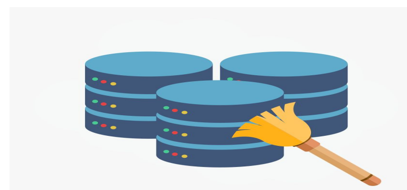
Hi everyone and welcome to our capstone presentation in STAT 451: Causal Inference. We are Yunyang Zhong and Charlotte Giang, guided by our professor Leslie Myint. Our research question is “Does sleeping time determine life satisfaction?” Spoiler alert: it does!

Data

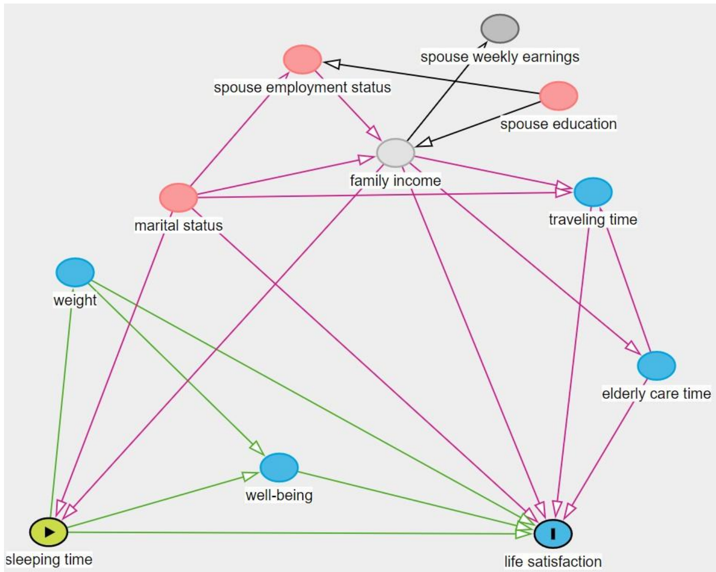
- Source: IPUMS ATUS
  - Years: 2003-2019
- Data cleaning:
  - Variable selection:
    - Sleeping time → Life satisfaction
    - Other: subject matter knowledge
  - Dropping non-responses to Life satisfaction
- Dimensions: 11,023 observations x 9 measured variables
- Unmeasured variable: Family income
- Selection bias?



Source: our data is from IPUMS time use, Annual American Time Use Survey (ATUS) data from 2003-2019. This data includes a variety of variables. We went through them and chose 11 that have the most entries available and are the most relevant to our project.

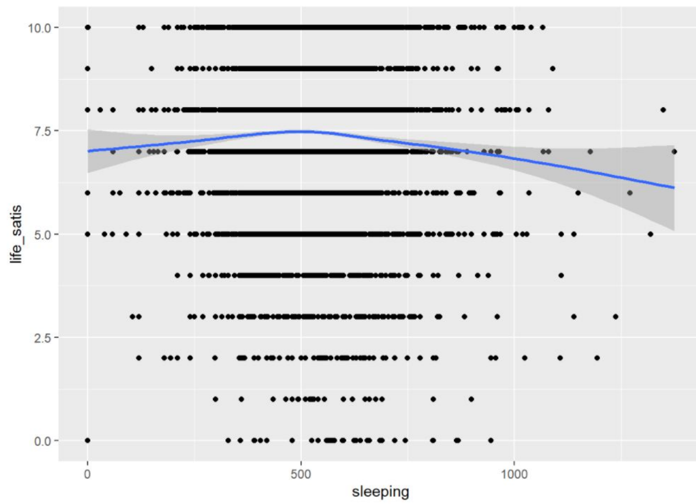
Sleeping time (in minutes per day) is our treatment and life satisfaction (from 1 to 10, 10 being the highest satisfaction) is our outcome. We drop all non-responses to life satisfaction because we want all entries in our outcome to be valid although dropping may lead to selection bias on those who are willing to share their life satisfaction and may be more satisfied with life overall. After dropping, we find that there are several outliers in partner education and partner employment status that are much larger than the average value and we filter out those observations. One of our confounders, family income, becomes unmeasured (all entries are NAs) as a result of data cleaning. Because we need to condition on family income to achieve conditional exchangeability, we add a proxy of it, partner income, so that it is still possible for us to hold it.

Initial Domain Knowledge Causal Graph



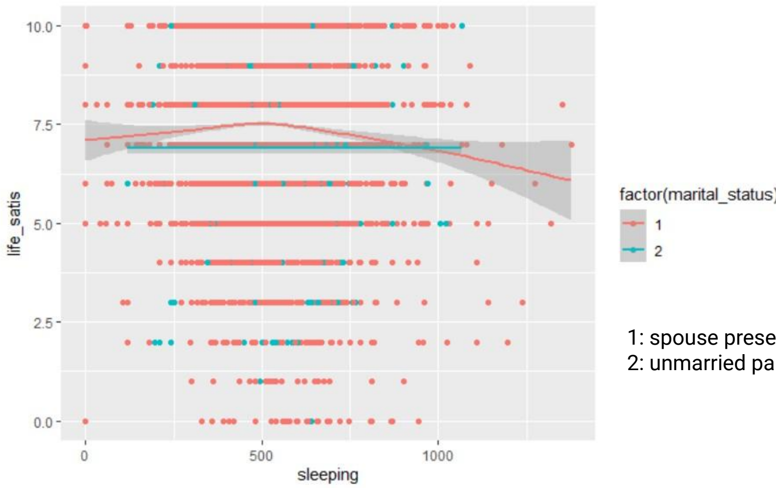
Weight, marital status, well-being, and family income are the four confounders that are directly connected with both treatment and outcome. We assume that sleeping affects life satisfaction through well-being and weight and that the other two are causes of both treatment and outcome.

Exploratory Data Analysis



From the graph, we can see that when sleeping time is lower than 500 minutes, there is a positive association between sleeping and life satisfaction: as sleeping time increases, life satisfaction increases. However, after the point where sleeping = 500, this positive relationship turns into a negative one: life satisfaction decreases as sleeping time increases. There seems to be a strong association between the treatment and the outcome.

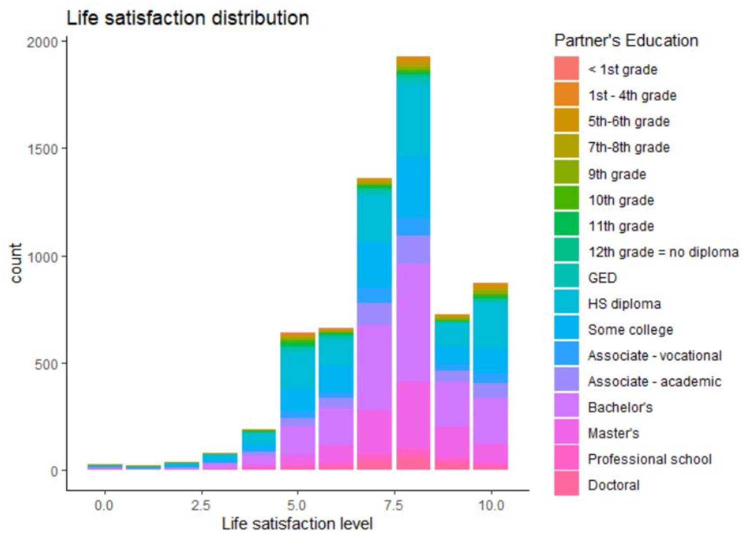
Exploratory Data Analysis



1: spouse present
2: unmarried partner

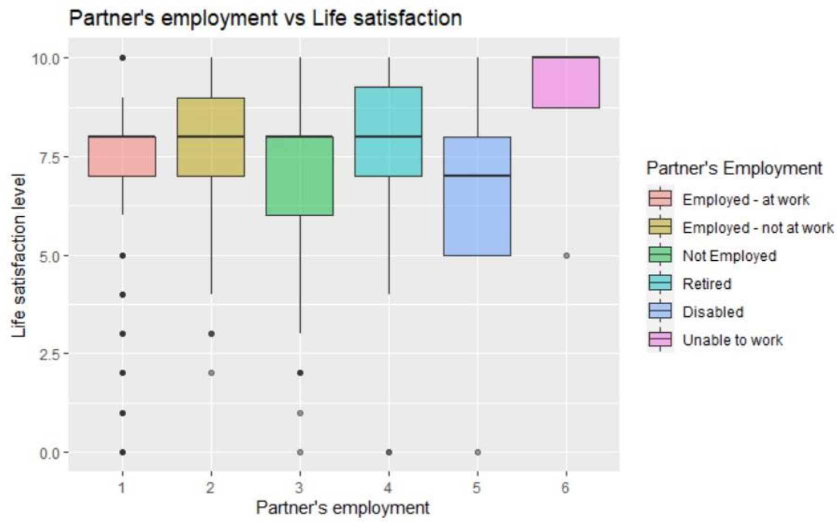
When we add marital status as a confounder, it is clear that different marital status groups have totally different results. For group one, the relationship is the same as above: a positive one before sleeping=500 and a negative one after. However, as the horizontal green line for those with unmarried partners shows, life satisfaction seems to be independent of sleeping in group 2.

Exploratory Data Analysis



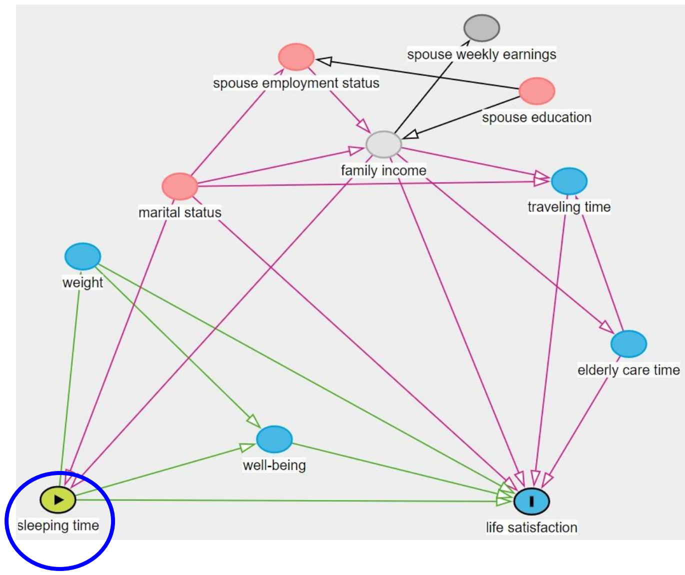
This graph shows the distribution of life satisfaction. The majority of the data are higher than 5.0, indicating that most of the participants are relatively satisfied with their lives - note that this is after we drop all non-responses, which could lead to a selection bias for happier persons. Different colors represent different education levels of their partners. The percentage of purple and red gets larger as life satisfaction increases. Thus there might be a positive relationship between education level and life satisfaction.

Exploratory Data Analysis



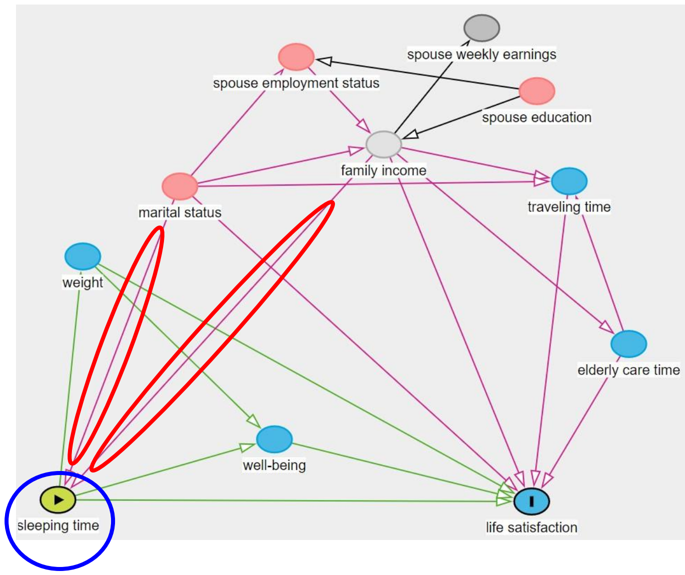
This graph explores how the employment status of the partners influences the association between sleeping and life satisfaction. Bars of each color have no overlap, indicating that each employment status has a specific range of sleeping time. Except for those who are disabled or unable to work, the median levels of life satisfaction are about the same though the range varies. For disabled partners, life satisfaction is noticeably lower; for people unable to work, life satisfaction is noticeably higher - we expect employment status to be a confounder.

Inverse Probability Weighting



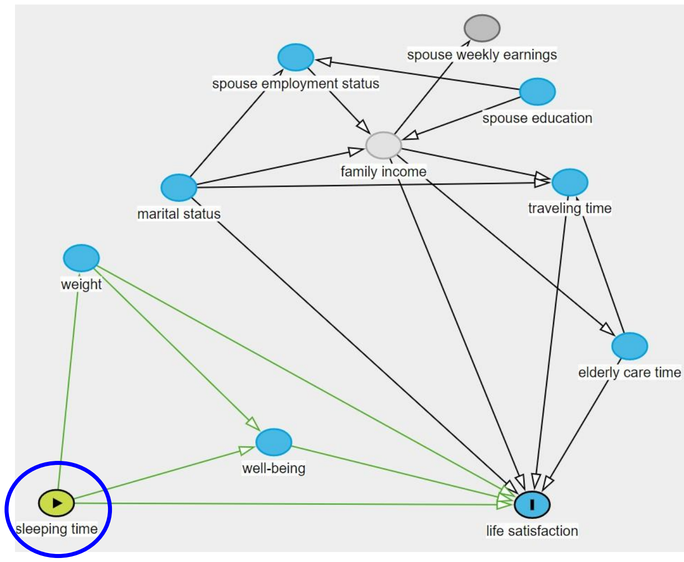
We estimated the causal effect of sleeping time on life satisfaction using inverse probability weighting. In inverse probability weighting, each subject is reweighted so that they receive every treatment value available. The **weight** of each subject is equal to the inverse of their probability to receive treatment, thus the name “inverse probability weighting”. The reason that we make everyone receive all treatment values is that, we want to remove arrows pointing **towards** the treatment, which is the circled node “sleeping time” in this graph.

Inverse Probability Weighting



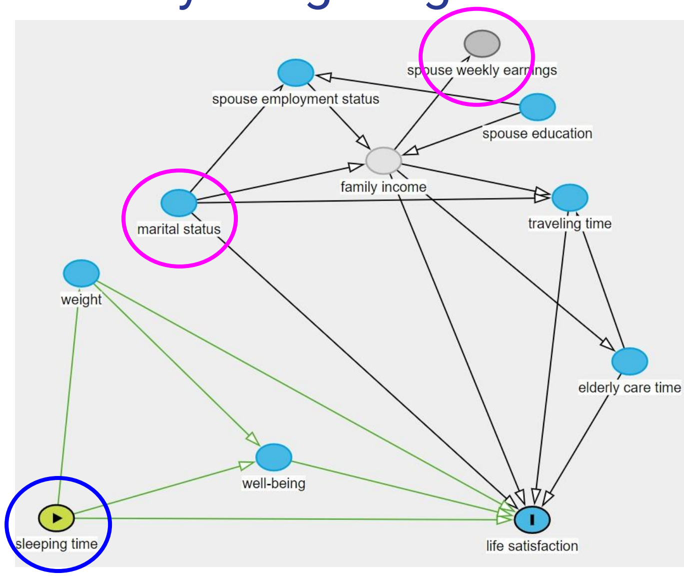
These are the 2 arrows pointing towards the treatment, “sleeping time”.

Inverse Probability Weighting



By performing inverse probability weighting, we removed those 2 arrows, and now the treatment is not an effect of any variables. Everyone receives all values of treatment available. Now that those 2 arrows are removed, 2 non-causal paths that carry misleading association between sleeping time and life satisfaction are blocked.

Inverse Probability Weighting



To block the rest of the non-causal paths, we condition on (or fix) marital status and partner's weekly earnings. Now the only open paths are ones that carry causal association between sleeping time and life satisfaction, and we can correctly estimate the causal effect between them.

Inverse Probability Weighting

```
Call:
svyglm(formula = life_satis ~ sleeping_time, design = design,
    data = time_use)

Survey design:
svydesign(ids = ~0, weights = time_use$ip_weight, data = time_use)
```

Here is the code for inverse probability weighting that used an R package called "survey". The **weight** of each subject, or the inverse of their probability to receive treatment, is included in this model.

Inverse Probability Weighting

```
Call:
svyglm(formula = life_satis ~ sleeping_time, design = design,
    data = time_use)

Survey design:
svydesign(ids = ~0, weights = time_use$ip_weight, data = time_use)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.105e+01  1.133e+00  -36.22   <2e-16 ***
sleeping_time  3.492e-02  8.237e-04   42.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    2.5 %       97.5 %
(Intercept)   -43.26835479 -38.82603707
sleeping_time   0.03330262   0.03653166
```

The result indicates that each additional minute of sleeping time within 24 hours will increase life satisfaction level by 0.035, conditioning on marital status and partner's weekly earnings. The p-value for this coefficient is statistically significant.

The 95% confidence interval of the average causal effect is between 0.033 and 0.036.

Inverse Probability Weighting

```
Call:
svyglm(formula = life_satis ~ sleeping_time, design = design,
    data = time_use)

Survey design:
svydesign(ids = ~0, weights = time_use$ip_weight, data = time_use)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -4.105e+01  1.133e+00  -36.22   <2e-16 ***
sleeping_time  3.492e-02  8.237e-04   42.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = life_satis ~ sleeping_time, data = time_use, weights = time_use$ip_weight)
```

We can also estimate the causal effect using regression with the famous “lm” model. Here, the weight of each subject is also included in the “lm” model.

Inverse Probability Weighting

```
Call:
svyglm(formula = life_satis ~ sleeping_time, design = design,
    data = time_use)

Survey design:
svydesign(ids = ~0, weights = time_use$ip_weight, data = time_use)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.105e+01  1.133e+00  -36.22   <2e-16 ***
sleeping_time  3.492e-02  8.237e-04   42.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = life_satis ~ sleeping_time, data = time_use, weights = time_use$ip_weight)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.105e+01  6.584e-02  -623.4   <2e-16 ***
sleeping_time  3.492e-02  4.792e-05   728.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We got the same result in the "lm" model as when we used the "survey" package for inverse probability weighting. This shows the two results are consistent, which is to be expected.

Inverse Probability Weighting

```
svyglm(formula = life_satis ~ sleeping_time * num_partner, design = design,
    data = time_use)

Survey design:
svydesign(ids = ~0, weights = time_use$ip_weight, data = time_use)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -4.105e+01  1.128e+00  -36.39   <2e-16 ***
sleeping_time              3.492e-02  8.201e-04   42.58   <2e-16 ***
num_partner2               4.428e+01  1.745e+00   25.37   <2e-16 ***
sleeping_time:num_partner2 -2.903e-02  1.510e-03  -19.22   <2e-16 ***
```

Here, we include an interaction between sleeping time and marriage status. Conditioning on partner’s weekly earnings, if a subject has an unmarried partner, the cause effect of sleeping on life satisfaction level will change by 0.029. The results are statistically significant.

Inverse Probability Weighting

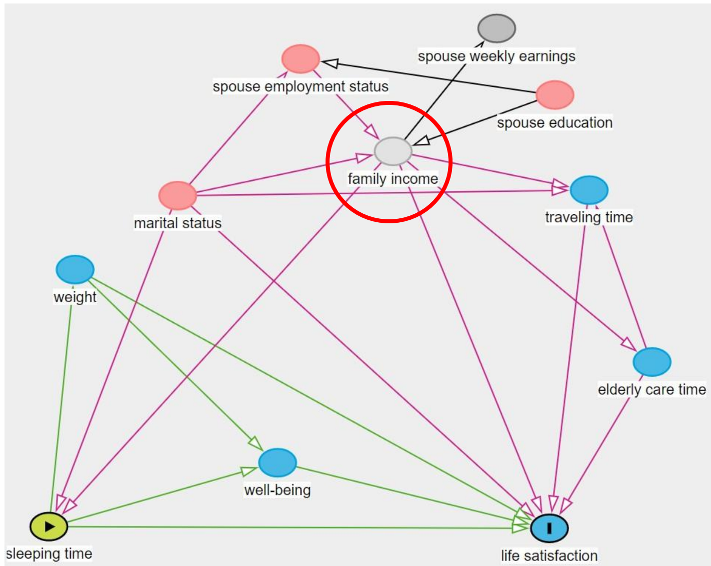
```
svyglm(formula = life_satis ~ sleeping_time * partner_employ,
    design = design, data = time_use)

Survey design:
svydesign(ids = ~0, weights = time_use$ip_weight, data = time_use)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4.111e+01  1.085e+00  -37.89   <2e-16 ***
sleeping_time                 3.496e-02  7.888e-04   44.33   <2e-16 ***
partner_employ2               4.592e+01  1.566e+00   29.32   <2e-16 ***
partner_employ3               4.959e+01  1.651e+00   30.04   <2e-16 ***
partner_employ4               5.053e+01  1.360e+00   37.16   <2e-16 ***
partner_employ5               4.633e+01  1.195e+00   38.78   <2e-16 ***
partner_employ6               5.622e+01  1.123e+00   50.04   <2e-16 ***
sleeping_time:partner_employ2 -3.031e-02  1.392e-03  -21.78   <2e-16 ***
sleeping_time:partner_employ3 -3.802e-02  1.346e-03  -28.24   <2e-16 ***
sleeping_time:partner_employ4 -4.462e-02  1.837e-03  -24.29   <2e-16 ***
sleeping_time:partner_employ5 -3.154e-02  2.530e-03  -12.47   <2e-16 ***
sleeping_time:partner_employ6 -4.502e-02  8.402e-04  -53.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

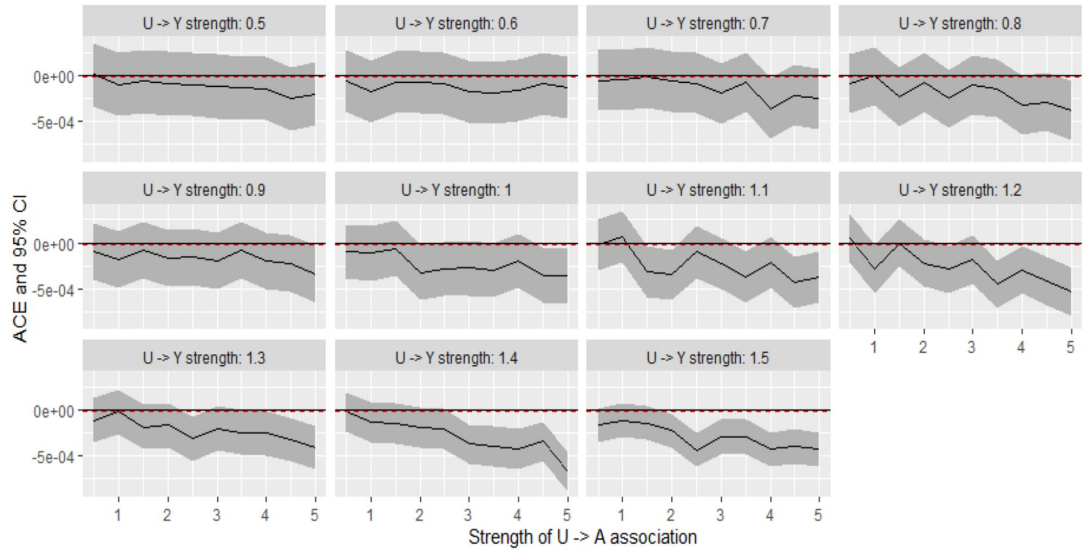
In another model, we explore the interaction between sleeping time and partner’s employment status, because we just discovered the importance of partner’s employment status in our exploratory analysis. So conditioning on partner’s weekly earnings, if the partner is not at work, not employed, retired, disabled or unable to work, the causal effect of sleeping minute per day on life satisfaction level will change by -0.03 to -0.045. Again, the results are statistically significant.

Sensitivity Analysis w/ Unmeasured Confounder



In the beginning, we mention an unmeasured variable, family income. Now how does family income affect our results above? Since that variable is unmeasured, there's no way to know for sure. However, we can see how much a **range** of values of family income affects our results. That method is called "sensitivity analysis".

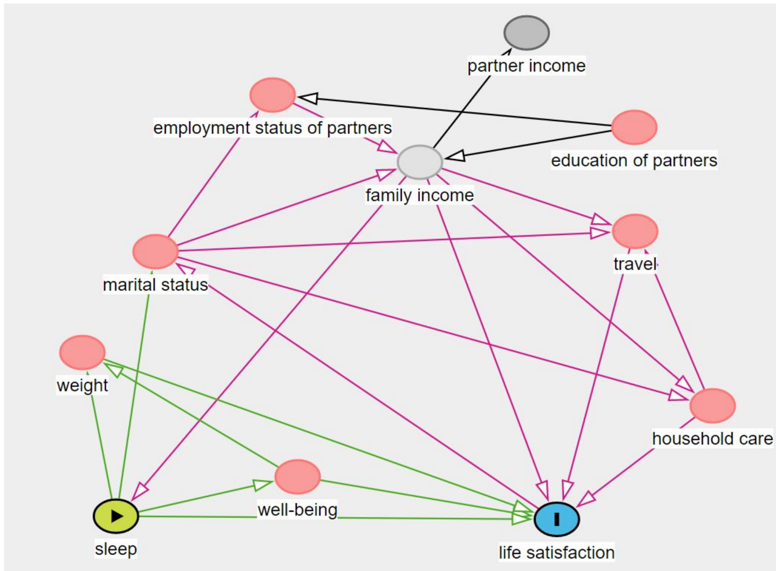
Sensitivity Analysis w/ Unmeasured Confounder



Here, we vary the strength of the associations between family income & sleeping time (U & A) and between family income & life satisfaction (U & Y). The black lines represent the average causal effect and the grey areas represent its 95% confidence interval. As the U & Y association increases from 0.5 to 5, the average causal effect decreases from around 0 to at most -0.0005. As the U & A association increases from 0.5 to 1.5, the average causal effect decreases at a faster rate. When the U & Y association is greater than 0.8, as the U & A association increases, the 95% confidence interval decreases from positive to negative. It is worth noting that the causal effect also changes sign from positive to negative when we include sensitivity analysis with the unmeasured confounder.

So as the associations between the unmeasured family income and each of the treatment & outcome increases, the average causal effect on life satisfaction only changes by a minimal amount of at most 0.0005. Therefore, our result is robust.

Adapted Domain Knowledge Causal Graph



The above is our new causal graph implementing results from EDA. Weight, marital status, well-being, and family income are still the four confounders that are directly connected with both treatment and outcome. But now weight and marital status become two colliders.

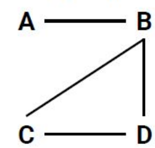
Causal Discovery

```
mod1 <- lm(life_satis~sleeping, data = data)
summary(mod1)
```

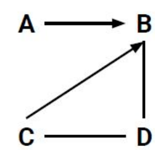
```
##
## Call:
## lm(formula = life_satis ~ sleeping, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6163 -1.3673  0.5484  1.5356  2.8020
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.6163314  0.0777024  98.019  < 2e-16 ***
## sleeping    -0.0003921  0.0001460  -2.686  0.00725 **
```

Overall structure of causal discovery

Step 1: Determine graph skeleton.



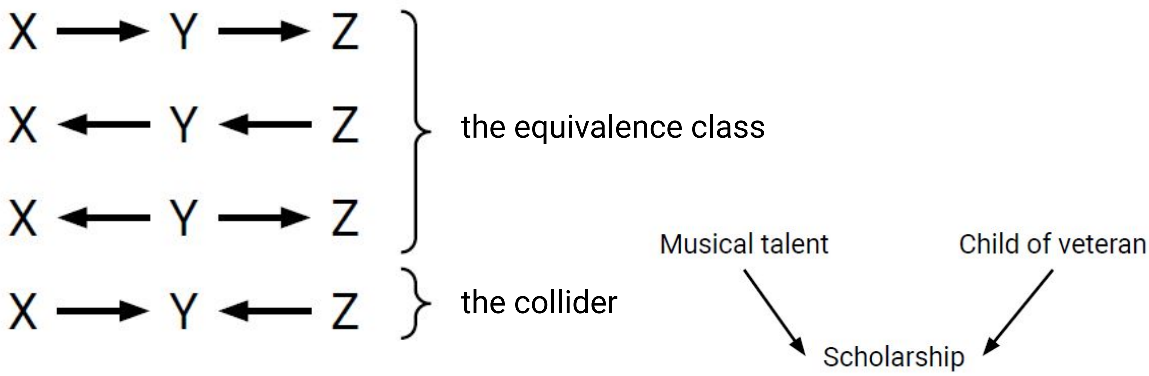
Step 2: Orient edges.



(picture credit to Prof. Leslie’s youtube video)

Causal discovery helps us go from data to graph and examine the relationship between potentially associated variables. For example, if we have four variables A, B, C, and D, the first step is to determine the graph skeleton - whether there exist associations between every two nodes - by modeling building. In mod1 output above, we can see the association is significant and thus there should be an edge between them. But we are still not sure about the direction yet - this is step 2: orient edges.

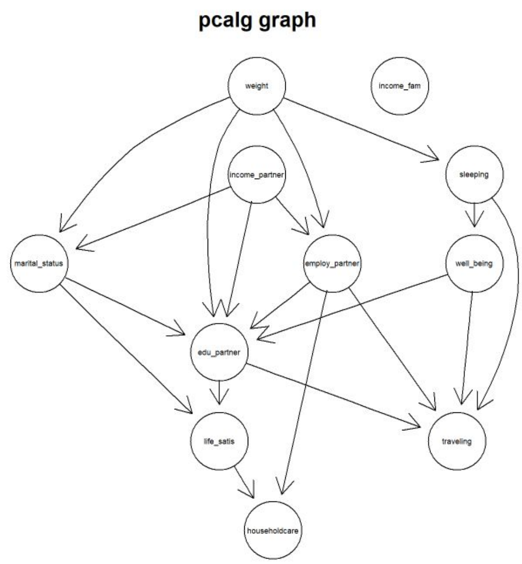
Causal Discovery



(picture credit to Prof. Leslie's youtube video)

There is a total of four potential situations in a three-variable group. The first three are called the equivalence group because we are not able to distinguish using data only. The fourth one is called a collider - this is the one we can examine using data and is the focus of our project. Here is an example of a collider: if only those who have musical talent or are child of veteran can get scholarship, then both of them have an arrow point to scholarship, which forms a collider. When we condition on scholarship (a student gets scholarship), then we know that this student either has musical talent or is a child of veteran - these two variables are not dependent on each other and should have a significant association. Thus, if we build a model with three variables, musical talent, child of veteran, and scholarship, we will have an association between musical talent and child of veteran.

Causal Discovery (**pcalg** package in R)



Using pcalg package in R, this new causal graph is presented. In this graph, our treatment and outcome, sleeping and life satisfaction, are not directly connected. However, as the output of mod1 indicates, the association between sleeping and life satisfaction is relatively significant. Therefore, we would like to go with our domain knowledge causal graph with some modifications instead of this pcalg causal discovery graph.

Note that family income (the node in the box) is not connected to any other node because it is an unmeasured variable.

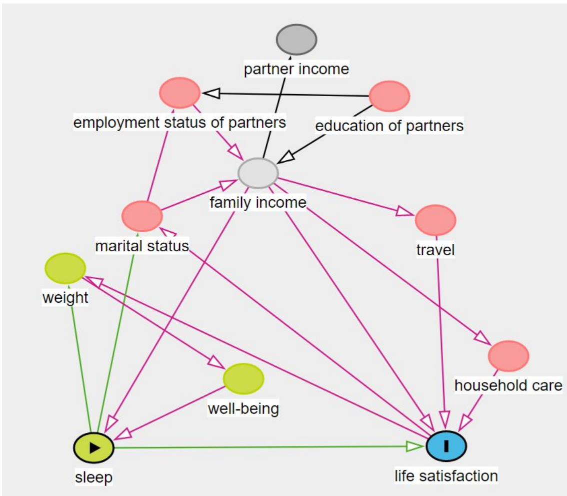
Causal Discovery

```
mod10 <- lm(traveling~householdcare, data = data)
summary(mod10)
```

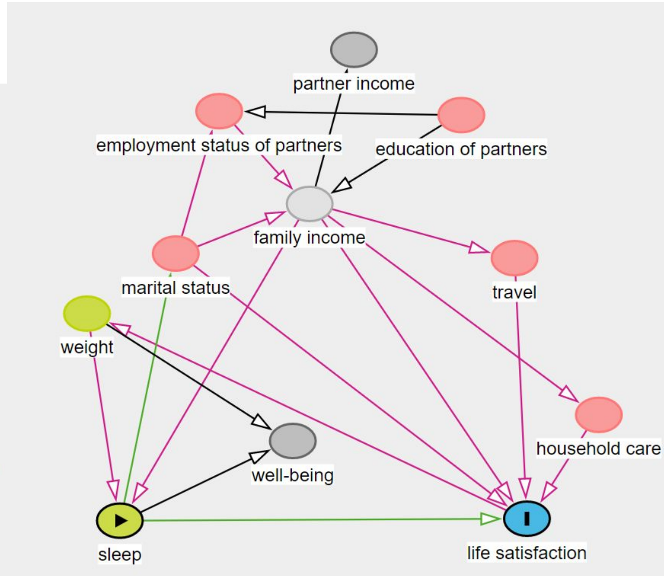
```
##
## Call:
## lm(formula = traveling ~ householdcare, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -82.37  -52.05  -14.05   25.95 1013.95
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   74.04856    0.78113  94.796   <2e-16 ***
## householdcare  0.01870    0.03786   0.494    0.621
```

We partly perform a causal discovery by hand to examine the association between every two nodes in the domain knowledge causal graph. There are four of the associations are not significant enough to include in our model. mod10, which tests the relationship between traveling and household care is one of them. As the p-value is 0.621, much higher than our chosen significance level 0.05, the arrow from household care to traveling should be removed. The other three are life satisfaction and well-being, marital_status and household care, and marital_status and traveling.

Causal Discovery



Insignificant associations removed



Arrow directions changed (based on pcalg)

The graph on the left is our new causal graph with insignificant associations removed. Though we would like to mainly go with this graph rather than the pcalg one, we still want to implement what we learn from the pcalg graph. Our main focus is on variables that are directly associated with either the treatment or the outcome. As the other associations are extra ones that we do not include in the original graph, we choose to ignore them. There are three direct associations including weight->sleep, sleep ->well-being, and marital status->life satisfaction, all of which are the opposite of what we have in the modified graph (on the left). Implementing these into the graph leads us to our finalized causal graph on the right.

Based on this newest graph, we might want to remove well-being from our causal graph - it is now presented in grey, which indicates that it is neither a cause of the treatment nor a cause of the outcome.

Conclusions

- Average causal effect: 0.035
    - Interactions with # partners and partner's employment status
- Generalizability: those who responded to IPUMS ATUS
- Future work:
    - More variables in the causal graph, e.g. physical exercise, diet, # children
    - Interactions with other variables, e.g. travelling time, elderly care time

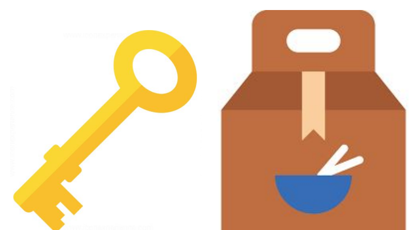
Using inverse probability weighting, we estimated the average causal effect between sleeping time (in minutes per day) and life satisfaction level (from 1 to 10, 10 being the highest satisfaction) is 0.003. Conditioning on marital_status and partner's weekly income as a proxy for family income, each additional minute of sleeping time increases life satisfaction level by 0.003. We also looked into the interaction between sleeping time & marital_status and sleeping time & employment status. According to our models, the average causal effect decreases (meaning more sleeping time decreases life satisfaction) if the partner is unmarried or not at work.

This result, however, is not necessarily generalizable to other demographics. In particular, IPUMS ATUS survey only contains Americans who responded to the survey from 2003 to 2019 and who provided their sleeping time, life satisfaction level, partners' education level and partners' employment status. This is related to the concept of selection bias mentioned at the beginning.

Time allowing, we would like to include more relevant variables in our causal analysis, for example, physical exercise, diet, and number of children. It'd also be interesting to look into interactions with other variables in our data set, including travelling time and elderly care time.

Key takeaways

- **Exploratory data analysis**
- **Regression models**
  - Interaction terms
- **Coefficient interpretation**
  - Interaction terms
- New concepts!
  - Correlation does not imply causation
  - Measuring causal effects



EDA: (for 155 students) what we have in the EDA section is no new materials - this is we learned in 155. Although you may think what you are learning in 155 is simple and may not be useful, they actually are! Though our modeling is more complex and we take more into consideration (probability weighting, unmeasured variables, mediators, etc), we still need these box plots and bar plots to visually present the relationship between variables.
Regression models: another thing we learned in 155! Similarly, we still depend on p-value to tell us whether an association is significant and we still need lm and glm to build models.
Coefficient interpretation: in 155, we interpret the coefficient as when holding all other variables constant, an increase in sleeping leads to an increase in life satisfaction - we are saying that there is a correlation but not causation. But now you can interpret in a causal language: it is the causal effect of sleeping on life satisfaction.
However, this does not mean that correlation implies causation: if we go back to the causal graph by pcalg, we can see that pcalg believes there is no direct arrow from sleeping to life satisfaction though they are associated. If we do not include indirect causal effects as causation, this would be an example of correlation but not causation.

This concludes the end of our presentation. Thank you for listening!

References

Sandra L. Hofferth, Sarah M. Flood, Matthew Sobek and Daniel Backman. American Time Use Survey Data Extract Builder: Version 2.8 [dataset]. College Park, MD: University of Maryland and Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D060.V2.8

https://remlapmot.github.io/cibookex-r/ip-weighting-and-marginal-structural-models.html#program-12.4