the effects ch 7 and 8

Emilio Horner

2023-02-16

Chapter 7

1. 1. You are making a simplified causal diagram to represent the data generating process of viewership for a TV show. Which of the following is true?

a. The diagram should include a variable for “number of celebrities in the cast”
b. The diagram should contain one variable for “show airs in the evening” and another for “show doesn’t air in the evening”
c. The diagram should not contain a variable for “show budget” because budgets are often secret and the researcher can’t measure them
d. The diagram should contain the variable “review score in the Jefferson Weekly,” which is the newspaper published by the students at Jefferson High School, with a readership of about 120 people.

I would include (a) because number of celebrities sounds relevant. I would not have shows that air in evening and shows that don’t air in the evening as two variables. You can have unmeasurable variables so I would include budget, and I wouldn’t include the student newspaper review because of its small circulation.

so answer (a)

2.
3. Draw a causal diagram for the research question “do long shift hours make doctors give lower-quality care?” that incorporates the following features (and only the following features):

a. Long shift hours affect both how tired doctors are, and how much experience they have, both of which affect the quality of care
b. How long shifts are is often decided by the hospital the doctor works at. There are plenty of other things about a given hospital that also affect the quality of care, like its funding level, how crowded it is, and so on
c. New policies that reduce shift times may be implemented at the same time (with the timing determined by some unobservable change in policy preferences) as other policies that also attempt to improve the quality of care

long shift hours (independent) and lower-quality care (output variable)

variables long shift hours = A tired= doesn’t need to be represented experience= B quality of care = C hospitalfunding = D hospitalcrowding =E Policies=F

```
library(ggdag)
```

```
##
## Attaching package: 'ggdag'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggplot2)
library(dagitty)
library(tidyverse)
```

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ✓ purrr   0.3.5
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks ggdag::filter(), stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

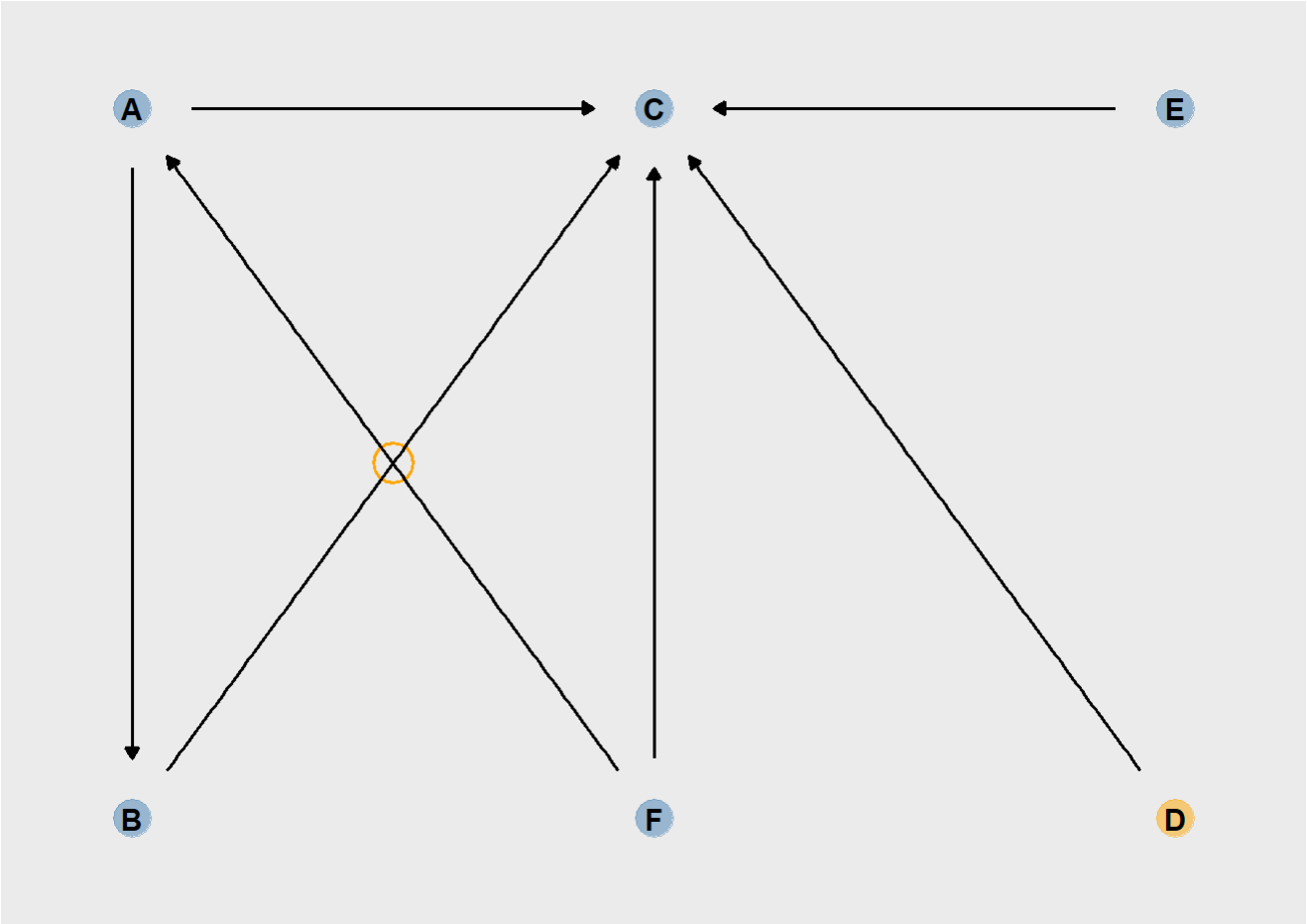
```
library(broom)
```

```
library(ggdag)

dag_coords <-
  tibble(name = c("A", "B", "C", "D", "E", "F"),
         x    = c(1, 1, 3, 5, 5, 3),
         y    = c(3, 1, 3, 1, 3, 1))

dagify(C ~ A,
       B ~ A,
       C ~ B,
       C ~ D,
       C ~ E,
       C ~ F,
       A ~ F,
       coords = dag_coords) %>%

    ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point(aes(color = name == "D"),
                 alpha = 1/2, size = 6.5, show.legend = F) +
  geom_point(x = 2, y = 2,
             size = 6.5, shape = 1, stroke = 1, color = "orange") +
  geom_dag_text(color = "black") +
  geom_dag_edges() +
  scale_color_manual(values = c("steelblue", "orange")) +
  scale_x_continuous(NULL, breaks = NULL, expand = c(.1, .1)) +
  scale_y_continuous(NULL, breaks = NULL, expand = c(.1, .1))
```



3.

Consider this research question: Does the funding level of public schools affect student achievement for students in your country? a. What is the treatment and what is the outcome of interest? b. Write down a list of relevant variables. c. Which of the variables in your list in part b are causes of both treatment and outcome? d. Why might we want to pay extra attention to the variables listed in part c? e. Draw a causal diagram of the variables listed in part b. f. Simplify the diagram from part e.

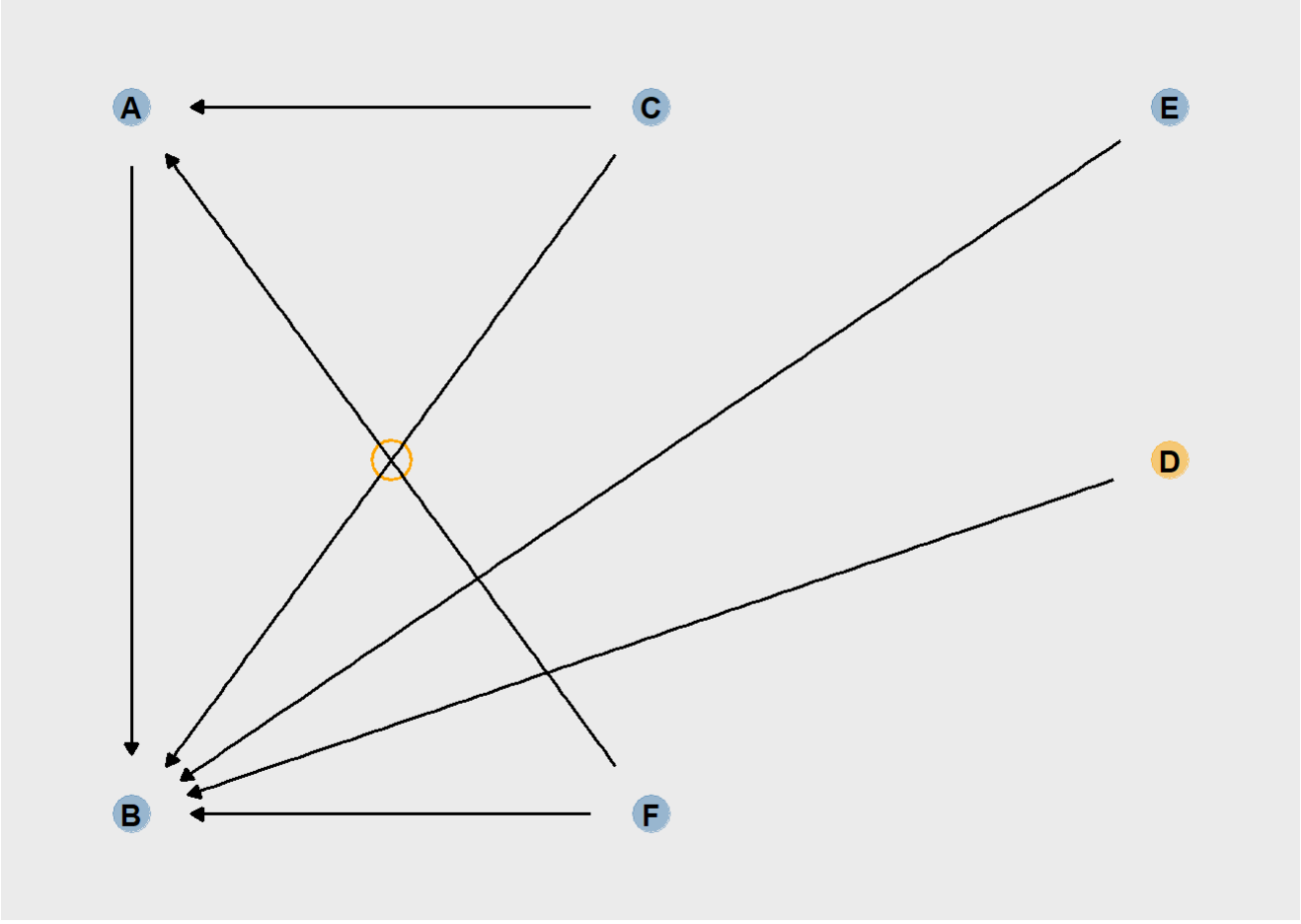
a. treatment variable: funding level outcome variable: student achievement
b. funding, student achievement, ses,gender, race, teacher quality
c. ses, maybe teacher quality (depending on if there's some link in test scores to teacher funding)
d. If they impact both they are essentially being controlled for.
e.

Funding =A Student Achievement = B SES = C Gender = D Race = E Teacher Quality = F

```
dag_coords2 <-
  tibble(name = c("A", "B", "C", "D", "E", "F"),
         x    = c(1, 1, 3, 5, 5, 3),
         y    = c(3, 1, 3, 2, 3, 1))

dagify(B ~ A,
       B ~ C,
       B ~ D,
       B ~ E,
       B ~ F,
       A ~ C,
       A ~ F,
       coords = dag_coords2) %>%

    ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point(aes(color = name == "D"),
                 alpha = 1/2, size = 6.5, show.legend = F) +
  geom_point(x = 2, y = 2,
             size = 6.5, shape = 1, stroke = 1, color = "orange") +
  geom_dag_text(color = "black") +
  geom_dag_edges() +
  scale_color_manual(values = c("steelblue", "orange")) +
  scale_x_continuous(NULL, breaks = NULL, expand = c(.1, .1)) +
  scale_y_continuous(NULL, breaks = NULL, expand = c(.1, .1))
```



f.

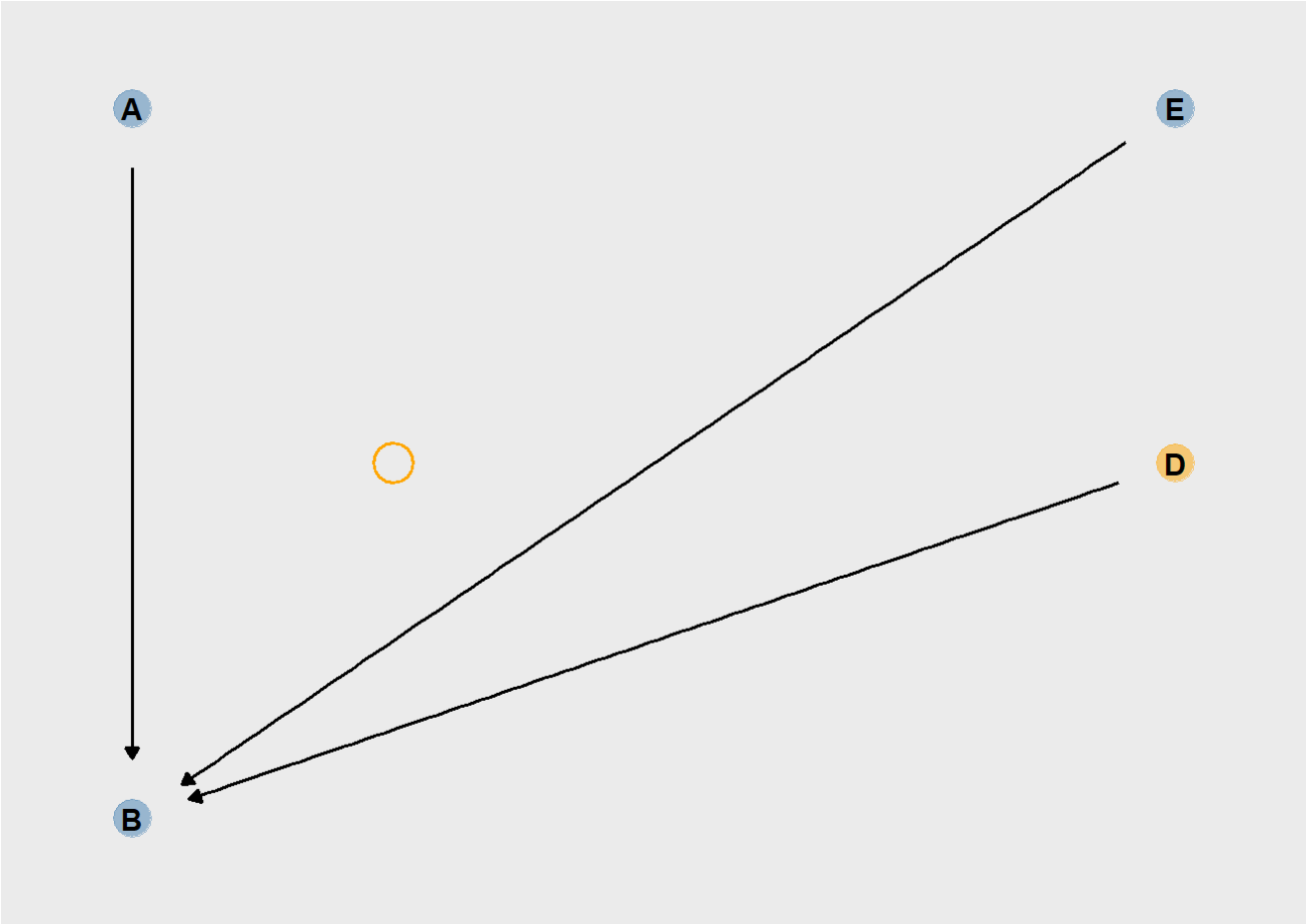
```
dag_coords2 <-
  tibble(name = c("A", "B", "D", "E"),
         x    = c(1, 1, 5, 5),
         y    = c(3, 1, 2, 3))

dagify(B ~ A,
       B ~ C,
       B ~ D,
       B ~ E,
       B ~ F,
       A ~ C,
       A ~ F,
       coords = dag_coords2) %>%

    ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point(aes(color = name == "D"),
                 alpha = 1/2, size = 6.5, show.legend = F) +
  geom_point(x = 2, y = 2,
             size = 6.5, shape = 1, stroke = 1, color = "orange") +
  geom_dag_text(color = "black") +
  geom_dag_edges() +
  scale_color_manual(values = c("steelblue", "orange")) +
  scale_x_continuous(NULL, breaks = NULL, expand = c(.1, .1)) +
  scale_y_continuous(NULL, breaks = NULL, expand = c(.1, .1))
```

```
## Warning: Removed 1 rows containing missing values (`geom_dag_point()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_dag_text()`).
```



4.

Describe the kinds of situations that each of the following could be applied to in order to simplify a causal diagram. a. Unimportance b. Redundancy c. Mediators d. Irrelevance

a. This describes a variable that is largely not significant. Hair color on income could be an example.
b. This refers to variables that occupy the same space. An example could be race and gender which reduce to demographic.
c. A variable that exists only to impact another variable. For example if it was impact of dinner on hunger we wouldn't need a variable "eating" in the middle of the dag.
d. This can be a variable that is important to the data overall, but not on a path from the treatment to outcome variable at hand. An example could be diet on lifespan and the irrelevant variable could be Earthquakes.

5. How can a causal diagram be modified so as to avoid cyclic relationships?

A way to modify the dag to avoid cyclic relationships is to add the element of time. This shows that the second variable didn't cause the first since it happened after.

6.

Think of a research question in your field of interest. a. What is the cause variable and what is the outcome variable? b. Write down a list of between 5 and 10 relevant variables in the data generating process. c. Draw a causal diagram incorporating all the variables from part b. d. Stop working on this problem for fifteen minutes and do something else. Then come back, look at your causal diagram from part c again, and describe one bad assumption you think it made, or something it left out.

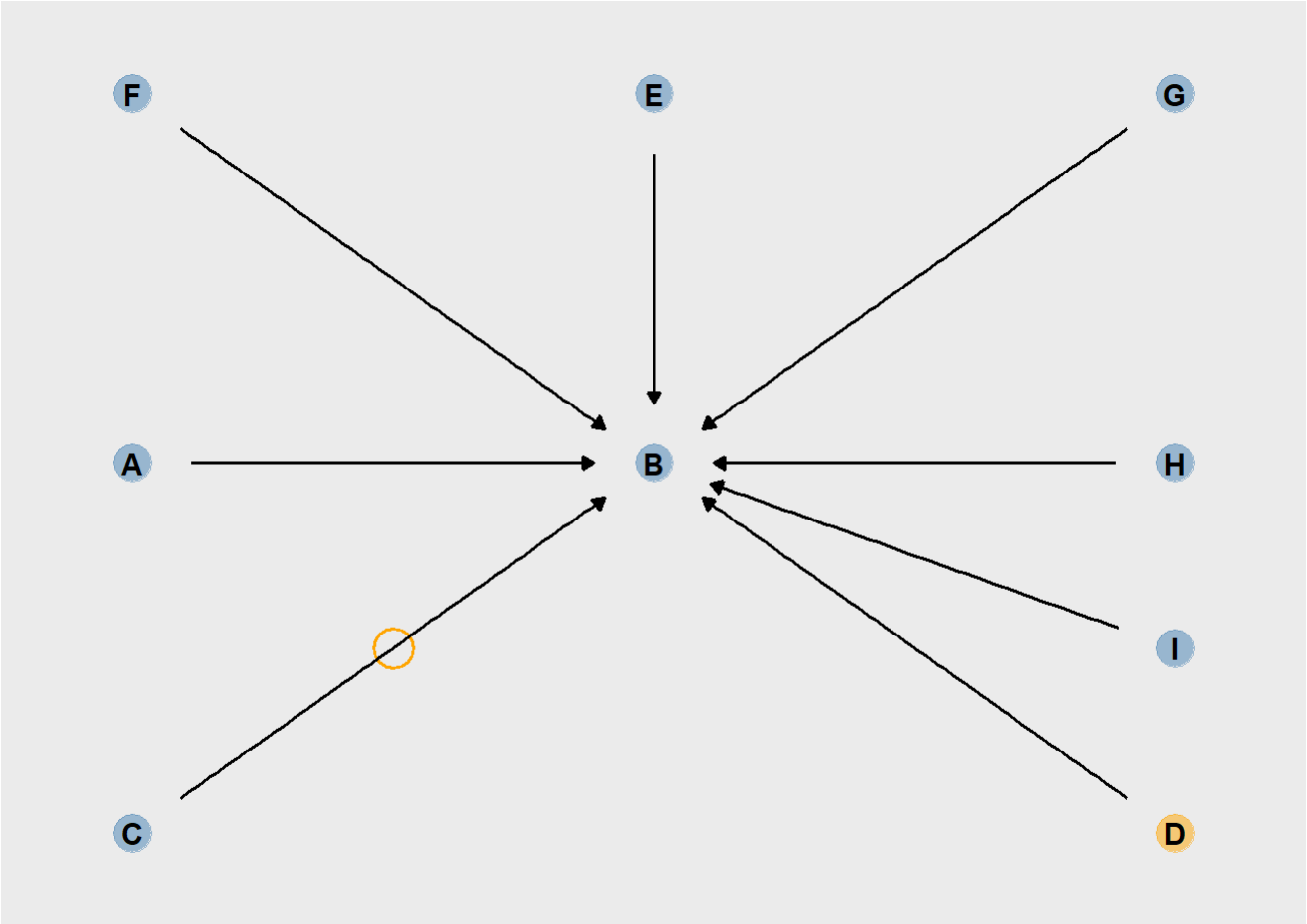
Does quality of how someone dresses for an interview cause someone to have better chances of getting the job? a. cause variable: QualityOfDress outcome variable: GotTheJob b. Resume quality, Job Experience, Race, Gender, Age, Intelligence, Parents Jobs c.

QualityofDress = A GotTheJob= B Resume Quality = C Job Experience = D Race = E Gender = F Age = G Intelligence = H Parents Jobs = I

```
dag_coords2 <-
  tibble(name = c("A", "B", "C", "D", "E", "F", "G", "H",
                  "I"),
         x    = c(1, 3, 1, 5, 3, 1, 5, 5, 5 ),
         y    = c(3, 3, 1, 1, 5, 5, 5, 3, 2))

dagify(B ~ A,
       B ~ C,
       B ~ D,
       B ~ E,
       B ~ F,
       B ~ G,
       B ~ H,
       B ~ I,
       coords = dag_coords2) %>%

    ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point(aes(color = name == "D"),
                 alpha = 1/2, size = 6.5, show.legend = F) +
  geom_point(x = 2, y = 2,
             size = 6.5, shape = 1, stroke = 1, color = "orange") +
  geom_dag_text(color = "black") +
  geom_dag_edges() +
  scale_color_manual(values = c("steelblue", "orange")) +
  scale_x_continuous(NULL, breaks = NULL, expand = c(.1, .1)) +
  scale_y_continuous(NULL, breaks = NULL, expand = c(.1, .1))
```



d. I think some of the variables could also impact quality of dress. One's job experience, intelligence, and parent's jobs could impact one's cultural knowledge of how to dress at interviews.

7. Consider the diagram below. It depicts a cyclical relationship between student achievement and motivation. If students achieve more (i.e., score well on exams), then their motivation goes up, and if their motivation goes up, they achieve more. Change the diagram so that the relationship is not cyclic anymore.

Student AchievementMonday————> Motivation——-> StudentAchievementTuesday

Chapter 8.

1. Assuming that a path has no colliders on it, what is the difference between a path being Open and Closed?

An open path has variation in all the variables along the path. A closed path has at least one variable with no variation.

2.

a. X —-> A —-> Y X <——B <——D —-> Y X <——B ——> Y X——> C <—-D——>Y X ——> C <——D <——B ——>Y

b. front door paths X —-> A —-> Y

c. open back door paths X <——B <——D —-> Y X <——B ——> Y

d. Control for B and D.

3.

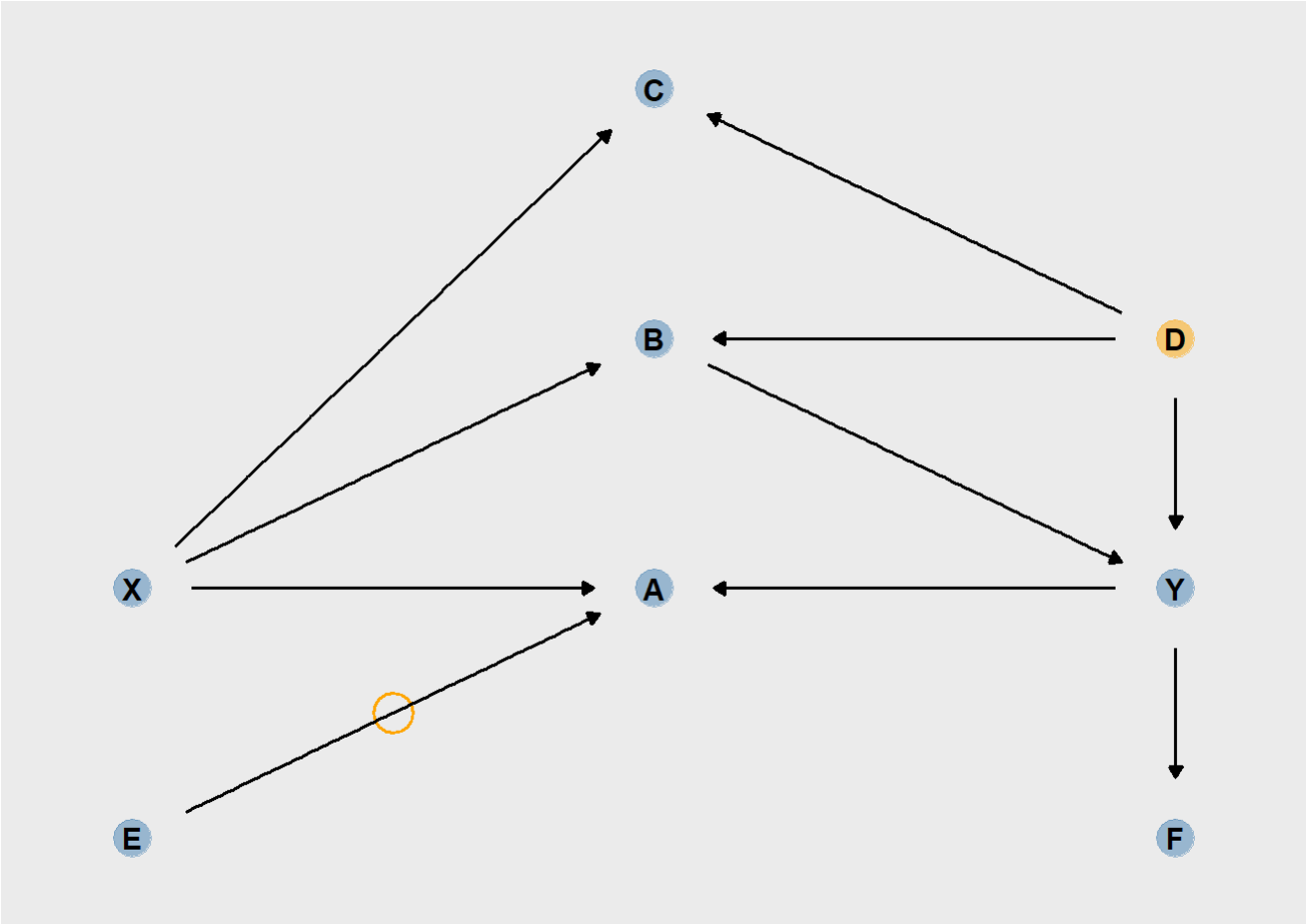
Consider the research question: Does having higher income cause better health? a. Draw a causal diagram depicting the data generating process for this relationship with 5-10 variables on it. b. Identify the Front Door paths. c. Identify the Back Door paths. d. Identify the paths that represent direct effects. e. Identify the Good Paths and the Bad Paths.

a.X = Income Y = Health E = Race A = Gender C = Genetics D = Environmental Toxins Y = Age F = War

```
dag_coords <-
  tibble(name = c("E", "X", "A", "B", "C", "D", "Y", "F"),
         x    = c(1, 1, 3, 3, 3, 5, 5, 5),
         y    = c(1, 3, 3, 5, 7, 5, 3, 1))

dagify(A ~ E,
       A ~ X,
       C ~  X,
       C ~ D,
       B ~ D,
       B ~ X,
       Y ~ B,
       A ~ Y,
       Y ~ D,
       F ~ Y,
       coords = dag_coords) %>%

    ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point(aes(color = name == "D"),
                 alpha = 1/2, size = 6.5, show.legend = F) +
  geom_point(x = 2, y = 2,
             size = 6.5, shape = 1, stroke = 1, color = "orange") +
  geom_dag_text(color = "black") +
  geom_dag_edges() +
  scale_color_manual(values = c("steelblue", "orange")) +
  scale_x_continuous(NULL, breaks = NULL, expand = c(.1, .1)) +
  scale_y_continuous(NULL, breaks = NULL, expand = c(.1, .1))
```



b. X ——> B ——> Y

c. X —-> A <——Y X ——> B <—— D —-> Y X ——> C<——D——> Y

d. On Health D —-> Y B —-> Y

e. good path: X ——> B ——> Y bad paths: X —-> A <——Y X ——> B <—— D —-> Y X ——> C<——D——> Y

4. Which of the following describes a causal path where all the arrows point away from the treatment?

c. Front Door Path

5.

a. a collider variable
b. if you control for a collider it will open the path up, but in a way that will give inaccurate data.

6.

a. Lockdown —-> Recession Lockdown —-> Unemployment —-> Recession Lockdown —->Unemployment <——Prior Economy —-> Recession Lockdown <—-Prior Economy ——> Unemployment —-> Recession Lockdown <—-Prior Economy ——> Recession
b. List all of the paths that are Front Door Paths.

Lockdown —-> Unemployment —-> Recession

c. What would happen if we controlled for unemployment?

It would be harmful for this path Lockdown —->Unemployment <——Prior Economy —-> Recession because unemployment is a collider variable on this path.

d. Is it possible to measure each of the variables adequately?

Yes, there's values for unemployment (unemployment rate), recession, and prior economy. For lockdown it could be a dummy variable of did lockdown of didn't do lockdown.

e. Can you think of any variables and paths not depicted in the diagram that may be relevant to identify the answer to the research question? List at least one and no more than three.

The impact of the virus itself is not depicted on the chart. If there's a lot of sick people it will impact the labor force.

7.

Consider the question: Does obtaining a higher education improve income? Think of a couple of examples of Bad Paths in a causal diagram depicting the data generating process for this research question.

higher education <—— parents income ——> connections —> income

higher education <—— ability—-> income

In both these examples it is not higher levels of education that led to higher income but actually parents income or ability.