Workshop 3 - Path Analysis

JPS

23/12/2021

Introduction

We are going to practice designing causal diagrams (using the package *ggdag*) and implementing them (using the package *mediation*). We will focus on the exploration of mediating effects (path analysis), but we will also pay attention to potential confounding and collider 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 insitution and the subjects under its authority) is also an important factor explaining 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 condition on 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 capturing Salaries of academic staff from a given college in the US.

Exercise 3: The Salaries data only captures a few variables and is restricted to one particular employer. In this exercise you will be asked to undertake a more ambitious study of the gender gap using the Labour Force Survey.

Exercise 1. The Procedural Justice Model

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

```
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"
```

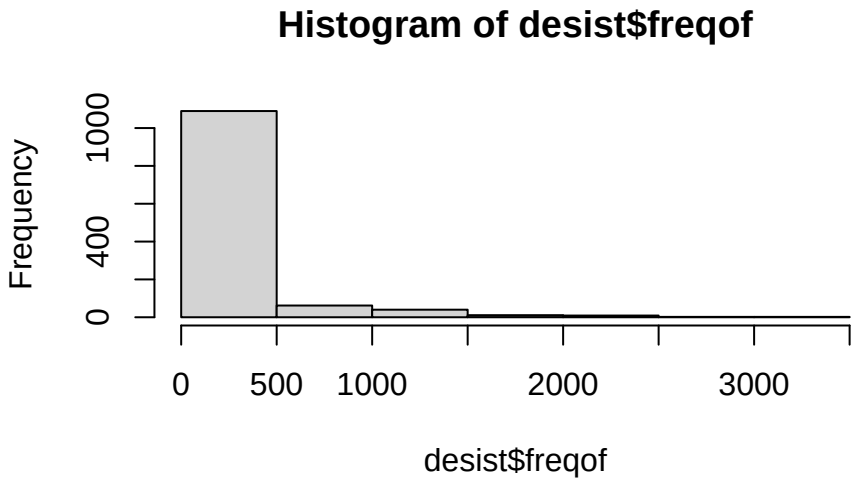
```
summary(desist)
```

```
##       age            ethn               gend               pjcop
##  Min.   :14.00   Length:1217        Length:1217        Min.   :1.390
##  1st Qu.:15.00   Class :character   Class :character   1st Qu.:2.397
##  Median :16.00   Mode  :character   Mode  :character   Median :2.710
##  Mean   :16.03                                         Mean   :2.760
##  3rd Qu.:17.00                                         3rd Qu.:3.090
##  Max.   :19.00                                         Max.   :4.490
##                                                        NA's   :1
##      legit           freqof
##  Min.   :1.000   Min.   :   1.0
##  1st Qu.:1.910   1st Qu.:   4.0
##  Median :2.270   Median :  17.0
##  Mean   :2.283   Mean   : 169.2
##  3rd Qu.:2.640   3rd Qu.: 110.0
##  Max.   :4.000   Max.   :3493.0
##  NA's   :1
```

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) respectively.

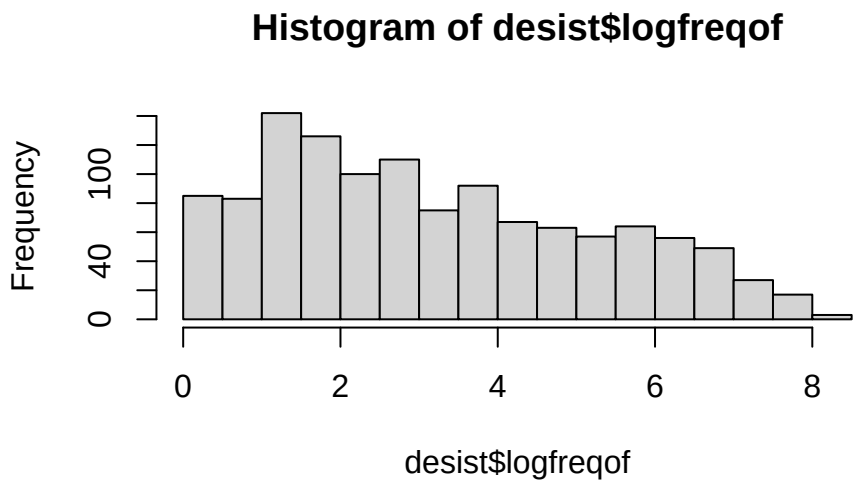
From the exploratory analysis we can identify a couple of issues. There are a few missing cases in some of the variables, probably due to non-response. Since the proportion of missing cases to the sample size is tiny we can simply drop them from our study. In addition, 'freqof' seems to be affected by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

```
desist = desist[complete.cases(desist), ] #This is to drop all 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.

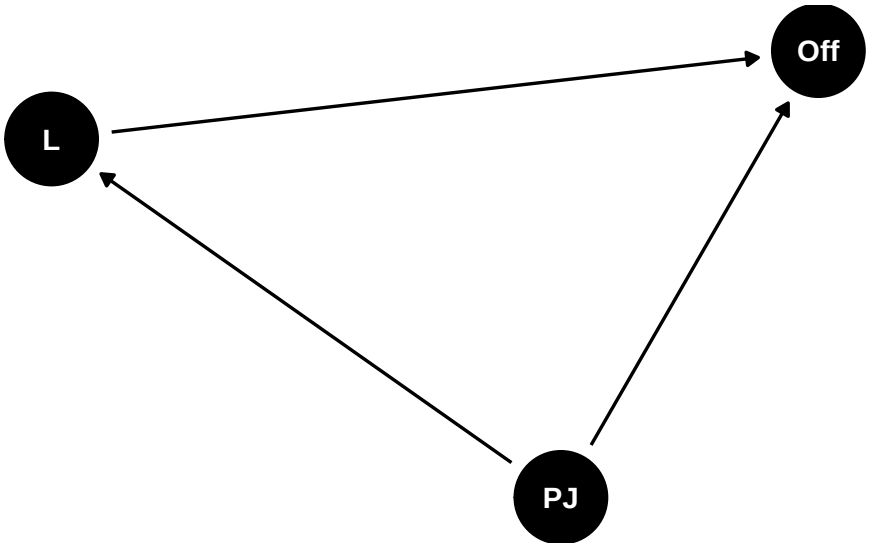
```
desist$logfreqof = log(desist$freqof)
hist(desist$logfreqof)
```



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. At this point the model is quite simple, but for more complex models representing them visually will help 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. It is not a surprise to learn that there is an R package to do exactly this, draw DAGs - this is one of the reasons why we love R so much, there is a package for everything. The graph below represents the procedural justice model that we seek to explore, with legitimacy (L) as a mediator of procedural justice (PJ).

```
library(ggdag)  #This is to draw DAGs
library(dplyr)  #This is to use %>%
dag1 = dagify(Off~PJ, Off~L, L~PJ)
ggdag(dag1) + theme_dag_blank()  #theme_dag_blank() provides a white background and no axes
```




```
#Alternatively we could simplify the above code as follows:
#dagify(Off~PJ, Off~L, L~PJ) %>% ggdag + theme_dag_blank()
```

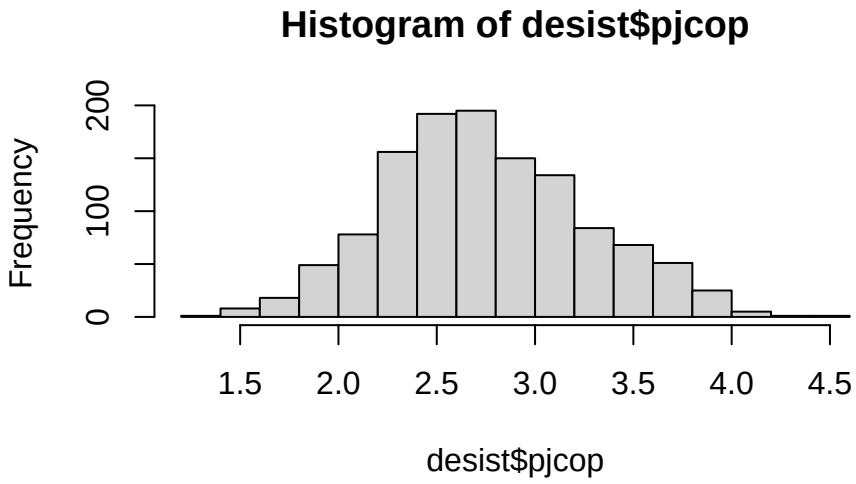
We are now going to estimate the model depicted in the figure above. To understand the mediating effect of legitimacy completely we should break this process up into three steps:

1. Estimate the effect of procedural justice on offending
2. Estimate the effect of procedural justice on offending while controlling for legitimacy
3. Estimate the effect of procedural justice on legitimacy

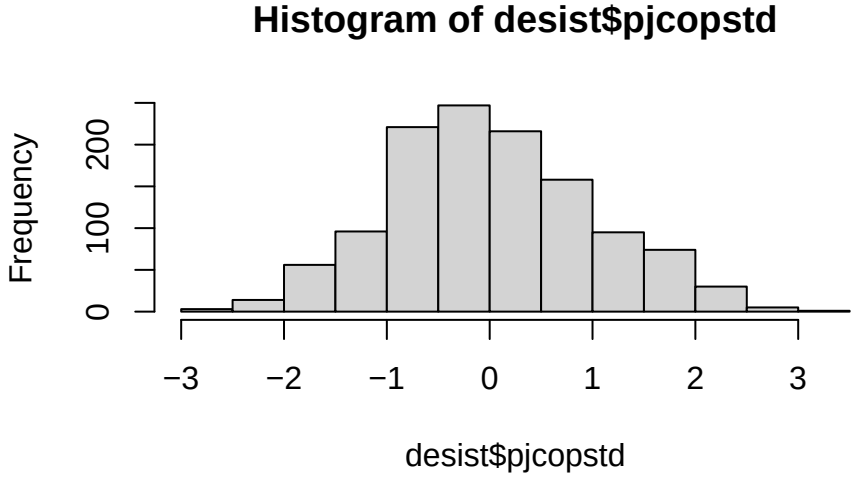
To facilitate the interpretation of the different coefficients of procedural justice and legitimacy we could standardise these variables. This way the coefficient of each of those variables could be interpreted not like the change in the outcome variable after the explanatory variable increases in one unit, but as the change in the outcome variable as 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.

Mathematically, 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 transformations.

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




```
desist$pjcopstd = (desist$pjcop - mean(desist$pjcop))/sd(desist$pjcop)
hist(desist$pjcopstd)
```




```
desist$legitstd = (desist$legit - mean(desist$legit))/sd(desist$legit)
```

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(logfreqof~pjcopstd, data=desist)
summary(model1)
```

```
##
## Call:
## lm(formula = logfreqof ~ pjcopstd, data = desist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1291 -1.6455 -0.2987  1.4812  5.0842
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.16355    0.05837  54.201  < 2e-16 ***
## pjcopstd    -0.38396    0.05839  -6.576 7.18e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.035 on 1214 degrees of freedom
## Multiple R-squared:  0.03439,    Adjusted R-squared:  0.0336
## F-statistic: 43.24 on 1 and 1214 DF,  p-value: 7.177e-11
```

Procedural justice exerts a significant influence in the frequency of offending. Specifically, we can estimate that for those that report procedural justice to be one standard deviation above average, the number of offences committed goes down by roughly seven. To do this we need to back-transform the regression coefficients using exponents to reflect that the outcome variable was log-transformed. The exponential transformation of the intercept will 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 will represent the 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)
pj = coef(model1)[2]  #This is to record the coefficient of procedural justice
b = exp(int+pj)
table1 = c(a, b)
names(table1) = c("offences committed by the average offender", "offences commited by offenders reporti
table1
```

```
##                                       offences committed by the average offender
##                                                                         23.65432
## offences commited by offenders reporting one std. dev. higher procedural justice
##                                                                         16.11233
```

Let's now test whether perceptions of procedural justice influence perceptions of legitimacy.

```
model2 = lm(legitstd~pjcopstd, data=desist)
summary(model2)
```

```
##
## Call:
## lm(formula = legitstd ~ pjcopstd, data = desist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60456 -0.58729  0.00244  0.56008  2.93480
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.289e-16  2.456e-02    0.00        1
## pjcopstd     5.171e-01  2.457e-02   21.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8563 on 1214 degrees of freedom
## Multiple R-squared:  0.2674, Adjusted R-squared:  0.2668
## F-statistic: 443.1 on 1 and 1214 DF,  p-value: < 2.2e-16
```

They do. In fact, procedural justice alone explains over a quarter of the variability in the 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(logfreqof~pjcopstd+legitstd, data=desist)
summary(model3)
```

```
##
## Call:
## lm(formula = logfreqof ~ pjcopstd + legitstd, data = desist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4244 -1.5776 -0.2319  1.4041  5.6154
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.16355    0.05768  54.847  < 2e-16 ***
## pjcopstd    -0.19269    0.06742  -2.858  0.00433 **
## legitstd    -0.36989    0.06742  -5.487 4.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.011 on 1213 degrees of freedom
## Multiple R-squared:  0.05778,    Adjusted R-squared:  0.05622
## F-statistic: 37.19 on 2 and 1213 DF,  p-value: < 2.2e-16
```

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 (roughly half the size) than what it was in Model 1, which tell us that the direct effect of procedural justice on offending described in Model 1 had been overestimated. Yet, since procedural justice still has a significant effect on offending (even after controlling for legitimacy), and given that 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 (mediated through legitimacy) effect 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 have not relied on 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 offending
indirect = coef(model3)[3]*coef(model2)[2] #The effect of legitimacy on offending times the effect of p
total = direct + indirect
table2 = c(direct, indirect, total)
names(table2) = c("direct", "indirect", "total")
table2
```

```
##     direct   indirect      total
## -0.1926885 -0.1912724 -0.3839609
```

If you want to obtain standard errors for the above effects we can use the *mediation* package, which estimates them using 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, e.g. when combining estimates from different models. 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 this 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 all get the same random draws when using bootstrap
table3 = mediate(model2, model3, treat='pjcopstd',
                 mediator='legitstd', 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  <2e-16 ***
## ADE              -0.193       -0.308        -0.06  <2e-16 ***
## Total Effect     -0.384       -0.477        -0.28  <2e-16 ***
## Prop. Mediated    0.498        0.310         0.81  <2e-16 ***
## ---
## 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.

This is all pretty exciting, don't you think? 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. See for example Pina-Sánchez et al. 2018, where we estimate the total effect of 'Step-One' sentencing guideline factors (those determining harm caused and offender culpability) on the final sentence imposed using path analysis. 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 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 omitted relevant variables, i.e. whether there are important confounding factors that we are not controlling 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 irrelevant factors and colliders (i.e. those which are affected by the outcome).

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(legitstd~pjcopstd+gend+ethn+age, data=desist)
summary(model2b)
```

```
##
## Call:
## lm(formula = legitstd ~ pjcopstd + gend + ethn + age, data = desist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66746 -0.54138  0.01864  0.55649  3.16072
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.03860    0.34878   2.978 0.002961 **
## pjcopstd          0.49138    0.02460  19.975  < 2e-16 ***
## gend(2) Female    0.20421    0.07148   2.857 0.004352 **
## ethn(2) Black    -0.24552    0.06599  -3.721 0.000208 ***
## ethn(3) Hispanic  0.02476    0.06755   0.367 0.713996
## ethn(4) Other     0.07942    0.12498   0.636 0.525218
## age              -0.06106    0.02159  -2.828 0.004755 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8418 on 1209 degrees of freedom
## Multiple R-squared:  0.2948, Adjusted R-squared:  0.2913
## F-statistic: 84.25 on 6 and 1209 DF,  p-value: < 2.2e-16
```

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

```
##              GVIF Df GVIF^(1/(2*Df))
## pjcopstd 1.037452  1        1.018554
## gend     1.007085  1        1.003536
## ethn     1.023343  3        1.003853
## age      1.026773  1        1.013298
```

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(logfreqof~legitstd+pjcopstd+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.

We claimed that Tyler's procedural justice model is corroborated, these results increase the robustness of such claim. However, in reporting these results we should make our assumptions as explicit as possible. We should include caveats pointing at the potential bidirectional causal paths, or the range of potential confounding factors that we have not been able to control.

Exercise 2. The Gender Gap (academic salaries)

Let's now explore the gender pay gap using data from a US college and the techniques we have practised in the previous exercise. The data is stored in the 'car' package.

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

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

```
summary(Salaries)
```

```
##         rank     discipline yrs.since.phd    yrs.service        sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266              Median :21.00   Median :16.00
##                             Mean   :22.31   Mean   :17.61
##                             3rd Qu.:32.00   3rd Qu.:27.00
##                             Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  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).

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

```
##           [,1]      [,2]
## [1,] 1.0000000 0.9096491
## [2,] 0.9096491 1.0000000
```

```
plot(Salaries$yrs.service, Salaries$yrs.since.phd)
```

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 as well as possible. 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'.

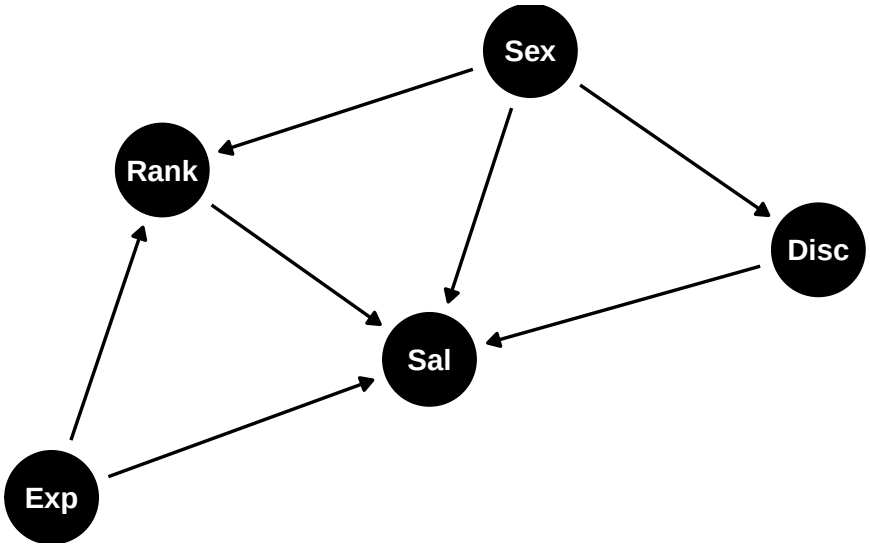
```
Salaries$yrs.service = NULL
```

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)
Salaries$rankrec = factor(Salaries$rankrec, levels = c(0,1), labels=c("NotFullProf","Full Prof"))
Salaries$rank = NULL
```

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.

```
dagify(Sal~Exp, Sal~Rank, Sal~Disc, Sal~Sex,
       Rank~Sex, Disc~Sex, Rank~Exp) %>% ggdag() + 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 dicipline mediate the effect of sex on salary, since there is evidence that female workers do not ask 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.

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  21.001  < 2e-16 ***
## sexMale        14088       5065   2.782  0.00567 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30030 on 395 degrees of freedom
## Multiple R-squared:  0.01921,    Adjusted R-squared:  0.01673
## F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667
```

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.

```
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)        73025.5     4279.8  17.063  < 2e-16 ***
## rankrecFull Prof   37437.2     3278.0  11.421  < 2e-16 ***
## disciplineB        14208.3     2371.4   5.992 4.71e-09 ***
## yrs.since.phd        184.6      122.1   1.511    0.132
## sexMale             4156.6     3918.6   1.061    0.289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22920 on 392 degrees of freedom
## Multiple R-squared:  0.4332, Adjusted R-squared:  0.4275
## F-statistic: 74.91 on 4 and 392 DF,  p-value: < 2.2e-16
```

```
VIF(model2)
```

```
##       rankrec    discipline yrs.since.phd           sex
##      1.795522      1.054304      1.867856      1.028101
```

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.

```
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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0075  -0.4305   0.1138   0.4938   1.6962
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.93065    0.58289  -6.743 1.55e-11 ***
```

```
## sexMale        0.67773    0.46912   1.445    0.149
## yrs.since.phd  0.23025    0.02381   9.671  < 2e-16 ***
## ---
## 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: 264.61
##
## 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 using the exponent, $e^{\beta_{sex}}$, 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
## 1.969398
```

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)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.254  -1.254   1.102   1.102   1.113
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1542     0.3212   0.480    0.631
## sexMale       0.0251     0.3383   0.074    0.941
##
## (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: 551.26
##
## Number of Fisher Scoring iterations: 3
```

ok, so this other potential mediating effect is definitely refuted. The gender effect on 'discipline' is not only statistically non-significant, it is also really small, making it substantively non-significant.

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 probably not big enough to detect some 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. We can do better and take the exploration of the gender gap three notches up by making use of the Labour Force Survey.

Exercise 3. The Gender Gap (Labour Force Survey)

Consider this as a 'take-home' exercise if we run out of time in the workshop. You are required to explore the gender pay gap in the UK using the Labour Force Survey. To do so start by designing a causal model based on no more than five to six variables (this is to avoid making it too long, a thorough study on this topic should be more complex than that). Identify important confounders, avoid including colliders, mind problems of multicollinearity, and if possible provide the direct, indirect, and total effect of gender of salaries. We will update this script with our own gender gap model before the end of the week. Remember that this will not be the 'right model', but only one that of the many that could make sense to us.

Accessing the data and keeping variables of interest. 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).

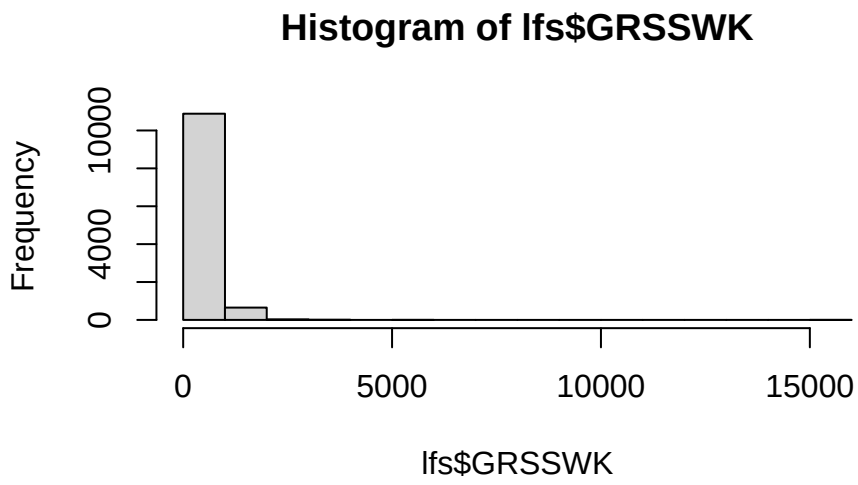
```
load("C:/Users/JPS/Dropbox/Leeds/LAW3287 Quantitative Social Research II/Week17.Nonlinear effects/lfs.r
vars = c("SEX","AGE","SC10MMJ","TTUSHR","QUAL_1","GRSSWK")
lfs = lfs[vars]
summary(lfs)
```

Now I trim down the number of cases in the dataset by removing those with missing data in any of the variables.

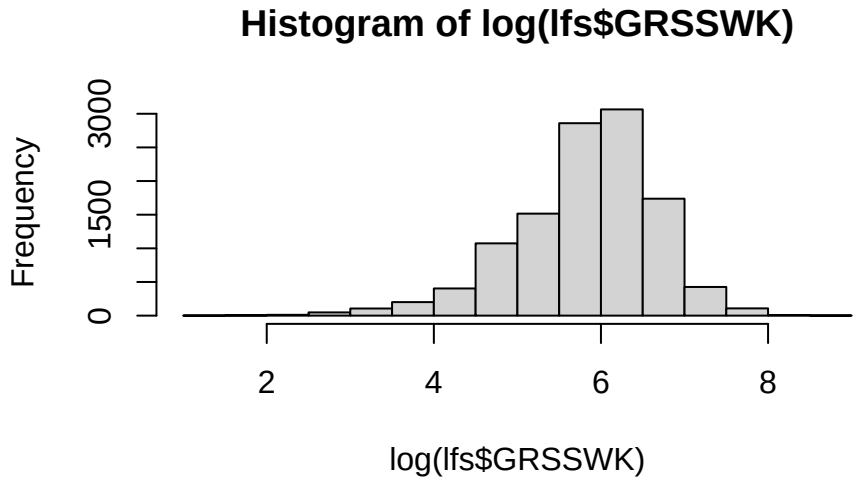
```
table(lfs$GRSSWK, useNA="ifany")
lfs = lfs[which(lfs$GRSSWK>-1),]
table(lfs$SC10MMJ, useNA="ifany")
lfs = lfs[which(lfs$SC10MMJ!="Does not apply"),]
table(lfs$TTUSHR, useNA="ifany")
lfs = lfs[which(lfs$TTUSHR>-1),]
```

I explore the outcome variable in more detail. This helps me identify one outlier, which I proceed to remove. I can also see that salaries are not normally distributed, they are right-skewed. As a result I proceed to explore whether a logarithmic transformation of the outcome variable seems more appropriate. I also explore the distribution of 'TTUSHR' since I suspect it might be mediating the effect of gender on salaries. I won't transform this one since it seems approximately normal.

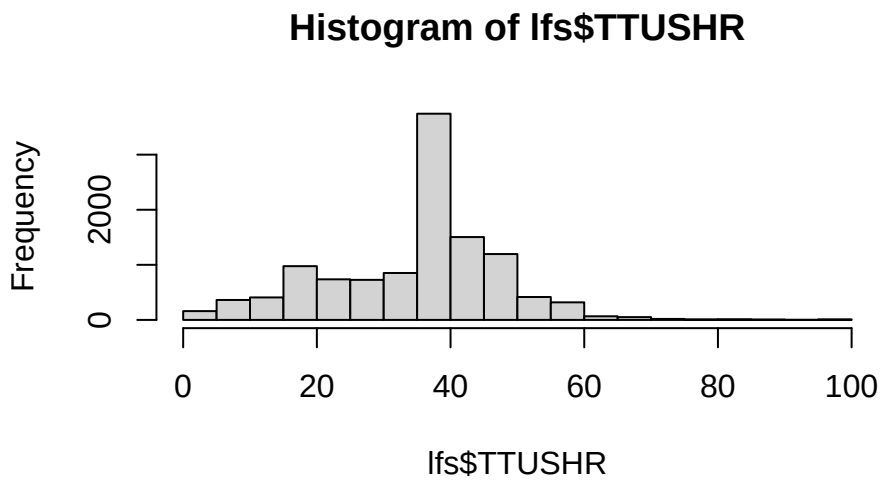
```
hist(lfs$GRSSWK)
```




```
lfs = lfs[which(lfs$GRSSWK<10000),]  #I get rid of one outlier
hist(log(lfs$GRSSWK))
```

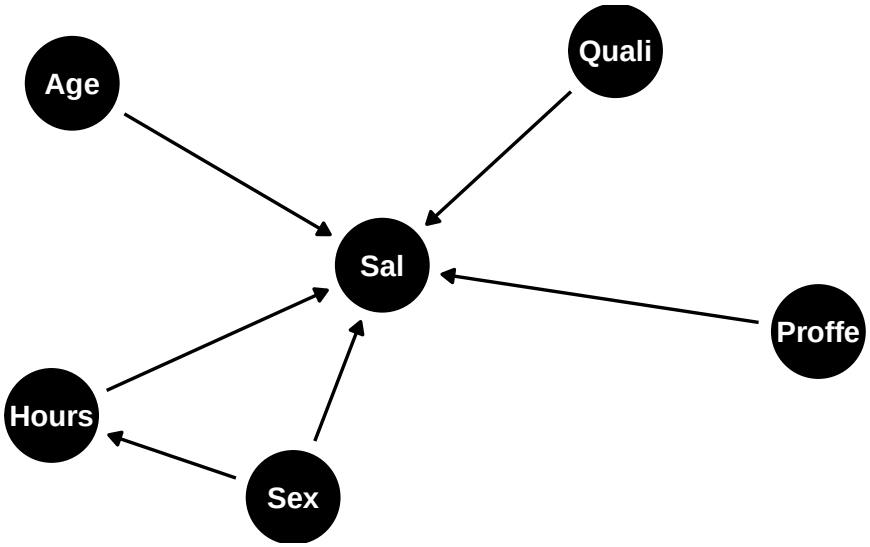



```
hist(lfs$TTUSHR)
```



I am going to explore now the gender gap in three stages (as we did above), first looking at the effect of gender on salaries without any controls, then a second model where potential confounders are controlled, and third I will explore the possibility of the number of total hours worked mediating the effect of gender on salaries. 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. The full causal model can be represented as follows:

```
dagify(Sal~Proffe, Sal~Quali, Sal~Age, Sal~Sex, Sal~Hours, Hours~Sex) %>% ggdag() + theme_dag_blank()
```




```
lm1 = lm(log(lfs$GRSSWK)~SEX, data=lfs)
summary(lm1)
```

```
##
## Call:
## lm(formula = log(lfs$GRSSWK) ~ SEX, data = lfs)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.4755 -0.4150  0.0888  0.5395  2.7005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.08492    0.01082  562.50   <2e-16 ***
## SEXFemale   -0.51076    0.01492  -34.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8016 on 11572 degrees of freedom
## Multiple R-squared:  0.09194,    Adjusted R-squared:  0.09186
## F-statistic:  1172 on 1 and 11572 DF,  p-value: < 2.2e-16
```

```
lm2 = lm(log(lfs$GRSSWK)~SEX+AGE+SC10MMJ+TTUSHR+QUAL_1, data=lfs)
summary(lm2) #Evidence of the hypothesised direct effect
```

```
##
## Call:
## lm(formula = log(lfs$GRSSWK) ~ SEX + AGE + SC10MMJ + TTUSHR +
##     QUAL_1, data = lfs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1788 -0.2186  0.0331  0.2725  2.6070
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       4.3930655  0.0293819
## SEXFemale                                        -0.1062803  0.0101934
## AGE                                               0.0067629  0.0003535
## SC10MMJProfessional Occupations                   0.0760164  0.0180639
## SC10MMJAssociate Professional and Technical Occupations -0.0244356  0.0189864
## SC10MMJAdministrative and Secretarial Occupations -0.2450236  0.0199567
## SC10MMJSkilled Trades Occupations                 -0.3310595  0.0219088
## SC10MMJCaring, Leisure and Other Service Occupations -0.4792226  0.0213244
## SC10MMJSales and Customer Service Occupations     -0.5196091  0.0222015
## SC10MMJProcess, Plant and Machine Operatives      -0.4723624  0.0232921
## SC10MMJElementary Occupations                     -0.6693554  0.0207665
## TTUSHR                                            0.0388346  0.0003919
## QUAL_1Yes                                         0.1853952  0.0116615
##                                                   t value Pr(>|t|)
## (Intercept)                                       149.516  < 2e-16 ***
## SEXFemale                                         -10.426  < 2e-16 ***
## AGE                                                19.131  < 2e-16 ***
## SC10MMJProfessional Occupations                     4.208 2.59e-05 ***
## SC10MMJAssociate Professional and Technical Occupations  -1.287    0.198
## SC10MMJAdministrative and Secretarial Occupations -12.278  < 2e-16 ***
## SC10MMJSkilled Trades Occupations                 -15.111  < 2e-16 ***
## SC10MMJCaring, Leisure and Other Service Occupations -22.473  < 2e-16 ***
## SC10MMJSales and Customer Service Occupations     -23.404  < 2e-16 ***
## SC10MMJProcess, Plant and Machine Operatives      -20.280  < 2e-16 ***
## SC10MMJElementary Occupations                     -32.232  < 2e-16 ***
## TTUSHR                                             99.102  < 2e-16 ***
## QUAL_1Yes                                          15.898  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4721 on 11561 degrees of freedom
## Multiple R-squared:  0.6854, Adjusted R-squared:  0.685
## F-statistic:  2099 on 12 and 11561 DF,  p-value: < 2.2e-16
```

```
VIF(lm2) #No multicollinearity
```

```
##             GVIF Df GVIF^(1/(2*Df))
## SEX     1.345570  1        1.159987
## AGE     1.049864  1        1.024629
## SC10MMJ 1.995857  8        1.044138
## TTUSHR  1.356302  1        1.164604
## QUAL_1  1.428995  1        1.195406
```

```
lm3 = lm(TTUSHR~SEX, data=lfs)
summary(lm3) #Evidence of the hypothesised indirect effect
```

```
##
## Call:
## lm(formula = TTUSHR ~ SEX, data = lfs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.873  -5.873   0.127   7.127  65.859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.8731     0.1633   250.3   <2e-16 ***
## SEXFemale    -9.7323     0.2253   -43.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.1 on 11572 degrees of freedom
## Multiple R-squared:  0.1389, Adjusted R-squared:  0.1388
## F-statistic:  1866 on 1 and 11572 DF,  p-value: < 2.2e-16
```

These findings suggest the presence of a gender gap, which takes the form of a direct effect, female workers earning less than their male counterparts after controlling for age, profession, number of hours worked and whether having a degree. In addition, I have also found evidence that this gap could be reinforced by an indirect effect through a potential impediment for women to work as many hours as men.