Q      Max
-120.539  -21.539   -4.539   19.461   78.461

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  123.000      3.036   40.52   <2e-16 ***
group         -2.461      3.282   -0.75    0.454
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.98 on 766 degrees of freedom
Multiple R-squared:  0.0007336, Adjusted R-squared:  -0.0005709
F-statistic: 0.5624 on 1 and 766 DF,  p-value: 0.4535

> confint(linear.model2)
                2.5 %     97.5 %
(Intercept) 117.040971 128.959029
group        -8.903969   3.981595
```

LINEAR REGRESSION—GLUCOSE ~ PREGNANCY (5)

$$Glucose = \beta_0 + \beta_1(1) + \varepsilon$$
$$Glucose = \beta_0 + \beta_1(0) + \varepsilon$$

$$\Delta Glucose_1 = \beta_1(1)$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  123.000      3.036   40.52   <2e-16 ***
group         -2.461      3.282   -0.75    0.454
```

$$\beta_0 = 123.00$$
$$\beta_1 = -2.46$$

Glucose level for subjects with pregnancy

$$Glucose = 123.00 - 2.46(1)$$
$$120.54 = 123.00 - 2.46(1)$$

Glucose level for subjects without pregnancy

$$Glucose = 123.00 - 2.46(0)$$
$$123.00 = 123.00$$

$$120.54 - 123.00 = -2.46$$

$$\Delta Glucose_1 = -2.46$$

LINEAR REGRESSION—GLUCOSE ~ AGE + PREGNANCY (1)

Is Age associated with Glucose level controlling for Pregnancy history?

Does a 1-unit change in Age significantly associated with the Glucose level controlling for Pregnancy history?

The null hypothesis is, there is no statistically significant association between Age and Glucose level controlling for Pregnancy history

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$E[Glucose|Age, Pregnancy] = \beta_0 + \beta_1(Age) + \beta_2(Pregnancy) + \varepsilon$$

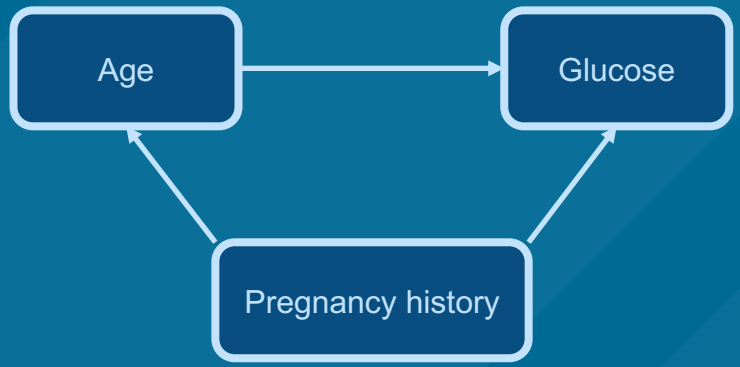
This is an example of a multivariable linear regression model

LINEAR REGRESSION—GLUCOSE ~ AGE + PREGNANCY (2)

Is Age associated with Glucose level controlling for Pregnancy history?

Does a 1-unit change in Age significantly associated with the Glucose level controlling for Pregnancy history?

The null hypothesis is, there is no statistically significant association between Age and Glucose level controlling for Pregnancy history



LINEAR REGRESSION—GLUCOSE ~ AGE + PREGNANCY (3)

Group represents Pregnancy and is a dichotomous variable
(0 = Not pregnant and 1 = Pregnant)

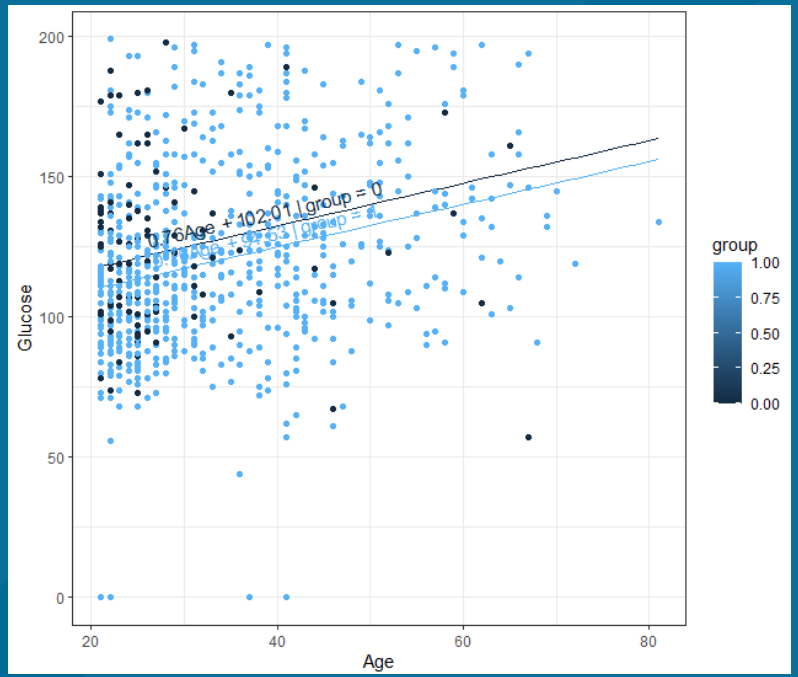
```
Call:
lm(formula = Glucose ~ Age + group, data = diabetes.data)

Residuals:
     Min       1Q   Median       3Q      Max
-125.715  -20.546   -2.991   17.316   87.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.00752    3.95100  25.818  < 2e-16 ***
Age           0.76050    0.09638   7.891 1.04e-14 ***
group        -7.47264    3.22137  -2.320   0.0206 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.77 on 765 degrees of freedom
Multiple R-squared:  0.07594,   Adjusted R-squared:  0.07352
F-statistic: 31.43 on 2 and 765 DF,  p-value: 7.592e-14
```

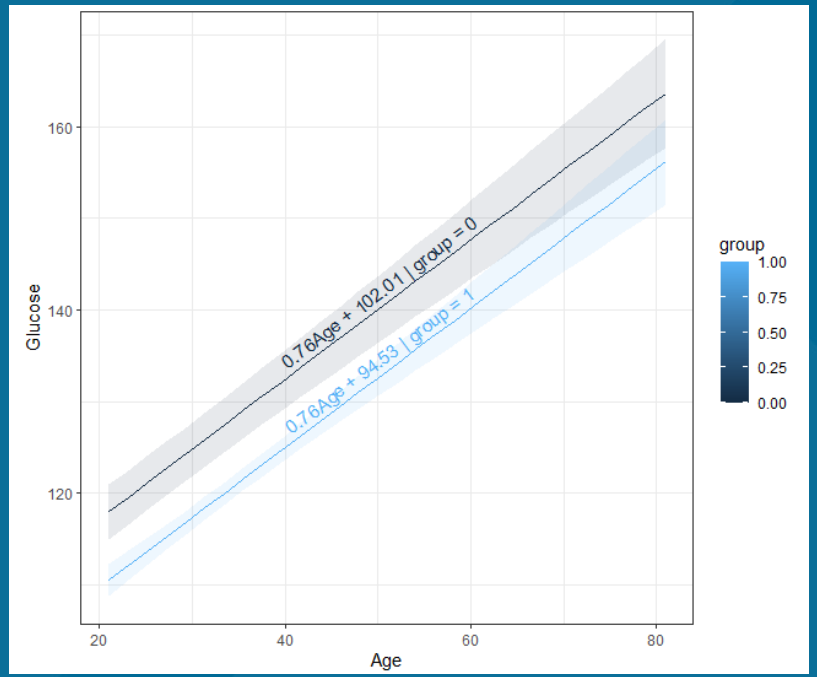
```
> confint(linear.model3)
                2.5 %      97.5 %
(Intercept) 94.2514334 109.7636012
Age          0.5712954   0.9497004
group      -13.7964201  -1.1488593
```

$\beta_0 = 102.01$
$\beta_1 = 0.76$
$\beta_2 = -7.47$

LINEAR REGRESSION—GLUCOSE ~ AGE + PREGNANCY (4)

```
#### Visual the linear regression model
ggPredict(linear.model3, digits = 2)
ggPredict(linear.model3, digits = 2, show.point = FALSE, se = TRUE, xpos = 0.5)
```





LINEAR REGRESSION—GLUCOSE ~ AGE + PREGNANCY (5)

```
Call:
lm(formula = Glucose ~ Age + group, data = diabetes.data)

Residuals:
     Min       1Q   Median       3Q      Max
-125.715  -20.546   -2.991   17.316   87.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.00752    3.95100  25.818  < 2e-16 ***
Age           0.76050    0.09638   7.891 1.04e-14 ***
group        -7.47264    3.22137  -2.320   0.0206 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.77 on 765 degrees of freedom
Multiple R-squared:  0.07594,   Adjusted R-squared:  0.07352
F-statistic: 31.43 on 2 and 765 DF,  p-value: 7.592e-14
```

```
> confint(linear.model3)
                  2.5 %      97.5 %
(Intercept)  94.2514334 109.7636012
Age           0.5712954   0.9497004
group       -13.7964201  -1.1488593
```

According to the linear regression model, subjects who were older by 1 year in Age had, on average, a 0.76-point increase in glucose levels (95% CI: 0.57, 0.95) controlling for Pregnancy history.

LINEAR REGRESSION—WHAT DOES ‘CONTROLLING’ OR ‘ADJUSTING’ MEAN? (1)

According to the linear regression model, subjects who were older by 1 year in Age had, on average, a 0.77-point increase in glucose levels (95% CI: 0.57, 0.95) controlling for Pregnancy history.

What does “controlling” mean?

Table 3. Linear regression model output with confounder (Glucose ~ Age + Pregnancy History)

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	102	94, 110	<0.001
Age	0.76	0.57, 0.95	<0.001
pregnancy.history	-7.5	-14, -1.1	0.021

[1] CI = Confidence Interval

$$E[Glucose|Age, Pregnancy] = \beta_0 + \beta_1(Age) + \beta_2(Pregnancy) + \varepsilon$$

$$Glucose = 102 + 0.76(Age) - 7.5(Pregnancy) + \varepsilon$$

glucose	age	Preg.hx
117.3	30	1
121.1	35	1
124.9	40	1
glucose	age	Preg.hx
124.8	30	0
128.6	35	0
132.4	40	0

3.8 = 5*0.76

3.8 = 5*0.76

LINEAR REGRESSION—WHAT DOES ‘CONTROLLING’ OR ‘ADJUSTING’ MEAN? (2)

According to the linear regression model, subjects who were older by year in Age had, on average, a 0.77-point increase in glucose levels (95% CI: 0.57, 0.95) controlling for Pregnancy history.

What does “controlling” mean?

Table 3. Linear regression model output with confounder (Glucose ~ Age + Pregnancy History)

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	102	94, 110	<0.001
Age	0.76	0.57, 0.95	<0.001
pregnancy.history	-7.5	-14, -1.1	0.021

[1] CI = Confidence Interval

$$E[Glucose|Age, Pregnancy] = \beta_0 + \beta_1(Age) + \beta_2(Pregnancy) + \varepsilon$$

$$Glucose = 102 + 0.76(Age) - 7.5(Pregnancy) + \varepsilon$$

glucose	age	Preg.hx
117.3	30	1
121.1	35	1
124.9	40	1
glucose	age	Preg.h
124.8	30	0
128.6	35	0
132.4	40	0

-7.5 = -7.5

-7.5 = -7.5

MODEL COMPARISONS

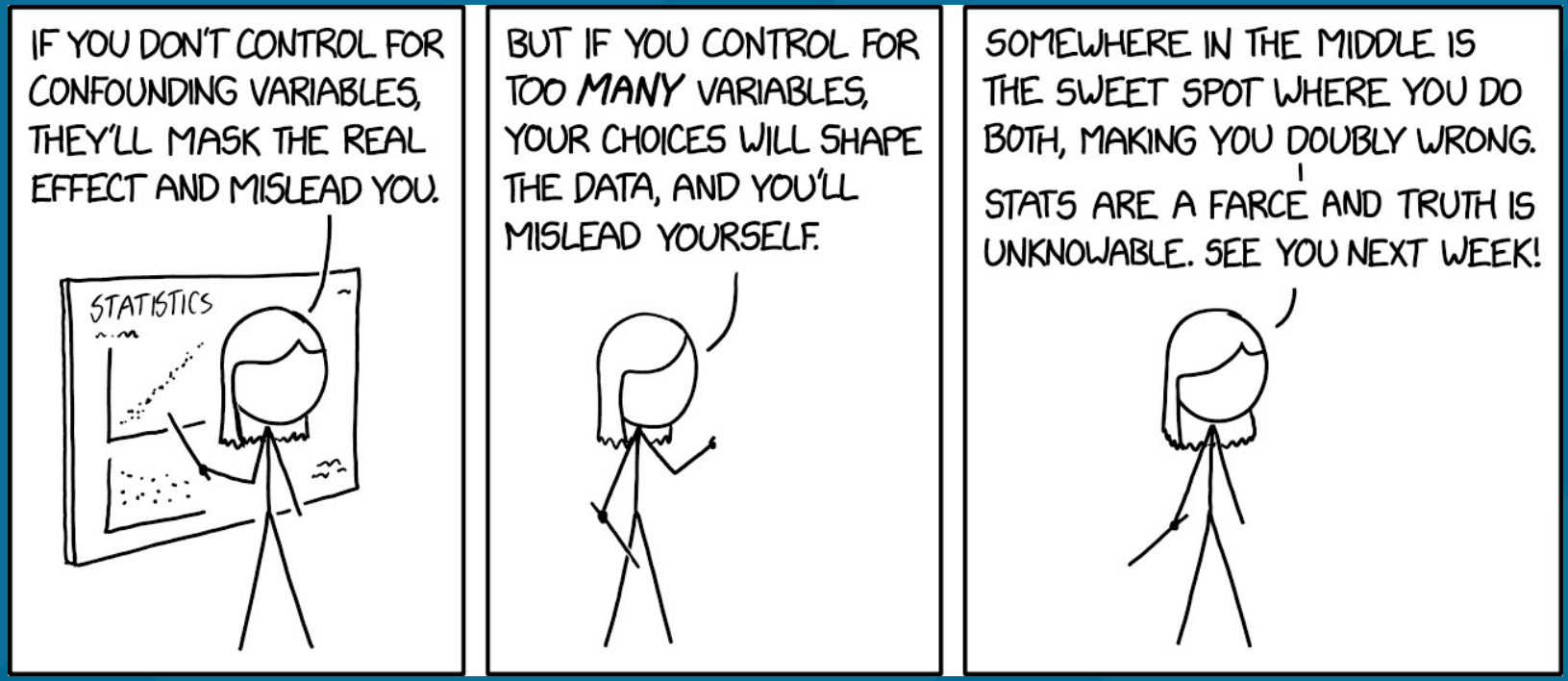
Parameters	Model 1 (Glucose ~ Age)	Model 2 (Glucose ~ Pregnancy)	Model 3 (Glucose ~ Pregnancy + Age)
Y-intercept (β_0)	97.08	123.00	102.01
Age coefficient	0.72	---	0.76
Pregnancy coefficient	---	-2.46	-7.47
R2	0.068	-0.001	0.074

$$Model\ 1\text{: } E[Glucose|Age] = \beta_0 + \beta_1(Age) + \varepsilon$$

$$Model\ 2\text{: } E[Glucose|Pregnancy] = \beta_0 + \beta_1(Pregnancy) + \varepsilon$$

$$Model\ 3\text{: } E[Glucose|\ Age, Pregnancy] = \beta_0 + \beta_1(Age) + \beta_2(Pregnancy) + \varepsilon$$

CONTROLLING FOR CONFOUNDING

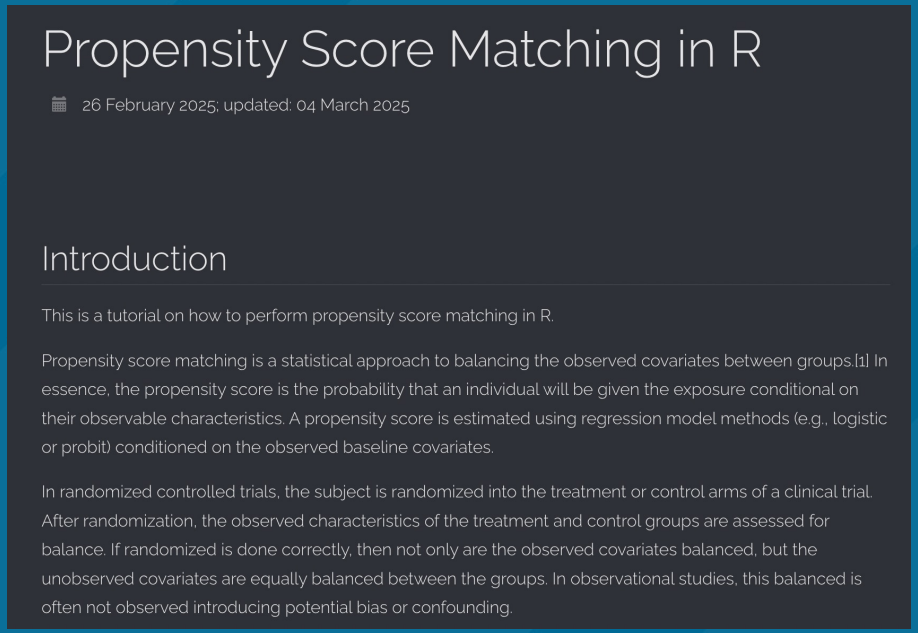


https://www.explainxkcd.com/wiki/index.php/2560:_Confounding_Variables

ANOTHER METHOD TO ADDRESS CONFOUNDING

Propensity score matching

(R Tutorial on propensity score matching methods)



Propensity Score Matching in R

26 February 2025; updated: 04 March 2025

Introduction

This is a tutorial on how to perform propensity score matching in R.

Propensity score matching is a statistical approach to balancing the observed covariates between groups.[1] In essence, the propensity score is the probability that an individual will be given the exposure conditional on their observable characteristics. A propensity score is estimated using regression model methods (e.g., logistic or probit) conditioned on the observed baseline covariates.

In randomized controlled trials, the subject is randomized into the treatment or control arms of a clinical trial. After randomization, the observed characteristics of the treatment and control groups are assessed for balance. If randomized is done correctly, then not only are the observed covariates balanced, but the unobserved covariates are equally balanced between the groups. In observational studies, this balanced is often not observed introducing potential bias or confounding.

ADDITIONAL RESOURCES

Linear Regression Chapter: Learning Statistics with R (book)

Linear Regression in R tutorial

Propensity score matching in R tutorial