Causal Inference and SEM: Lab 3 - Applications

Noémi Schuurman, Oisín Ryan, and Ellen Hamaker

March 2022

1. Simpsons Paradox, Berksons Paradox, Selection Bias

In the following exercises you will practice recognizing/analyzing/visualizing variables in the context of Simpsons Paradox and Berksons Paradox (Berksons Bias). All of the data were simulated.

The data can be found in `DataEx1Lab3.Rdata` .

▶ Load the data.

```
load('dataEx1Lab3.Rdata')
```

〈 Collider example 〉

1.1 Observational Data 1 - “High IQ Makes Students Lazy!?!”

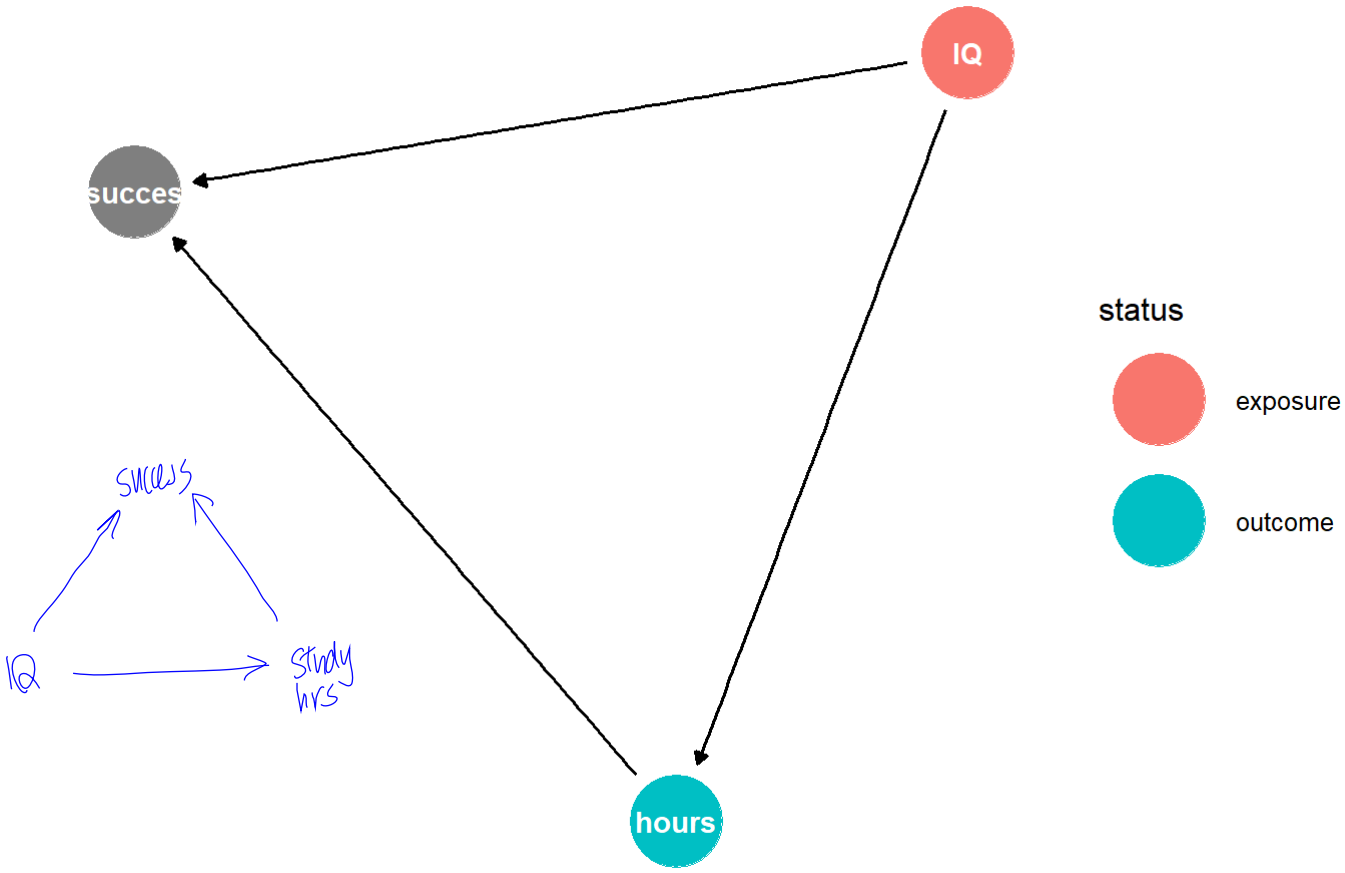
Researchers have measured three variables for university students: Their study success after three bachelor years (higher scores indicate more success), study hours (amount of hours put into study per week on average), and their IQ scores.

They want to evaluate if the IQ score of students affects how many of hours they study on average per week.

▶ Based on the description above, what do you believe is the most reasonable causal model (DAG) for these data?

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

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




```
## I have chosen a DAG where succes is caused by both the hours students put in per week and their IQ. Next to this, IQ also causes the number of hours a student puts in. The arrow of hours to succes can only go in one direction because of how these variables were measures (succes after 3 years), but the other arrows may be reversed. I actually think it is probable that the amount of hours studying put in may also affect IQ, but this would not be possible to include in the DAG (acyclic) without considering repeated measures of these variables. Here I have chosen IQ as the cause given that researchers tend to view this as a relatively stable characteristic. You may have chosen different most reasonable DAGs, which is fine, what matters is having some argumentation for your preference.
```

▶ Take a look at the dataframe `study_succes_data` . In the following analyses, use hours of study as your dependent variable.

1.1.1 Simpsons Paradox

▶ Use linear regression to evaluate marginal association between hours of study and IQ scores.

```
studyh_marg <- lm(studyhours~IQ,data=study_succes_data)
summary(studyh_marg)
```

```
##
## Call:
## lm(formula = studyhours ~ IQ, data = study_succes_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3849  -3.2664  -0.1033   3.2087  15.3351
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.503053   3.343853   9.122   <2e-16 ***
## IQ          -0.007098   0.029100  -0.244    0.807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.978 on 598 degrees of freedom
## Multiple R-squared:  9.948e-05,  Adjusted R-squared:  -0.001573
## F-statistic: 0.05949 on 1 and 598 DF,  p-value: 0.8074
```

▶ How can the regression coefficient you obtain be interpreted? What would your conclusion be if you interpreted these results in a causal manner?

```
## People with one IQ-point higher are expected to have studied approximately 0.007 hour
s less than people with one IQ-point lower. The effect is non-significant; we'd conclude
we have no evidence that indicates IQ is a cause of study hours.
```

▶ Use linear regression to evaluate the association between hours of study and IQ scores, conditional on study success.

```
studyh_cond <- lm(studyhours~IQ+studysucces,data=study_succes_data)
summary(studyh_cond)
```

```
##
## Call:
## lm(formula = studyhours ~ IQ + studysucces, data = study_succes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9836 -2.0394  0.0952  2.0404  8.9153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.26619    2.35996  -1.384    0.167
## IQ          -0.10562    0.01844  -5.727 1.62e-08 ***
## studysucces  0.49773    0.01625  30.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.107 on 597 degrees of freedom
## Multiple R-squared:  0.6113, Adjusted R-squared:   0.61
## F-statistic: 469.4 on 2 and 597 DF,  p-value: < 2.2e-16
```

▶ How can the regression coefficient you obtain be interpreted? What would your conclusion be if you interpreted these results in a causal manner?

```
## People with one IQ-point higher are expected to have studied approximately 0.11 hours less per week than people with one IQ-point lower, given that studysuccess is equal to zero (that is, CONDITIONAL on studysucces). The effect is significant; the causal conclusion would be that people's IQ causes how many hours they study, specifically, higher IQs result in lower hours of study.
```

▶ How do you explain the results of the above two analyses? Given your initial DAG and the results of the analyses, what do you think is the most reasonable conclusion?

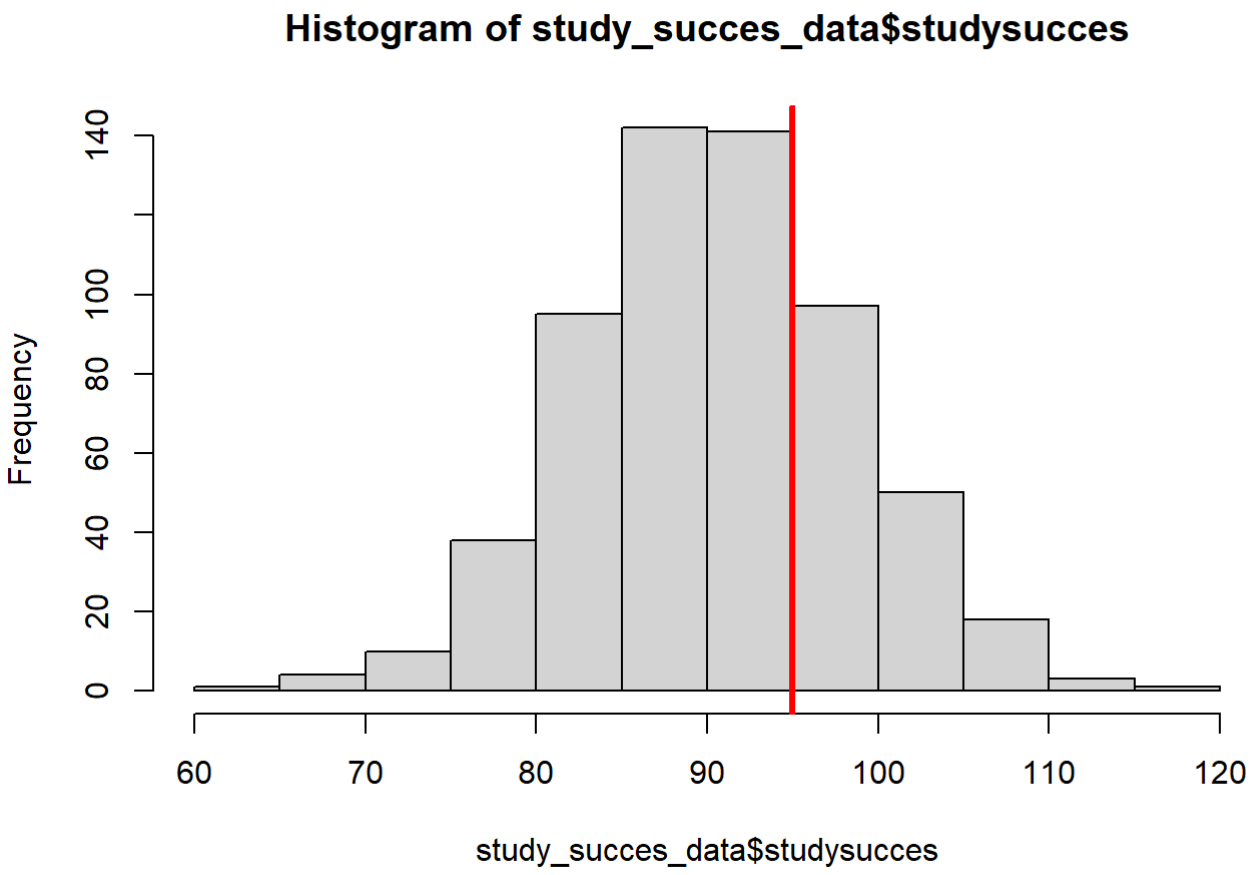
```
## We see that the marginal relationship and conditional relationship are different - th
ese generally do not have to be the same. Based on my initial DAG, I believe the variabl
e we condition (succes) on is a collider variable, with the result here that IQ and hour
s are conditionally dependent, but marginally independent.
##
##     Without considering the causal DAG, both causal conclusions may seem reasonable -
people with higher IQs may need to put in less hours; Or IQ does not affect the amount o
f hours put in at all (maybe 'motivation')  would be a better cause, for example. Howeve
r, based on my chosen DAG earlier, success is a collider variable, and hence I should no
t control for it to obtain the causal effect. Hence the marginal relationship is most ap
propriate, and my conclusion should be that I have no evidence for a causal effect of IQ
on the amount of hours put into their study on average per week.
##
##     You may have chosen a somewhat different DAG and hence have a different final con
clusion here.
```

1.1.2 Berksons Bias

Imagine a research team collecting the same data - but instead of the population of students as a whole they focused on students that are in the running for ‘honors’ or even a ‘cum laude’ diploma. These are students that have a higher study success than 95. They only directly measured IQ scores and the average hours of study per week.

▶ Select a sample from the original dataset with only students in the running for a honors or cum laude diploma.

```
hist(study_succes_data$studysucces)
abline(v=95, col="red", lwd=3)
```




```
honors_data = study_succes_data[which(study_succes_data$studysucces>95),]
```

▶ Use linear regression to evaluate the association between hours of study and IQ scores. Interpret the results, and compare them to the previous two analyses.

```
studyh_honors <- lm(studyhours~IQ,data=honors_data)
summary(studyh_honors)
```

```
##
## Call:
## lm(formula = studyhours ~ IQ, data = honors_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6753 -2.3922  0.1916  2.3506 11.3520
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.76814    4.58928   9.537   <2e-16 ***
## IQ          -0.07944    0.03953  -2.009   0.0461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.608 on 167 degrees of freedom
## Multiple R-squared:  0.02361,    Adjusted R-squared:  0.01776
## F-statistic: 4.038 on 1 and 167 DF,  p-value: 0.04611
```

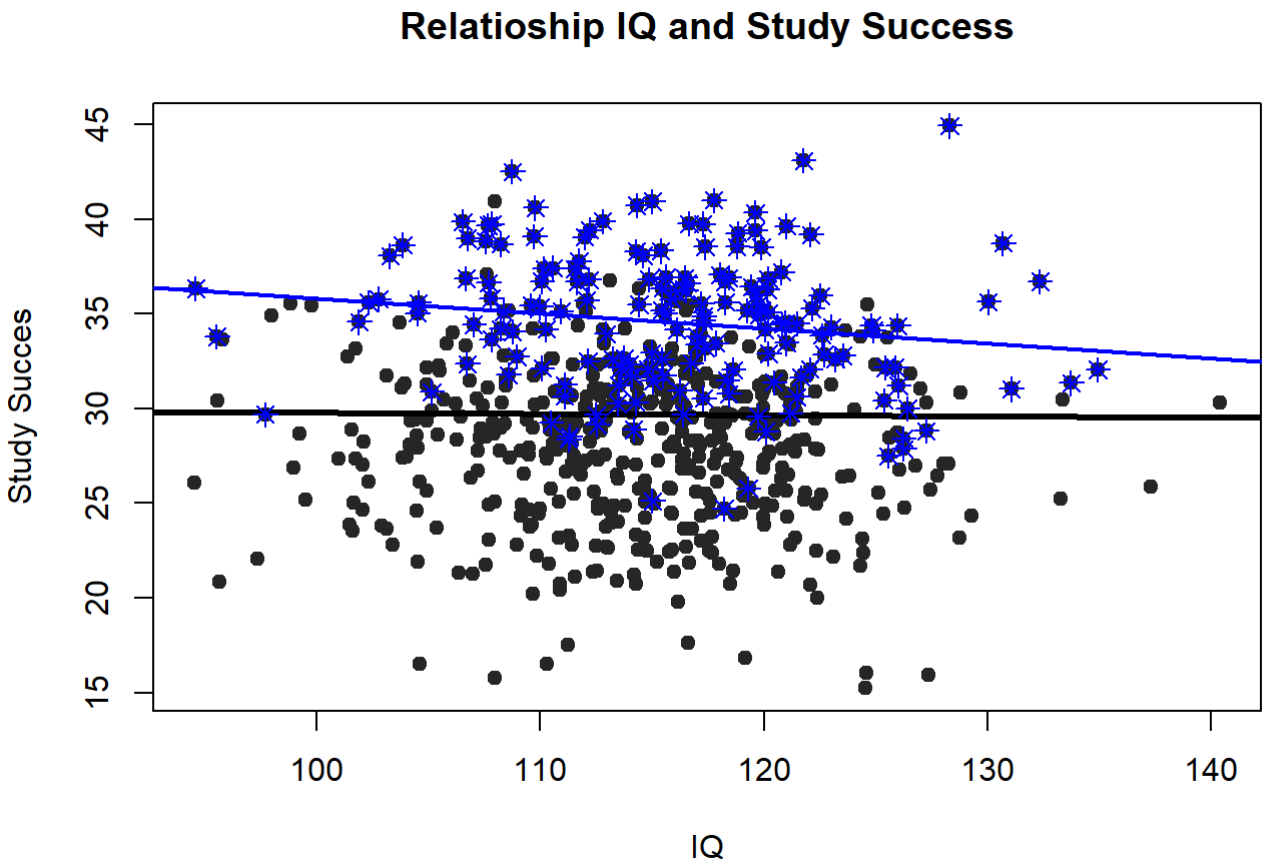
```
## People with one IQ-point higher are expected to have studied approximately 0.08 hours less per week than people with one IQ-point lower, given that studysuccess is higher than 95 (that is, CONDITIONAL on studysucces). Note that the regression we specified looks like it is estimating a marginal relationship, but due to how we sampled we are in fact conditioning on studysuccess. The estimated coefficient is accordingly also more similar to our initial conditional regression than the initial marginal regression (note that this also happens because we have fully linear relationship here, see later exercise.)
##
##     Note that if this would simply be the sample we collected, there would be no straightforward way to recover the actual marginal effect with that dataset (aside from collecting more data for a wider range of studysuccess).
```

1.1.3 Visualize

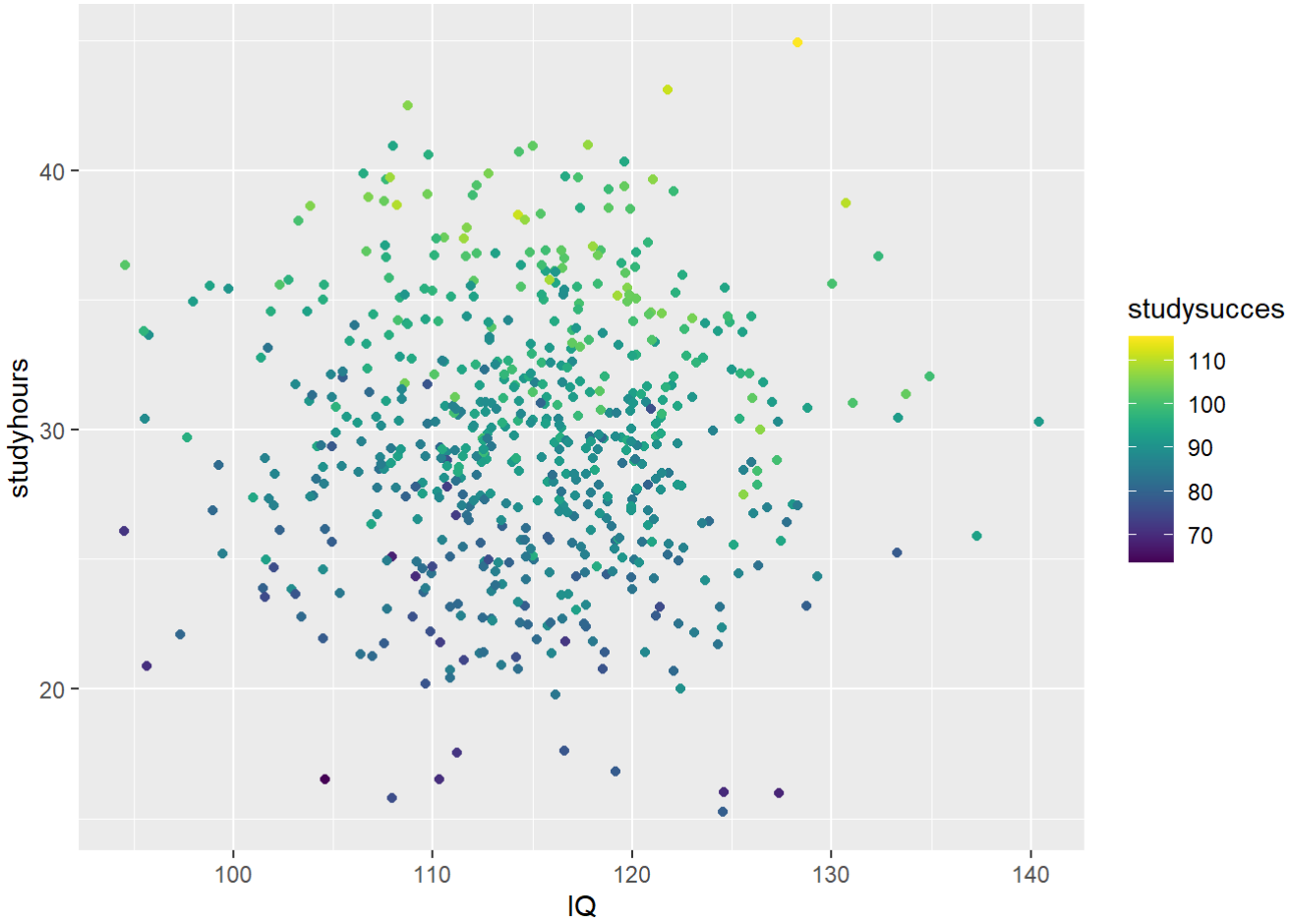
▶ Make a scatterplot in which you visualize the marginal relationship between hours of study and IQ scores, for the whole student population (e.g., not conditioning on study success in any way).

▶ Make another scatterplot in which you visualize the relationship between hours of study and IQ scores, for the potential cum laude students. (or add this to the previous scatterplot)

```
plot(study_succes_data$IQ, study_succes_data$studyhours, main = "Relatioship IQ and Stud
y Success", ylab="Study Succes", xlab="IQ", pch=19, col="grey15")
abline(a = studyh_marg$coefficients[1] , b = studyh_marg$coefficients[2], col = "black",
lwd=3)

points(honors_data$IQ, honors_data$studyhours, pch=8,col="blue",cex=1.2)
abline(a = studyh_honors $coefficients[1] , b = studyh_honors $coefficients[2], col = "b
lue", lwd=2)
```




```
####

#Or, to visualize the effect of conditioning on Study Succes in more of a continuous fas
hion
library(ggplot2)
library(viridis)
# Gradient color
ggplot(study_succes_data, aes(x = IQ, y = studyhours, colour = studysucces)) +
  geom_point()+
  scale_color_viridis(option = "D")
```




```
## In the first plot, you see how the relationship changes as soon as we select a specif
ic part of our sample (we condition on studysuccess). You can see a similar effect in a
more continuous way with the second plot: If you select (condition on) a specific (set o
f) colors in this graph, and imagine you draw a regression line to those specifically co
lored data, you may imagine how the regression effects will differ depending on how you
conditioned.
```

1.1.4 The true average causal effect

The data was simulated in the following way:

```
set.seed(113221701) # set the seed for comparison
n <- 600          # total sample size
IQ <- rnorm(n,115,7)
studyhours<- rnorm(n,30,5)
studysucces<- 25 + .25*IQ + 1.25*studyhours + rnorm(n,0,5)
```

► What is the true average causal effect? Calculate this by mimicking an intervention: Determine the expected value of the dependent variable while you fix the cause variable to zero, and determine the expected value of the dependent variable while you fix the cause variable to one. Then calculate the ACE.

```
###IQ set to 1###
IQ <- 1
EVstudyhours_IQ1 <- 30
EVstudysucces <- 25 + .25*IQ + 1.25*EVstudyhours_IQ1

###IQ set to 1###
IQ <- 0
EVstudyhours_IQ0<- 30
EVstudysucces <- 25 + .25*IQ + 1.25*EVstudyhours_IQ0

EVstudyhours_IQ1-EVstudyhours_IQ0
```

Average value of study hours

```
## [1] 0
```

```
## In the `true' DAG/simulated data, there is no causal effect at all of IQ on hours (or vice versa)!
```

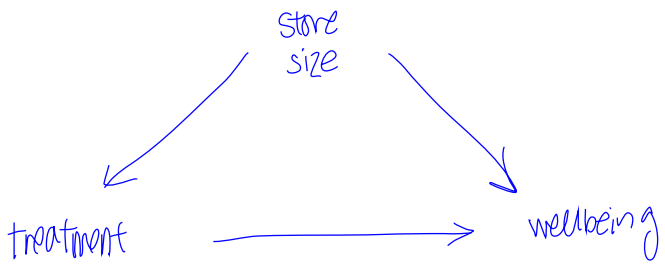
Confounding example.

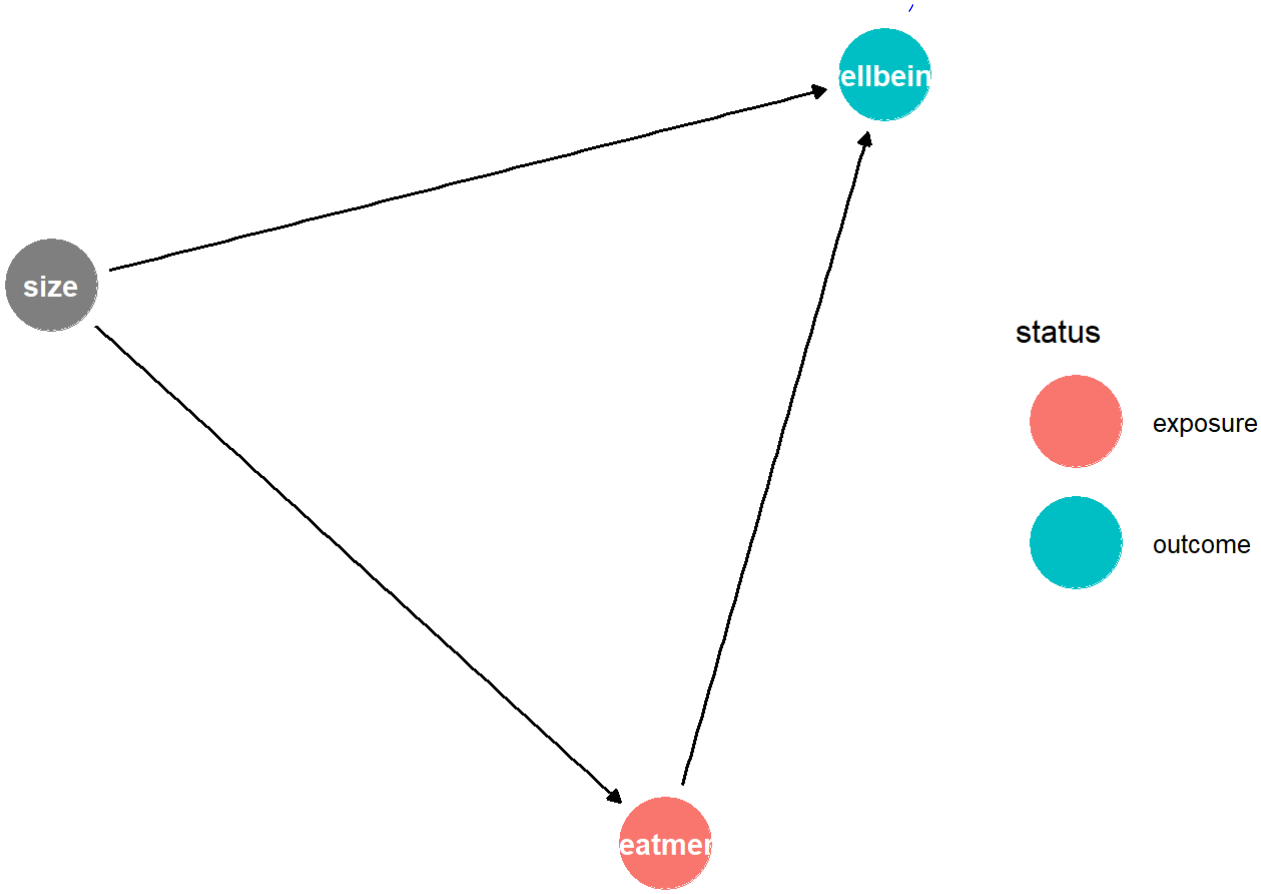
1.2 Observational Data 2 - “Does Size Matter?”

In this exercise we loosely mimick a classic (real research) example of simpsons paradox, where two treatments (cause variables of interest) for kidney stones were evaluated. The outcome variable of interest is a measurement recovery/health/wellbeing, and a potential covariate to consider is kidney stone size. It is observational data: People come in with kidney stone problems, and are assigned treatment. After treatment, their wellbeing is measured.

항상 시간이 흐르는지 check → 화살표 arrow 방향 결정하기 ★

▶ Based on the description above, what do you believe is the most reasonable causal model (DAG) for these data?






```
## I have chosen a DAG where stonesize determines the treatment that someone gets, and b
oth stone size and treatment affect wellbeing. To me reversed arrows seem less reasonabl
e, because I don't immediately see how, after someone comes in with stone problems the t
reatment would cause the size. Wellbeing is mentioned to be measured AFTER treatment. No
te I (again) ignored potential unobserved confounders here.
```

▶ Take a look at the dataframe `kidneystone_data` .

1.2.1 Simpsons Paradox or Berksons Bias

▶ Use linear regression to uncover either an example of Simpsons Paradox or Berksons Bias (your choice). Discuss the results.

```
#simpsons# ~ marginal & conditional relation ship
wellb_marg <- lm(wellbeing~treatment,data=kidneystone_data)
summary(wellb_marg)
```

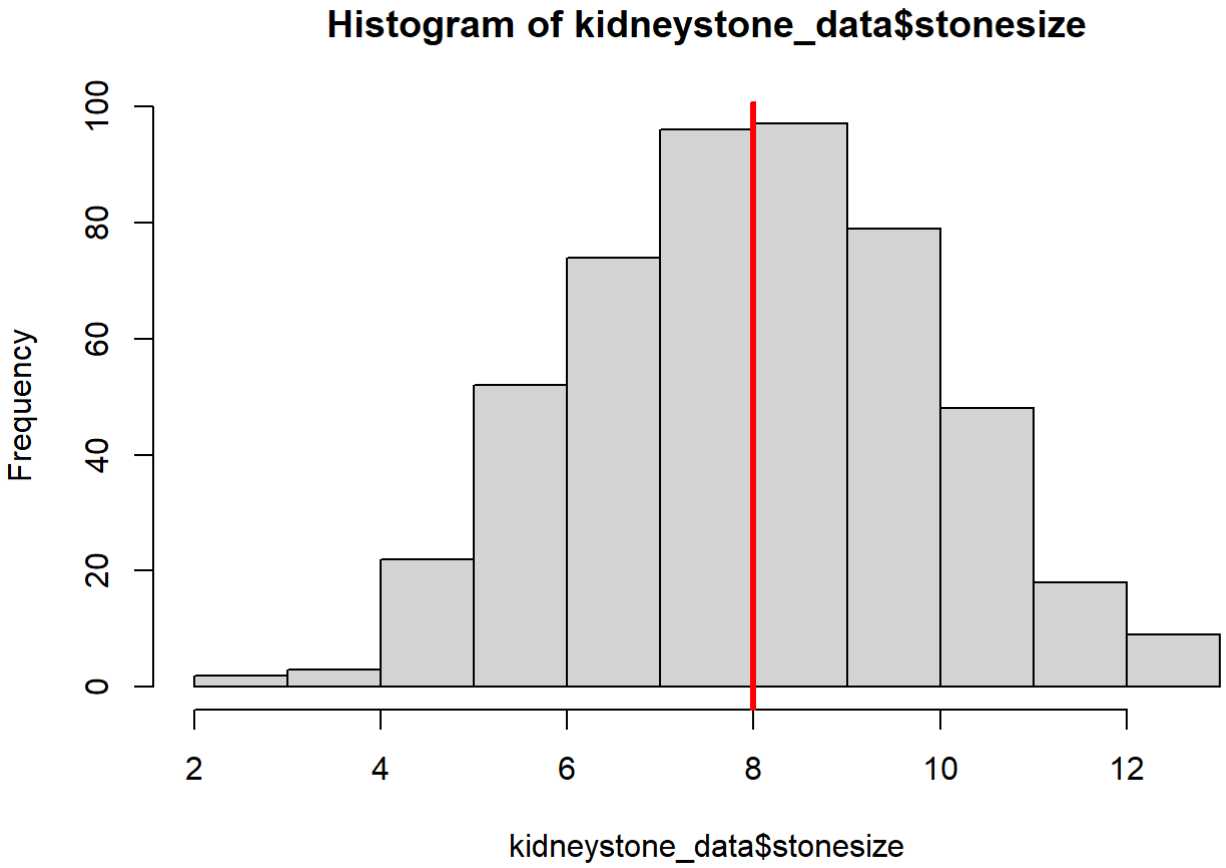
```
##
## Call:
## lm(formula = wellbeing ~ treatment, data = kidneystone_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6558 -1.8821 -0.4212  1.9130 11.2518
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.9355     0.1401  113.77   <2e-16 ***
## treatment    -4.4220     0.2524  -17.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.605 on 498 degrees of freedom
## Multiple R-squared:  0.3813, Adjusted R-squared:  0.3801
## F-statistic:   307 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
wellb_cond <- lm(wellbeing~treatment+stonesize,data=kidneystone_data)
summary(wellb_cond)
```

```
##
## Call:
## lm(formula = wellbeing ~ treatment + stonesize, data = kidneystone_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60895 -0.75667  0.00616  0.71325  3.03285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.76371    0.27639  107.69   <2e-16 ***
## treatment    1.81421    0.15845   11.45   <2e-16 ***
## stonesize   -1.96950    0.03855  -51.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.043 on 497 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.9007
## F-statistic:  2263 on 2 and 497 DF,  p-value: < 2.2e-16
```

conditioning by selecting specific subgroup.

```
#Berksons. In the example below, we select for people that have kidneystones larger than
8 mm (but you may select in different ways, of course, and this may alter the results).

hist(kidneystone_data$stonesize)
abline(v=8, col="red", lwd=3)
```




```
larges_data = kidneystone_data[which(kidneystone_data$stonesize>8),]

wellb_larges <- lm(wellbeing~treatment,data=larges_data)
summary(wellb_larges)
```

```
##
## Call:
## lm(formula = wellbeing ~ treatment, data = larges_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6558 -1.0791  0.0669  1.2510  4.2217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.1346     0.1830  71.766  < 2e-16 ***
## treatment    -1.6212     0.2337  -6.938 3.41e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.803 on 249 degrees of freedom
## Multiple R-squared:  0.162,  Adjusted R-squared:  0.1586
## F-statistic: 48.14 on 1 and 249 DF,  p-value: 3.414e-11
```

Simpsons

Berkson

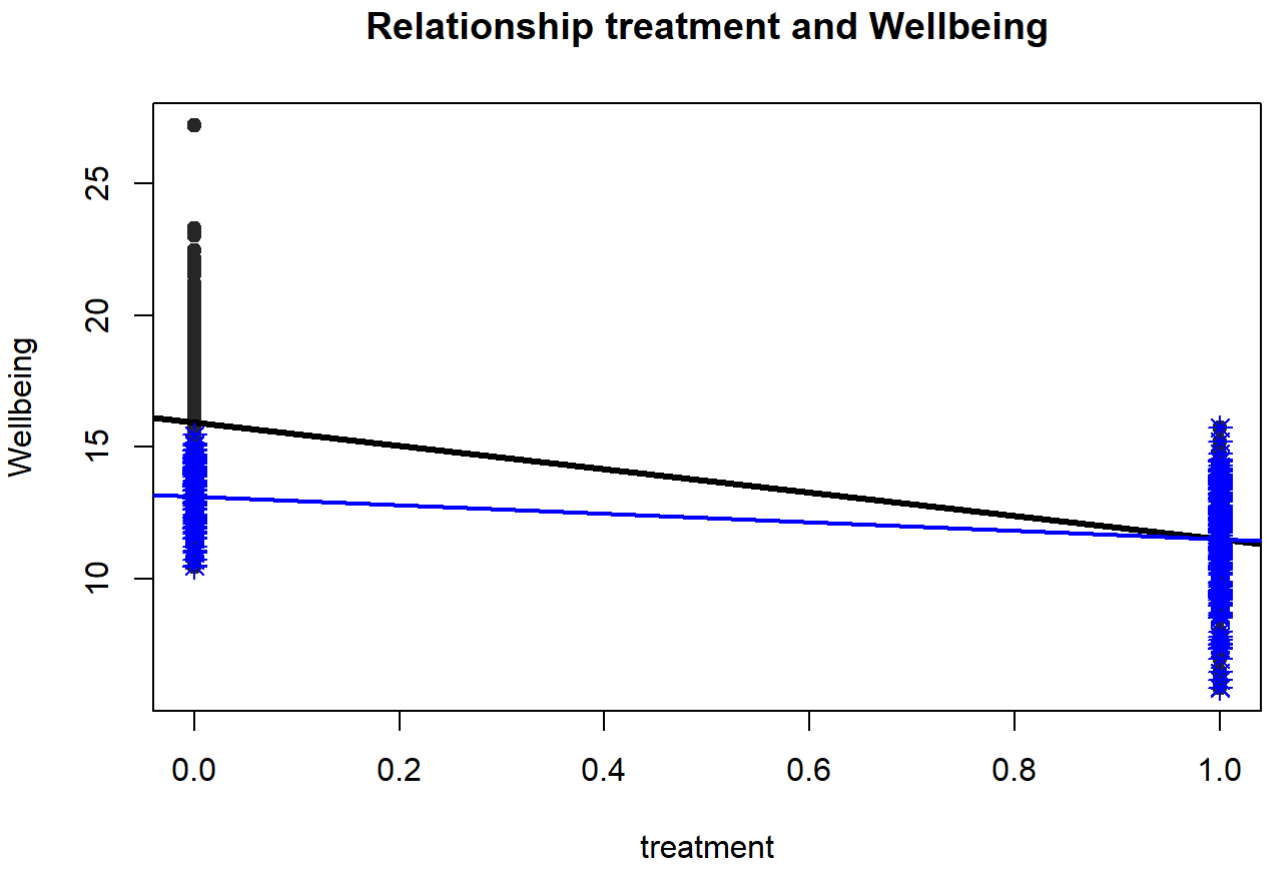
```
##  We see that the marginal relationship and conditional relationship are different - t
hese generally do not have to be the same. Marginally, the treatment effect is negative
(-4.4), meaning that treatment 1 is worse than treatment 0 for wellbeing. Conditionally
on stonesize, the conclusion is the opposite - there is a positive effect of treatment
(1.8), meaning that treatment 1 is better in terms of wellbeing than treatment 0. When w
e condition by selecting only people with stonesizes larger than 8, we see a negative ef
fect of treatment but a much weaker one than the marginal relationship (an effect somewh
ere between our initial marginal and conditional relationship).
##
## Based on my initial DAG, I believe the variable we condition (size) on is a confoundi
ng variable, hence I should control for it to obtain the appropriate causal effect. Henc
e the conditional relationship is most appropriate, and my conclusion should be that tre
atment 1 is better for people's wellbeing than treatment 0 (there is a causal effect of
treatment).
```

→ treatment causes higher well-being,

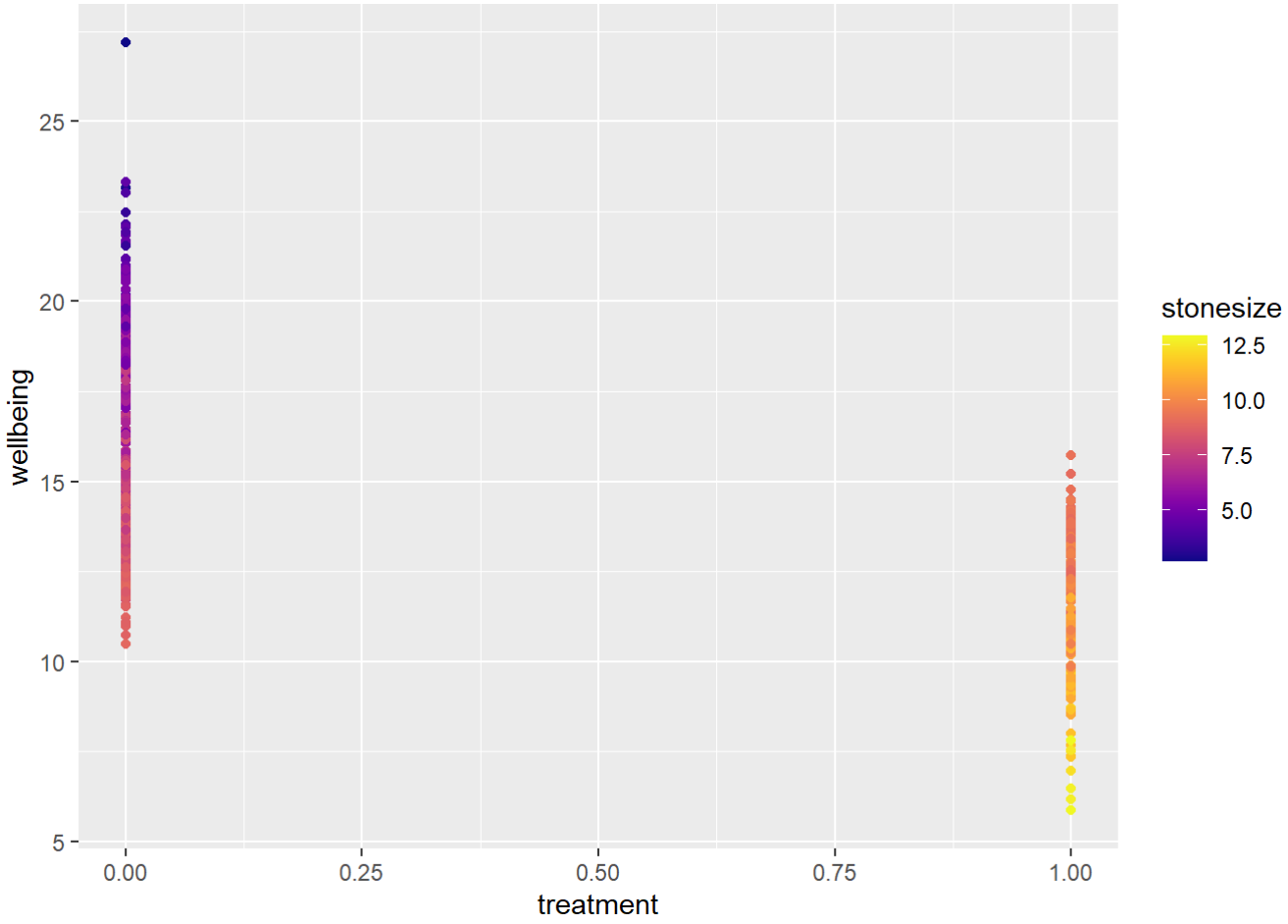
1.2.2 Visualize

▶ Use scatterplots to illustrate the Paradox.

```
plot(kidneystone_data$treatment, kidneystone_data$wellbeing, main = "Relationship treatm
ent and Wellbeing", ylab="Wellbeing", xlab="treatment", pch=19, col="grey15")
abline(a = wellb_marg$coefficients[1] , b = wellb_marg$coefficients[2], col = "black",lw
d=3)

points(larges_data$treatment, larges_data$wellbeing, pch=8,col="blue",cex=1.2)
abline(a = wellb_larges$coefficients[1] , b = wellb_larges$coefficients[2], col = "blue"
, lwd=2)
```




```
#Optional, to visualize the effect of conditioning on Study Succes in more of a continuo
us fashion
library(ggplot2)
library(viridis)
# Gradient color
ggplot(kidneystone_data, aes(x = treatment, y = wellbeing, colour = stonesize)) +
  geom_point()+
  scale_color_viridis(option = "C")
```




```
## In the first plot, you see how the relationship changes as soon as we select a specif
ic part of our sample (we condition on studysuccess).
##
##     In the second plot, you see very nicely that small stonesizes appear only among t
he people with treatment 0. For treatment 1, there are only people with relatively large
stones! You can imagine here what selecting only specific colors here would do to the re
gression effects.
```

1.2.3 The true average causal effect

The data was simulated in the following way:

```
set.seed(133222032) # set the seed for comparison
n <- 500          # total sample size
stonesize <- rnorm(n,8,2) # averge stone size of 8 mm (yikes)
treatment=rep(0,n)
treatment[which(stonesize>9)] <- 1  #Treatment is assigned deterministically (no probabi
lity involved here) completely based on stone size. People with large stones (>9) get tr
eatment '1'.
wellbeing <- 30 + 2*treatment - 2*stonesize + rnorm(n,0,1)
```

▶ Calculate the true causal effect by mimicking an intervention. Discuss the results.

① Set the causal variable to 1 & 0
② take the average value for the other covariates.
③ Define the OV as per SCM equation, but remove the error term,

```
###Treatment = 1###
stonesize_EV <- 8 # averge stone size of 8 mm (yikes)
treatment=1
wellbeing_t1_EV <- 30 + 2*treatment - 2*stonesize_EV

###Treatment = 0###
stonesize_EV <- 8 # averge stone size of 8 mm (yikes)
treatment=0
wellbeing_t0_EV <- 30 + 2*treatment - 2*stonesize_EV

wellbeing_t1_EV-wellbeing_t0_EV
```

```
## [1] 2
```

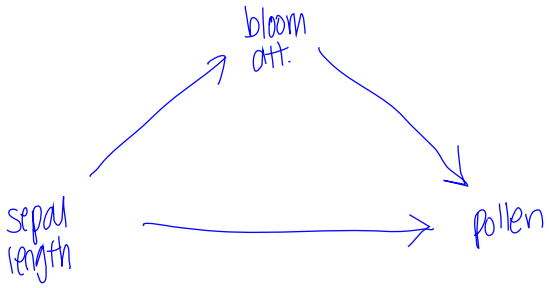
```
## Note that in the simulated data, treatment is deterministically determined by stone s
ize - that is, no probabilities there, simply everyone with a large stone gets treatment
1. Because size is a confounder, this means that people with poor wellbeing outcomes wil
l tend to get treatment 1, while, as we see from the true causal effect, treatment 1 act
ually produces better wellbeing in general.
```

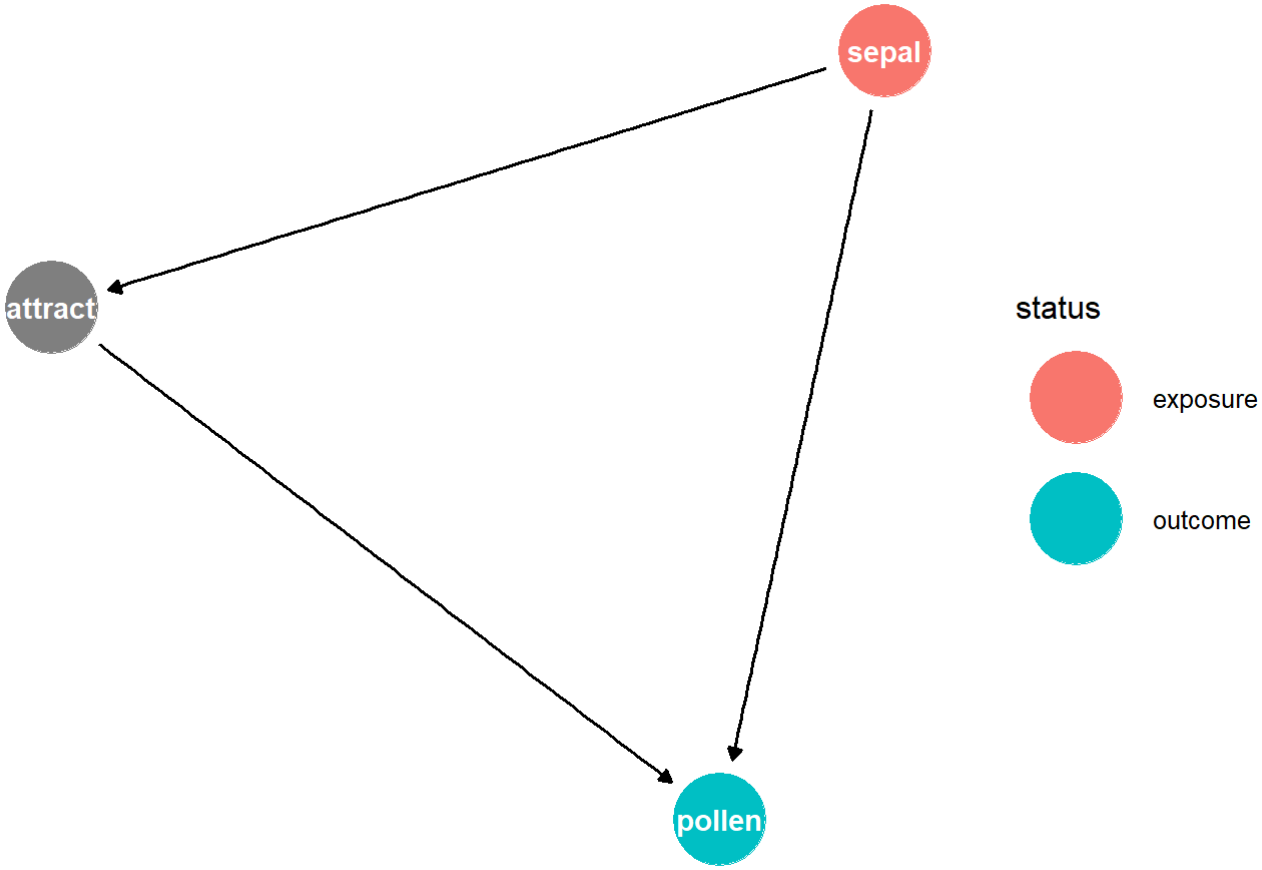
< mediator example >

1.3 Observational Data 3 - “Does length matter?”

In this exercise consider a (simulated) study on the amount of pollen flowers receive (outcome of interest) from pollinators, given their sepal length (potential cause of interest) and a general measure of bloom attractiveness (third variable). It is observational data.

▶ Based on the description above, what do you believe is the most reasonable causal model (DAG) for these data?






```
## I have chosen a DAG where sepal length determins how attractive a bloom is (I imagine
larger petals will produce more attractive blooms), and how attractive a bloom is determ
ines how much pollen it receives, given that more pollinators may visit is. I also imagi
ne sepal length has a direct effect on pollen, because it may be harder for pollinators
to get out of flowers with longer petals, hence more opportunity to leave more pollen. Y
ou may have different ideas about this and hence a different DAG. Note I (again) ignored
potential unobserved confounders here.
```

▶ Take a look at the dataframe `pollen_data` .

1.3.1 Simpsons Paradox or Berksons Bias

▶ Use linear regression to uncover either an example of Simpsons Paradox or Berksons Bias (your choice). Discuss the results.

```
#simpsons#
pollen_marg <- lm(pollen~sepal_length,data=pollen_data)
summary(pollen_marg)
```

```
##
## Call:
## lm(formula = pollen ~ sepal_length, data = pollen_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41156 -0.07373  0.00051  0.07480  0.34329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.524290   0.025516   20.55   <2e-16 ***
## sepal_length 0.106500   0.005029   21.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1132 on 498 degrees of freedom
## Multiple R-squared:  0.4739, Adjusted R-squared:  0.4728
## F-statistic: 448.5 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
pollen_cond <- lm(pollen~sepal_length+bloom_attract,data=pollen_data)
summary(pollen_cond)
```

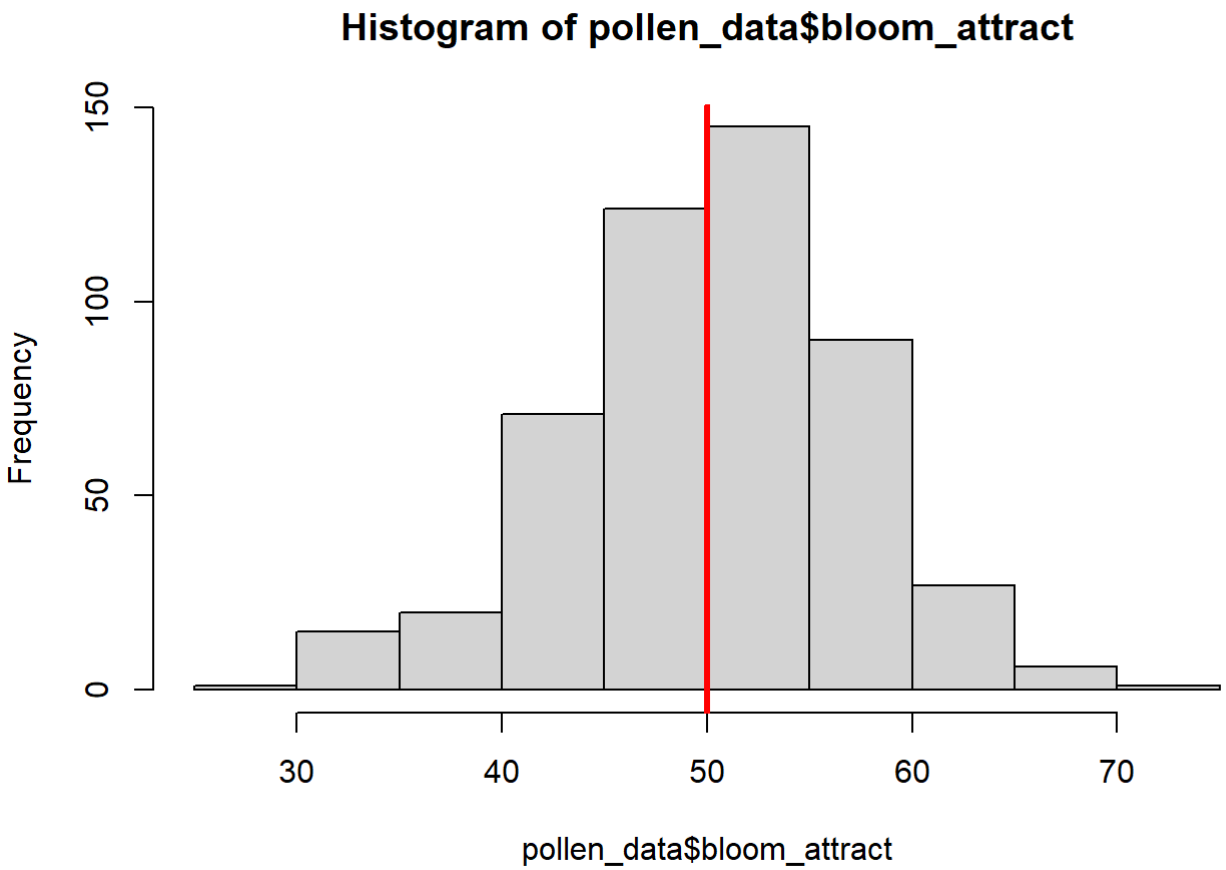
```
##
## Call:
## lm(formula = pollen ~ sepal_length + bloom_attract, data = pollen_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17273 -0.03699 -0.00053  0.03640  0.18137
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0051951  0.0169807   0.306   0.7598
## sepal_length  0.0071709  0.0032970   2.175   0.0301 *
## bloom_attract 0.0201787  0.0004733  42.632   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05252 on 497 degrees of freedom
## Multiple R-squared:  0.887,  Adjusted R-squared:  0.8866
## F-statistic:  1951 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
#Berksons. In the example below, we select for flowers that have bloom attractiveness hi
gher than 50 (but you may select in different ways, of course, and this may alter the re
sults).

quantile(pollen_data$bloom_attract, probs = seq(0, 1, 0.1), na.rm = FALSE,
         names = TRUE, type = 7)
```

```
##       0%      10%      20%      30%      40%      50%      60%      70%
## 29.28626 40.98337 44.51708 47.12024 49.03087 50.64113 52.18181 54.07859
##      80%      90%     100%
## 55.82669 58.50389 70.03212
```

```
hist(pollen_data$bloom_attract)
abline(v=50, col="red", lwd=3)
```

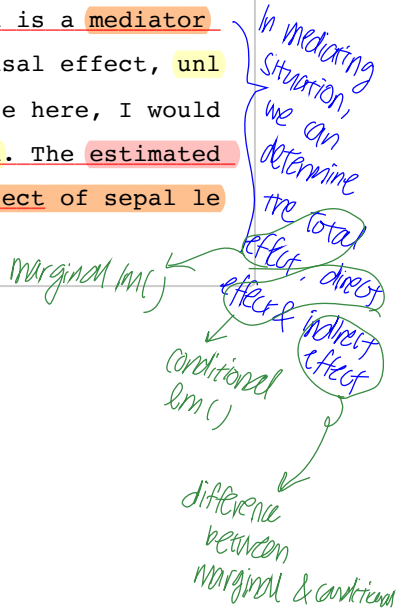
" at approximately median.,




```
highatt_data = pollen_data[which(pollen_data$bloom_attract>50),]

pollen_highatt <- lm(pollen~sepal_length,data=highatt_data)
summary(pollen_highatt)
```

```
##
## Call:
## lm(formula = pollen ~ sepal_length, data = highatt_data)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.269887 -0.054109 -0.004626  0.050222  0.255858
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.85303    0.03574  23.871  < 2e-16 ***
## sepal_length  0.05609    0.00643   8.724 2.95e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08601 on 267 degrees of freedom
## Multiple R-squared:  0.2218, Adjusted R-squared:  0.2189
## F-statistic:  76.1 on 1 and 267 DF,  p-value: 2.954e-16
```

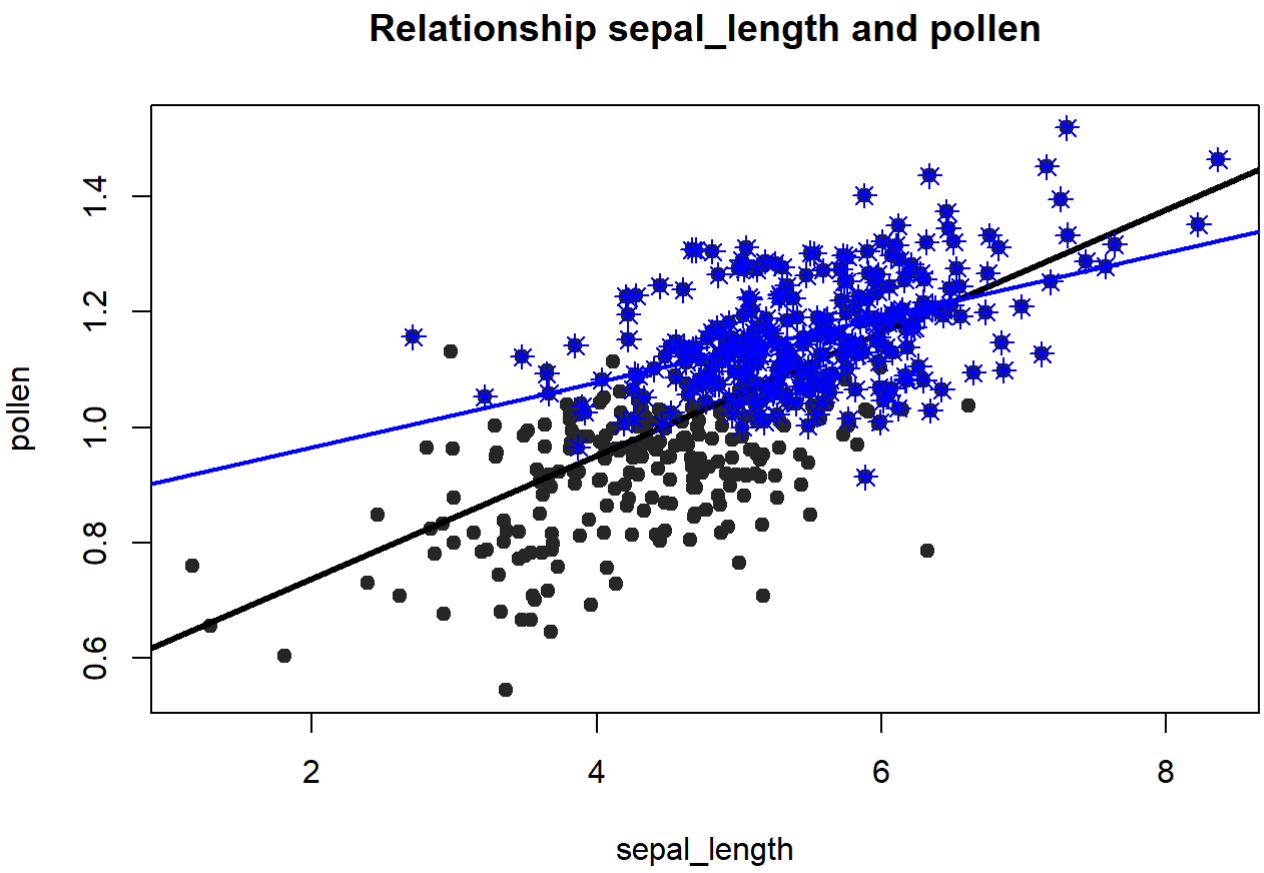
```
## We see that the marginal relationship and conditional relationship are different - th
ese generally do not have to be the same. Marginally, the treatment effect is significan
tly positive (.11), meaning flowers with a larger sepal length are expected to have a hi
gher amount of pollen received. Conditionally on bloom attractiveness, the effect of sep
al length is still positive and significant, but the effect has decreased in size (.00
7).  When we condition on attractiveness by selecting only flowers with attractiveness h
igher than 50 (the median approximately), we see an effect between the initial condition
al and marginal relationship of 0.056. again somewhere in between...
##
## Based on my initial DAG, I believe the variable we condition (size) on is a mediator
variable, hence I should not control for it to obtain the appropriate causal effect, unl
ess I want to specifically estimate the direct causal effect. In each case here, I would
conclude that there is a positive causal effect of sepal length on pollen. The estimated
total effect being .11, and the direct effect .007. Hencem the direct effect of sepal le
ngth here is estimated to be smaller than the indirect effect.
```



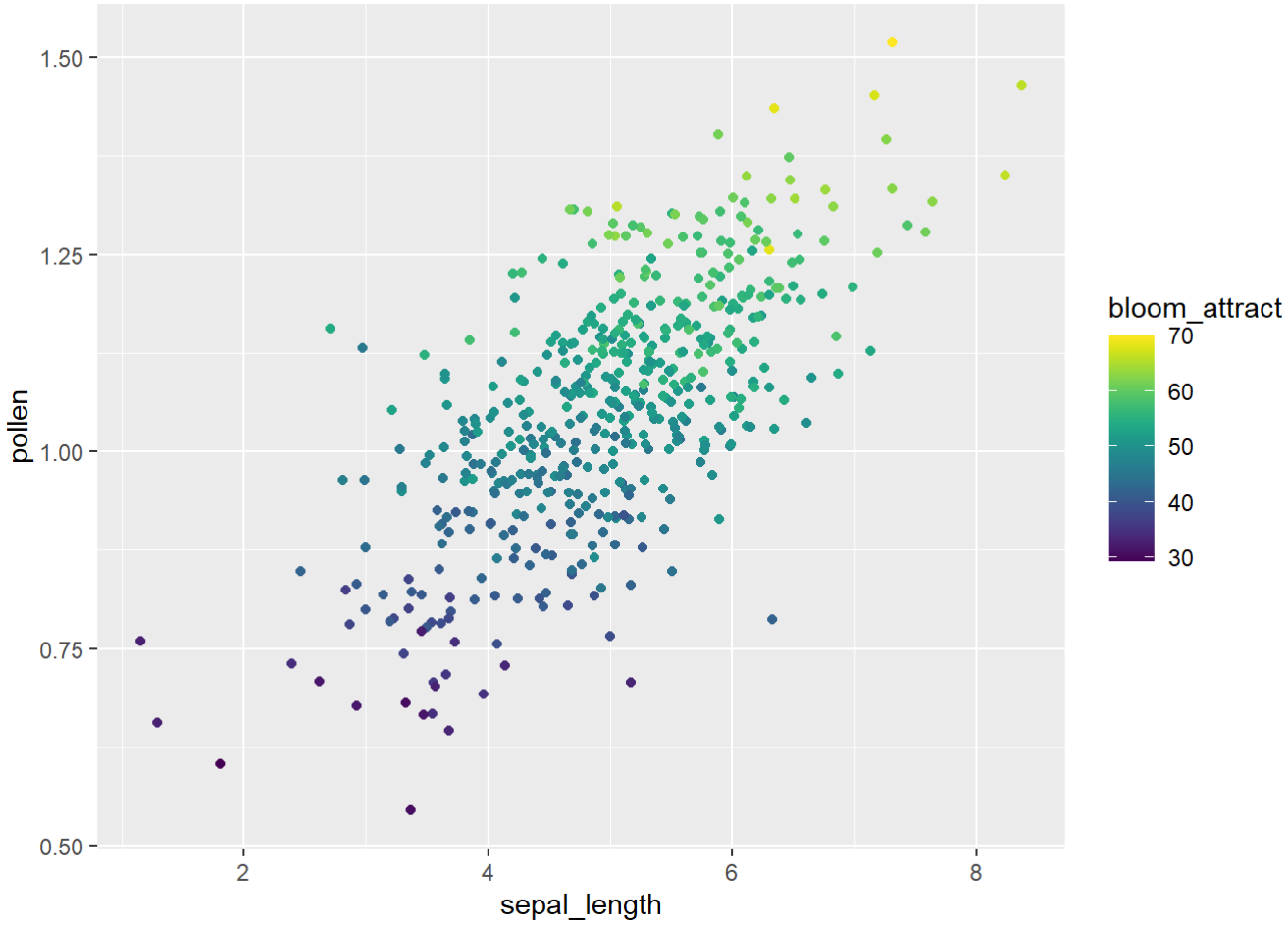
1.3.2 Visualize

▶ Use scatterplots to illustrate the Paradox.

```
plot(pollen_data$sepal_length, pollen_data$pollen, main = "Relationship sepal_length and
pollen", ylab="pollen", xlab="sepal_length", pch=19, col="grey15")
abline(a = pollen_marg$coefficients[1] , b = pollen_marg$coefficients[2], col = "black",
lwd=3)

points(highatt_data$sepal_length, highatt_data$pollen, pch=8,col="blue",cex=1.2)
abline(a = pollen_highatt$coefficients[1] , b = pollen_highatt$coefficients[2], col = "b
lue", lwd=2)
```




```
#Optional, to visualize the effect of conditioning on Study Succes in more of a continuo
us fashion
library(ggplot2)
library(viridis)
# Gradient color
ggplot(pollen_data, aes(x = sepal_length, y = pollen, colour = bloom_attract)) +
  geom_point()+
  scale_color_viridis(option = "D")
```

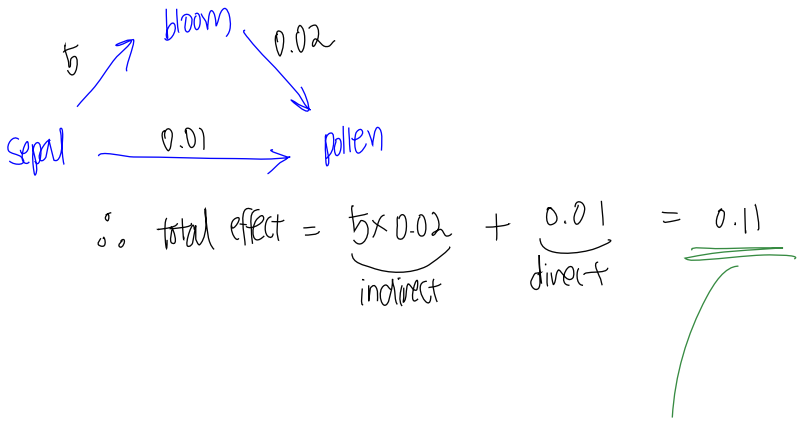


1.3.3 The true average causal effect

The data was simulated in the following way:

```
set.seed(133222058) # set the seed for comparison
n <- 500          # total sample size
sepal_length <- rnorm(n,5,1)
bloom_attract= 25+5*sepal_length + rnorm(n,0,5)
pollen <- 0 + .02*bloom_attract + .01*sepal_length + rnorm(n,0,.05)
```

▶ Calculate the true causal effect by mimicking an intervention. Discuss the results.



```
###sepal_length = 1###
sepal_length <-1
bloom_attract_EV_sl1= 25+5*sepal_length
pollen_EV_sl1 <- 0 + .02*bloom_attract_EV_sl1 + .01*sepal_length

###sepal_length = 0###
sepal_length <-0
bloom_attract_EV_sl0= 25+5*sepal_length
pollen_EV_sl0 <- 0 + .02*bloom_attract_EV_sl0 + .01*sepal_length

pollen_EV_sl1-pollen_EV_sl0
```

```
## [1] 0.11
```

```
## From the linear structural causal model you see that there is indeed a partial mediation going on, where sepal length affects pollen both directly and indirectly via bloom attractiveness. The total effect is 5*.02 (indirect) plus .01 (direct).
```

2. Controlling for Pretests, ANCOVA, Change Scores

This exercise will help you to better understanding of the effects of controlling for pre-tests and the relation between ANCOVA approaches and Change-score approaches. In this exercise you will practice with different scenarios (or data generating mechanisms) that give rise to data in which there is an outcome variable Y2, a (binary) treatment variable X, and a covariate Y1 that is a previous measurement of the same variable as the outcome variable.

The scenarios under which we generate data are:

- a randomized controlled trial (RCT), in which treatment assignment X is independent of pretest Y1
- a scenario in which treatment groups X already existed prior to the pretest Y1, such that the pretest is a mediator
- a scenario in which treatment assignment X is entirely based on the pretest score Y1, such that the pretest is a confounder
- a scenario in which treatment assignment X is partly based on the pretest score Y1, such that the pretest is a confounder

In each scenario, the data for the outcome variables Y2 are simulated using the regression model: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_i$. The *key difference* between these scenarios is in *how* and *why* treatment X and pretest Y1 are related.

For each dataset, you will then estimate four models:

- an ANCOVA model with the posttest score Y2 as the outcome: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_i$
- an ANCOVA model with the gain score G=Y2-Y1 as the outcome (in other words, a change score model where we control for pre-test): $Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_i$
- a change score model with only X as a predictor: $Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + e_i$
- a marginal model in which Y1 is not included: $Y_{2i} = \phi_0 + \phi_1 X_i + e_i$

2.1 RCT

We begin with simulating data according to an RCT. This implies that treatment X and pretest Y1 are independent of each other, while the posttest Y2 may depend on both.

The DAG for this model looks like this:

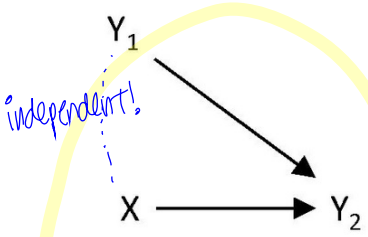


Fig.1a: DAG of the RCT model

The DAG for this model can be extended to include the gain score G=Y2-Y1:

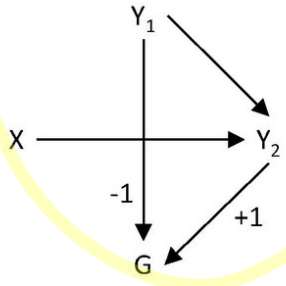


Fig.1b: DAG of the RCT model with change score included

2.1.1 Simulate data

We first need to simulate X (treatment) and Y1 (pretest). Note that in an RCT these will be unrelated (as is also clear from the DAG).

▶ Use the following code to simulate the data of an RCT.

```r
set.seed(482) # set the seed for comparison

N <- 500          # total sample size

# For Y1
mY1 <- 115    # mean on pretest
sdY1 <- 20    # sd on pretest
Y1 <- rnorm(N,mY1,sdY1)   # simulate Y1

# For X
X <- rbinom(N,1,0.5)      # simulate treatment

# For Y2
b0 <- 70          # intercept
b1 <- 20          # causal effect
b2 <- .4          # covariate effect
sdE2 <- 5         # within-group residual sd
Y2 <- b0 + b1*X + b2*Y1 + rnorm(N,0,sdE2)

dat1 <- data.frame(Y1,Y2,X)
```

▶ Consider the correlations between Y1, Y2, and X; what can you say about these?

```r
round(cor(dat1),3)
```

```
##       Y1    Y2     X
## Y1 1.000 0.635 0.019
## Y2 0.635 1.000 0.703
## X  0.019 0.703 1.000
```

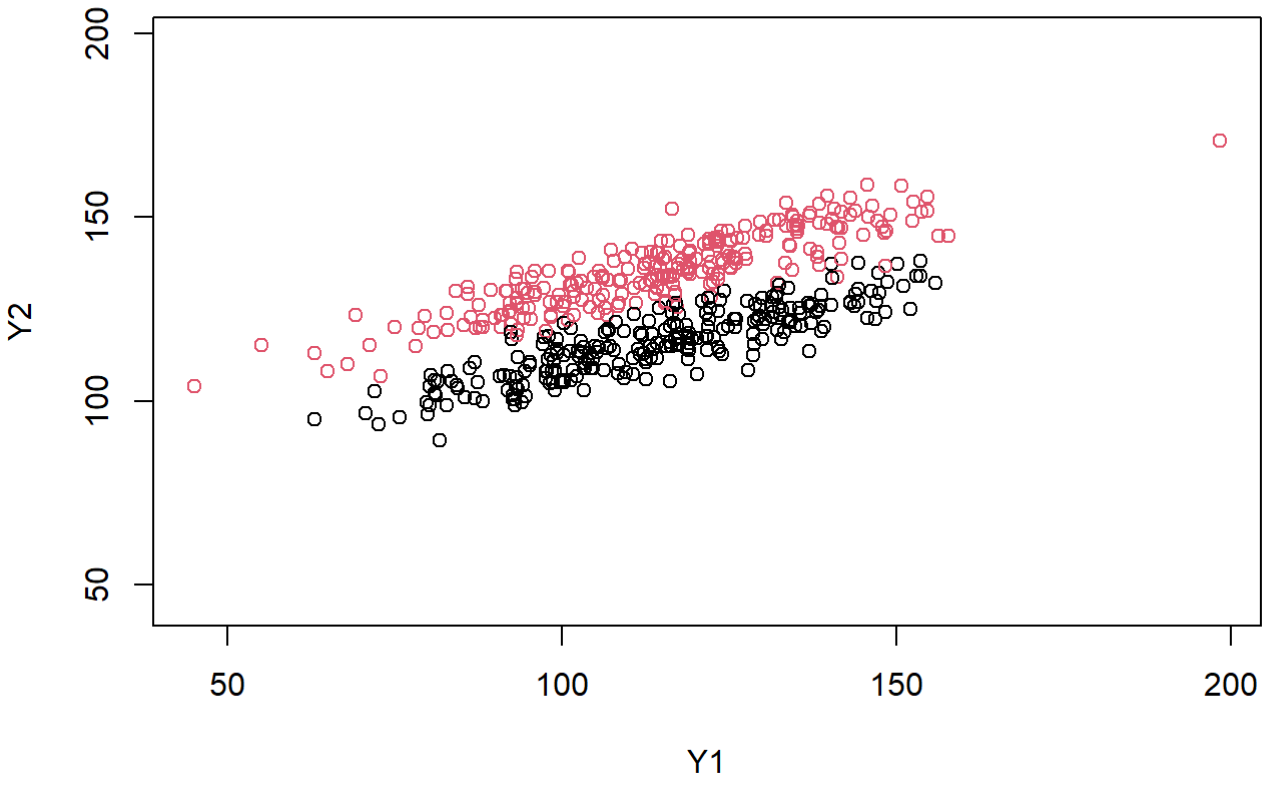
```r
# Note that Y1 and Y2 are continuous variables, while X is dichotomous.
# To determine the correlation between these three variables, we can
# use Pearson's correlations.
#
# It shows that:
#   a) Y1 and Y2 are strongly positively related
#   b) X and Y2 are strongly positively related
#   c) Y1 and X are not related (as expected)
```

2.1.2 Plot the data

Now we will make plots of the data like the ones used in the lecture (and the literature).

▶ First create a scatter plot with Y1 on the x-axis and Y2 on the y-axis, and with separate colors for the two treatment groups. Indicate for all the parameters defined above (mY1, sdY1, b0, b1, b2, sdE2) what feature in the plot they represent.

```
# For plotting it may be useful to find the overall minimum
# and maximum across the pretest and posttest, to make the two
# axes comparable
minY <- min(c(Y1,Y2))
maxY <- max(c(Y1,Y2))

plot(x=dat1$Y1, y=dat1$Y2, col = (dat1$X+1),
        xlim=c(minY,maxY),ylim=c(minY,maxY),
        xlab="Y1", ylab="Y2")
```



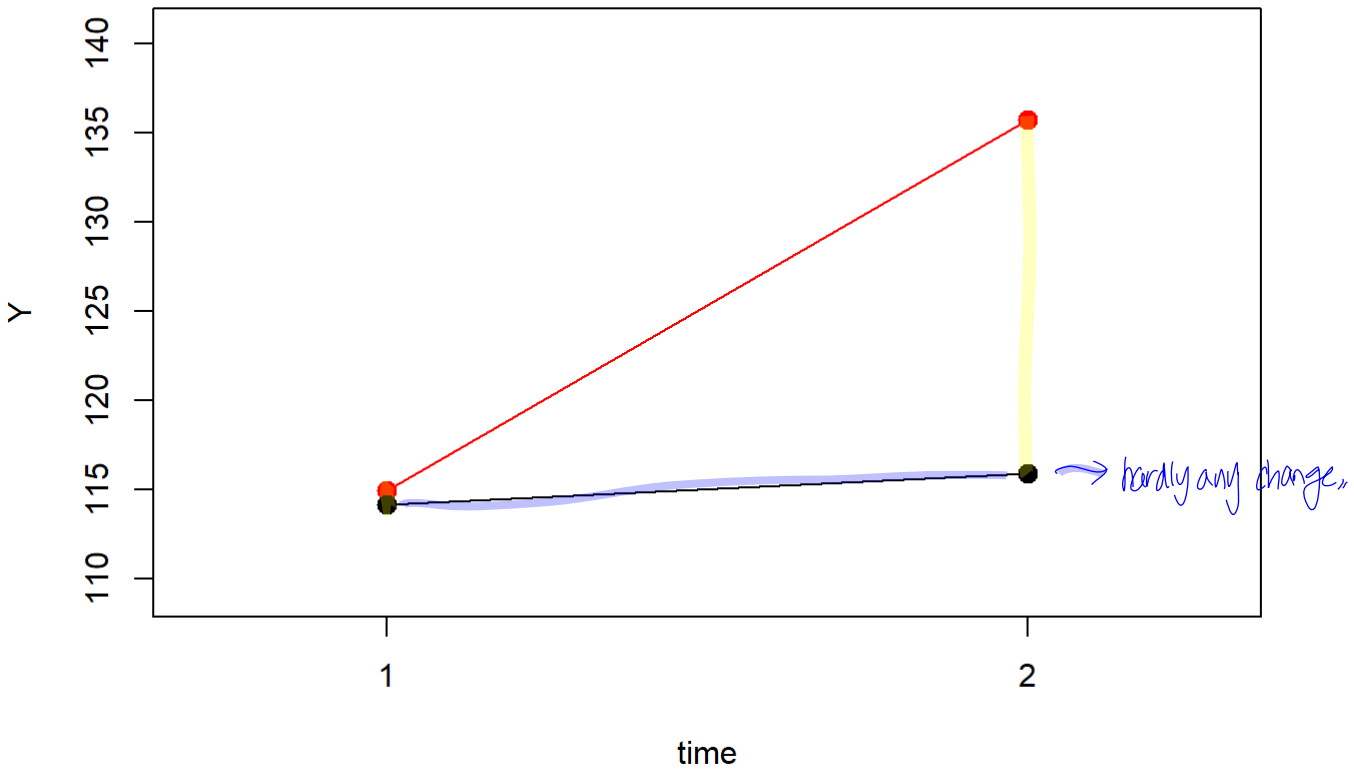
```
# mY1 is the mean on the x-axis
# sdY1 is the variability on the x-axis
# b0 is the intercept of the regression line (not shown)
# of the no treatment group (when X=0; in  black)
# b1 is the difference in intercepts of regression lines
# (not shown) of the the treatment group (X=1; in red)
# and the non-treatment group (X=0); it is equal to the ACE
# b2 is the slope of the regression line in each group
# (i.e., the effect of pretest on posttest)
# sdE2 is the residual variability around these regression lines
```

▶ Which part of this plot is reflecting the causal effect from the ANCOVA model?

```
# The difference in intercept between the two groups, or:
# the distance between the parallel regression lines of the two groups
```

▶ Second, make a plot of the means of the pretest scores and post test scores of each group. In the plot, place the two time points on the x-axis and the means of the two groups separately at each time point on the y-axis. Connect the means that belong to the same group using a line plot, and use different colors for each group.

```
# Means of:
# - pretest non-treatment group (mY10)
# - pretest treatment group (mY11)
# Note that both should be  about mY1=115 (value used in simulation)
mY10 <- mean(Y1[X==0])
mY11 <- mean(Y1[X==1])

# Means of:
# - posttest non-treatement group (mY20)
# - posttest treatment group (mY21)
# Note that the first should be about:
# mY20 = b0 + b2*mY10 = 70 + 0.4*115 = 116
# and the second should be about:
# mY21 = b0 + b1 + b2*mY11 = 70 + 20 + 0.4*115 = 136
mY20 <- mean(Y2[X==0])
mY21 <- mean(Y2[X==1])

# Gather means on both occasions per treatment condition
mYX0 <- c(mY10,mY20)
mYX1 <- c(mY11,mY21)

minY <- min(mYX0,mYX1)
maxY <- max(mYX0,mYX1)

plot(c(1,2), mYX0, type="l",
             xlim=c(0.7,2.3),
             ylim=c(minY-5,maxY+5),xaxt="n",
             xlab="time",
             ylab="Y")
lines(c(1,2),mYX1,col="red")
points(c(1,2),mYX1,pch=19,cex=1.3,col="red")
points(c(1,2),mYX0,pch=19,cex=1.3)
axis(side=1, at=seq(1, 2, by=1))
```



▶ How can this plot be used to determine whether in a change score model we would find evidence for a causal effect?

```
# A causal effect of X on the gain score (G=Y2-Y1) shows up as a
# different distance between the two lines (groups) at the two occasions;
# this is referred to as differenc-in-differences.
# Here there is no difference at first time point (due to random assignment!),
# whereas there is a 20 point difference at the second time point; this
# indicates there is a causal effect of X on Y2.
# Note furher that when no treatment is given, there is hardly any
# change between the means on pretest and posttest.
```

2.1.3 Analyze the data

Next, we analyze the data using the four models.

▶ First, estimate an ANCOVA (e.g., using the function `lm()`), with Y2 as the outcome, X as the grouping variable, and Y1 as the covariate (that is: X and Y1 are both predictors in your regression). Report the parameter (estimate, test, p-value) that is relevant for the causality question.

```
ANCOVA <- lm(Y2 ~ X + Y1, data=dat1)
summary(ANCOVA)
```

```
##
## Call:
## lm(formula = Y2 ~ X + Y1, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.368  -3.164  -0.228   3.411  15.822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.83802    1.26455   52.85   <2e-16 ***
## X           19.50102    0.43838   44.48   <2e-16 ***
## Y1           0.42946    0.01074   40.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 497 degrees of freedom
## Multiple R-squared:  0.8801, Adjusted R-squared:  0.8796
## F-statistic:  1824 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X represents the difference in
# intercept between the two treatment groups; hence, the estimated
# causal effect here is significantly different from zero: 19.50
# (se=0.44), p<0.0001.
```

▶ Second, estimate an ANCOVA with the gain score (i.e., Y2-Y1) as the outcome variable, X as the grouping variable and Y1 as the covariate (that is: X and Y1 as predictors). Compare the results regarding the estimated causal effect of X from this model to those obtained above.

```
ANCOVA <- lm((Y2-Y1) ~ X + Y1, data=dat1)
summary(ANCOVA)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X + Y1, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.368  -3.164  -0.228   3.411  15.822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.83802    1.26455   52.85   <2e-16 ***
## X           19.50102    0.43838   44.48   <2e-16 ***
## Y1          -0.57054    0.01074  -53.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 497 degrees of freedom
## Multiple R-squared:  0.9046, Adjusted R-squared:  0.9042
## F-statistic:  2356 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X is EXACTLY the same across the two analyses.
```

▶ Third, estimate the change score model, by regressing the gain score (i.e., Y2-Y1) on X; note this is a regression model with the gain score as the outcome, and only treatment X as its predictor (it may also be recognized by some as an ANOVA on the gain score). Report the parameter (estimate, test, p-value) that is relevant for the causality question.

```
CSA <- lm((Y2-Y1) ~ X, data=dat1)
summary(CSA)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.265  -8.378  -0.492   8.649  39.228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7270     0.8066   2.141   0.0328 *
## X            19.0498     1.1318  16.832   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.65 on 498 degrees of freedom
## Multiple R-squared:  0.3626, Adjusted R-squared:  0.3613
## F-statistic: 283.3 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
# The causal effect is here estimated with the regression coefficient for X.
# Hence, there is evidence that X has a causal effect on the change, as it is
# estimated to be 19.05 (SE=1.13, p<0.0001).
```

▶ Fourth, estimate the marginal model in which Y2 is regressed on X (hence, Y1 is not included in any way). Report the parameter (estimate, test, p-value) that is relevant for the causal question.

```
Y2onX <- lm((Y2) ~ X, data=dat1)
summary(Y2onX)
```

```
##
## Call:
## lm(formula = (Y2) ~ X, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.699  -7.331  -0.099   7.600  35.119
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 115.8487     0.6410  180.73   <2e-16 ***
## X            19.8407     0.8993   22.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.05 on 498 degrees of freedom
## Multiple R-squared:  0.4943, Adjusted R-squared:  0.4932
## F-statistic: 486.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
# The causal effect is here estimated with the regression coefficient for X.
# Hence, there is evidence that X has a causal effect on the change, as it is
# estimated to be 19.84 (SE=0.90, p<0.0001).
```

2.1.4 Conclusion

When comparing the results, what is your conclusion about the causal effect of treatment on the outcome?

is reg. coef. of X for all analyses!

Model	Equation	Causal parameter	Estimate (SE)	p-value
ANCOVA	$Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_i$	β_1	19.50 (0.44)	<0.0001
ANCOVA on change score	$Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_i$	β_1	19.50 (0.44)	<0.0001
Change score analysis	$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + e_i$	γ_1	19.05 (1.13)	<0.0001
Marignal analysis	$Y_{2i} = \phi_0 + \phi_1 X_i + e_i$	ϕ_1	19.84 (0.90)	<0.0001

First, we see that doing an ANCOVA with the post-test (Y2) and an ANCOVA with the gain score (Y2-Y1) as the outcome both lead to the exact same estimate of the causal effect. This is in agreement with the expressions for the ANCOVA model that we found in the lecture.

Second, the other two approaches (change score analysis, and marginal model), lead to slightly different estimates (from the ANCOVA and each other). But the substantive conclusions are all the same. This is because Y1 and X are unrelated (as would be the case in an RCT due to random assignment).

2.2 Existing treatment groups ~ Y1 : mediator situation

A second scenario that we consider, is when there are pre-existing treatment groups. One possible example of such a scenario is the one described by Lord, where the groups were male versus female students.

The DAG that represents such a scenario looks like this:

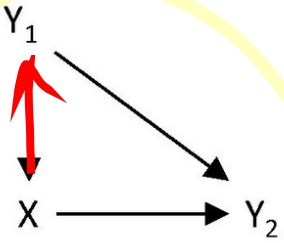


Fig.2a: DAG of pre-existing groups

The DAG for this scenario which also includes the gain score looks like this:

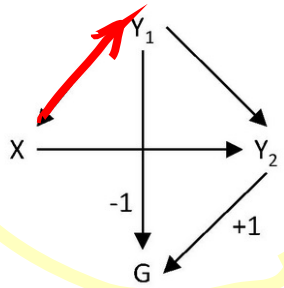


Fig.2b: DAG of pre-existing groups with change score included

2.2.1 Simulate data

Again, to simulate Y2, we make use of the regression model: $Y_{2i} = b_0 + b_1 X_i + b_2 Y_{1i} + e_{2i}$.

However, Y1 is now simulated in a different way: Its values will depend on X.

▶ Simulate the data using the code below.

```
set.seed(268) # set the seed for comparison

N <- 500          # total sample size

# For X:
X <- rbinom(N,1,0.5)  # simulate treatment

# For Y1:
mY10.true <- 100      # mean of non-treatment group at pretest
mY11.true <- 130      # mean of treatment group at pretest
sdY1 <- 15            # within-group sd at pretest (both groups)

# Note that when X=1, the first part will be zero; if X=0, the second
# part is zero; so this ensures that for each individual only one part
# remains when simulating Y1:
Y1 <- (1-X)*rnorm(N, mY10.true, sdY1) + X*rnorm(N, mY11.true, sdY1)

# As a result, the two groups now have different means (mY10, mY11) on the pretest.

# For Y2:
b0 <- 70          # intercept
b1 <- 20          # causal effect
b2 <- .4          # covariate effect
sdE2 <- 5         # within-group residual sd

Y2 <- b0 + b1*X + b2*Y1 + rnorm(N,0,sdE2)

dat2 <- data.frame(Y1,Y2,X)
```

▶ Look at the correlations between X, Y1 and Y2; what can you say about these?

```
round(cor(dat2),3)
```

```
##        Y1    Y2     X
## Y1 1.000 0.873 0.725
## Y2 0.873 1.000 0.904
## X  0.725 0.904 1.000
```

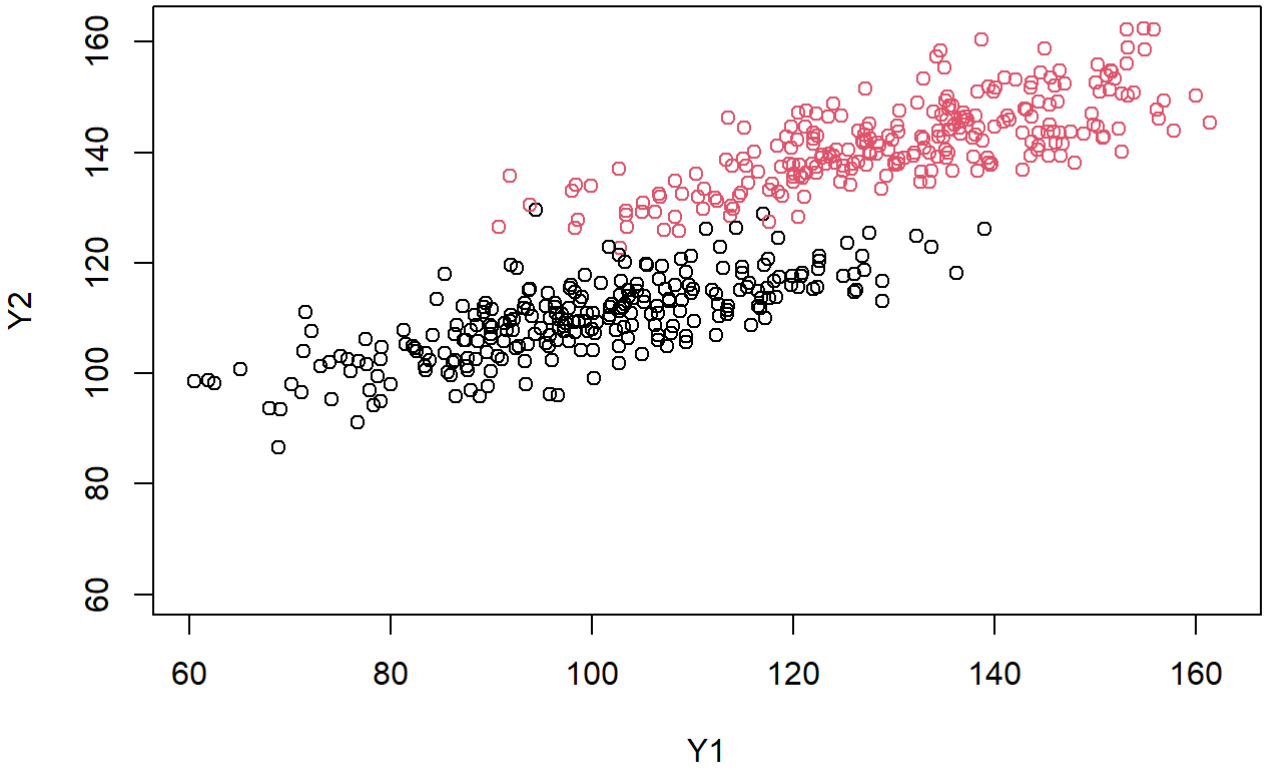
```
# Note that Y1 and Y2 are continuous variables, while X is dichotomous.
# To determine the correlation between these three variables, we can
# use Pearson's correlations.
#
# It shows that:
# Y1 and Y2 are strongly positively related
# X and Y2 are strongly positively related
# Y1 and X are strongly positively related
```
In RCT example, they weren't!

2.2.2 Plot the data

We will consider the same two kind of plots (scatterplot with regression lines, and plot of group means at both measurement occasions).

▶ First, plot the post-test Y2 against the pre-test Y1. Use different colors for the two groups. What part of this plot informs you on whether there is evidence for a causal effect when using the ANCOVA model?

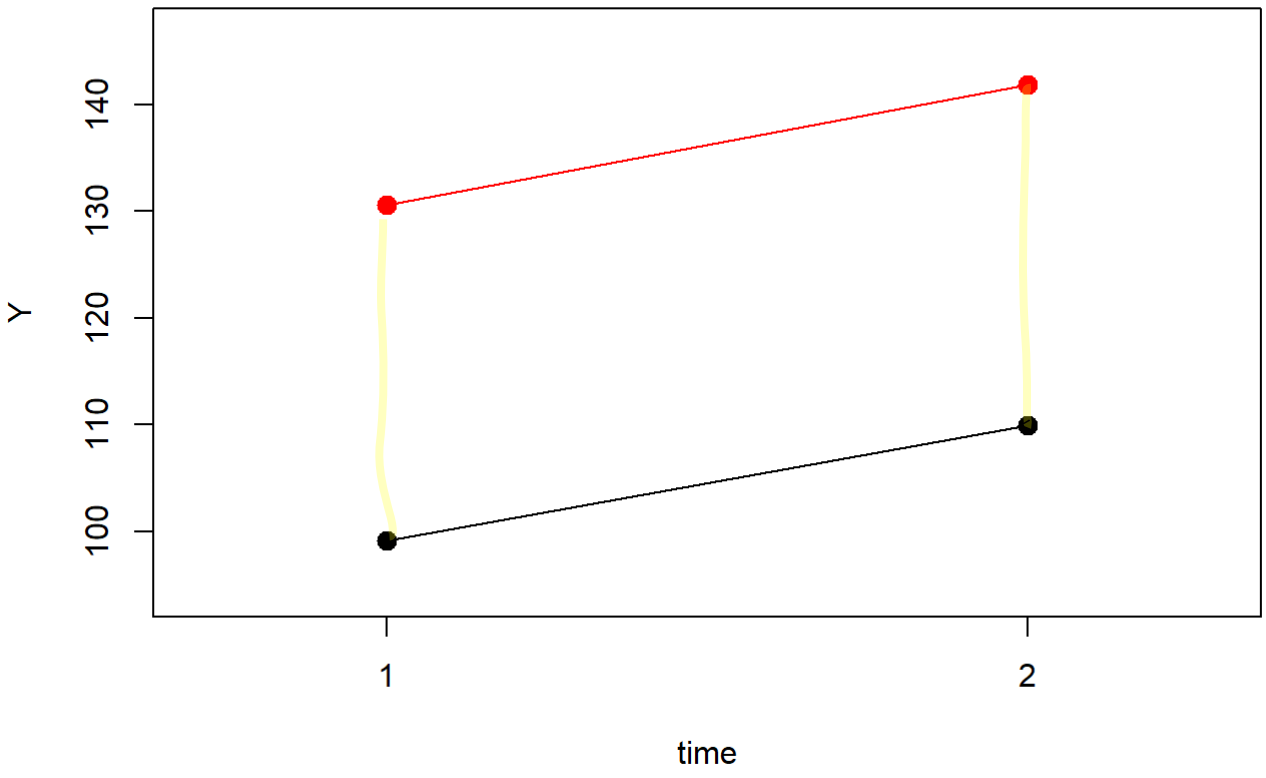
```
minY <- min(c(Y1,Y2))
maxY <- max(c(Y1,Y2))
plot(x=dat2$Y1, y=dat2$Y2, col = (dat2$X+1),
    xlim=c(minY,maxY),ylim=c(minY,maxY),
    xlab="Y1", ylab="Y2")
```




```
# A causal effect of X on Y in the ANCOVA model would
# show up as a difference in intercepts of the groups.
# This difference is actually b1 used in simulating the data.
```

▶ Second, make a plot in which you have the two time points on the x-axis and the means of the two groups separately (i.e., the plot related to the changes score approach). How can this plot be used to determine whether a change score analysis would provide evidence for a causal effect?

```
# Means of:
# - pretest non-treatment group (mY01)
# - pretest treatment group (mY11)
mY10 <- mean(Y1[X==0])
mY11 <- mean(Y1[X==1])

# Means of:
# - posttest non-treatement group (mY20)
# - posttest treatment group (mY21)
mY20 <- mean(Y2[X==0])
mY21 <- mean(Y2[X==1])

# Gather means on both occasions per treatment condition
mYX0 <- c(mY10,mY20)
mYX1 <- c(mY11,mY21)

minY <- min(mYX0,mYX1)
maxY <- max(mYX0,mYX1)

plot(c(1,2), mYX0, type="l",
            xlim=c(0.7,2.3),
            ylim=c(minY-5,maxY+5),xaxt="n",
            xlab="time",
            ylab="Y")
lines(c(1,2),mYX1,col="red")
points(c(1,2),mYX1,pch=19,cex=1.3,col="red")
points(c(1,2),mYX0,pch=19,cex=1.3)
axis(side=1, at=seq(1, 2, by=1))
```




```
# A causal effect of X on Y would show up as a different distance between
# the two lines at the two occasions. Here there is a very small effect,
# as the two lines are a bit further apart at the second measurement occasion.
```

2.2.3 Analyze the data

Again, we analyze the data with four models.

▶ First, estimate the ANCOVA with Y2 as the outcome, and X and Y1 as the predictors. Report the parameters (estimate, test, p-value) that are relevant for the causality question.

```
ANCOVA <- lm(Y2 ~ X + Y1, data=dat2)
summary(ANCOVA)
```

```
##
## Call:
## lm(formula = Y2 ~ X + Y1, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9743  -3.3461  -0.1449   3.0778  21.4051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.90937    1.55533   46.88   <2e-16 ***
## X           20.22580    0.66666   30.34   <2e-16 ***
## Y1           0.37341    0.01534   24.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.135 on 497 degrees of freedom
## Multiple R-squared:  0.9163, Adjusted R-squared:  0.916
## F-statistic:  2721 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X represents the difference in
# intercept between the two treatment groups; hence, the estimated
# causal effect here is significantly different from zero:
# 20.23 (se=0.67), p<0.0001.
```

▶ Second, estimate the ANCOVA on on the gain score (i.e., Y2-Y1), with X and Y1 as the predictors. Compare the results to the results obtained with an ANCOVA on Y2.

```
ANCOVA.CS <- lm((Y2-Y1) ~ X + Y1, data=dat2)
summary(ANCOVA.CS)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X + Y1, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9743  -3.3461  -0.1449   3.0778  21.4051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.90937    1.55533   46.88   <2e-16 ***
## X           20.22580    0.66666   30.34   <2e-16 ***
## Y1          -0.62659    0.01534  -40.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.135 on 497 degrees of freedom
## Multiple R-squared:  0.7705, Adjusted R-squared:  0.7696
## F-statistic: 834.4 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# As we also saw in the RCT scenario, the parameter estimate
# for the effect of X on Y2 and on Y2-Y1 is estimated to be
# exactly the same.
# The model fitted here is: Y2 - Y1 = b0 + b1 X + (b2-1) Y1 + e2
# that is, it gives th same b0 and b1 as the ANCOVA model above
# the ony difference is in the coefficient for Y1; here it is
# b2-1 (where b2 is the b2 from the previous model)
```

▶ Third, estimate the change score model by regressing the gain score (Y2-Y1) on X. Report the parameter (estimate, test, p-value) that is relevant for the causality question.

```
CSA <- lm((Y2-Y1) ~ X, data=dat2)
summary(CSA)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.831  -6.951   0.476   7.137  32.663
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.7877     0.6757  15.964   <2e-16 ***
## X             0.4922     0.9576   0.514    0.607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.71 on 498 degrees of freedom
## Multiple R-squared:  0.0005302,  Adjusted R-squared:  -0.001477
## F-statistic: 0.2642 on 1 and 498 DF,  p-value: 0.6075
```

```
# The causal effect is here estimated with the regression coefficient for X.
# It is not significantly different from zero:
# 0.49 (SE=0.96), p=0.61
```

▶ Fourth, estimate the marginal model by regressing Y2 on X. Report the parameter (estimate, test, p-value) that is relevant for the causality question.

```
Y2onX <- lm(Y2 ~ X, data=dat2)
summary(Y2onX)
```

```
##
## Call:
## lm(formula = Y2 ~ X, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3303  -4.9666   0.0988   5.0648  20.4291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.9308     0.4793  229.33   <2e-16 ***
## X            31.9861     0.6793   47.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.594 on 498 degrees of freedom
## Multiple R-squared:  0.8166, Adjusted R-squared:  0.8162
## F-statistic:  2217 on 1 and 498 DF,  p-value: < 2.2e-16
```

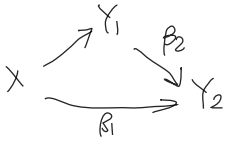
```
# The causal effect is estimated with the regression coefficient for X.
# It is significant and positive:
# 31.99 (SE=0.68), p<0.0000001
```

2.2.4 Conclusion

When comparing the results, what is your conclusion about the causal effect of treatment on the outcome, and what model should be preferred?

Model	Equation	Causal parameter	Estimate (SE)	p-value
ANCOVA	$Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_i$	β_1	20.23 (0.67)	0.0001
ANCOVA on change score	$Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_i$	β_1	20.23 (0.67)	0.0001
Change score analysis	$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + e_i$	γ_1	0.49 (0.96)	0.61
Marignal analysis	$Y_{2i} = \phi_0 + \phi_1 X_i + e_i$	ϕ_1	31.99 (0.68)	0.0001

Clearly, the two ANCOVAs lead to the same estimate of the causal effect; it is estimated to be about 20, which is the parameter used in the regression equation to simulate the data.



In contrast, the change score model is leading to the conclusion that treatment has no effect, while the marginal model (Y2 regressed on X) results in a causal effect of about 30.

Based on the DAGs, we know that X has a direct effect on Y2, which is obtained with the ANCOVA approach. But there is also an indirect effect, as X affects Y1 which in turn affects Y2 (for the sake of completeness, the latter is equal to b2*(mY11-mY01)).

The marginal model is helping us to estimate that. Hence, if we are interested in the total causal effect of X on Y2, we should consider the last model. But we can add also the estimate of the direct effect, to make the story richer.

2.3 Assignment excl. based on pre-test

In the previous scenario, we started with creating different groups and then simulated the pre-test. Here, we will start with creating a pretest score Y1, and then create two groups X. The groups will be based on a mean-split, and are thus *fully determined* by the pre-test score.

An example of such a scenario is when people do a test (e.g., a test of their soccer skills), and all indiduals that score above a certain threshold are then given the treatment (i.e., assigned to X=1, which consists of two additional training sessions per week), whereas the people who score below the threshold are not (i.e., they are assigned to X=0); then, at the end of the year, you measure them agian (i.e., their soccer skills), and you want to determine whether the treatment (additional trainings) had a beneficial effect.

The DAG of this scenario looks like this:

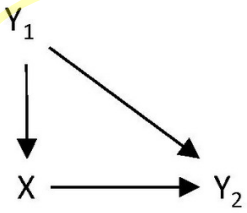


Fig.3a: DAG of groups based on pre-test score

The DAG can be extended with the gain score in it to look like this:

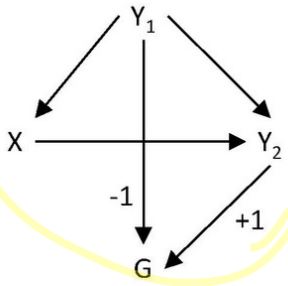


Fig.3b: DAG of groups based on pre-test score

2.3.1 Simulate data

We start with simulating the data. For this, we first need to simulate Y1, which then determines X. Then we can simulate Y2. To make it comparable to the previous example, we use the mean and standard deviation from the Y1 in the previous scenario.Furthermore, when creating Y2, we will make use of the same b0, b1, b2 and residual standard deviation as in the scenario above.

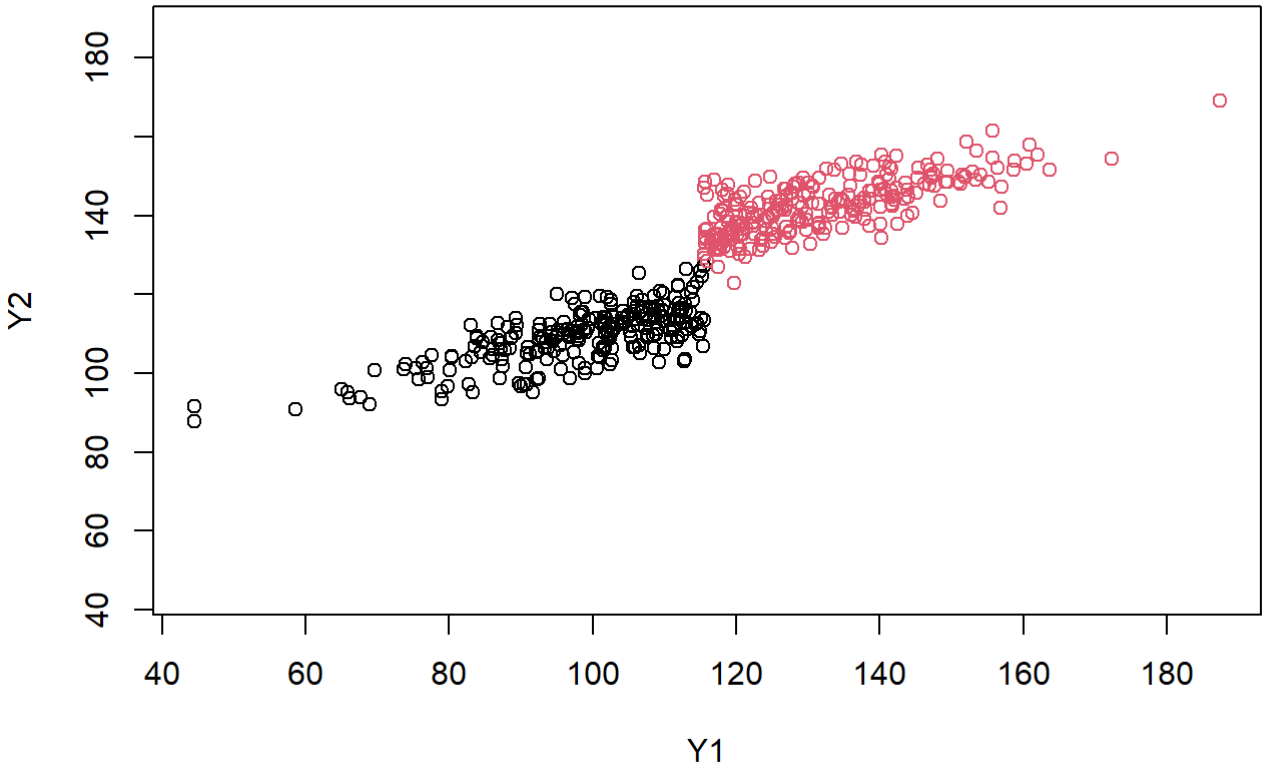
▶ Use the following code to simulate the data.

```
sd.pre <- sd(Y1)                    # determine total sd of Y1 in simulations above
mean.pre <- mean(Y1)               # determine total mean of Y1 in sims above

# create a new pre-test score with the same mean and variance
Y1 <- rnorm(N,mean.pre,sd.pre)

# Use a mean split for Y1 to create two groups:
X <- rep(0,N)
X [Y1>mean(Y1)] <- 1          # Treatment groups based on mean split of pretest

# Simulate Y2
Y2 <- b0 + b1*X + b2*Y1 + rnorm(N,0,sdE2)

dat3 <- data.frame(Y1,Y2,X)
```

2.3.2 Plot the data

▶ First, make the pre-post test plot again and describe what this teaches us.

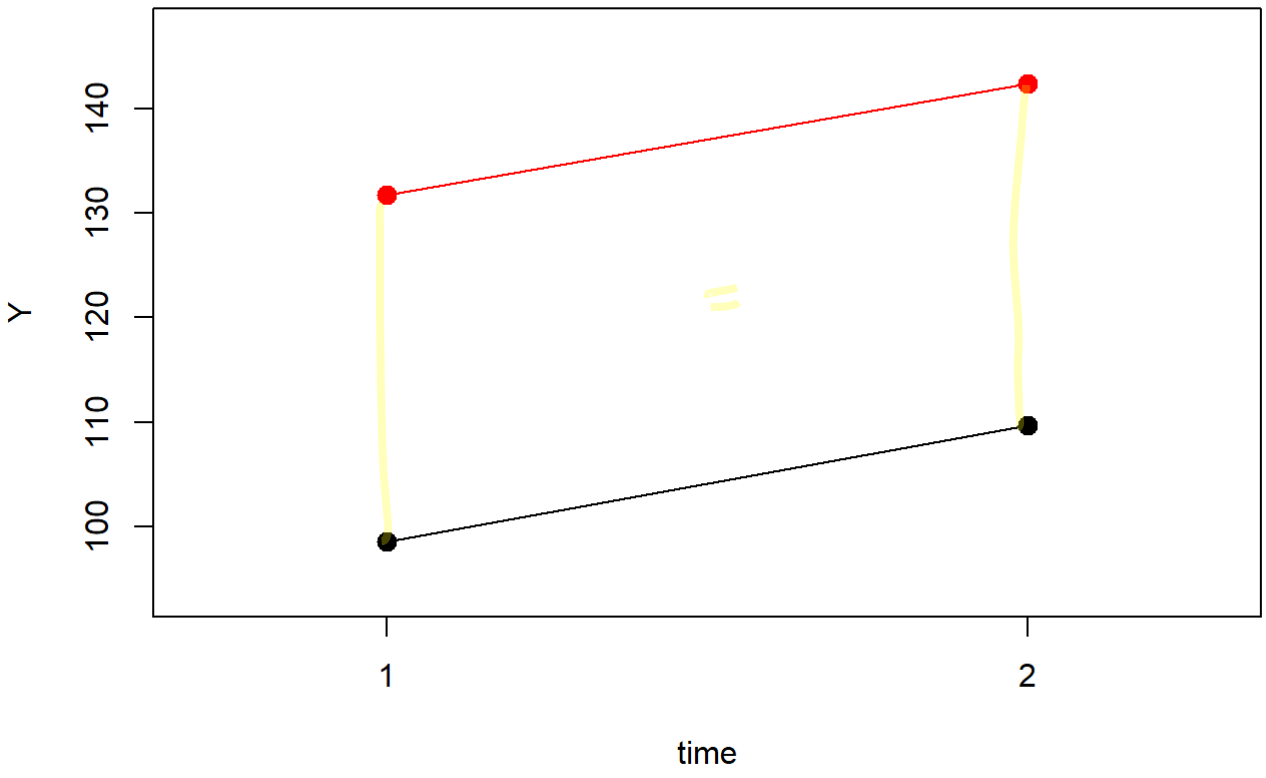
```
minY <- min(c(Y1,Y2))
maxY <- max(c(Y1,Y2))
plot(x=dat3$Y1, y=dat3$Y2, col = (dat3$X+1),
    xlim=c(minY,maxY),ylim=c(minY,maxY),
    xlab="Y1", ylab="Y2")
```




```
# A causal effect of X on Y according to an ANCOVA model shows up
# as a difference in intercept for the two groups.
# This difference is actually b1 used in simulating the data.
#
# Note further that the stark difference on the pretest score
# does not mean there are different groups on the pretest; in fact,
# when the data were created, there were no groups yet; hence,
# seeing a difference on the pretest, does not inform us whether
# Y1 caused X or X caused Y1!
```

▶ Second, make the plot of the means.

```
# Determine means at each occasion for each group
mY10 <- mean(Y1[X==0])
mY11 <- mean(Y1[X==1])
mY20 <- mean(Y2[X==0])
mY21 <- mean(Y2[X==1])

# Gather means on both occasions per treatment condition
mY0 <- c(mY10,mY20)
mY1 <- c(mY11,mY21)

minY <- min(mY0,mY1)
maxY <- max(mY0,mY1)

plot(c(1,2), mY0, type="l",
            xlim=c(0.7,2.3),
            ylim=c(minY-5,maxY+5),xaxt="n",
            xlab="time",
            ylab="Y")
lines(c(1,2),mY1,col="red")
points(c(1,2),mY1,pch=19,cex=1.3,col="red")
points(c(1,2),mY0,pch=19,cex=1.3)
axis(side=1, at=seq(1, 2, by=1))
```




```
# This plot helps to see what we will get from running a change score model:
# A causual effect of X on Y would show up as a different distance between
# the two lines at the two occasions. Here there the lines seem parallel, which
# would imply there is no causal effect of X on the gain.
```

2.3.3 Analyze the data

▶ First, estimate the ANCOVA model with Y2 as the outcome, and X and Y1 as its predictors. What conclusion can you draw about the causal effect based on this analysis?

```
ANCOVA <- lm(Y2 ~ X + Y1, data=dat3)
summary(ANCOVA)
```

```
##
## Call:
## lm(formula = Y2 ~ X + Y1, data = dat3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6433  -3.1055   0.1386   2.9358  12.8257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.39985    1.72669   40.19   <2e-16 ***
## X           19.23627    0.71780   26.80   <2e-16 ***
## Y1           0.40816    0.01723   23.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 497 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9258
## F-statistic:  3116 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X represents the difference in
# intercept between the two treatment groups; hence, the estimated
# causal effect here is significantly different from zero and positive:
# 19.24 (SE=0.72), p<0.000001
# which is close to the b1=20 with which we simulated.
```

▶ Second, estimate the ANCOVA with teh gain score Y2-Y1 as the outcome and Y1 and X as the predictors. What conclusion can you draw about the causal effect based on this analysis?

```
ANCOVA.CS <- lm((Y2-Y1) ~ X + Y1, data=dat3)
summary(ANCOVA.CS)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X + Y1, data = dat3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6433  -3.1055   0.1386   2.9358  12.8257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.39985    1.72669   40.19   <2e-16 ***
## X           19.23627    0.71780   26.80   <2e-16 ***
## Y1          -0.59184    0.01723  -34.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 497 degrees of freedom
## Multiple R-squared:  0.7037, Adjusted R-squared:  0.7025
## F-statistic: 590.2 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X in this model is
# exactly the same as that in the previous model
# (as we would expect based on the analytic results
# we discussed in the lecture).
```

▶ Third, estimate the change score model by regressing Y2-Y1 on X. What conclusion can you draw based on this analysis?

```
CSA <- lm((Y2-Y1) ~ X, data=dat3)
summary(CSA)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X, data = dat3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.913  -5.955  -0.023   5.850  35.935
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0677     0.5711  19.378   <2e-16 ***
## X            -0.3735     0.7982  -0.468     0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.921 on 498 degrees of freedom
## Multiple R-squared:  0.0004395,  Adjusted R-squared:  -0.001568
## F-statistic: 0.219 on 1 and 498 DF,  p-value: 0.64
```

```
# The causal effect is here estimated with the regression coefficient for X;
# it is not significantly different from zero:
#  -0.37 (SE=0.80), p=0.64
# which would lead us to conclude that the treatment has
# no effect on the change.
```

▶ Fourth, estimate the marginal model with Y2 as the outcome and X as the predictor. What conclusion can you draw based on this analysis?

```
Y2onX <- lm(Y2 ~ X, data=dat3)
summary(Y2onX)
```

```
##
## Call:
## lm(formula = Y2 ~ X, data = dat3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7100  -5.0590   0.2801   4.8511  26.7645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.6276     0.4537  241.63   <2e-16 ***
## X            32.7598     0.6341   51.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.087 on 498 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.8425
## F-statistic:  2669 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
# The causal effect is estimated with the regression coefficient for X;
# here it is significant and positive:
# 32.76 (SE=0.63), p<0.000001
# Hence, treatment has a positive effect on Y2, according to this analysis.
```

2.3.4 Conclusion

When comparing the results, what is your conclusion about the causal effect of X, and which analysis should be preferred?

Model	Equation	Causal parameter	Estimate (SE)	p-value
ANCOVA	$Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_i$	β_1	19.24 (0.72)	<0.0001
ANCOVA on change score	$Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_i$	β_1	19.24 (0.72)	<0.0001
Change score analysis	$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + e_i$	γ_1	-0.37 (0.80)	0.64
Marignal analysis	$Y_{2i} = \phi_0 + \phi_1 X_i + e_i$	ϕ_1	32.76 (0.63)	0.0001

Again, the first two models (ANCOVA and ANCOVA on the change score) lead to the exact same estimate. The other two models lead to different estimates: the third model leads to the conclusion that there is no causal effect, whereas the last model leads to the conclusion that there is a larger positive effect than that obtained with the ANCOVA model. In fact, these results are very similar to the results we obtained with the data in the previous scenario, which illustrates that they are not so informative by themselves.

To decide which analysis is informative if we are interested in obtaining an estimate of the ACE, we should use the DAGs: then, we would conclude that the ANCOVA model is the correct approach here, because Y1 is a confounder (X <- Y1 -> Y2) for which we need to adjust. Note however that it is a very specific scenario, in which there is no overlap between the groups on the covariate (which, btw, would be considered a violation of positivity in the Rubin causal framework!).

2.4 Assignment partly based on pre-test

In the previous scenario, we created different groups using a cut-off for the pre-test. As a result, there was no overlap between the groups on the pretest, which was clearly visible in the first plot (Y2 plotted against Y1) that we made.

Now, we will create data in which the group assignment is only *partly* based on the pre-test score. Hence, there should be partial overlap between the two groups on the pre-test. The DAGs for this scenario are the same as the ones for the previous scenario.

2.4.1 Simulate data

We start again with creating the pre-test score as we did in the previous scenario. Then we use this variable to assign people to one of two treatment conditions, but now the assignment is not deterministic: There should be overlap between the groups on the pretest score.

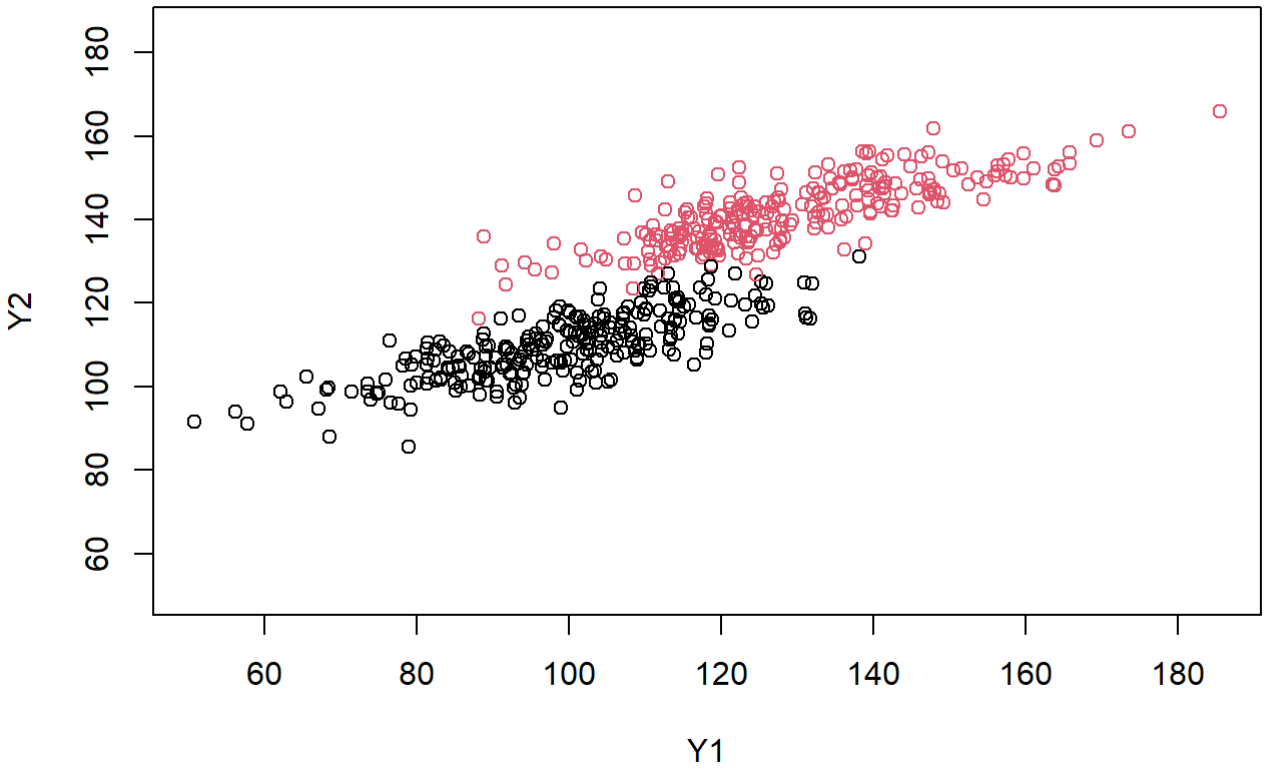
▶ We do this using a logistic regression model, using the following code:

```
Y1 <- rnorm(N, mean.pre, sd.pre)  # pre-test

# To create a treatment variable based on this pretest
# but the probability of being treated only partly depends on the pretest
z <- 0.15*(Y1-mean(Y1))           # linear combination
pr <- 1/(1+exp(-z))               # pass through an inv-logit function
X <- rbinom(N,1,pr)               # Bernoulli treatment variable

# Now we can create the post-test, like we have done before
Y2 <- b0 + b1*X + b2*Y1 + rnorm(N,0,sdE2)

dat4 <- data.frame(Y1,Y2,X)
```

2.4.2 Plot the data

▶ First, make the pre-post test plot.

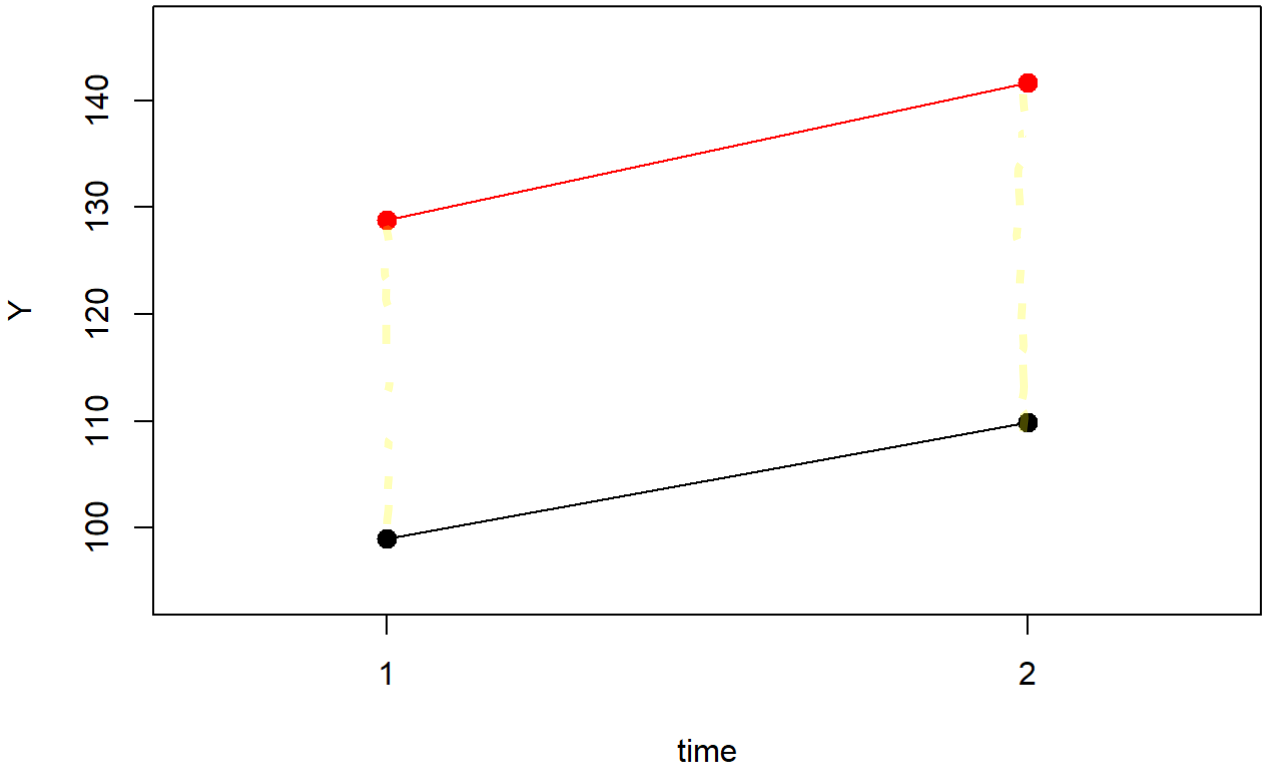
```
minY <- min(c(Y1,Y2))
maxY <- max(c(Y1,Y2))
plot(x=dat4$Y1, y=dat4$Y2, col = (dat4$X+1),
    xlim=c(minY,maxY),ylim=c(minY,maxY),
    xlab="Y1", ylab="Y2")
```




```
# A causal effect of X on Y in the ANCOVA model would show up as a difference in interce
pt
# for the two groups. This difference is actually b1 used in simulating the data.
```

▶ Second, make a plot of the means.

```
# Compute means at first and second occasion
mY10 <- mean(Y1[X==0])
mY11 <- mean(Y1[X==1])
mY20 <- mean(Y2[X==0])
mY21 <- mean(Y2[X==1])

# Gather means on both occasions per treatment condition
mY0 <- c(mY10,mY20)
mY1 <- c(mY11,mY21)

minY <- min(mY0,mY1)
maxY <- max(mY0,mY1)

plot(c(1,2), mY0, type="l",
            xlim=c(0.7,2.3),
            ylim=c(minY-5,maxY+5),xaxt="n",
            xlab="time",
            ylab="Y")
lines(c(1,2),mY1,col="red")
points(c(1,2),mY1,pch=19,cex=1.3,col="red")
points(c(1,2),mY0,pch=19,cex=1.3)
axis(side=1, at=seq(1, 2, by=1))
```




```
# A causal effect of X on Y in the change score model would show up as a
# different distance between the two lines at the two occasions.
# Here there is a very small effect, as the two lines are a bit further
# apart at the second measurement occasion.
```

2.4.3 Analyze the data

▶ First, estimate the ANCOVA with Y2 as the outcome, and X and Y1 as its predictors. What conclusion can you draw about the causal effect based on this analysis?

```
ANCOVA <- lm(Y2 ~ X + Y1, data=dat4)
summary(ANCOVA)
```

```
##
## Call:
## lm(formula = Y2 ~ X + Y1, data = dat4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2914  -3.3854   0.0857   3.3269  13.7512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.19772    1.42827   49.15   <2e-16 ***
## X           19.88475    0.61232   32.48   <2e-16 ***
## Y1           0.40056    0.01408   28.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.986 on 497 degrees of freedom
## Multiple R-squared:  0.9223, Adjusted R-squared:  0.922
## F-statistic:  2949 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X represents the difference in
# intercept between the two treatment groups; hence, the estimated
# causal effect here is significantly different from zero and positive:
# 19.88 (SE=0.61, p<0.000001)
# which is close to the b1=20 with which we simulated.
```

▶ Second, estimate the ANCOVA with Y2-Y1 as the outcome and Y1 and X as the predictors. What conclusion can you draw about the causal effect based on this analysis?

```
ANCOVA.CS <- lm((Y2-Y1) ~ X + Y1, data=dat4)
summary(ANCOVA.CS)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X + Y1, data = dat4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2914  -3.3854   0.0857   3.3269  13.7512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.19772    1.42827   49.15   <2e-16 ***
## X           19.88475    0.61232   32.48   <2e-16 ***
## Y1          -0.59944    0.01408  -42.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.986 on 497 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.7859
## F-statistic: 916.9 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# The regression coefficient for X in this model is
# exactly the same as that in the previous model
# (as we would expect based on the analytic results
# we discussed in the lecture).
```

▶ Third, estimate the change score model by regressing Y2-Y1 on X. What conclusion can you draw based on this analysis?

```
CSA <- lm((Y2-Y1) ~ X, data=dat4)
summary(CSA)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X, data = dat4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.394  -6.751   0.659   6.931  34.154
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.8441     0.6712  16.156   <2e-16 ***
## X             2.0245     0.9608   2.107   0.0356 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.74 on 498 degrees of freedom
## Multiple R-squared:  0.008835,   Adjusted R-squared:  0.006845
## F-statistic: 4.439 on 1 and 498 DF,  p-value: 0.03562
```

```
# The causal effect is estimated with the regression coefficient for X;
# it is significantly different from zero, but it is much smaller than
# the estimate from the ANCOVA model above:
# 2.02 (SE=0.96, p=0.0356)
# It would lead us to conclude that the treatment has a small positive
# effect on the change.
```

▶ Fourth, estimate the marginal model with Y2 as the outcome and X as the predictor. What conclusion can you draw based on this marginal?

```
Y2onX <- lm(Y2 ~ X, data=dat4)
summary(Y2onX)
```

```
##
## Call:
## lm(formula = Y2 ~ X, data = dat4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4075  -5.8867  -0.1505   6.1577  24.2936
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.8590     0.5048  217.63   <2e-16 ***
## X            31.8193     0.7226   44.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.077 on 498 degrees of freedom
## Multiple R-squared:  0.7956, Adjusted R-squared:  0.7952
## F-statistic:  1939 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
# The causal effect is estimated with the regression coefficient for X;
# here it is significant and positive:
# 31.82 (SE=0.72, p<0.000001)
# Hence, treatment has a large positive effect on Y2, according to this model.
```

2.4.4 Conclusion

When comparing the results, what is your conclusion? How can a DAG help to decide which model to use?

Model	Equation	Causal parameter	Estimate (SE)	p-value
ANCOVA	$Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_i$	β_1	19.88 (0.61)	<0.0001
ANCOVA on change score	$Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_i$	β_1	19.88 (0.61)	<0.0001
Change score analysis	$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + e_i$	γ_1	2.02 (0.96)	0.0356 sig!!
Marignal analysis	$Y_{2i} = \phi_0 + \phi_1 X_i + e_i$	ϕ_1	31.82 (0.72)	0.0001

Again, the two ANCOVAs (with Y2 or Y2-Y1 as the outcome) lead to the same estimate of the causal effect. The change score model (with only X as a predictor of Y2-Y1) gives a much smaller, yet significant positive effect. Finally, the marginal model (With Y2 regressed on X) results in a very large significant positive effect.

Based on the DAG, we know that Y1 is a common cause of X and Y2, and therefore we should control for it (i.e., condition on it by including it as a predictor). Hence, the two ANCOVA approaches should be preferred over the other two approaches.

2.5 Overall conclusion

What we have seen in this exercise is that the critical distinction between the four scenarios is the relation between X and Y1. In the RCT they are independent of each other, and in that case, all four analyses lead to the same conclusion.

In the second scenario, we first created X and based on X we created Y1; hence Y1 is a mediator on the indirect path from X to Y2. Controlling for it (as is done in the ANCOVAs) blocks this path, which results in estimating the direct effect rather than the total effect of X on Y2. The change score analysis (i.e., regressing Y2-Y1 on X) results in estimating the total effect of treatment on change, while the marginal model (i.e., regressing Y2 on X) gives the total effect of treatment on Y2.

The third and fourth scenario are based on creating Y1 first, and then creating X based on this. In these scenarios, Y1 is a confounder or common cause of X and Y2. This means we should definitely control for it if we want to estimate the causal effect of X on Y2. Hence, the ANCOVA would be the analysis of choice here.

Note that in the current scenarios, the parameters were chosen such that the ANCOVA model resulted in a positive effect of about 20, whereas the change score model tended to result in a non-significant or vary small effect, and the marginal model resulted in a very large positive effect. However, it is also possible to get very different patterns across these models (e.g., see the examples in the lecture for this).

3. Changes Scores & Unobserved Time-Invariant Confounding

The change score model has been advocated as a useful approach when there is unobserved time-invariant confounding. **Kim and Steiner (2019)** discussed this (see this week's reading materials). To get a sense of how this works, you will simulate data with such a time-invariant confounder, and then analyze the data *without* this confounder (as if it is unobserved!).

Examples of such an unmeasured time-invariant confounder are for instance ability, when we measure something like math achievement or language skills. But an unmeasured time-invariant confounder can also consist of personality, or genetic or physical factors that influence our repeated measures of for instance attitude, motivation, interests, mood, symptoms (e.g., of depression or cancer), behavior, social interactions, and so on.

3.1 Simulate data

We will consider the following model to simulate data:

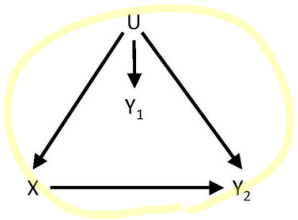


Fig.4: DAG of repeated measures and unmeasured confounding

▶ To begin, simulate the time-invariant confounder U using a standard normal distribution (i.e., mean of zero and standard deviation of 1) for a sample of 5000 people. (Note that we use a (relatively) large sample size here, to decrease the sampling variance.)

```
set.seed(934) # set the seed for comparison
N <- 5000 # sample size
U <- rnorm(N)  # Unmeasured time-invariant confounder
```

▶ Next, create the treatment variable that is dependent on this U.

```
# Create a treatment variable that is dependent on the unmeasured confounder
z <- -1.1*(U)           # linear combination with noise
pr <- 1/(1+exp(-z))              # pass through an inv-logit function
X <- rbinom(N,1,pr)              # bernoulli treatment variable
cor(X,U)                        # check the correlation
```

```
## [1] -0.4338681
```

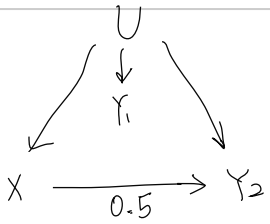
```
# Create the pre-test and the post-test
Y1 <- 10 + 0.7*U + rnorm(N)
Y2 <- 11 + 0.7*U + 0.5*X + rnorm(N)

dat5 <- data.frame(U,X,Y1,Y2)
round(cor(dat5),2)
```

```
##        U     X    Y1    Y2
## U   1.00 -0.43  0.57  0.50
## X  -0.43  1.00 -0.25 -0.05
## Y1  0.57 -0.25  1.00  0.29
## Y2  0.50 -0.05  0.29  1.00
```

▶ Now create Y1 and Y2; both are dependent on U. Furthermore, Y2 also depends on treatment X.

```
# Create the pre-test and the post-test
Y1 <- 10 + 0.7*U + rnorm(N)
Y2 <- 11 + 0.7*U + 0.5*X + rnorm(N)

dat5 <- data.frame(U,X,Y1,Y2)
round(cor(dat5),2)
```

```
##         U     X    Y1    Y2
## U    1.00 -0.43  0.56  0.49
## X   -0.43  1.00 -0.26 -0.03
## Y1   0.56 -0.26  1.00  0.27
## Y2   0.49 -0.03  0.27  1.00
```



▶ What is the size of the direct effect of X on Y2? And what is the size of the backdoor path?

```
# There is a direct effect from X to Y2 of size 0.5.
# There is also a backdoor path, X <- U -> Y2; the size of this path
# is more difficult to determine, because X is not linearly related to U
# as it is a binary variable. Hence, we cannot just multiply the
# path coefficients along this path, like Pearl (2016) and Kim and Steiner (2019)
# do in their examples. Nevertheless, given the specific parameter values, we
# can see it is a negative relation (as higher values on U result a smaller a
# probability of treatment, while U has a positive effect on Y2).
```

(neg correlation between X & U = −0.43)
pos. corr. between Y & U = 0.49

▶ Based on your answer to the previous answer, what do you expect to happen when you fail to account for the backdoor path, and just regress Y2 on X?

```
# This will result in a biased estimate of the causal effect of X on Y2
# and the bias will be negative, meaning: we will underestimate the effect
# of the treatment X on the outcome Y2. Whether the effect of X on Y2 is
# then estimated to be positive or negative, depends on whether the size
# of the backdoor path is larger or smaller in absolute value than the
# true causal effect of X on Y2 (which is 0.5).
```

3.2 Analyze the data

We will analyze these data in a variety of ways. Note however that in *none* of the analyses will we be able to include U, as this variable represents the unobserved confounder. That is, we should think of our observed data to consist *only* of the variables: treatment X, pre-test Y1, and post-test Y2.

3.2.1 Marginal model

We start with the marginal model, by which we estimate the prima facie effect (see Week 2); this implies we just regress Y2 on X, and assume there is no confounding.

▶ What is the marginal effect of X on Y2? How does this compare to the true effect with which you simulated the data?

```
Y2onX <- lm(Y2 ~ X, data=dat5)
summary(Y2onX)
```

```
##
## Call:
## lm(formula = Y2 ~ X, data = dat5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6682 -0.7977  0.0014  0.7994  4.1755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.28813    0.02355  479.39   <2e-16 ***
## X           -0.07359    0.03329   -2.21   0.0271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.177 on 4998 degrees of freedom
## Multiple R-squared:  0.0009767,  Adjusted R-squared:  0.0007768
## F-statistic: 4.886 on 1 and 4998 DF,  p-value: 0.02712
```

```
# The effect is estimated to be -0.11 (SE=0.03, p<0.00001).
# This means that on average individuals who received treatment are estimated to
# score 0.11 points lower on Y2 than individuals who did not receive treatment.
# This is quite different from the true causal effect of X, which is 0.5.
# The negative bias in estimating this effect is the result of the backdoor path
# X <- U -> Y2, which is negative, and that was not blocked in this analysis.
```

3.2.2 ANCOVA model

Run an ANCOVA model with X as the predictor of interest, and Y1 as covariate. As Y1 is a proxy of the unmeasured confounder U, including it as a covariate may help to block the backdoor path $X \leftarrow U \rightarrow Y2$ to some extent.

▶ What is the effect of X on Y2 in this approach, how would you describe it to others, and how does this compare to the true value?

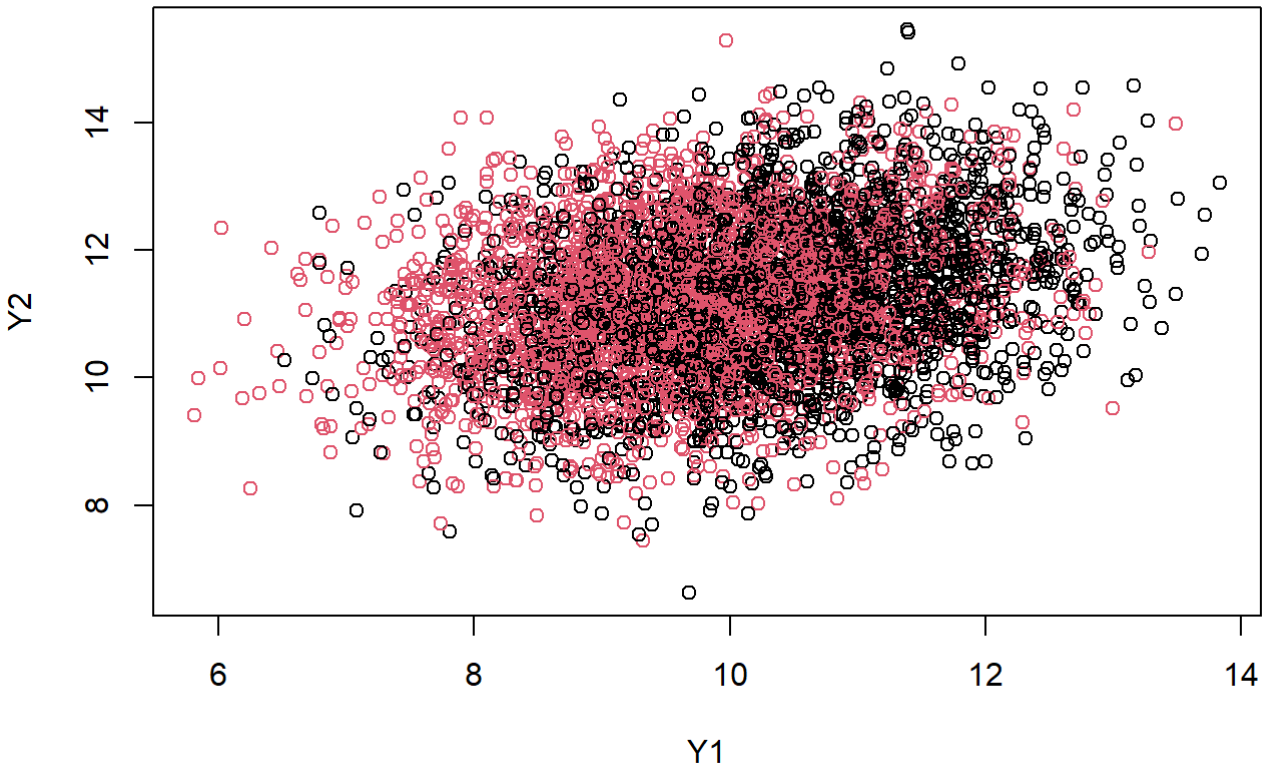
```
ANCOVA <- lm(Y2 ~ X + Y1, data=dat5)
summary(ANCOVA)
```

```
##
## Call:
## lm(formula = Y2 ~ X + Y1, data = dat5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5020 -0.7586 -0.0080  0.7739  4.0066
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.53842    0.14166  60.274  < 2e-16 ***
## X            0.09283    0.03318   2.798  0.00516 **
## Y1           0.26691    0.01357  19.664  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.134 on 4997 degrees of freedom
## Multiple R-squared:  0.07273,    Adjusted R-squared:  0.07236
## F-statistic:   196 on 2 and 4997 DF,  p-value: < 2.2e-16
```

```
# After conditioning on Y1, the effect of X on Y2 is
# 0.09 (SE=0.03, p=0.005
# which would lead us to conclude X has no causal effect on Y2.
# More specifically, we would say that, even though there is a mean
# difference between the treatment groups on Y2 (see previous analysis),
# there is no difference between the groups on Y2 after conditioning on
# pretest score Y1. Hence, if we compare individuals with the same Y1,
# there is no difference between the groups.
# Compared to the true value, the estimated causal effect is too small;
# yet is is closer than the estimate of the prima facie effect (marginal model).
```

▶ Given the result above, what would you expect the scatter plot of the data (with Y2 plotted against Y1 with different colors for the two groups) to look like? Make this plot.

```
plot(x=dat5$Y1, y=dat5$Y2, col = (dat5$X+1),
        xlab="Y1", ylab="Y2")
```




```
# Note that in the plot the treatment group (X=1) is in red, and the
# control group (no treatment; X=0) is in black.
```

```
# We would expect to see a slight positive relation between Y2 and Y1 in
# each group, as the effect of Y1 on Y2 is estimated to be positive.
# If we would add a regression line for each group to that plot, we would
# NOT expect to see a difference in the intercepts of the groups (as the
# effect of X represents the difference in intercept, and it is very close
# to zero).
```

3.2.3 Change score model

According to Kim and Steiner (2016), and many others, this model should result in an unbiased estimate of the effect of X on Y2. The DAG for this approach is shown below.

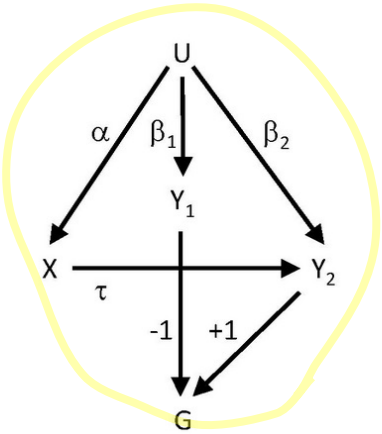


Fig.4: DAG of repeated measures and unmeasured confounding

▶ Use the DAG to explain why the causal effect of X on the gain score G is the same as the effect of the X on Y2.

```
# The effect of X on Y2 is tau.
# The effect of X on G is tau*1=tau; hence, it is the same.
```

▶ Run the change score model, with the gain score G=Y2-Y1 as the outcome variable and X as the predictor. What effect do you find, and how does this compare to the true causal effect of X on Y2?

```
CSA <- lm((Y2-Y1) ~ X, data=dat5)
summary(CSA)
```

```
##
## Call:
## lm(formula = (Y2 - Y1) ~ X, data = dat5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0186 -0.9556  0.0082  0.9700  4.8137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98598    0.02855   34.54   <2e-16 ***
## X            0.54993    0.04037   13.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.427 on 4998 degrees of freedom
## Multiple R-squared:  0.03581,    Adjusted R-squared:  0.03561
## F-statistic: 185.6 on 1 and 4998 DF,  p-value: < 2.2e-16
```

control group's change score (Y_2-Y_1) will increase by 0.99 ;

treatment group's change score will increase by 0.55 + 0.99

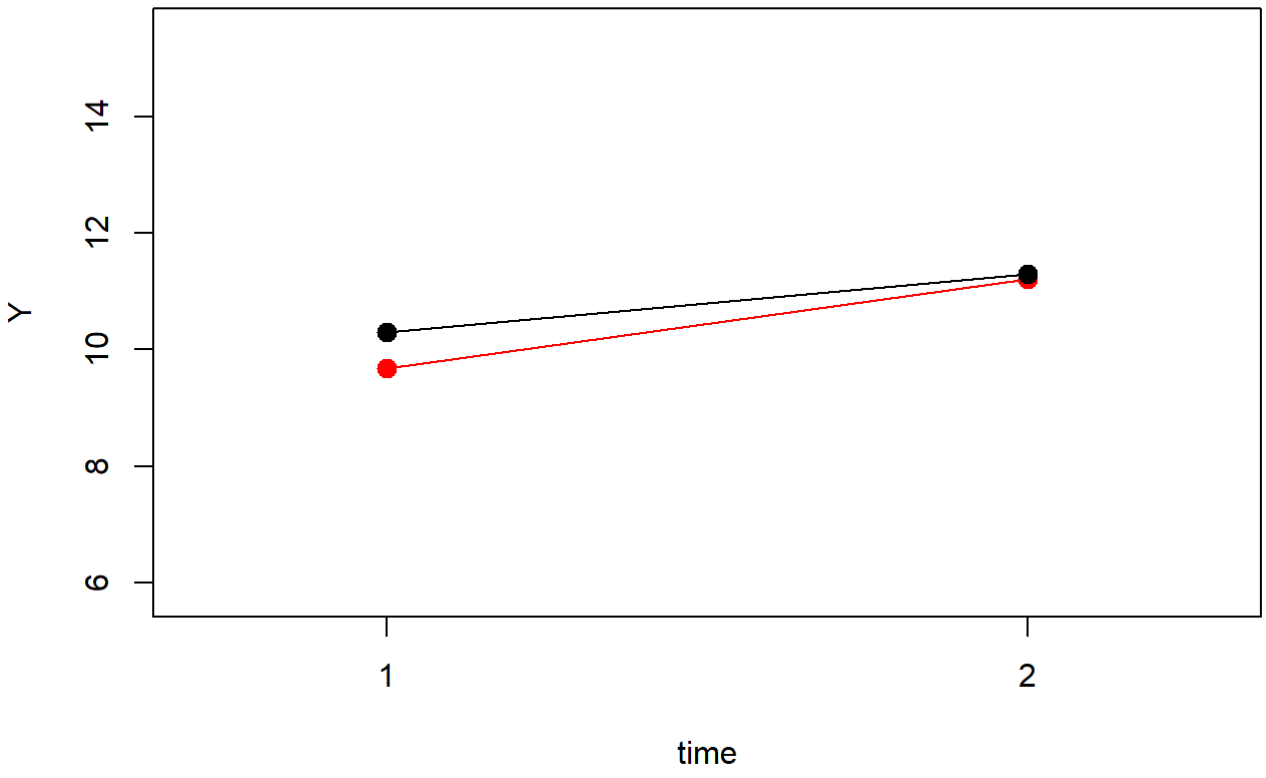
they'll not be parallel!

```
# The causal effect of X on G (and thus on Y2) is estimated to be
# 0.49 (SE=0.04, p<0.000001)
# meaning it is significantly different from zero, and positive.
# It is actually very close to the true value of 0.5 used in simulating
# the data. This confirms the analytical derivation presented by
# Kim and Steiner, that showed that in this approach, the two paths
# X <- U -> Y1 -> G
# X <- U -> Y2 -> G
# if the effect of U on Y1 is the same as the effect of U on Y2
# (referred to as "common trend or time-invariant confounding assumption").
```

▶ Based on the results, do you know what to expect if you would make a plot of the means of the groups at pre-test and at post-test? Make the plot, and compare this to your expectations.

```
# X has a positive effect on Y2; this means that the two lines will not
# run parallel (there will be a difference in the differences between the
# two groups at the two occasions).
# Whether the control group increases, decreases, or remains stable over time
# can be seen from the intercept of the gain score, which is:
# 1.01 (SE=0.03, p<0.000001)
# which indicates that there is (on average) an increase of 1.01 in the
# control group from the pretest to the posttest.
# Hence, we know that the treatment group will increase by 1.01 + 0.49 = 1.50
# from the pre-test to the post-test.
```

```r
# Compute means at first and second occasion
mY10 <- mean(Y1[X==0])
mY11 <- mean(Y1[X==1])
mY20 <- mean(Y2[X==0])
mY21 <- mean(Y2[X==1])

# Gather means on both occasions per treatment condition
mY0 <- c(mY10,mY20)
mY1 <- c(mY11,mY21)
minY <- min(Y1,Y2)
maxY <- max(Y1,Y2)

# Gather means on both occasions per treatment condition
plot(c(1,2), mY0, type="l",
            xlim=c(0.7,2.3),
            ylim=c(minY,maxY),xaxt="n",
            xlab="time",
            ylab="Y")
lines(c(1,2),mY1,col="red")
points(c(1,2),mY1,pch=19,cex=1.3,col="red")
points(c(1,2),mY0,pch=19,cex=1.3)
axis(side=1, at=seq(1, 2, by=1))
```




```
# Note again in this plot, the treatment group (X=1) is in red,
# and the control group (no treatment; X=0) is in black.
```

▶ Suppose that Y1 and Y2 are measures of math achievement, U is the unobserved math ability, and X is participation in a math camp. How would you explain the results we found above in substantive terms then?

```
# The (marginal model) first analysis shows that on the post-test, those who did not
# participate in the math camp (X=0) score higher on math achievement
# afterwards (Y2) than those who did participate (X=1). (∵ β1 = -0.07)
# However, we also see that individuals who participated in the math
# camp (X=1, red group) scored lower on math achievement to begin with
# (Y1; you can see this from the means in the last plot at the first
# measurement occasion).
# We also see that both groups score higher on math achievement on the
# second occasion than at the first. Moreover, we see that those who
# participated in the math camp (X=1, red group) actually increased more
# over time than the group that did not participate (X=0, black group).
# Hence, the difference between the groups was reduced as a result of
# the intervention.
```

3.3 Conclusion

In this exercise you have gained more experience with the gain score model, and the specific scenario in which it should be preferred over doing an ANCOVA: When there is an unmeasured time-invariant confounder that is a common cause of X, Y1, and Y2, and that has a stable effect on Y1 and Y2. Other assumptions are that there are no other relations between the pre-test Y1 and X, or between Y1 and Y2.

We have seen that in this scenario, the marginal model (regressing Y2 on X) leads to a biased estimated of the causal effect of X on Y2, as it does not account for the backdoor path X <- U -> Y2.

We have also seen that including Y1 as a proxy of the unmeasured confounder (regressing Y2 on X and Y1; i.e., ANCOVA), only partly removes this effect.

Here, using a change score model (regression of gain score Y2-Y1 on X) leads to an unbiased estimate, as the two backdoor paths from X to the gain score G cancel each other out. Kim and Steiner refer to this technique as *off-setting* these paths, rather than *blocking* them (which would require conditioning on a variable along the path).

While this illustrates the elegance of this approach very nicely, we have to realize that the assumptions (mentioned above) are quite strong. Using simulations like these, we could furhter investigate what happens when the effect of U is not stable over time, or when Y1 affects X and/or Y2, or Y1 is affected by X. Such an analysis could help to establish how robust a conclusion (about the strength and sign of the effect of X) is against violations of each of these assumptions.