Workshop 3 - Path Analysis (full answers)

JPS

Introduction

We are going to practice the design of causal diagrams (using the package *ggdag*) and their implementation (using the package *mediation*). We will focus on the exploration of mediating effects (path analysis), but we will also pay attention to potential confounding effects. In so doing, we will practice model building strategies driven by theory. These are the type of model building strategies that we should consider when we seek to *explain* as opposed to simply predict. We are going to practice this using two influential theoretical models in the Social Sciences, the procedural justice model and the gender gap model, which we will explore using three different datasets.

Exercise 1: The procedural justice model was formulated by Tyler (1990), and has become one of the most influential theories explaining compliance with the law (i.e. law abiding behaviour). This model builds upon the classical work from Weber (1968) pointing at individual perceptions of institutional legitimacy as a key precursor of voluntary compliance, and upon Thibaut and Walker (1975), who indicated that procedural justice (the fairness in the interactions between an institution and the subjects under its authority) is also an important predictor of compliance. Tyler (1990) argued that the causal effect of procedural justice on compliance takes the form of a direct effect, but also of an indirect effect mediated through legitimacy. We will explore this model using a trimmed version of the first wave of the longitudinal study Pathways to Desistance. Specifically, we will test whether the procedural justice model can be used to explain criminal behaviour among young offenders in the US.

Exercise 2: The gender gap in salaries is one of the most challenging research questions in modern Social Sciences. At the population level women are systematically found to be earning less than men doing the same work. However, at the heart of this debate resides the problem of confounding effects. In order to make fair comparisons and ascertain truly discriminatory practices in the labour market we need to be able to control for relevant confounding factors. More insightful studies indicate that to understand the gender gap we should also focus on the different choices made by men and women with regards to training and type of industry, or how women face multiple barriers throughout their lives, which ends up impacting their careers, childcare being the most visible one. We will design a causal model to test some of these hypothesis using data from the Labour Force Survey, and you will be requested to test the hypothetical direct and indirect effects on your own. In the full version I have also added another **Bonus Exercise** exploring the gender gap based on a smaller dataset capturing Salaries of academic staff from a given college in the US. This is a guided exercise and you can take a look at it if you want to keep practicing path analysis.

Exercise 1. The Procedural Justice Model

Let's access the trimmed version of the Pathways to Desistance data and run some simple exploratory analyses.

```
options(scipen=999, digits=5) #This is to remove scientific notation
desist = read.csv("w1desistance.csv")
#Make sure you provide the direction to the folder where you saved the dataset.
names(desist)
```

```
## [1] "age"    "ethn"   "gend"   "pjcop"  "legit"  "freqof"
```

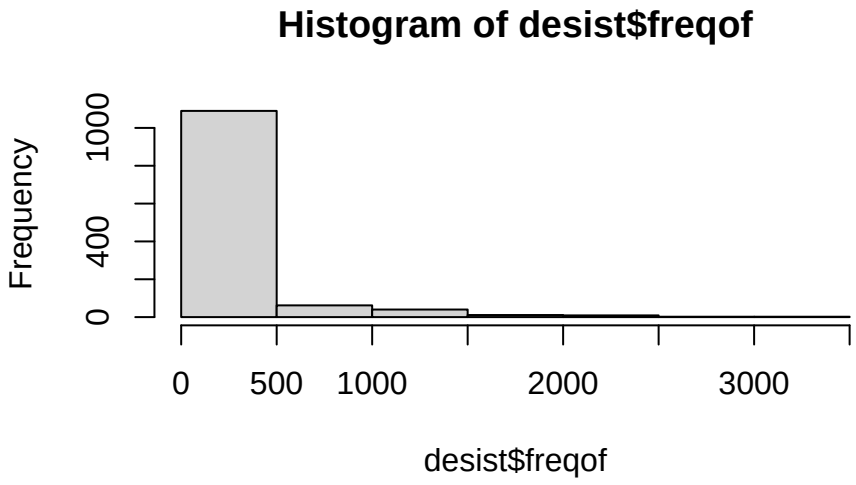
```r
summary(desist)
```

```
##       age          ethn               gend               pjcop          legit
##  Min.   :14   Length:1217        Length:1217        Min.   :1.39   Min.   :1.00
##  1st Qu.:15   Class :character   Class :character   1st Qu.:2.40   1st Qu.:1.91
##  Median :16   Mode  :character   Mode  :character   Median :2.71   Median :2.27
##  Mean   :16                                         Mean   :2.76   Mean   :2.28
##  3rd Qu.:17                                         3rd Qu.:3.09   3rd Qu.:2.64
##  Max.   :19                                         Max.   :4.49   Max.   :4.00
##                                                     NA's   :1      NA's   :1
##      freqof
##  Min.   :   1
##  1st Qu.:   4
##  Median :  17
##  Mean   : 169
##  3rd Qu.: 110
##  Max.   :3493
##
```

We have six variables, the first three are self-explanatory demographic factors of the participant. The last three represent indexes created after aggregating responses to different questions: 'freqof' represents the sum of 24 questions asking about how frequently different types of offences were committed by the participant in the last 12 months; 'pjcop' and 'legit' represent the mean to 19 and 11 likert scale questions (coded from one to five) on perceptions of procedural justice (how fairly were the participants treated by the police in their interactions) and legitimacy (in relation to the whole criminal justice system).

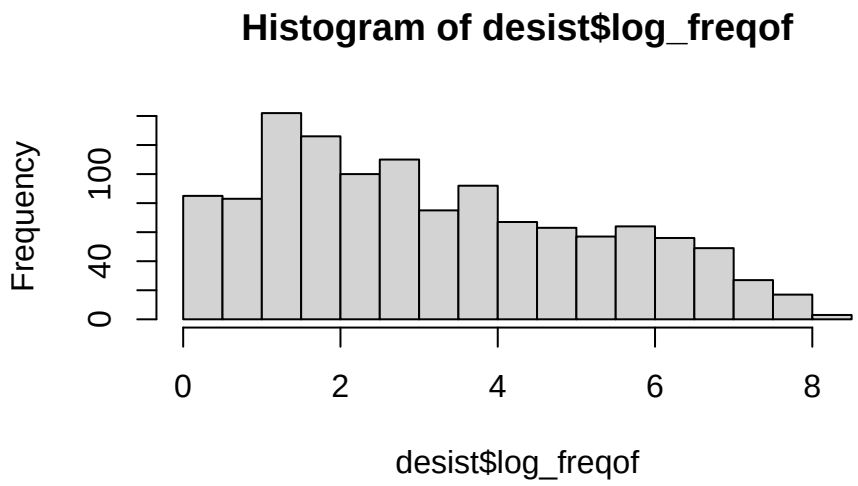
From the exploratory analysis we can identify a couple of issues. There are a few missing cases (NA's) in some of the variables, probably due to non-response. Since the proportion of missing cases to the sample size is tiny we can simply drop them from our study. In addition, 'freqof' seems to be affected by some strong outliers (Max.: 3493). Let's explore this using a plot.

```r
desist = desist[complete.cases(desist), ] #This is to drop missing values.
hist(desist$freqof)
```



It seems that the problem is not just with a few outliers (extreme values), but a distribution that is heavily right skewed. Since we are going to use this variable as an outcome for some of our models we proceed to log-transform it to make it more normally distributed.

```r
desist$log_freqof = log(desist$freqof)
hist(desist$log_freqof) #As expected, taking logs makes the distribution less-skewed.
```



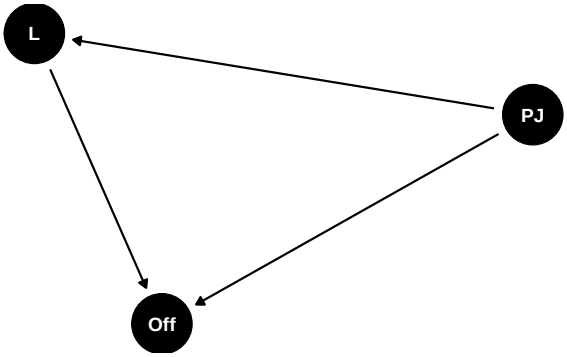

```r
desist$freqof = NULL #This is to drop the original variable for offending since we won't
#be using it anymore.
```

It is not exactly normal, but it will do the job. To model variables like this one in their orginal form, we could employ generalised linear models that reflect the (0,inf) range seen in count and duration data more accurately, like Poisson or Exponential models. Sadly, we do not have time to cover these models in our module.

Let's now test the procedural justice model using this data. Specifically, we want to assess whether the causal effect of procedural justice on offending is mediated by legitimacy. We can visualise our casual model using DAGs. This model is quite simple, but for more complex models having a visual representation helps in many ways: a) to make sense of the theory; b) to identify potential confounders, colliders and mediator effects; and c) to report our findings more clearly. We can start by simply designing our casual model using pen and paper, and once we are happy with it we can use powerpoint, word, latex, or any software we want to give them a more professional look. There is even an R package that you can use to do this, draw DAGs, which is one of the reasons why I like R so much, there is a package for everything.

The graph below represents Tyler's theoretical model, where procedural justice (PJ) is taken to affect offending (Off) directly, but also indirectly, mediated through legitimacy (L).

```r
library(ggdag)  #This is to draw DAGs using ggdag
dag1 = dagify(L~PJ, Off~PJ, Off~L)  #The causal relationships expected
ggdag(dag1) + theme_dag_blank()  #The DAG with a white background and no axes
```



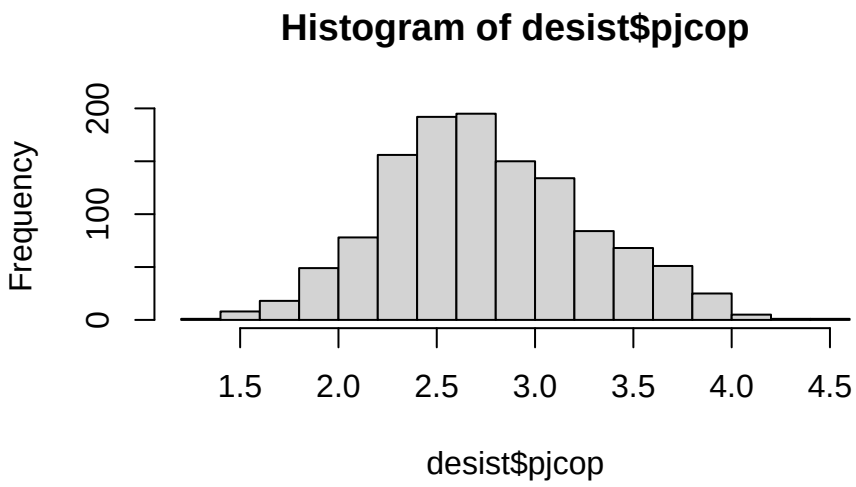
We are now going to estimate this theoretical model using path analysis. To disentangle the direct and indirect effects of procedural justice we should break this process up into three steps:

1. Estimate the effect of procedural justice on offending, $Off = \alpha + \beta PJ + e$. The β here gives us the total effect of PJ. You can take this as a sort of necessary condition. We estimate this model to check whether there is any kind of effect between PJ and Off, before we proceed to disentangle its direct and indirect part. If PJ is found non-significant then there really is no point in going forward using path analysis.
2. Estimate the effect of procedural justice on legitimacy, $L = \alpha + \beta PJ + e$. Here β gives us the first part of the indirect effect of PJ on Off through L.
3. Estimate the effect of procedural justice on offending while controlling for legitimacy, $Off = \alpha + \beta_1 PJ + \beta_2 L + e$. Here β_1 gives us the direct effect of PJ on Off, and β_2 gives us the second part of the potential indirect effect of PJ on Off through L.

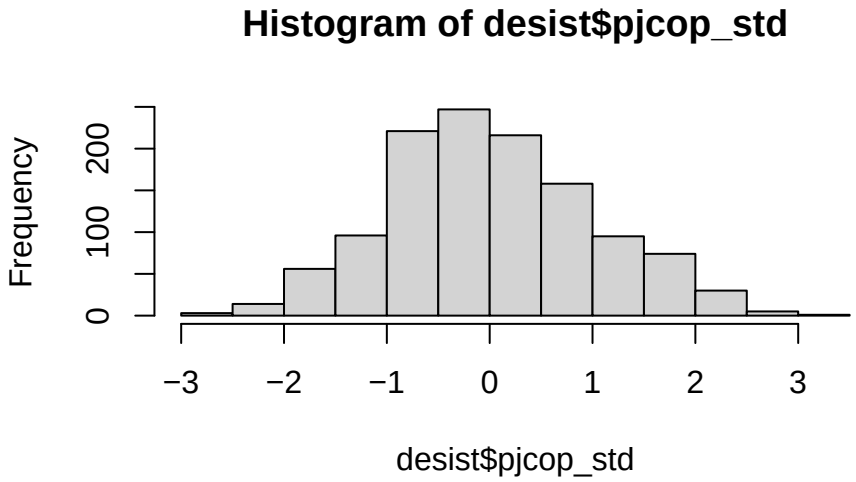
To facilitate the interpretation of the regression coefficients of procedural justice and legitimacy we could standardise these variables. If we do so their coefficients would not be interpreted like the change in the outcome variable after the explanatory variable increases in one unit, but as the change in the outcome variable when the explanatory variable goes up in one standard deviation. This way we do not care anymore about the scale used to measure procedural justice or legitimacy, which makes results more comparable across variables in our study but also across studies in the literature.

As we saw in Workshop 1, a given variable, X, can be standardised by subtracting its mean, μ, and dividing by its standard deviation, σ, such as: $X^* = (X - \mu)/\sigma$. This will transform the original variable X into X^*, which has a mean of 0 and standard deviation of 1. Notice how this is the case in the following transformation.

```
hist(desist$pjcop)
```




```
desist$pjcop_std = (desist$pjcop - mean(desist$pjcop))/sd(desist$pjcop)
hist(desist$pjcop_std) #Notice how for the mean of the standardised variable is 0.
```




```
desist$pjcop = NULL #This is to remove the original PJ variable from the dataset.
desist$legit_std = (desist$legit - mean(desist$legit))/sd(desist$legit)
desist$legit = NULL #As above.
```

Ok, let's now estimate the three models listed above sequentially to figure out whether and to what extent legitimacy mediates the effect of procedural justice on offending.

```
model1 = lm(log_freqof~pjcop_std, data=desist)
summary(model1)
```

```
##
## Call:
## lm(formula = log_freqof ~ pjcop_std, data = desist)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -4.129 -1.645 -0.299  1.481  5.084
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   3.1635     0.0584   54.20 < 0.0000000000000002 ***
## pjcop_std    -0.3840     0.0584   -6.58       0.000000000072 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.04 on 1214 degrees of freedom
## Multiple R-squared:  0.0344, Adjusted R-squared:  0.0336
## F-statistic: 43.2 on 1 and 1214 DF,  p-value: 0.0000000000718
```

Procedural justice exerts a significant influence in the frequency of offending (see p-value of 'pjcopstd'). It is good practice to check that this is the case before we carry out path analysis.

Question: Based on the above model, can you report what is the number of offences committed by the average participant in our sample? Hint1: Remember that we have standardised PJ, so this would be the predicted value of Off from your model when PJ=0. Hint2: Remember that we have log-transformed Off, so to retrieve any predictions in their original scale you need to back-transform them using *exp*. Also, can you report what is the effect of PJ on Off?

We can estimate that for offenders reporting levels of procedural justice to be one standard deviation above average, the number of offences committed goes down by roughly seven. We get this figure after back-transforming the regression coefficients using the exponential with base e to reflect that the outcome variable was log-transformed. The exponential transformation of the intercept represent the average offending in the sample (when procedural justice equals 0), while the exponential transformation of the intercept plus the regression coefficient for procedural justice represent the frequency of offending for a subject with perceptions of procedural justice one standard deviation higher than the average.

```
int = coef(model1)[1] #This is to record the coefficient of the intercept
a = exp(int)          #The average offending in the sample
pj = coef(model1)[2]  #This is to record the coefficient of procedural justice
b = exp(int+pj)       #The offending amount for someone with perceptions of PJ one
                      #standard deviation above average.
table1 = c(a, b)      #Grouping these two results in one table.
names(table1) = c("offences committed by the average offender",
"offences commited by offenders reporting one std. dev. higher procedural justice")
table1
```

```
##                                       offences committed by the average offender
##                                                                           23.654
## offences commited by offenders reporting one std. dev. higher procedural justice
##                                                                           16.112
```

Let's now test whether perceptions of procedural justice influence perceptions of legitimacy. This is a necessary condition to claim that the effect of procedural justice on offending is mediated through legitimacy.

```
model2 = lm(legit_std~pjcop_std, data=desist)
summary(model2)
```

```
##
## Call:
## lm(formula = legit_std ~ pjcop_std, data = desist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.6046 -0.5873  0.0024  0.5601  2.9348
##
## Coefficients:
##                           Estimate            Std. Error t value            Pr(>|t|)
## (Intercept) -0.00000000000000529  0.024555331589151484     0.0                   1
## pjcop_std    0.517106236939089636  0.024565434585501242    21.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.856 on 1214 degrees of freedom
## Multiple R-squared:  0.267,  Adjusted R-squared:  0.267
## F-statistic:  443 on 1 and 1214 DF,  p-value: <0.0000000000000002
```

```
#If you want to format your model tables take a look at the stargazer package.
```

They do. In fact, procedural justice alone explains over a quarter of the variability in offenders' perceived legitimacy of criminal justice authorities (see R^2). This, the fact that procedural justice and legitimacy are significantly associated, is a necessary condition to establish the role of legitimacy as a mediator, but it is not a sufficient condition. We also need to determine whether legitimacy has an effect on the frequency of offending.

```
model3 = lm(log_freqof~pjcop_std+legit_std, data=desist)
summary(model3)
```

```
##
## Call:
## lm(formula = log_freqof ~ pjcop_std + legit_std, data = desist)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.424 -1.578 -0.232  1.404  5.615
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   3.1635     0.0577   54.85 < 0.0000000000000002 ***
## pjcop_std    -0.1927     0.0674   -2.86               0.0043 **
## legit_std    -0.3699     0.0674   -5.49          0.00000005 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01 on 1213 degrees of freedom
## Multiple R-squared:  0.0578, Adjusted R-squared:  0.0562
## F-statistic: 37.2 on 2 and 1213 DF,  p-value: <0.0000000000000002
```

And it does, it is roughly twice as important as procedural justice. Notice as well how the effect of procedural justice is much smaller now (about half the size) than what it was in model 1. This is because from model 3 we derive only the direct effect, not the total effect, which is derived from model 1. Also, since procedural justice has a significant effect on legitimacy (model 2), while legitimacy has got a significant effect on offending (model 3), we can conclude that procedural justice has got both a direct and an indirect effect (mediated through legitimacy) on offending. In other words, we have corroborated Tyler's procedural justice model.

After establishing the mediating role of legitimacy we can proceed to estimate the total effect of procedural justice on offending. To do so we need to determine its direct and indirect effect first. Remember that the indirect effect (aka mediated effect) of procedural justice on offending is calculated as the effect of procedural justice on legitimacy times the effect of legitimacy on offending. As we see below, this indirect effect of procedural justice on offending is as important as its direct effect. If we had relied on a standard regression

model, i.e. if we had not used path analysis, we would not have been able to ascertain the full relevance of procedural justice.

```
direct = coef(model3)[2]    #The direct effect of procedural justice on log(offending)
indirect = coef(model3)[3]*coef(model2)[2] #The indirect effect mediated through
#legitimacy, calculated as the effect of legitimacy on offending times the effect
#of procedural justice on legitimacy.
total = direct + indirect
table2 = c(direct, indirect, total)
names(table2) = c("direct", "indirect", "total")
table2
```

```
##   direct indirect    total
## -0.19269 -0.19127 -0.38396
```

The following part of this exercise expands what was covered in the instructions uploaded in advance of the workshop.

To obtain standard errors for the above effects we can use the *mediation* package, which provides a range of different methods, one of them being *Bootstrap*. Bootstrap is a computationally intensive technique that can be used to obtain measures of uncertainty when these cannot be easily traced out algebraically. The method relies on replicating the analysis multiple times, using a slightly different subsample of the original sample each time. Measures of uncertainty are then derived from the observed variability in the results obtained across each iteration. Normally we use 1000 iterations or more, here I will just use 100 to speed the process up. We need to specify the *mediator* variable, and the explanatory variable causing both the mediator and the outcome, called *treat* (for treatment). In addition, we need to include the models where the indirect and direct effects are explored, in the specific order indicated below.

```
library(mediation)
set.seed(7)   #This is to ensure that we get the same random draws when using bootstrap
table3 = mediate(model2, model3, treat='pjcop_std',
                 mediator='legit_std', boot=TRUE, sims=100)
summary(table3)
```

```
##
## Causal Mediation Analysis
##
## Nonparametric Bootstrap Confidence Intervals with the Percentile Method
##
##                Estimate 95% CI Lower 95% CI Upper             p-value
## ACME             -0.191       -0.253        -0.13 <0.0000000000000002 ***
## ADE              -0.193       -0.308        -0.06 <0.0000000000000002 ***
## Total Effect     -0.384       -0.477        -0.28 <0.0000000000000002 ***
## Prop. Mediated    0.498        0.310         0.81 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 1216
##
##
## Simulations: 100
```

We get four different estimates: the indirect effect is reported as *ACME* (average causal mediation effects), the direct effect by *ADE* (average direct effects), *Prop. Mediated* reports the extent of the effect of procedural justice mediated by legitimacy.

As you can see, path analysis (the exploration of mediating effects) is an advanced technique that can be used to shed new light on lots of different complex causal problems in the Social Sciences. However, we should

never forget that we are making a series of assumptions when running these kinds of models. The most important of them all is that the causal path does not operate in reverse. We are drawing arrows, which imply a given causal direction, but that is entirely based on our theoretical assumptions. We are only testing whether the association between variables captured by the arrows is statistically significant, not the direction in which they occur. In Workshop 8 we will see how we can say more about this using longitudinal data.

Something we can and should do to assess the robustness of the causal interpretation made above is to explore whether our findings are sensitive to confounding factors that we are not controlling for in our models. The causal framework helps us theorise what potential confounding factors we might be missing. Namely, those which are simultaneously causing procedural justice and legitimacy or procedural justice and offending. We should also be mindful not to include additional mediating factors we are not interested on, and colliders.

For example, we could consider that potential confounders might be present amongst demographic factors (e.g. gender and age could be associated with defiant attitudes towards authorities and similarly associated with violence and crime, which for teenagers increases in the late teens), genetic factors (associated to impulsive, defiant and mistrusting behaviours), and possibly cultural and other socio-economic factors (such as different exposures to the media, inequality, etc.). We only have variables capturing the first group, we will also use ethnicity as a proxy for some of the socio-economic factors that we cannot control, but notice that we are missing lots of potentially relevant confounders. Importantly, none of the explanatory variables to be used can be considered colliders; i.e. age, gender, and ethnicity are - in principle - immutable factors and as such we can rule out that they are being affected by changes in the frequency of offending or perceptions of legitimacy.

```
model2b = lm(legit_std ~ pjcop_std + gend + ethn + age, data=desist)
summary(model2b)
```

```
##
## Call:
## lm(formula = legit_std ~ pjcop_std + gend + ethn + age, data = desist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6675 -0.5414  0.0186  0.5565  3.1607
##
## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        1.0386     0.3488    2.98              0.00296 **
## pjcop_std          0.4914     0.0246   19.98 < 0.0000000000000002 ***
## gend(2) Female     0.2042     0.0715    2.86              0.00435 **
## ethn(2) Black     -0.2455     0.0660   -3.72              0.00021 ***
## ethn(3) Hispanic   0.0248     0.0676    0.37              0.71400
## ethn(4) Other      0.0794     0.1250    0.64              0.52522
## age               -0.0611     0.0216   -2.83              0.00476 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.842 on 1209 degrees of freedom
## Multiple R-squared:  0.295,  Adjusted R-squared:  0.291
## F-statistic: 84.2 on 6 and 1209 DF,  p-value: <0.0000000000000002
```

```
library(regclass) #this is to use VIF
VIF(model2b)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## pjcop_std 1.0375  1          1.0186
## gend      1.0071  1          1.0035
```

```
## ethn      1.0233  3          1.0039
## age       1.0268  1          1.0133
```

We can see that the explanatory variables added are relevant but not really confounding the relationship between procedural justice and legitimacy (this is ascertained by noticing that the effect of procedural justice is very similar in 'model2b' and 'model2'). In addition, there is no evidence of multicollinearity (no $VIF > 5$), something worth checking once we start building up (i.e. complicating) the model.

```
model3b = lm(log_freqof ~ legit_std + pjcop_std + gend + ethn + age, data=desist)
summary(model3b)
VIF(model3b)
```

We find similar results for the model on offending. The demographic variables used are important but they are not really confounding the procedural justice and legitimacy relationships with offending. No evidence of multicollinearity either.

Before we claimed that Tyler's procedural justice model was corroborated, these additional results increase the robustness of such claim. However, we should still try to make our assumptions as explicit as possible when we get to report our findings. For example, in relation to our latest findings, we should include caveats pointing at a potential bidirectional causal path, and at the range of potential confounding factors that we have not been able to control.

Exercise 2. The Gender Gap

We are going to explore the gender pay gap in the UK using the Labour Force Survey. To do so I will give you a simple causal model based on six variables (a thorough study on this topic would normally be more complex than that), and you are requested to estimate the direct and indirect effects of gender on salary according to that causal model.

We start by importing the Labour Force Survey, and keeping just the variables of interest since this dataset is huge. The variables to be used are 'SEX', 'AGE', 'SC10MMJ' (major occupation group), 'TTUSHR' (total usual hours worked including overtime), 'QUAL_1' (degree level qualification), 'GRSSWK' (gross weekly pay in main job).

```
load("lfs.rda")  #Importing the data
vars = c("SEX","AGE","SC10MMJ","TTUSHR","QUAL_1","GRSSWK")
lfs = lfs[vars] #Keeping variables of interest
head(lfs)  #Taking a first look at the variables
summary(lfs)
```

We can see that there are some missing cases, coded as 'Does not apply' for major occupation group, or as -9 for weekly salary and total hours worked. This seems to point at participants in the survey who are not currently working, who fall outside the population of interest for our research (the gender gap in the labour market), so it is perfectly fine to simply drop these cases (in fact we have to do this).

```
table(lfs$GRSSWK, useNA="ifany") #This is to see what other values other than -9
#are used in this variable to represent missing cases, we see -8 is another one.
lfs = lfs[which(lfs$GRSSWK>-1),] #This is to remove all negative salaries, which
#are nonsensical and we should therefore treat as missing cases.
table(lfs$TTUSHR, useNA="ifany") #We do the same for total hours worked
lfs = lfs[which(lfs$TTUSHR>-1),]
table(lfs$SC10MMJ, useNA="ifany") #This is also to check that there are no other
#values used to record missing cases. It seems 'Does not apply' is the only one.
lfs = lfs[which(lfs$SC10MMJ!="Does not apply"),] #So we get rid of 'Does not apply'.
```

In addition, to facilitate the interpretation of the gender effect in models where we include controls, I proceed to 'de-mean' number of hours worked and age. That way, when looking at the coefficient of gender, we can

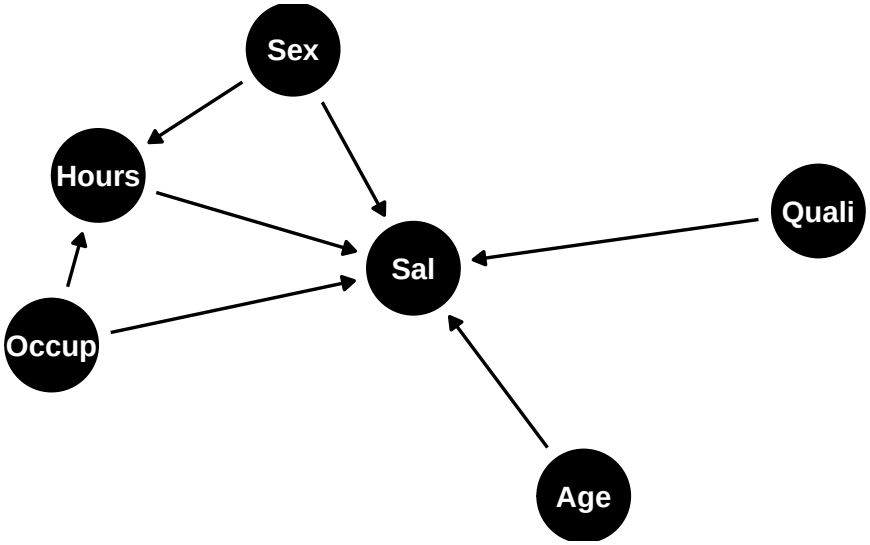
interpret this as the effect of gender on salaries for someone working the number of hours and aged just like the average British worker.

```
lfs$TTUSHR_d = lfs$TTUSHR - mean(lfs$TTUSHR) #I use '_d' to indicate 'de-meaned'
lfs$TTUSHR = NULL   #This is to drop the original variable from the dataset.
lfs$AGE_d = lfs$AGE - mean(lfs$AGE) #Same for age
lfs$AGE = NULL
```

We have imported and 'cleaned' the data, let's start considering the causal model that we will explore. Remember, we want to estimate the effect of 'sex' on 'salary', and to do so we will consider potential confounders and mediators. This requires careful theoretical reflection (you should dedicate a good amount of time to do this before you start the modelling process), which is clearly a subjective process, meaning that the causal model that I am suggesting here is not necessarily the right one. When you are doing this on your own I recommend that you use pen and paper, and only once you are happy with your model draw it more formally.

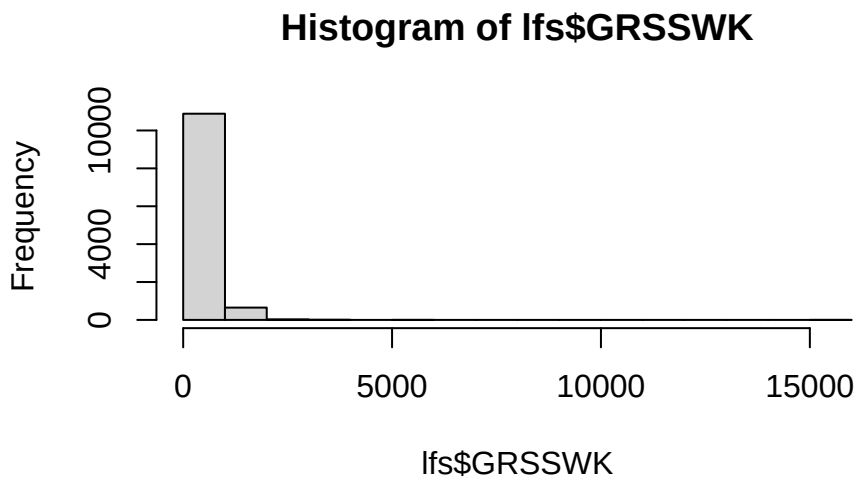
I have considered that sex, but also age, hours worked, degree qualification, and occupational group, have a causal effect on salary. In addition, I suspect that hours worked mediate the effect of sex on salary. The theoretical justification for this potential mediating effect stems from the much higher social pressure applied on women to carry out domestic and caring responsibilities, which I expect will be affecting the number of hours worked. Notice as well how I have hypothesised that the occupational group will also affect the number of hours worked. The full causal model can be represented as follows:

```
dag2 = dagify(Sal~Occup, Sal~Quali, Sal~Age, Sal~Sex, Sal~Hours, Hours~Sex, Hours~Occup)
ggdag(dag2) + theme_dag_blank()
```

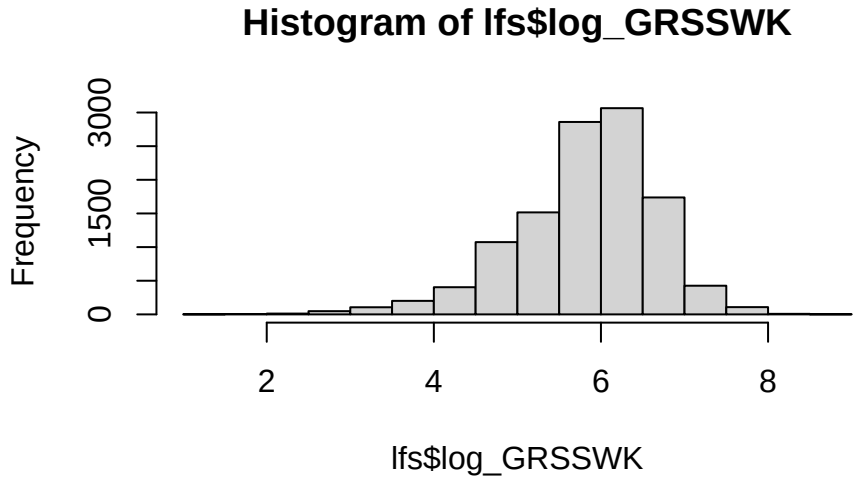


We need to undertake some more in depth exploratory analyses before we move to the modelling part. In particular, it is always good to look at the distribution of the outcome variable(s). We can do so with a histogram. This helps us see how extremely right-skewed the distribution of weekly salaries is. To avoid extreme outliers, I am taking the decision of dropping cases earning more than £10K per week. To 'normalise' that distribution I proceed to explore whether a logarithmic transformation seems more appropriate. We also need to explore the distribution of 'TTUSHR_d' since we want to test whether hours worked could be mediating the effect of gender on salaries, which means that we will have to run a model where total hours worked is used as the outcome variable.

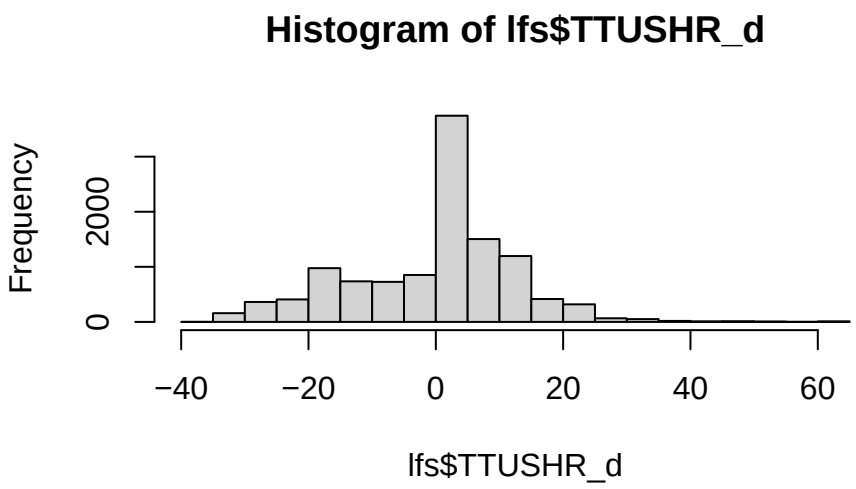
```
hist(lfs$GRSSWK)  #In its original form weekly salaries is very right-skewed.
```




```
lfs = lfs[which(lfs$GRSSWK<10000),]  #I get rid of one outlier.
lfs$log_GRSSWK = log(lfs$GRSSWK) #log-transforming weekly salary.
lfs$GRSSWK = NULL
hist(lfs$log_GRSSWK)  #This looks better.
```




```
hist(lfs$TTUSHR_d)  #This ditribution seems approximately normal.
```



Ok, at this point we have theorised our causal model, cleaned the data and undertaken a quick exploratory analysis. It is time to do some modelling. Question: Can you estimate the direct effect and indirect effect (mediated through hours worked) of gender on salary? Hint: you can borrow the same modelling process that we undertook in the previous exercise, i.e. build your model in three stages, first look at the effect of gender on salaries without any controls to determine if there are any disparities at all, then a second model controlling for potential confounders, and third, to test the expected mediating effect of hours worked you will need another model.

```
model1 = lm(lfs$log_GRSSWK ~ SEX, data=lfs) #Notice that the outcome variable is
#log-transformed to deal with the extreme right-skewness that we detected.
summary(model1)
```

```
##
## Call:
## lm(formula = lfs$log_GRSSWK ~ SEX, data = lfs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.476 -0.415  0.089  0.540  2.700
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   6.0849     0.0108   562.5 <0.0000000000000002 ***
## SEXFemale    -0.5108     0.0149   -34.2 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.802 on 11572 degrees of freedom
## Multiple R-squared:  0.0919, Adjusted R-squared:  0.0919
## F-statistic: 1.17e+03 on 1 and 11572 DF,  p-value: <0.0000000000000002
```

```
men = exp(summary(model1)$coefficients[1]) #The intercept captures the average weekly
#salary for men, we back-transform it so it is expressed in £, not log(£).
women = exp(summary(model1)$coefficients[1] + summary(model1)$coefficients[2])
#The women's average salary
men - women #The gender gap (without any controls)
```

```
## [1] 175.66
```

We find evidence of the expected gender pay gap, with women earning £176 less than men. We proceed to estimate model 2 including all the other factors that we thought could be explaining differences in salary.

```
model2 = lm(lfs$log_GRSSWK ~ SEX + TTUSHR_d + AGE_d + SC10MMJ + QUAL_1, data=lfs)
summary(model2)
```

```
##
## Call:
## lm(formula = lfs$log_GRSSWK ~ SEX + TTUSHR_d + AGE_d + SC10MMJ +
##     QUAL_1, data = lfs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.179 -0.219  0.033  0.273  2.607
##
## Coefficients:
##                                                    Estimate Std. Error t value
## (Intercept)                                        6.064992   0.016273  372.70
## SEXFemale                                         -0.106280   0.010193  -10.43
## TTUSHR_d                                           0.038835   0.000392   99.10
## AGE_d                                              0.006763   0.000354   19.13
## SC10MMJProfessional Occupations                    0.076016   0.018064    4.21
## SC10MMJAssociate Professional and Technical Occupations -0.024436   0.018986   -1.29
## SC10MMJAdministrative and Secretarial Occupations  -0.245024   0.019957  -12.28
## SC10MMJSkilled Trades Occupations                  -0.331060   0.021909  -15.11
## SC10MMJCaring, Leisure and Other Service Occupations -0.479223   0.021324  -22.47
## SC10MMJSales and Customer Service Occupations      -0.519609   0.022201  -23.40
## SC10MMJProcess, Plant and Machine Operatives       -0.472362   0.023292  -20.28
## SC10MMJElementary Occupations                      -0.669355   0.020767  -32.23
## QUAL_1Yes                                          0.185395   0.011661   15.90
##                                                             Pr(>|t|)
## (Intercept)                                        < 0.0000000000000002 ***
## SEXFemale                                          < 0.0000000000000002 ***
## TTUSHR_d                                           < 0.0000000000000002 ***
## AGE_d                                              < 0.0000000000000002 ***
## SC10MMJProfessional Occupations                                0.000026 ***
## SC10MMJAssociate Professional and Technical Occupations             0.2
## SC10MMJAdministrative and Secretarial Occupations  < 0.0000000000000002 ***
## SC10MMJSkilled Trades Occupations                  < 0.0000000000000002 ***
## SC10MMJCaring, Leisure and Other Service Occupations < 0.0000000000000002 ***
## SC10MMJSales and Customer Service Occupations      < 0.0000000000000002 ***
## SC10MMJProcess, Plant and Machine Operatives       < 0.0000000000000002 ***
## SC10MMJElementary Occupations                      < 0.0000000000000002 ***
## QUAL_1Yes                                          < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.472 on 11561 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.685
## F-statistic: 2.1e+03 on 12 and 11561 DF,  p-value: <0.0000000000000002
```

```
men = exp(summary(model2)$coefficients[1])
women = exp(summary(model2)$coefficients[1] + summary(model2)$coefficients[2])
men - women
```

```
## [1] 43.408
```

```
#This represents the gender gap for the average worker in the reference category
#(SC10MMJ='Managers, Directors and Senior Officials')
```

The gap has shrunk but there is still a significant gender effect that cannot be explained by gender differences in age, hours worked, occupation, and qualification, therefore, pointing at a direct effect of gender on salary - assuming our causal model is correct. Furthermore, as we expected, we see that salary is positively correlated with number of hours worked. To test whether the number of hours worked might be mediating the effect of gender on salary we need to run a third model with hours worked as the outcome variable, and gender and any other factor that we identified in our causal model as a precursor of hours worked as explanatory variables.

```
model3 = lm(TTUSHR_d ~ SEX + SC10MMJ, data=lfs)
summary(model3)
```

```
##
## Call:
## lm(formula = TTUSHR_d ~ SEX + SC10MMJ, data = lfs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.59  -6.12   0.43   7.00  68.49
##
## Coefficients:
##                                                    Estimate Std. Error t value
## (Intercept)                                          10.833      0.358   30.26
## SEXFemale                                            -7.885      0.231  -34.10
## SC10MMJProfessional Occupations                      -2.593      0.420   -6.17
## SC10MMJAssociate Professional and Technical Occupations  -4.394      0.449   -9.78
## SC10MMJAdministrative and Secretarial Occupations    -8.589      0.465  -18.48
## SC10MMJSkilled Trades Occupations                    -4.894      0.510   -9.59
## SC10MMJCaring, Leisure and Other Service Occupations -10.193      0.492  -20.71
## SC10MMJSales and Customer Service Occupations        -13.727      0.502  -27.34
## SC10MMJProcess, Plant and Machine Operatives         -3.995      0.545   -7.33
## SC10MMJElementary Occupations                        -14.138      0.467  -30.31
##                                                             Pr(>|t|)
## (Intercept)                                          < 0.0000000000000002 ***
## SEXFemale                                            < 0.0000000000000002 ***
## SC10MMJProfessional Occupations                         0.00000000071448 ***
## SC10MMJAssociate Professional and Technical Occupations < 0.0000000000000002 ***
## SC10MMJAdministrative and Secretarial Occupations    < 0.0000000000000002 ***
## SC10MMJSkilled Trades Occupations                    < 0.0000000000000002 ***
## SC10MMJCaring, Leisure and Other Service Occupations < 0.0000000000000002 ***
## SC10MMJSales and Customer Service Occupations        < 0.0000000000000002 ***
## SC10MMJProcess, Plant and Machine Operatives            0.00000000000024 ***
## SC10MMJElementary Occupations                        < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.2 on 11564 degrees of freedom
## Multiple R-squared:  0.258,  Adjusted R-squared:  0.258
## F-statistic:  447 on 9 and 11564 DF,  p-value: <0.0000000000000002
```

Women work on average 8 hours less per week than men. Hence, the gender gap can be documented through a direct effect on salaries, but also through an indirect effect via the number of hours worked.

```r
direct = coef(model2)[2]    #The direct effect of gender on log(salary).
indirect = coef(model3)[2]*coef(model2)[3] #The indirect effect mediated through
#hours worked. Specifically, the effect of gender on hours * the effect of hours
#on log(salary).
total = direct + indirect
table3 = c(direct, indirect, total)
names(table3) = c("direct", "indirect", "total")
table3
```

```
##   direct indirect    total
## -0.10628 -0.30620 -0.41248
```

It is interesting to note that the indirect effect is three times stronger than the direct effect. That is, in the instances when women are not able to work as much as they wished (e.g. when acting as carers), they might lose three times more income than what is lost as a result of what seems like direct discrimination.

Bonus. The Gender Gap (academic salaries)

I have included another exercise using path analysis and exploring the gender pay gap, but this time using data from a US college. The data is stored in the 'car' package.

```r
library(car)  #This is to access the dataset 'Salaries'
data(Salaries)
names(Salaries)
```

```
## [1] "rank"          "discipline"    "yrs.since.phd" "yrs.service"   "sex"
## [6] "salary"
```

```r
summary(Salaries)
```

```
##         rank     discipline yrs.since.phd   yrs.service       sex          salary
##  AsstProf : 67   A:181      Min.   : 1.0   Min.   : 0.0   Female: 39   Min.   : 57800
##  AssocProf: 64   B:216      1st Qu.:12.0   1st Qu.: 7.0   Male  :358   1st Qu.: 91000
##  Prof     :266              Median :21.0   Median :16.0                Median :107300
##                             Mean   :22.3   Mean   :17.6                Mean   :113706
##                             3rd Qu.:32.0   3rd Qu.:27.0                3rd Qu.:134185
##                             Max.   :56.0   Max.   :60.0                Max.   :231545
```

We have six variables: 'rank', reporting the level of academic seniority (from the more junior 'AsstProf' to the more senior 'Prof'); 'discipline' ('A' for academics working in a theoretical discipline and 'B' for those in an applied discipline); 'sex' and 'salary' which are self-explanatory; and 'yrs.since.phd' and 'yrs.service', which can be used as proxies for years of experience.

From this preliminary exploratory analysis we can also detect a couple of potential problems. First, the sample size for female academics is very low, this has obvious implications in terms of external validity (generalisability) but it can also have modelling implications, if we can predict those 39 cases from the set of explanatory variables to be used in our model we will have a problem of perfect collinearity. The second problem is also related to potential multicollinearity, which can be assessed exploring the correlation between the two variables capturing years of experience. This is worth investigating in more detail as part of the exploratory analysis (before start modelling).

```r
cor(cbind(Salaries$yrs.service, Salaries$yrs.since.phd))
```

```
##         [,1]    [,2]
## [1,] 1.00000 0.90965
## [2,] 0.90965 1.00000
```

```
#This is to calculate the correlation between yrs.service and yrs.since.phd
plot(Salaries$yrs.service, Salaries$yrs.since.phd)
```


```
#a scatterplot to visualise the correlation between the two variables
```

The two variables are very highly correlated, if we try to include both of them in our model it will certainly lead to multicollinearity, hence it is worth dropping one of them. Since the two variables are basically capturing the same variability, choosing one or another won't have huge repercussions, still, we might want to avoid arbitrary decisions. We should try to justify our decisions the best we can. In this case I think I know what 'yrs.since.phd' captures, but I am not sure about 'yrs.service'. Is it number of years working in the same institution, or number of years working in the academic sector? Since that information is not provided I will simply keep 'yrs.since.phd'.

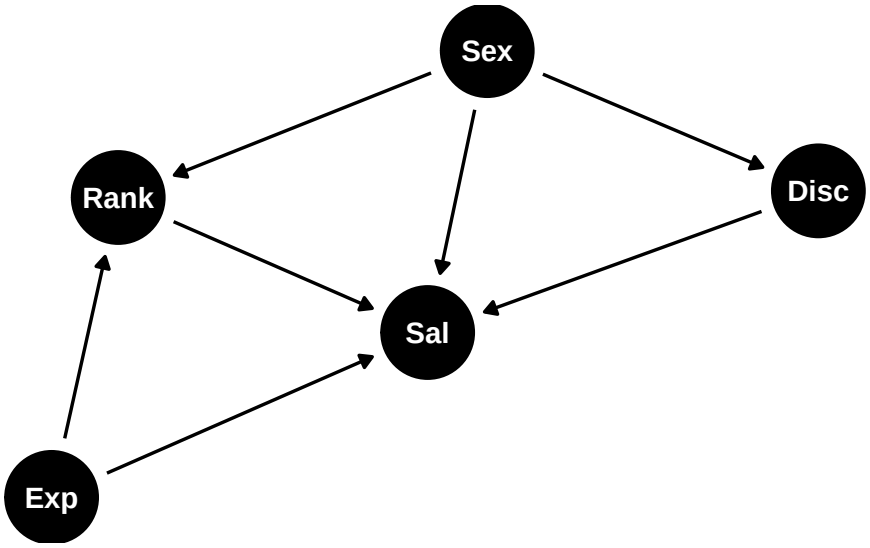
```
Salaries$yrs.service = NULL
```

Before proceeding we could also undertake one final modification to recode 'rank' into a (0,1) variable by aggregating the first two categories. That way the comparison will be more intuitive, 'full prof' vs 'not yet a full prof'. However, as you can see below, we are really doing this so we can use this variable as the outcome of a binary logistic model like the ones you saw last year. Ordinal variables with more than two categories, like 'rank' in its original form, can be specified using ordered logit models, but we have not covered that yet.

```
Salaries$rankrec = ifelse(Salaries$rank=="Prof", 1, 0) #The recoding
Salaries$rankrec = factor(Salaries$rankrec, levels = c(0,1),
                          labels=c("NotFullProf","Full Prof"))
#This is to make rankrec a categorical variable and to provide meaningful labels
Salaries$rank = NULL #To keep the dataset tidy I get rid of the original rank variable
```

At this point we can start designing our causal model. Remember, we want to estimate the effect of 'sex' on 'salary', and to do so we will consider potential confounders and mediators. This requires careful theoretical reflection (you should dedicate a good amount of time to do this before you start the modelling process), which is clearly a subjective process, which means that the causal model that I am suggesting below is not necessarily the right one, just a model that made sense to me. When you are doing this on your own I recommend that you use pen and paper, and only once you are happy with your model draw it more formally.

```
dag2 = dagify(Sal~Exp, Sal~Rank, Sal~Disc, Sal~Sex, Rank~Sex, Disc~Sex, Rank~Exp)
ggdag(dag2) + theme_dag_blank()
```



I have considered that sex, but also discipline, rank and experience, have a causal effect on salary. In addition, I suspect that rank and discipline mediate the effect of sex on salary, since there is evidence that female workers do not demand to be promoted (rank) as often as their male colleagues, and given that men tend to choose more applied disciplines. I have also included an effect of experience on rank since often people tend to get promoted simply as a function of time spent doing the same job, but that won't be explored here since the question of interest is the effect of gender on salary.

As we did in the previous exercise we start with a simple model where we look at the direct effect of sex on salary without any controls.

```
model1 = lm(salary~sex, data=Salaries)
summary(model1)
```

```
##
## Call:
## lm(formula = salary ~ sex, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57290 -23502  -6828  19710 116455
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   101002       4809   21.00 <0.0000000000000002 ***
## sexMale        14088       5065    2.78              0.0057 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30000 on 395 degrees of freedom
## Multiple R-squared:  0.0192, Adjusted R-squared:  0.0167
## F-statistic: 7.74 on 1 and 395 DF,  p-value: 0.00567
```

We find evidence of the expected gender pay gap, with female members of staff earning $14,088 less than male members of staff. We proceed by adding all the other factors that we thought could be explaining differences in salary.

```r
model2 = lm(salary~rankrec+discipline+yrs.since.phd+sex, data=Salaries)
summary(model2)
```

```
##
## Call:
## lm(formula = salary ~ rankrec + discipline + yrs.since.phd +
##     sex, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69758 -13836  -1953  11659  95704
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)          73026       4280   17.06 < 0.0000000000000002 ***
## rankrecFull Prof     37437       3278   11.42 < 0.0000000000000002 ***
## disciplineB          14208       2371    5.99          0.0000000047 ***
## yrs.since.phd          185        122    1.51                  0.13
## sexMale               4157       3919    1.06                  0.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22900 on 392 degrees of freedom
## Multiple R-squared:  0.433,  Adjusted R-squared:  0.427
## F-statistic: 74.9 on 4 and 392 DF,  p-value: <0.0000000000000002
```

```r
VIF(model2)
```

```
##       rankrec    discipline yrs.since.phd           sex
##        1.7955        1.0543        1.8679        1.0281
```

The gender effect is not significant anymore after controlling for rank and discipline. Two important issues need to be noted though. First, we still estimate a non-negligible gender differential of roughly $4,000, but we cannot claim that this is a statistically significant difference, probably because of the few women captured in our sample. However, just because a result is (or is not) statistically significant it does not mean that it is (or it is not) substantively significant, with a larger sample size we would probably found that difference to be statistically significant.

Second, although we do not find conclusive evidence of unwarranted gender disparities in this particular college (i.e. the observed disparities are explained by legitimate factors). However, we cannot yet conclude that gender discrimination is not present since it is possible that the gender effect on salary is fully mediated by the factor rank. As hypothesised, there is plenty of evidence in the literature that points at how female workers are less 'pushy' at applying for promotions. We proceed to explore that next.

```r
model3 = glm(rankrec~sex+yrs.since.phd, data=Salaries, family="binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = rankrec ~ sex + yrs.since.phd, family = "binomial",
##     data = Salaries)
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept)    -3.9306     0.5829   -6.74       0.000000000015 ***
## sexMale         0.6777     0.4691    1.44                 0.15
```

```
## yrs.since.phd   0.2303     0.0238    9.67 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 503.52  on 396  degrees of freedom
## Residual deviance: 258.61  on 394  degrees of freedom
## AIC: 264.6
##
## Number of Fisher Scoring iterations: 6
```

We cannot corroborate such hypothesis since the gender effect is not statistically significant. Again, this is probably due to the small number of women in our sample, since the estimated gender effect is not small. In fact, by transforming the coefficient for gender from log-odds to odds by exponentiating with base e, $e^{\hat{\beta}_{sex}}$, you can see how men are roughly two times more likely to be made full professors even after controlling for number of years since PhD.

```
exp(model3$coefficients[2])
```

```
## sexMale
##  1.9694
```

So, with the sample we have here we cannot claim that the effect of gender on salary has been mediated by rank. On the other hand, that is what could be happening with 'years.since.phd'. To determine this we should run a model for salary with 'years.since.phd' as the only explanatory variable, but that is not something relevant to our study of the gender gap, so we proceed to assess the second potential factor mediating the relationship between gender and salaries, 'discipline'.

Sidenote: Notice how the model we have just specified ('model3') is logistic, i.e. non-linear, whereas 'model2' is linear. This makes the calculation of total effects difficult since the regression coefficients are measured in different units, logs of salary in dollars, and log-odds of being a full professor. In these instances, and since this is only a first approximation to path analysis, we will not proceed to calculate the total effects. However, we can still determine the presence of mediating effects (partial and full), even if we cannot estimate their specific effect robustly.

```
model4 = glm(discipline~sex, data=Salaries, family="binomial")
summary(model4)
```

```
##
## Call:
## glm(formula = discipline ~ sex, family = "binomial", data = Salaries)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1542     0.3212    0.48     0.63
## sexMale       0.0251     0.3383    0.07     0.94
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 547.27  on 396  degrees of freedom
## Residual deviance: 547.26  on 395  degrees of freedom
## AIC: 551.3
##
## Number of Fisher Scoring iterations: 3
```

ok, so this other potential mediating effect is definitely refuted. The gender effect on 'discipline' is not

statistically significant, and its effect is really small, making it not meaningful at all.

In conclusion, we have not found evidence of direct or indirect gender discriminatory practices in the College studied. Specifically, we refute the gender gap to be explained by different disciplinary choices made by men and women. However, we have noted our sample size is not sufficiently big to detect gender disparities precisely enough. In addition, we have only explored a limited number of factors available in this sort of toy dataset that we have used.

Preparation for next week's workshop

Next week we will consider non-linear associations. Specifically, we will practice polynomial regression. To do so we will use the academics dataset from today's bonus exercise and the Labour Force Survey to explore how the relationship between age or experience with salaries is not constant, but rather it accelerates and then it disappears after a given point. As before, to prepare for the workshop you are requested to replicate the first exercise and to give a good try to the second exercise.