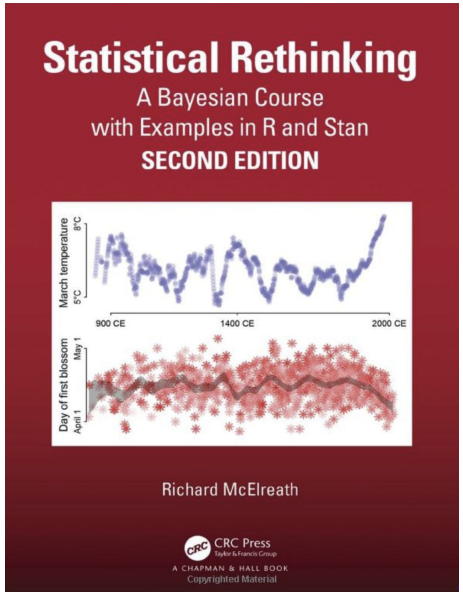
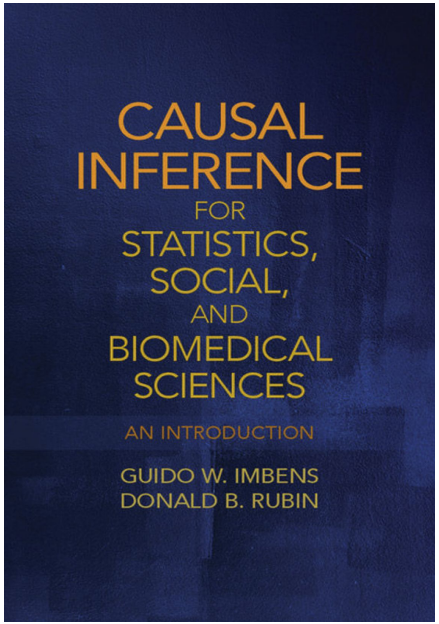
Introduction to Data Science

Causal Inference

Prof. Dr. Ralf Lämmel & **M.Sc. Johannes Härtel**
(johanneshaertel@uni-koblenz.de)



[McElreath20]

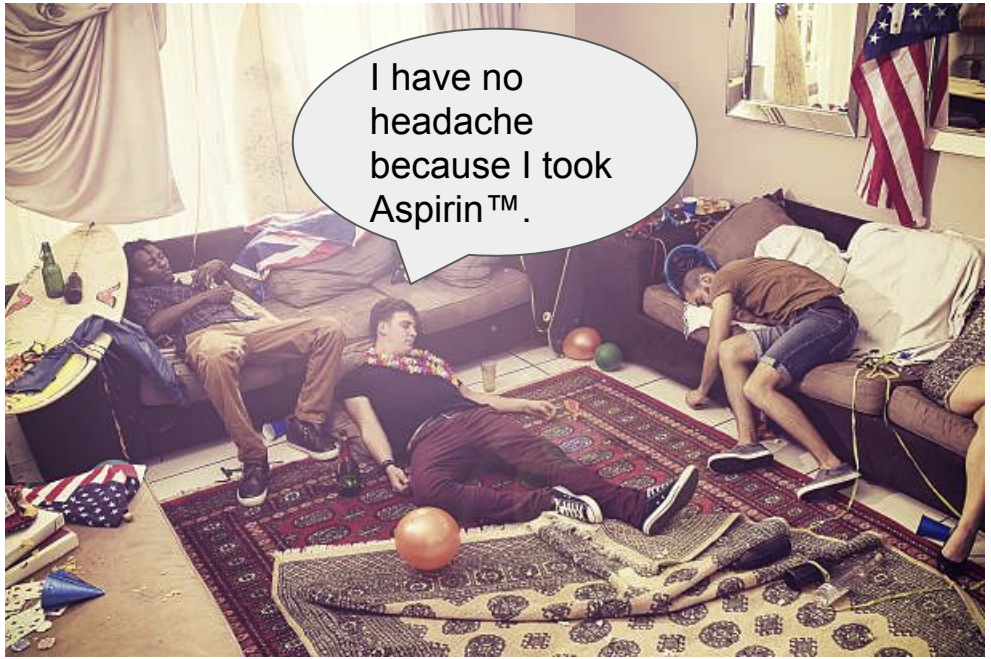


[ImbensR15]

The major source for this lecture.

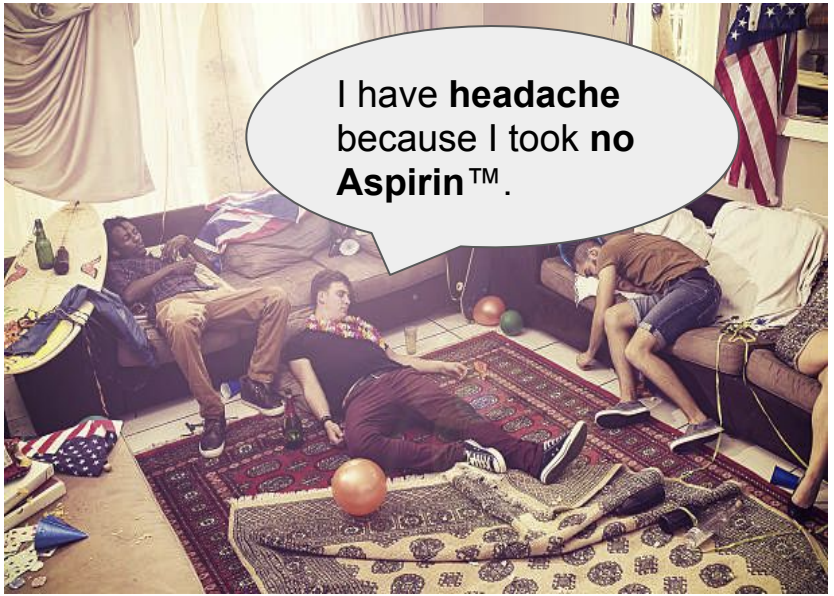
“In many applications of statistics, a large proportion of the questions of interest are fundamentally questions of causality rather than simply questions of description or association” [ImbensR15] (see Preface)

Example

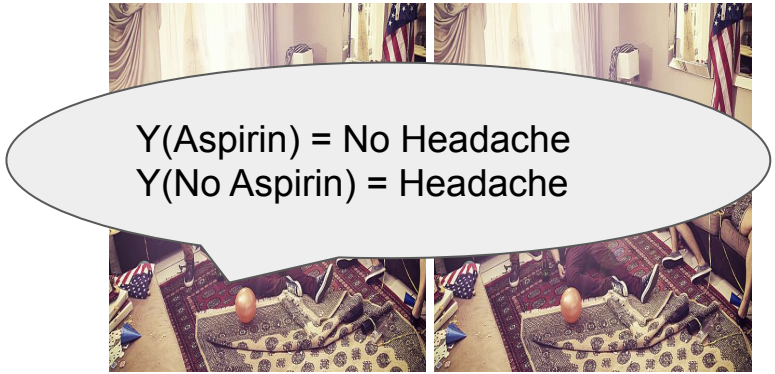


Defined causation comparing two potential outcomes:

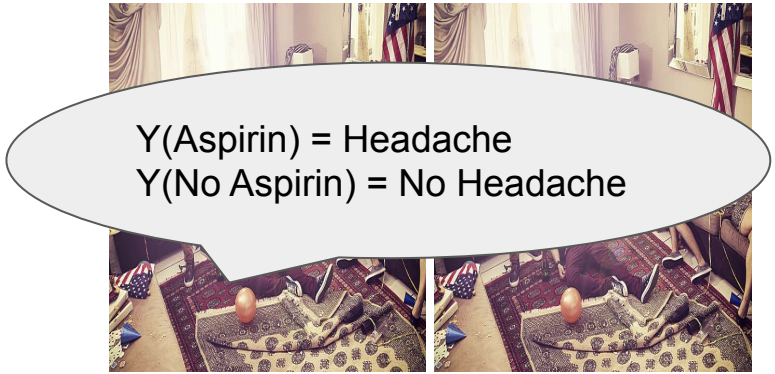
However, at most one outcome can be realized and observed (see [ImbensR15]).



Comparisons of potential outcomes



zero causal effect



zero causal effect

“… the fundamental problem of causal inference is the presence of **missing data”** [ImbensR15]

Defining Causation (according to [ImbensR15])

- We have **units** (subject or entities), **treatments** (modifications, manipulation, actions, or interventions), and **potential outcomes** for each treatment on a unit.
- **At most one** of the potential outcomes can be observed.
- We define causation as the **comparison of the potential outcome** for the **same unit**.
- This definition does not depend on **which of both outcome** we observe. However, this is a problem for estimation as we **miss data.**

Estimating causation:

Assumptions under which causation can be examined.

- We require **multiple units** and the **stable unit treatment value assumption** (STUVA).
  - Units do not interfere with each other.
  - No Hidden Variations of Treatments.
- We require an **assignment mechanism** that decides which units receive which treatment (formally a function of all covariates and of all potential outcomes):
  - **Individualistic assignment**: This limits the dependence of a particular unit's assignment probability on the values of covariates and potential outcomes for other units.
  - **Probabilistic assignment:** This requires the assignment mechanism to imply a nonzero probability for each treatment value, for every unit.
  - **Unconfounded assignment:** This disallows dependence of the assignment mechanism on the potential outcomes.

(parts copied from [ImbensR15])

Experimental and observation studies

> *“[…] we also make a distinction between **experiments**, where the **assignment mechanism** is both **known and controlled** by the researcher, and **observational studies**, where the **assignment mechanism** is **not known** to, or **not under the control** of, the researcher.”* [ImbensR15]

- **Classical randomized experiments f**ulfills all three restrictions on the assignment process by design.
- **In Observational studies,** the restrictions may hold but are assumptions, rather than satisfied by design.

Three ways of using **covariates variables** in an analysis

Covariates are not of direct interest, but we can use them:

- To make estimates **more precise** (explaining some variation in outcomes)
- Examine the causal effect of the treatment on **subgroups** (as defined by a covariate). This can be described by **interactions**.
- To account for their effect on the **assignment mechanism**.

See [ImbensR15] (Page 16)

Directed Acyclic Graphs (DAGs):
We will represent causal relationships among variables by DAGs.

- The **dag** describes process of causation on an abstract level, i.e., consequences of treatment (intervention, action, or modification) if the DAG is correct.
  - **Nodes:** Observed or unobserved variables. Unobserved variables are depicted by circles.
  - **Edges:** Directions of influence (say “directly influences”)
- Different sorts of effects:
  - Indirect effect: X -> Z -> Y
  - Direct effect: X -> Y
- We can describe conditional independence, if we have X -> Z -> Y but not X -> Y.

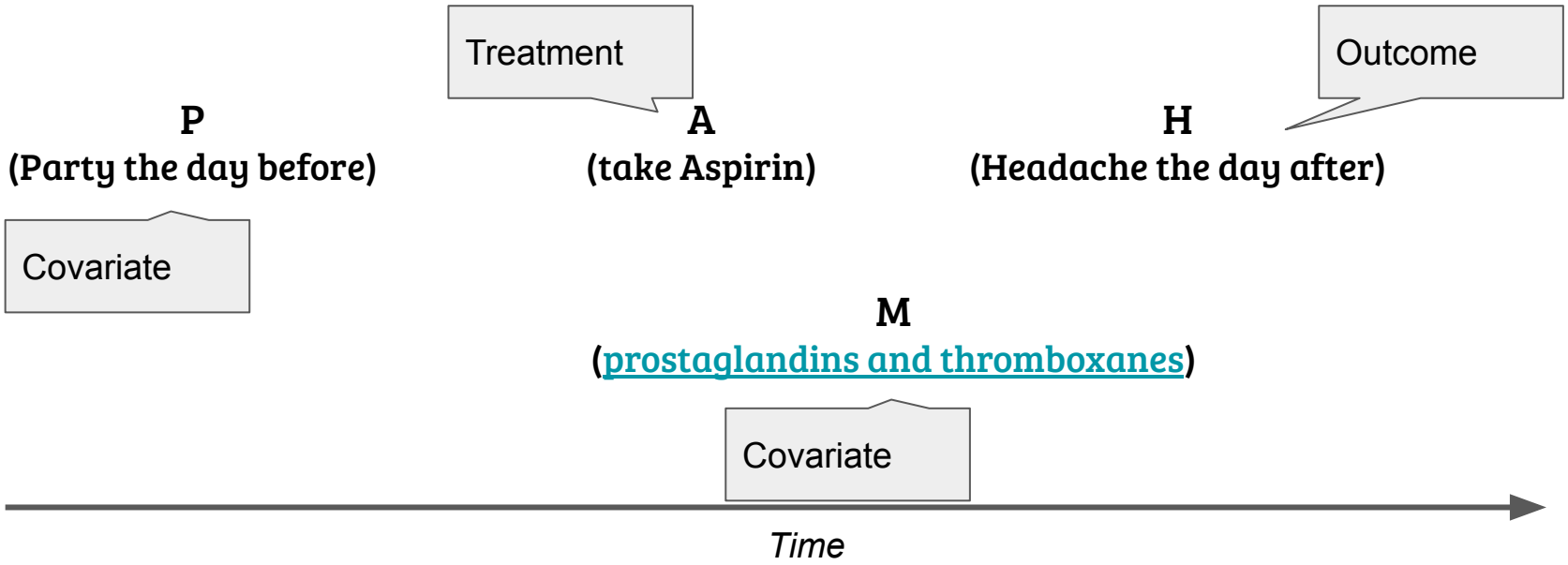
Simulations

- Simulations (how we will use them) **implement more concrete processes** that follow the DAG's structure describing causation.
- The simulations **exemplify the assumptions** under which causation can be examined.
- The simulations are **more concrete** than DAGs, coding specific relations between variables. There is a one-to-many relation between DAGs and simulations.
- IMPORTANT: Simulated data cannot be used to answer real questions. For real questions, we need real data.

Example

Example: Treatment, outcome and covariates

We have different variables and a temporal ordering how they have been recorded.



Scenarios 1: Randomized experiment

(Concrete problems and solutions)

Scenario 1 (dag): Randomized experiment

The basic case without any covariates.

Scenario 1 (sim): Randomized experiment

We first simulate the process underlying the DAG.

```
# Assignment mechanisms (Aspirin or not).
A <- rbinom(N, 1, 0.5)

# We simulate headache caused by taking no aspirin.
mu <- -0.4 - 0.2 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

Simulating the **random assignment** of N students.

In this simulation, we assume that aspirin **decreases headache** (by a factor of -0.2)

Simulating the **degree of headache**.

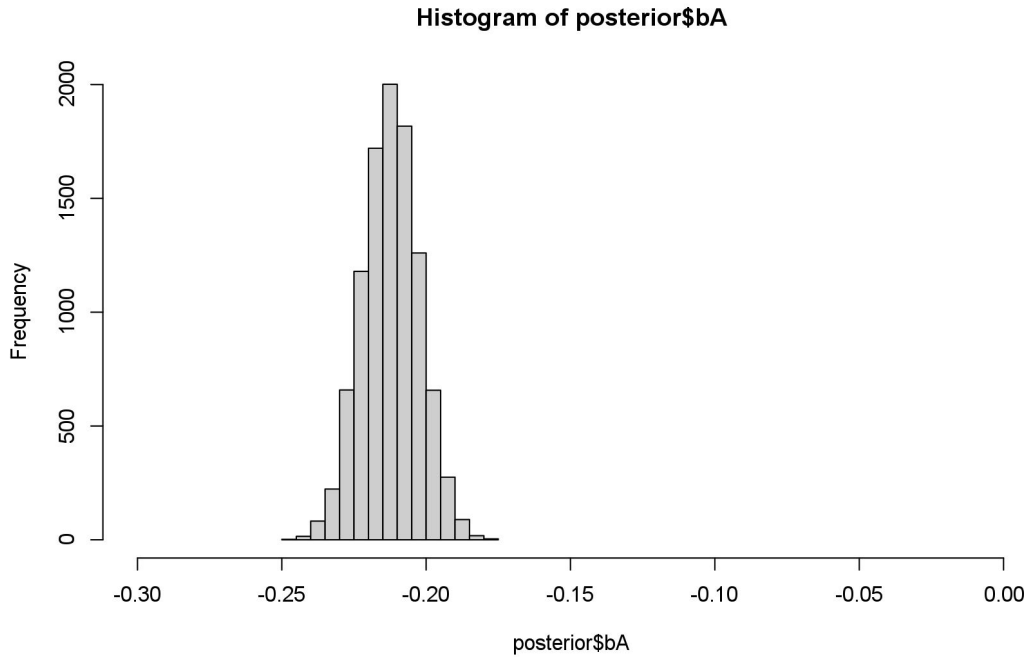
Scenario 1 (model): Randomized experiment

We can **estimate** the effect of aspirin A on headache H **as a parameter β_A** in a basic linear model.

$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0,1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

Scenario 1 (results): Randomized experiment

The posterior of parameter β_A comes close to the simulated effect of aspirin (≈ -0.2).

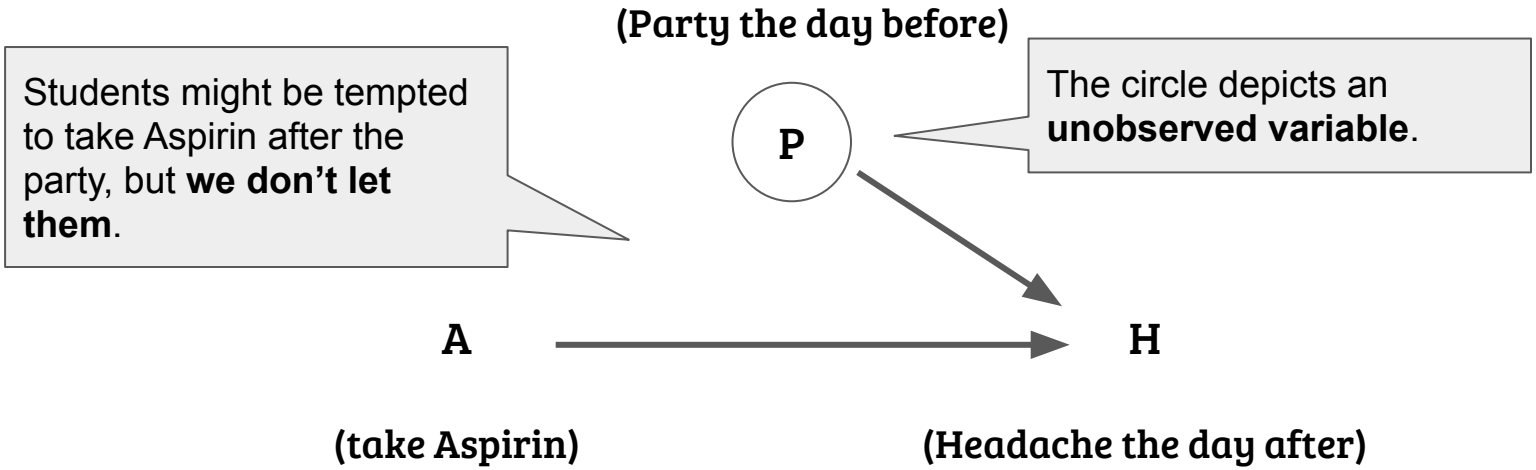


Scenarios 2: Randomized experiment and an unobserved covariate

(Concrete problems and solutions)

Scenario 2 (dag): Randomized experiment and an unobserved covariate

The **assignment mechanism** of a classical randomized experiment, **which we have control of,** protects us against the influence of the unobserved variable '*Party the day before' (P)*.



Scenario 2 (sim): Randomized experiment and an unobserved covariate

We first simulate the underlying process, adding the party P and an effect on the outcome headache H.

```
# Party or not.
P <- rbinom(N, 1, 0.5)

# Assignment mechanisms (Aspirin or not).
A <- rbinom(N, 1, 0.5)

# We simulate headache caused by taking no aspirin and party.
mu <- -0.3 + 0.3 * P - 0.2 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

Simulating the **unobserved covariate** indicating on Party P.

We see that there is no influence of P on A.

The party relates to the headache (with an effect of +0.3).

Scenario 2 (model): Randomized experiment and an unobserved covariate

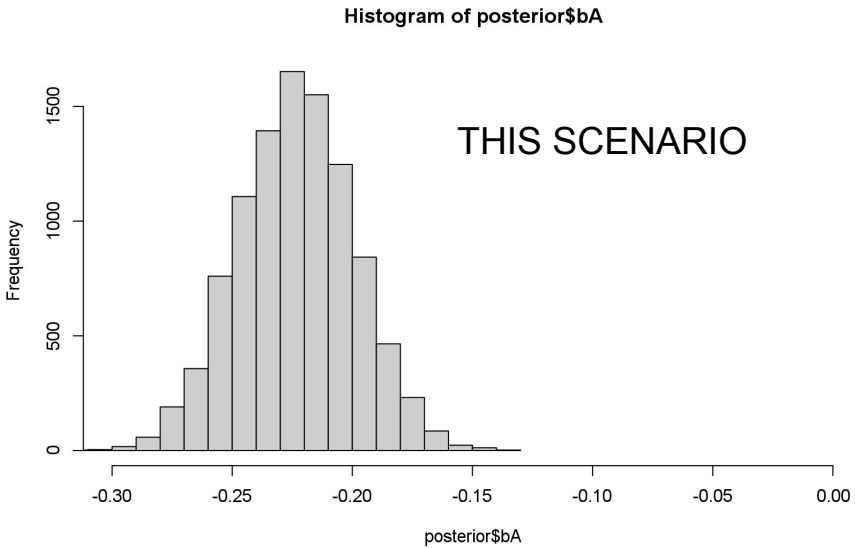
The model did not change, since we do no not observer the covariate P.

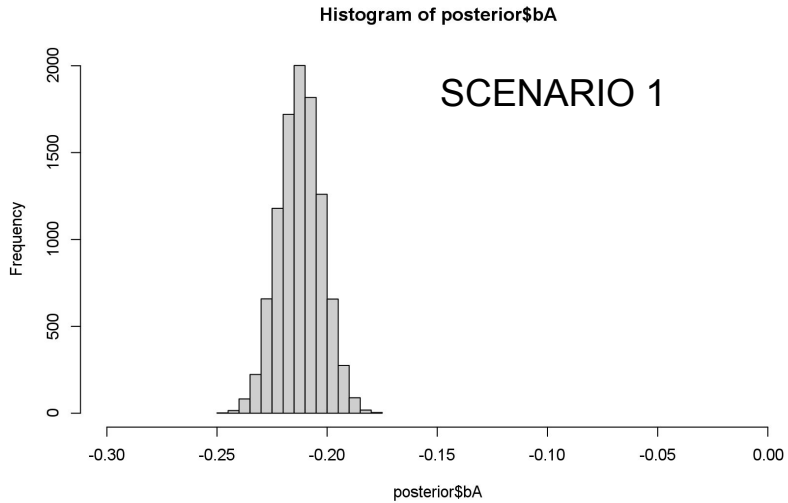
$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0, 1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

SAME AS IN SCENARIO 1

Scenario 2 (results): Randomized experiment and an unobserved covariate

Again, we estimate the effect of parameter β_A, but we also see that the estimate gets **less accurate** caused the new unobserved covariate P, not included in the model.



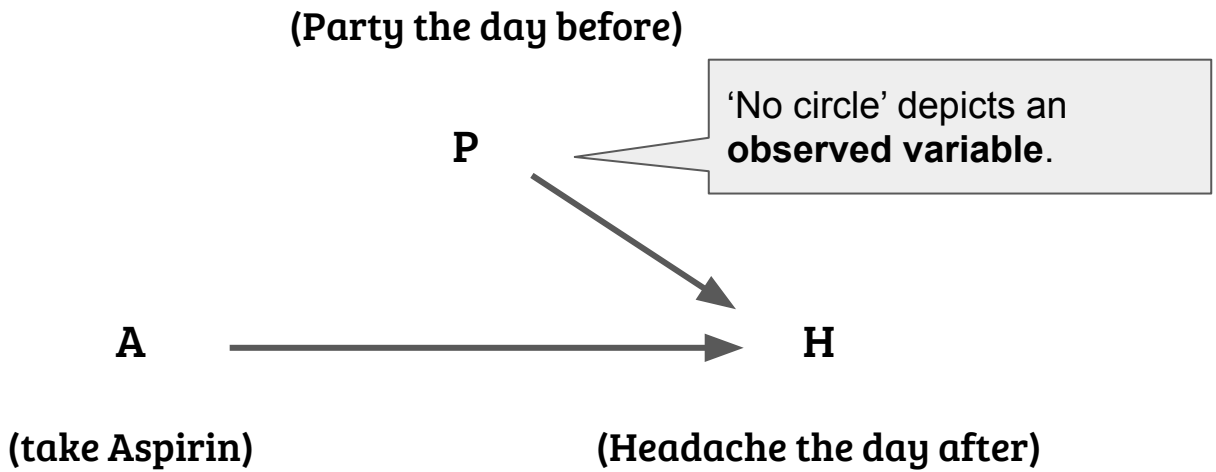


Scenarios 3: Randomized experiment and an observed covariate

(Concrete problems and solutions)

Scenarios 3 (dag): Randomized experiment and an observed covariate

If we observe P, we can use is to make the **estimate** of the effect of A **more precise**.



Scenario 3 (sim): Randomized experiment and an observed covariate

The simulation is the same as in scenario 2.

```
# Party or not.
P <- rbinom(N, 1, 0.5)

# Assignment mechanisms (Aspirin or not).
A <- rbinom(N, 1, 0.5)

# We simulate headache caused by taking no aspirin and party.
mu <- -0.3 + 0.3 * P - 0.2 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

SAME AS IN SCENARIO 2

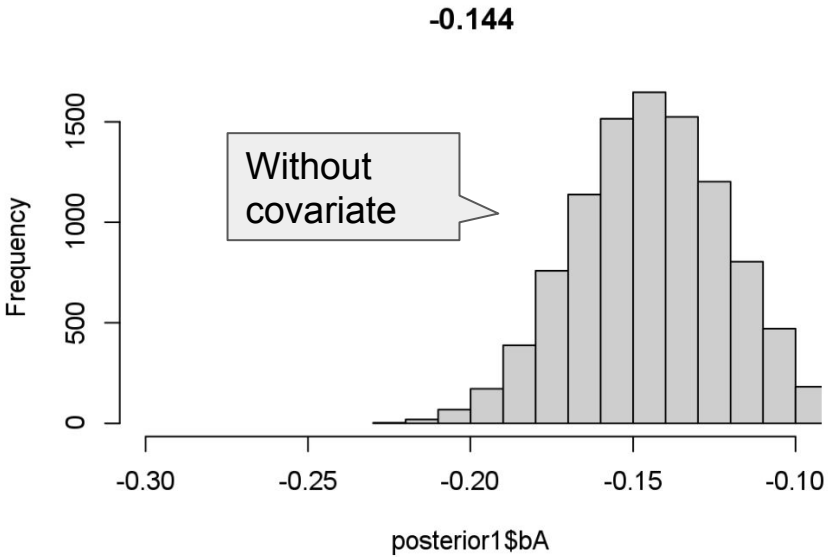
Scenario 3 (model): Randomized experiment and an observed covariate

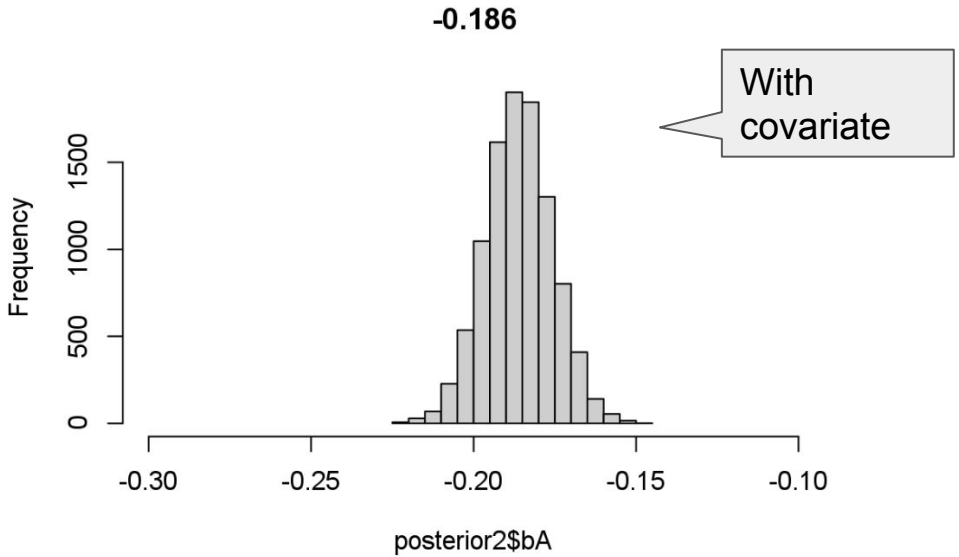
We now face the first **multiple regression model**, adding the covariate as a second predictor variable.

$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A + \beta_P P && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0,1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\beta_P &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

Scenario 3 (results): Randomized experiment and an observed covariate

Adding the covariate (right plot) **improves the estimate for parameter β_A,** compared to not adding the covariate (left plot).



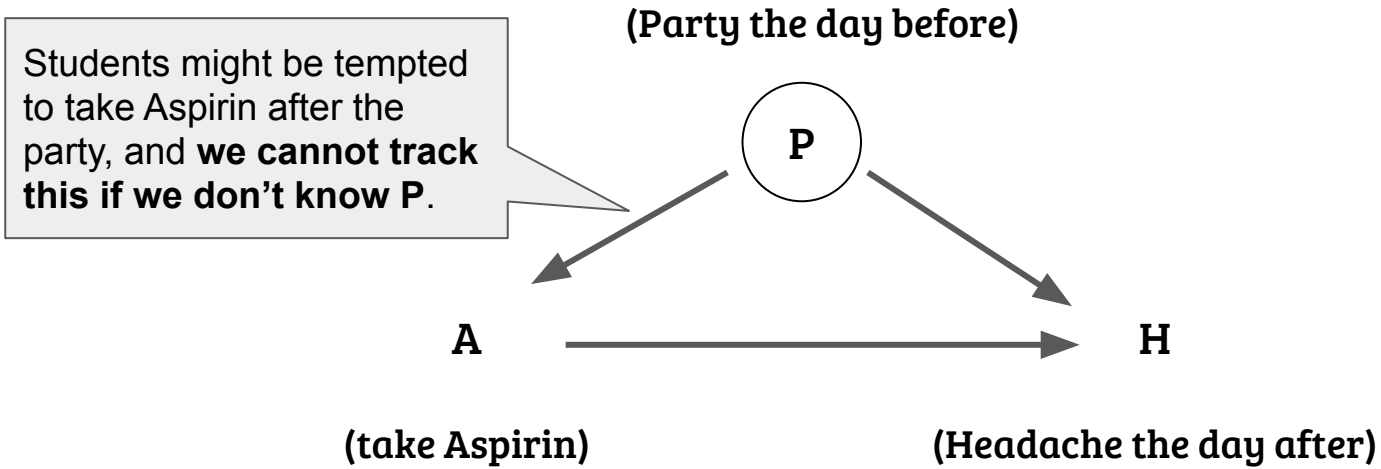


Scenarios 4: Observational study with an influenced assignment mechanism

(Concrete problems and solutions)

Scenario 4 (dag): Observational study with an influenced assignment mechanism

In this case, we **don't know** the variable P influencing the **assignment mechanisms**.



Scenario 4 (dag): Observational study with an influenced assignment mechanism

```r
# Party or not.
P <- rbinom(N, 1, 0.5)

# Assignment mechanisms now influenced by the party.
A <- rbinom(N, 1, prob = ifelse(P, 0.9, 0.1))

# We simulate headache caused by party and taking no aspirin.
mu <- -0.3 + 0.3 * P - 0.2 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

The influence of P on A.

Scenario 4 (model): Observational study with an influenced assignment mechanism

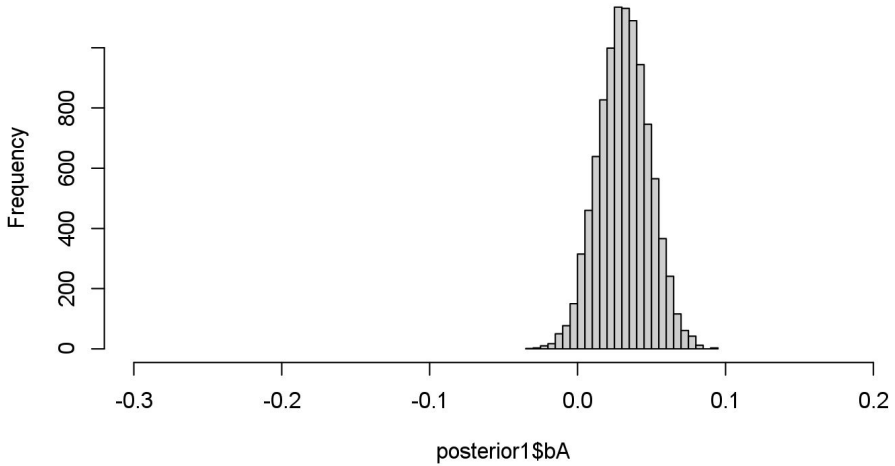
We don’t have P, so we may just examine the effect of A.

$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0, 1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

SAME AS IN SCENARIO 1

Scenario 4 (results): Observational study with an influenced assignment mechanism

We are getting a **spurious result**, with the posterior of β_A right to 0.0, suggesting that **aspirin causes headache**.

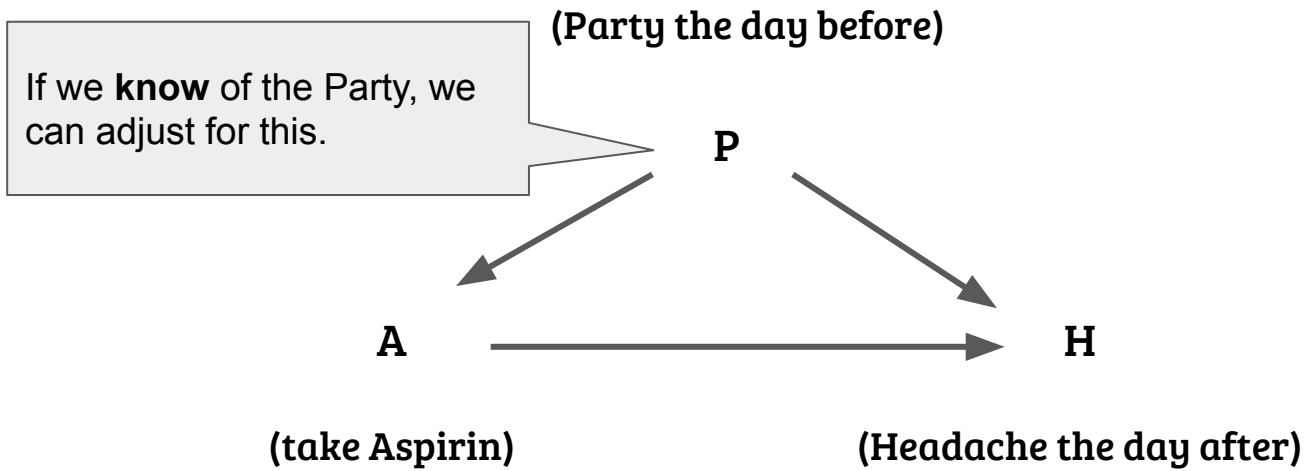


Scenario 5: Observational study with a known influenced assignment mechanism

(Concrete problems and solutions)

Scenario 5 (dag): Observational study with a known influenced assignment mechanism

If we have an idea how the assignment mechanism is working, we can fix this.



Scenario 5 (dag): Observational study with a known influenced assignment mechanism

```
# Party or not.
P <- rbinom(N, 1, 0.5)

# Assignment mechanisms now influenced by the party.
A <- rbinom(N, 1, prob = ifelse(P, 0.9, 0.1))

# We simulate headache caused by party and taking no aspirin.
mu <- -0.3 + 0.5 * P - 0.2 * A
sigma <- 0.27

H <- rnorm(N, mu, sigma)
```

SAME AS IN SCENARIO 4

Scenario 5 (model): Observational study with a known influenced assignment mechanism

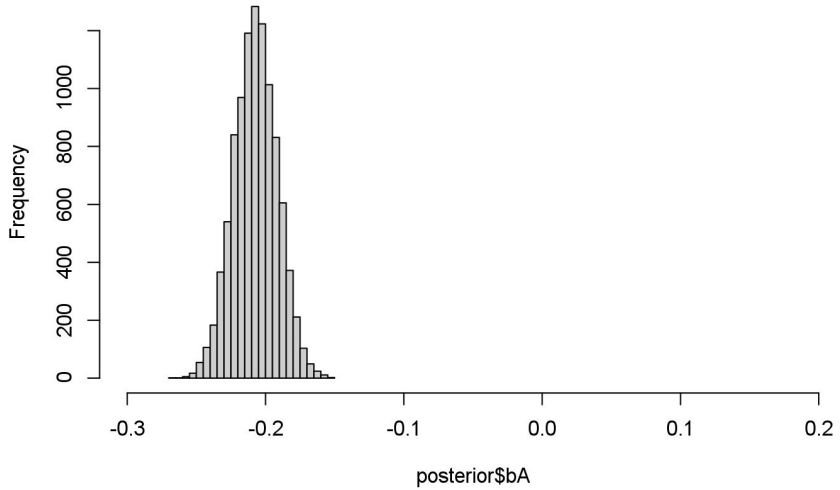
We again use a **multiple regression model**, adding P as a predictor to **adjust for the assignment mechanism**.

$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A + \beta_P P && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0, 1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\beta_P &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

SAME AS IN SCENARIO 3

Scenario 5 (results): Observational study with a known influenced assignment mechanism

We can adjust for the assignment mechanisms and again get the **correct β_A**(≈ -0.2).



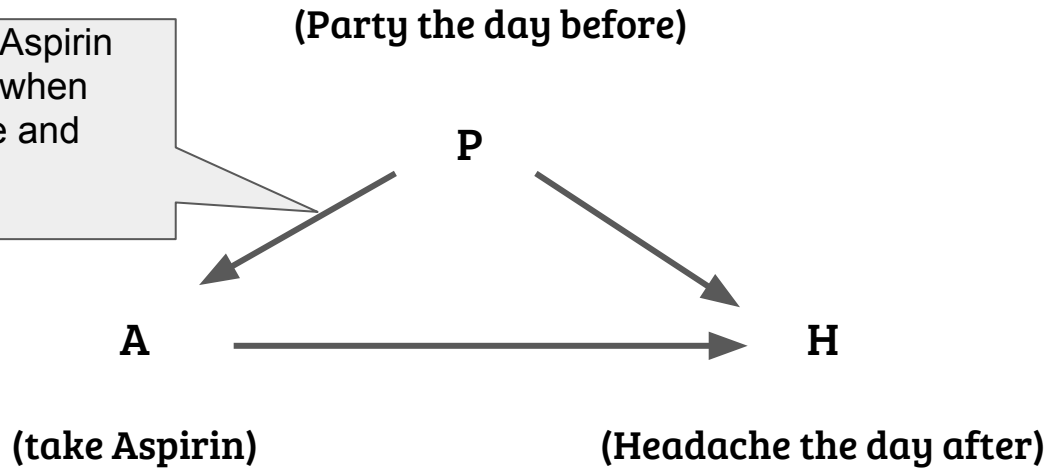
Scenario 6: Multicollinearity

(Concrete problems and solutions)

Scenario 6 (dag): Multicollinearity

Despite knowing the assignment process, we may face a problem with correlation in the data.

If student's always take Aspirin after a party, and never when having no party, we face and issue, referred to as **multicollinearity**.



Scenario 6 (sim): Multicollinearity

The two variables P and A will be highly correlated.

```r
# Party or not.
P <- rbinom(N, 1, 0.5)

# Assignment mechanisms influenced by the party.
A <- rbinom(N, 1, prob = ifelse(P, 0.99, 0.01))

# We simulate headache caused by party and taking no aspirin.
mu <- -0.3 + 0.3 * P - 0.2 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

We see that aspirin is almost certain after a party, and almost impossible without a party.

Scenario 6 (model): Multicollinearity

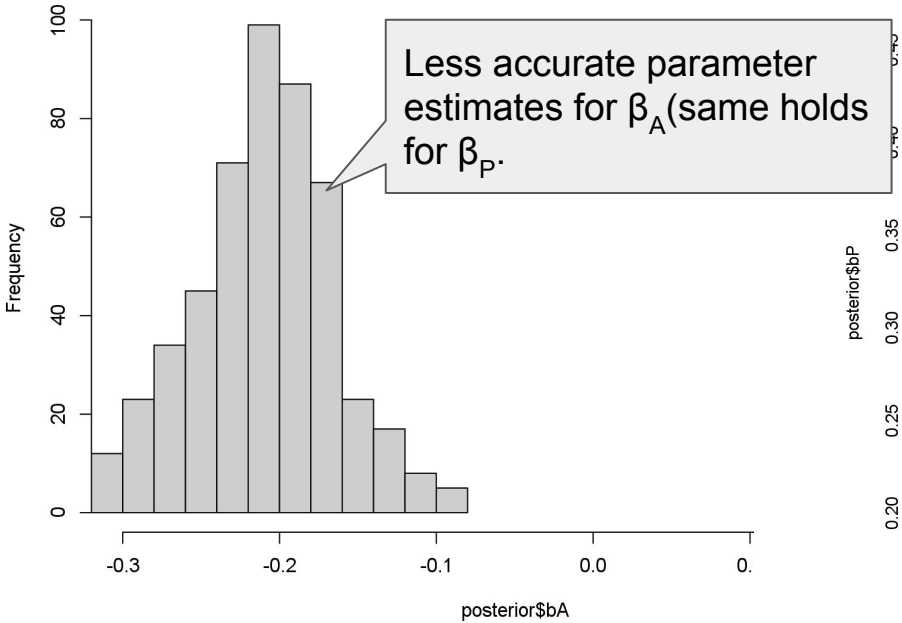
We still use a **multiple regression model**.

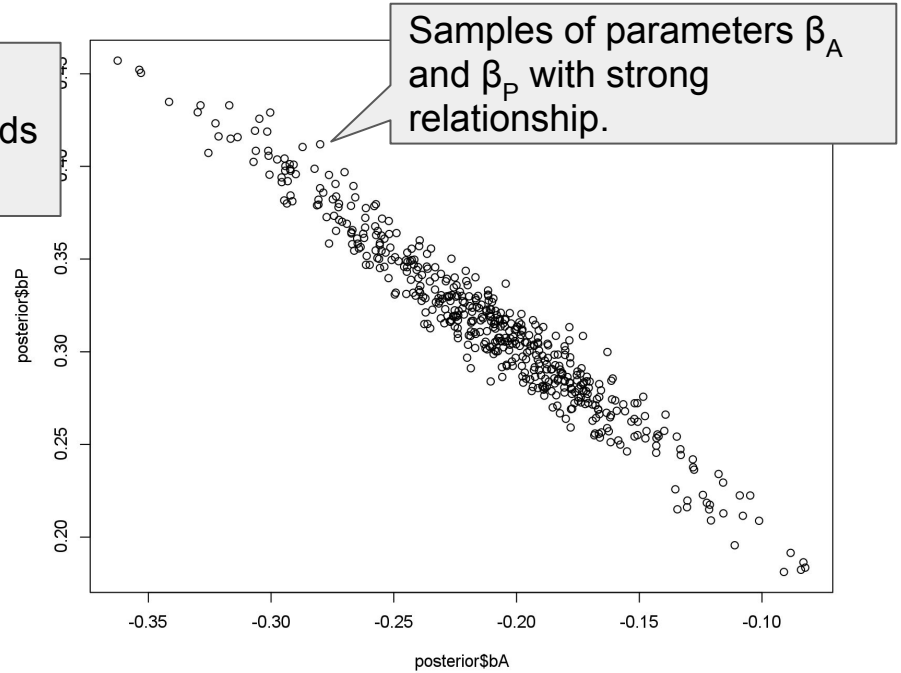
$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A + \beta_P P && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0, 1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\beta_P &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

SAME AS IN SCENARIO 3

Scenario 6 (results): Multicollinearity

We see that the posterior of β_A gets **less accurate**. The sampled parameters β_A and β_P are correlated in the posterior.



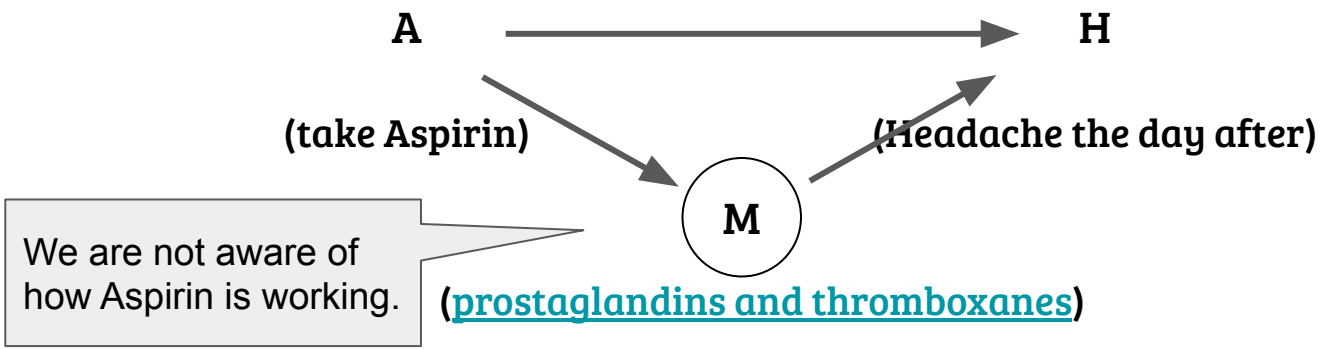


Scenario 7: Unobserved mediator

(Concrete problems and solutions)

Scenario 7 (dag): Unobserved mediator

Detailed examination of the mechanisms how aspiring is working (i.e., by decreasing prostaglandins and thromboxanes).



Scenario 7 (sim): Unobserved mediator

We simulate the mechanism by an unobserved variable M.

```
# Assignment mechanisms (Aspirin or not).
A <- rbinom(N, 1, 0.5)

# We simulate the mediator.
M <- rnorm(N, -0.1 * A, 0.1)

mu <- -0.4 + M - 0.1 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

Scenario 7 (model): Unobserved Mediator

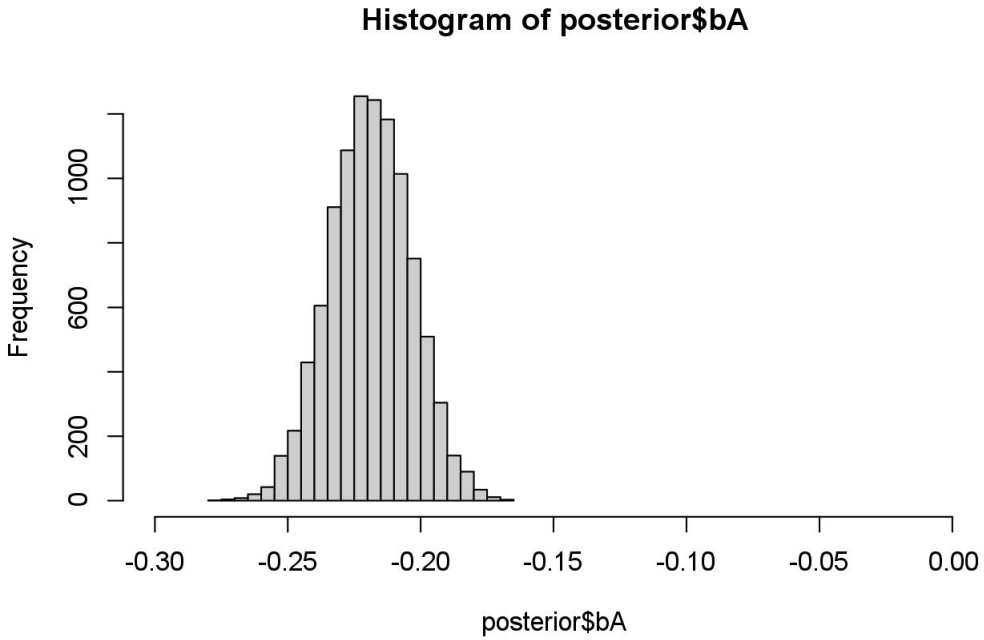
We can estimate the effect of aspirin A on headache H as a parameter β_A in a basic linear model. Since we don't know of the mediator M, we don't include it.

$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0, 1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

SAME AS IN SCENARIO 1

Scenario 7 (results): Unobserved mediator

We still estimate the decrease in headache to be caused by aspirin β_A(≈ -0.2).

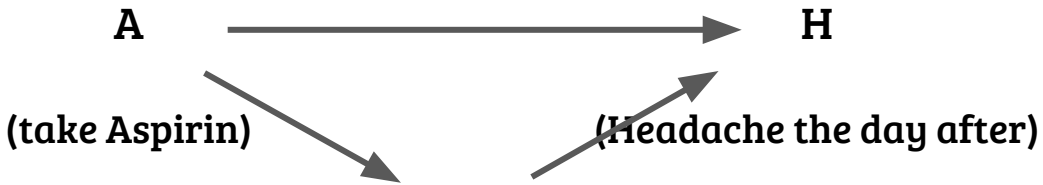


Scenario 8: Observed mediator

(Concrete problems and solutions)

Scenario 8 (dag): Observed mediator

However, we can also separate the indirect influence going over M.



If we add M, we can examine how much of the effect of Aspirin is does through M exclusively (indirect effect).

Scenario 8 (sim): Observed mediator

We simulate the mechanism by an unobserved variable M.

```
# Assignment mechanisms (Aspirin or not).
A <- rbinom(N, 1, 0.5)

# We simulate the mediator.
M <- rnorm(N, -0.1 * A, 0.1)

mu <- -0.4 + M - 0.1 * A
sigma <- 0.07

H <- rnorm(N, mu, sigma)
```

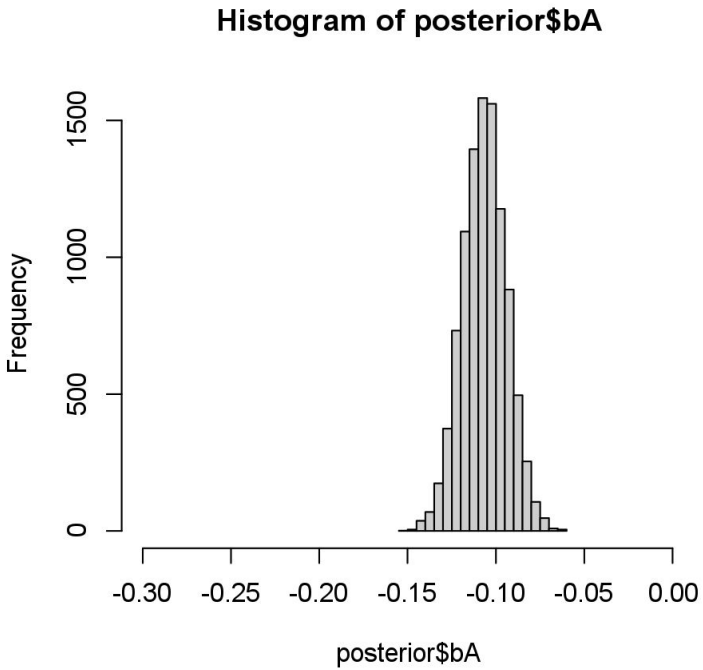
SAME AS IN SCENARIO 7

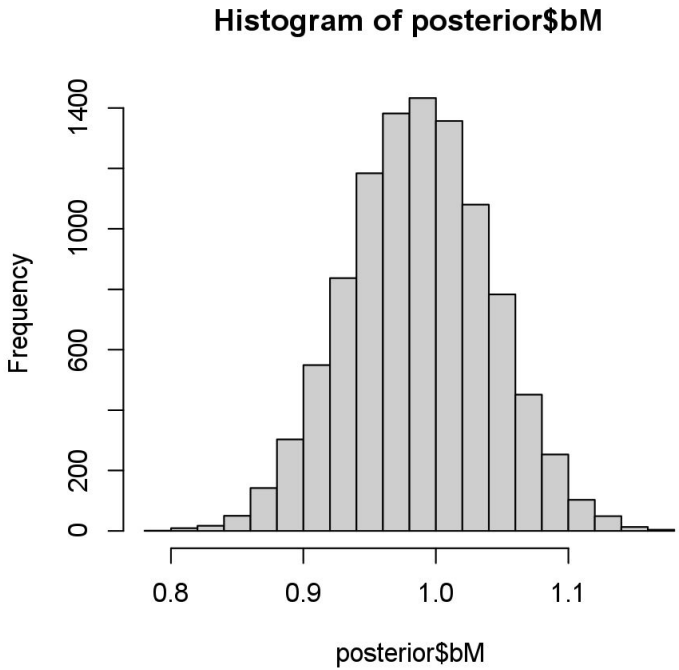
Scenario 8 (model): Observed mediator

We still use a **multiple regression model** to examine the effect of aspirin A and the mediator variable M.

$$
\begin{aligned}
H_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta_A A + \beta_M M && \text{[linear model]} \\
\\
\alpha &\sim \text{Normal}(0,1) && \text{[α prior]} \\
\beta_A &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\beta_M &\sim \text{Normal}(0, 1) && \text{[β prior]} \\
\sigma &\sim \text{Uniform}(0, 3) && \text{[σ prior]}
\end{aligned}
$$

Scenario 8 (results): Observed mediator





Summary

- The missing data problem.
- We can only examine causation with **additional assumptions**.
- DAGs, simulations and how to examine causation in concrete cases.