# Analysis of Male-Female Earnings Differentials Using 2018 Canadian Income Survey

Bahareh Aghababaei

2023-02-25

## Data Preparation

```
library(haven)
mydata <- read_dta("C:/Users/bahar/Documents/MSc courses/ECON 562/Assignment 4/cis_2018_en.dta")
nrow(mydata)
```

```
## [1] 94336
```

```
mydata_women <- subset(mydata, sex==2)
nrow(mydata_women)
```

```
## [1] 48054
```

```
mydata_men <- subset(mydata, sex==1)
nrow(mydata_men)
```

```
## [1] 46282
```

## Data Preparation-Provincial Restriction

```
mydata_2018 <- subset(mydata, prov==35 )
#women
mydata_2018_women <- subset(mydata, prov==35 & sex==2 )
nrow(mydata_2018_women)
```

```
## [1] 13004
```

```
#men
mydata_2018_men <- subset(mydata, prov==35 & sex==1 )
nrow(mydata_2018_men)
```

```
## [1] 12415
```

## Data Preparation-Age Restriction

```
mydata_2018 <- subset(mydata_2018, agegp==13 | agegp==12 | agegp==11 |agegp==10 | agegp=="09" | agegp==
nrow(mydata_2018)
```

```
## [1] 15355
```

```
mydata_2018_women <- subset(mydata_2018, sex==2)
nrow(mydata_2018_women)
```

```
## [1] 7862
```

```
mydata_2018_men <- subset(mydata_2018, sex==1)
nrow(mydata_2018_men)
```

```
## [1] 7493
```

## Data Preparation-Handling Missing values

```
na.omit(mydata_2018)
```

```
## # A tibble: 0 x 192
## # i 192 variables: year <dbl>, pumfid <dbl>, personid <chr>, fweight <dbl>,
## #   prov <chr>, uszgap <chr>, mbmregp <chr>, agegp <chr>, sex <chr>,
## #   marstp <chr>, cmphi <chr>, HLEV2G <chr>, studtfp <chr>, fllprtp <chr>,
## #   fworked <chr>, scsum <chr>, alfst <chr>, wksem <dbl+lbl>, wksuem <dbl+lbl>,
## #   wksnlf <dbl+lbl>, ushrwk <dbl>, alhrwk <dbl+lbl>, fpdwk <chr>, fsemp <chr>,
## #   funfw <chr>, immst <chr>, yrimmg <chr>, alimo <dbl+lbl>, alip <dbl+lbl>,
## #   atinc <dbl+lbl>, capgn <dbl+lbl>, ccar <dbl+lbl>, chfed <dbl+lbl>, ...
```

```
mydata_2018$fworked <- ifelse(mydata_2018$fworked==1,1,0)
mydata_2018 <- subset(mydata_2018 , marstp<10 & alfst<08 & HLEV2G<5 & yrimmg<6 )
nrow(mydata_2018)
```

```
## [1] 2955
```

```
mydata_2018_women <- subset(mydata_2018, sex==2)
nrow(mydata_2018_women)
```

```
## [1] 1571
```

```
mydata_2018_men <- subset(mydata_2018, sex==1)
nrow(mydata_2018_men)
```

```
## [1] 1384
```

## Descriptive Analysis

```r
# Earning
value_earning_women<- mean(mydata_2018_women$cfearng)
value_earning_men<- mean(mydata_2018_men$cfearng)

summary (mydata_2018_men$cfearng)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -40650   44925   87500  103095  139500 1370000
```

```r
#age
value_age_women<- mean(as.numeric(mydata_2018_women$agegp))
value_age_men<- mean(as.numeric(mydata_2018_men$agegp))


#education

value_edu_women<- mean(as.numeric(mydata_2018_women$HLEV2G))
value_edu_men<- mean(as.numeric(mydata_2018_men$HLEV2G))



#marital status

value_mrt_women<- mean(as.numeric(mydata_2018_women$marstp))
value_mrt_men<- mean(as.numeric(mydata_2018_men$marstp))


#immigrant status
value_img_women<- mean(as.numeric(mydata_2018_women$yrimmg))
value_img_men<- mean(as.numeric(mydata_2018_men$yrimmg))


#working status
value_ws_women<- mean(as.numeric(mydata_2018_women$fworked))
value_ws_men<- mean(as.numeric(mydata_2018_men$fworked))


#family size
value_fs_women<- mean(as.numeric(mydata_2018_women$cfsize))
value_fs_men<- mean(as.numeric(mydata_2018_men$cfsize))


#family composition
value_fc_women<- mean(as.numeric(mydata_2018_women$cfcomp))
value_fc_men<- mean(as.numeric(mydata_2018_men$cfcomp))


#years of schooling
```

```r
mydata_2018$sy <- ifelse(mydata_2018$HLEV2G==2, 12,
       ifelse(mydata_2018$HLEV2G==1, 10,
              ifelse(mydata_2018$HLEV2G==4, 16,15)))

#converting age to continuous variable for computing years of experience

range_age <- c(0,0,0,0,21,27,32,37,42,47,52,57,62)

for (i in 1:2955){
  mydata_2018$age[i]= range_age[as.numeric(mydata_2018$agegp[i])]

  mydata_2018$expr[i]=as.numeric(mydata_2018$age[i])-as.numeric(mydata_2018$sy[i])-5
}
mydata_2018_women <- subset(mydata_2018, sex==2)
mydata_2018_men <- subset(mydata_2018, sex==1)

value_expr_women<- mean(as.numeric(mydata_2018_women$expr))
value_expr_men<- mean(as.numeric(mydata_2018_men$expr))
```

```r
library(dplyr);
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(knitr);

summary_table <- data.frame(
  Variable = c("Earnings", "Age", "Education", "Marital Status",
               "Immigrant Status", "Working Status", "Family Size",
               "Family Composition", "Years of Experience"),
  Women = c(mean(mydata_2018_women$cfearng, na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$agegp), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$HLEV2G), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$marstp), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$yrimmg), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$fworked), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$cfsize), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$cfcomp), na.rm = TRUE),
            mean(as.numeric(mydata_2018_women$expr), na.rm = TRUE)),
  Men = c(mean(mydata_2018_men$cfearng, na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$agegp), na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$HLEV2G), na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$marstp), na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$yrimmg), na.rm = TRUE),
```

```r
          mean(as.numeric(mydata_2018_men$fworked), na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$cfsize), na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$cfcomp), na.rm = TRUE),
          mean(as.numeric(mydata_2018_men$expr), na.rm = TRUE))
)

# Print table in a readable format
kable(summary_table, caption = "Descriptive Statistics by Gender")
```

Table 1: Descriptive Statistics by Gender

| Variable | Women | Men |
|---|---:|---:|
| Earnings | 9.350013e+04 | 1.030948e+05 |
| Age | 9.591343e+00 | 9.596098e+00 |
| Education | 3.128580e+00 | 3.135838e+00 |
| Marital Status | 1.686824e+00 | 1.746387e+00 |
| Immigrant Status | 2.648631e+00 | 2.634393e+00 |
| Working Status | 7.371101e-01 | 8.937861e-01 |
| Family Size | 3.094844e+00 | 3.189306e+00 |
| Family Composition | 3.844685e+00 | 3.726156e+00 |
| Years of Experience | 2.538065e+01 | 2.538584e+01 |

## Regression Analysis: Probit model and IMR for Men and Women

```r
# Run the Probit model for men
library(sampleSelection)
```

```
## Loading required package: maxLik

## Loading required package: miscTools

##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

```r
probit_men <- glm(fworked~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg , data = mydata_2018_men, fami
summary(probit_men)
```

```
##
## Call:
## glm(formula = fworked ~ HLEV2G + agegp + marstp + cfsize + cfcomp +
##     yrimmg, family = binomial(link = "probit"), data = mydata_2018_men)
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.750049   0.465840   1.610 0.107376
## HLEV2G2      0.374924   0.187237   2.002 0.045242 *
## HLEV2G3      0.535622   0.190833   2.807 0.005004 **
## HLEV2G4      0.515276   0.177779   2.898 0.003751 **
## agegp06      0.500319   0.297747   1.680 0.092889 .
## agegp07     -0.133975   0.301721  -0.444 0.657017
## agegp08      0.270407   0.309313   0.874 0.382001
## agegp09      0.089626   0.319895   0.280 0.779344
## agegp10     -0.194548   0.291184  -0.668 0.504052
## agegp11     -0.201233   0.304020  -0.662 0.508032
## agegp12     -0.387180   0.312390  -1.239 0.215193
## agegp13     -0.880420   0.317684  -2.771 0.005582 **
## marstp02     0.079142   0.289988   0.273 0.784920
## marstp03    -0.719953   0.306312  -2.350 0.018754 *
## marstp04    -0.583752   0.237604  -2.457 0.014017 *
## cfsize       0.064946   0.073405   0.885 0.376284
## cfcomp02    -0.478660   0.292693  -1.635 0.101973
## cfcomp03    -0.090963   0.275707  -0.330 0.741457
## cfcomp04    -0.073786   0.332353  -0.222 0.824305
## cfcomp05    -0.303736   0.300400  -1.011 0.311966
## cfcomp06    -0.249113   0.441710  -0.564 0.572771
## cfcomp07     0.007126   0.372233   0.019 0.984726
## cfcomp08     0.714369   0.475891   1.501 0.133324
## cfcomp09    -0.475342   0.576892  -0.824 0.409956
## yrimmg2      0.333559   0.138058   2.416 0.015689 *
## yrimmg3      0.510205   0.158841   3.212 0.001318 **
## yrimmg4      0.804065   0.207606   3.873 0.000107 ***
## yrimmg5      0.379033   0.186652   2.031 0.042286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 937.04  on 1383  degrees of freedom
## Residual deviance: 828.47  on 1356  degrees of freedom
## AIC: 884.47
##
## Number of Fisher Scoring iterations: 6
```

```r
# Compute the inverse Mills ratio
IMR_MEN <- invMillsRatio(probit_men)

# Add the inverse Mills ratio to the dataset
mydata_2018_men$IMR_MEN <- IMR_MEN

summary(IMR_MEN)
```

```
##       IMR1               delta1             IMR0            delta0
##  Min.   :0.007838   Min.   :0.02204   Min.   :0.481   Min.   :-0.03585
##  1st Qu.:0.098180   1st Qu.:0.17669   1st Qu.:1.590   1st Qu.: 4.24388
##  Median :0.151313   Median :0.24179   Median :1.893   Median : 6.32398
##  Mean   :0.197720   Mean   :0.26713   Mean   :1.854   Mean   : 6.35946
##  3rd Qu.:0.258894   3rd Qu.:0.34665   3rd Qu.:2.112   3rd Qu.: 8.05171
```

```
##  Max.   :1.181965   Max.   :0.74040   Max.   :3.102   Max.   :18.32094
```

```r
# Run the Probit model for women
probit_women <- glm(fworked~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg, data = mydata_2018_women, fa
summary(probit_women)
```

```
##
## Call:
## glm(formula = fworked ~ HLEV2G + agegp + marstp + cfsize + cfcomp +
##     yrimmg, family = binomial(link = "probit"), data = mydata_2018_women)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.06520    0.38716  -0.168 0.866259
## HLEV2G2      0.48171    0.16108   2.990 0.002785 **
## HLEV2G3      0.82302    0.16154   5.095 3.49e-07 ***
## HLEV2G4      0.93941    0.15727   5.973 2.32e-09 ***
## agegp06     -0.19815    0.24530  -0.808 0.419223
## agegp07     -0.20712    0.23279  -0.890 0.373621
## agegp08     -0.33250    0.24144  -1.377 0.168466
## agegp09     -0.10586    0.24276  -0.436 0.662777
## agegp10     -0.04843    0.24735  -0.196 0.844761
## agegp11     -0.10927    0.24868  -0.439 0.660364
## agegp12     -0.45561    0.26013  -1.751 0.079862 .
## agegp13     -0.87523    0.26346  -3.322 0.000893 ***
## marstp02     0.52613    0.25039   2.101 0.035618 *
## marstp03     0.17536    0.19513   0.899 0.368817
## marstp04     0.19581    0.18782   1.043 0.297148
## cfsize      -0.06797    0.05110  -1.330 0.183495
## cfcomp02    -0.14064    0.26243  -0.536 0.592007
## cfcomp03     0.10868    0.22744   0.478 0.632766
## cfcomp04     0.08310    0.25857   0.321 0.747921
## cfcomp05    -0.08304    0.25782  -0.322 0.747380
## cfcomp06    -0.25777    0.21596  -1.194 0.232623
## cfcomp07     0.18766    0.24762   0.758 0.448541
## yrimmg2      0.42177    0.10451   4.035 5.45e-05 ***
## yrimmg3      0.67187    0.11946   5.624 1.86e-08 ***
## yrimmg4      0.45798    0.14563   3.145 0.001662 **
## yrimmg5      0.40629    0.14414   2.819 0.004822 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1810.0  on 1570  degrees of freedom
## Residual deviance: 1651.4  on 1545  degrees of freedom
## AIC: 1703.4
##
## Number of Fisher Scoring iterations: 4
```

```r
# Compute the inverse Mills ratio
IMR_WOMEN <- invMillsRatio(probit_women)
```

```r
# Add the inverse Mills ratio to the dataset
mydata_2018_women$IMR_WOMEN <- IMR_WOMEN

summary(IMR_WOMEN)
```

```
##       IMR1              delta1             IMR0           delta0
##  Min.   :0.06294   Min.   :0.1258   Min.   :0.211   Min.   :-0.2145
##  1st Qu.:0.28402   1st Qu.:0.3674   1st Qu.:1.084   1st Qu.: 1.6306
##  Median :0.40303   Median :0.4524   Median :1.305   Median : 2.6434
##  Mean   :0.44349   Mean   :0.4549   Mean   :1.298   Mean   : 2.8278
##  3rd Qu.:0.55065   3rd Qu.:0.5350   3rd Qu.:1.533   3rd Qu.: 3.8977
##  Max.   :1.71013   Max.   :0.8255   Max.   :2.316   Max.   : 9.8492
```

```r
# Run the Probit model with dummy variable for sex
probit_gender<- glm(fworked~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg+sex, data = mydata_2018, fam
summary(probit_gender)
```

```
##
## Call:
## glm(formula = fworked ~ HLEV2G + agegp + marstp + cfsize + cfcomp +
##     yrimmg + sex, family = binomial(link = "probit"), data = mydata_2018)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.511410   0.287855   1.777 0.075630 .
## HLEV2G2      0.403183   0.117732   3.425 0.000616 ***
## HLEV2G3      0.689414   0.118973   5.795 6.84e-09 ***
## HLEV2G4      0.754385   0.114051   6.614 3.73e-11 ***
## agegp06      0.068815   0.182464   0.377 0.706067
## agegp07     -0.162826   0.176562  -0.922 0.356424
## agegp08     -0.128190   0.180167  -0.712 0.476771
## agegp09     -0.006278   0.183391  -0.034 0.972693
## agegp10     -0.045567   0.182608  -0.250 0.802946
## agegp11     -0.071139   0.186026  -0.382 0.702155
## agegp12     -0.344103   0.193430  -1.779 0.075247 .
## agegp13     -0.779003   0.196317  -3.968 7.25e-05 ***
## marstp02     0.337166   0.187840   1.795 0.072659 .
## marstp03     0.043222   0.158038   0.273 0.784477
## marstp04    -0.061969   0.141771  -0.437 0.662036
## cfsize      -0.033172   0.040641  -0.816 0.414384
## cfcomp02    -0.172338   0.193136  -0.892 0.372226
## cfcomp03     0.170685   0.168596   1.012 0.311351
## cfcomp04     0.197273   0.195133   1.011 0.312034
## cfcomp05    -0.033460   0.189424  -0.177 0.859788
## cfcomp06    -0.079862   0.176389  -0.453 0.650720
## cfcomp07     0.266963   0.200896   1.329 0.183893
## cfcomp08     0.407532   0.435586   0.936 0.349482
## cfcomp09    -0.580497   0.555236  -1.045 0.295793
## yrimmg2      0.363749   0.081447   4.466 7.97e-06 ***
## yrimmg3      0.592414   0.093614   6.328 2.48e-10 ***
## yrimmg4      0.523567   0.114743   4.563 5.04e-06 ***
## yrimmg5      0.356362   0.112652   3.163 0.001559 **
## sex2        -0.642746   0.060954 -10.545  < 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: 2869.4  on 2954  degrees of freedom
## Residual deviance: 2542.3  on 2926  degrees of freedom
## AIC: 2600.3
##
## Number of Fisher Scoring iterations: 5
```

```
# Compute the inverse Mills ratio
IMR <- invMillsRatio(probit_gender)

# Add the inverse Mills ratio to the dataset
mydata_2018$IMR <- IMR
```

## Regression Analysis on Human Capital Earning with IMR

```
# Run the Regression model for men without IMR
Reg_men <- lm(log10(cfearng+0.01)~  HLEV2G + expr  , data = mydata_2018_men )
summary(Reg_men)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr, data = mydata_2018_men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7728  0.0690  0.3675  0.5693  1.5310
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.265388   0.191130  22.317  < 2e-16 ***
## HLEV2G2      0.259497   0.181519   1.430  0.15306
## HLEV2G3      0.512748   0.178650   2.870  0.00417 **
## HLEV2G4      0.538708   0.170709   3.156  0.00164 **
## expr        -0.005354   0.003327  -1.610  0.10773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.49 on 1372 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.01454,    Adjusted R-squared:  0.01166
## F-statistic: 5.059 on 4 and 1372 DF,  p-value: 0.0004765
```

```
# Run the Regression model for men with IMR
Reg_men_IMR <- lm(log10(cfearng+0.01)~  HLEV2G + expr+ IMR_MEN$IMR1  , data = mydata_2018_men )
summary(Reg_men_IMR)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr + IMR_MEN$IMR1,
##     data = mydata_2018_men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1091 -0.0336  0.2585  0.5508  2.2723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.224782   0.207145  25.223   <2e-16 ***
## HLEV2G2      -0.121849   0.179085  -0.680    0.496
## HLEV2G3      -0.125990   0.183446  -0.687    0.492
## HLEV2G4      -0.080026   0.175576  -0.456    0.649
## expr          0.001252   0.003274   0.382    0.702
## IMR_MEN$IMR1 -2.999831   0.295191 -10.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.437 on 1371 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.08357,    Adjusted R-squared:  0.08022
## F-statistic:    25 on 5 and 1371 DF,  p-value: < 2.2e-16
```

```r
# Run the Regression model for women without IMR
Reg_women <- lm(log10(cfearng+0.01)~  HLEV2G + expr , data = mydata_2018_women)
summary(Reg_women)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr, data = mydata_2018_women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7923  0.1137  0.4867  0.7715  2.1426
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.576172   0.239736  14.917  < 2e-16 ***
## HLEV2G2      0.721241   0.219981   3.279  0.00107 **
## HLEV2G3      0.946706   0.217655   4.350 1.45e-05 ***
## HLEV2G4      1.216161   0.212845   5.714 1.32e-08 ***
## expr        -0.008303   0.003807  -2.181  0.02932 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.795 on 1560 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.03635,    Adjusted R-squared:  0.03388
## F-statistic: 14.71 on 4 and 1560 DF,  p-value: 8.418e-12
```

```r
# Run the Regression model for women with IMR
Reg_women_IMR <- lm(log10(cfearng+0.01)~  HLEV2G + expr + IMR_WOMEN$IMR1 , data = mydata_2018_women)
summary(Reg_women_IMR)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr + IMR_WOMEN$IMR1,
##     data = mydata_2018_women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1637  0.0521  0.4167  0.7599  2.9197
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.288680   0.330103  16.021  < 2e-16 ***
## HLEV2G2         0.003711   0.236966   0.016    0.988
## HLEV2G3        -0.147318   0.259983  -0.567    0.571
## HLEV2G4         0.091066   0.258544   0.352    0.725
## expr           -0.001224   0.003863  -0.317    0.751
## IMR_WOMEN$IMR1 -2.094974   0.282731  -7.410 2.06e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.765 on 1559 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.06913,    Adjusted R-squared:  0.06615
## F-statistic: 23.16 on 5 and 1559 DF,  p-value: < 2.2e-16
```

```r
# Run the Regression model with dummy variable for sex without IMR
Reg_gender<- lm(log10(cfearng+0.01)~  HLEV2G +expr +sex, data = mydata_2018)
summary(Reg_gender)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr + sex, data = mydata_2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8891  0.0825  0.4228  0.6835  2.0989
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.060371   0.156455  25.952  < 2e-16 ***
## HLEV2G2      0.478700   0.143293   3.341 0.000846 ***
## HLEV2G3      0.717830   0.141432   5.075 4.11e-07 ***
## HLEV2G4      0.872983   0.136737   6.384 1.99e-10 ***
## expr        -0.007377   0.002548  -2.895 0.003824 **
## sex2        -0.235043   0.061381  -3.829 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.661 on 2936 degrees of freedom
```

```
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.02983,    Adjusted R-squared:  0.02818
## F-statistic: 18.05 on 5 and 2936 DF,  p-value: < 2.2e-16
```

```r
# Run the Regression model with dummy variable for sex with IMR
Reg_gender_IMR<- lm(log10(cfearng+0.01)~  HLEV2G + expr +sex+ IMR$IMR1, data = mydata_2018)
summary(Reg_gender_IMR)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr + sex + IMR$IMR1,
##     data = mydata_2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0972  0.0241  0.3321  0.6384  3.3927
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.422046   0.184264  29.425  < 2e-16 ***
## HLEV2G2     -0.225807   0.149339  -1.512   0.1306
## HLEV2G3     -0.406144   0.162100  -2.506   0.0123 *
## HLEV2G4     -0.266589   0.158896  -1.678   0.0935 .
## expr         0.001741   0.002574   0.677   0.4988
## sex2         0.524353   0.083220   6.301  3.4e-10 ***
## IMR$IMR1    -3.110867   0.237617 -13.092  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.614 on 2935 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.08336,    Adjusted R-squared:  0.08149
## F-statistic: 44.48 on 6 and 2935 DF,  p-value: < 2.2e-16
```

# Heckman two-step method for Men to correct the non-random selected samples

In this part, I compute selection-corrected earnings estimates using Heckman's two-step method. To do so, I used two methods including Heckit function in R and semi-parametric method. The results of both are as following:

```r
library(sampleSelection)

#selection model for men

heckit_model_men <- heckit( fworked ~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg,
   log(cfearng+0.01) ~  HLEV2G +expr,data=mydata_2018_men, method = '2step' )

summary(heckit_model_men)
```

```
## --------------------------------------------
```

```
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 1377 observations (147 censored and 1230 observed)
## 36 free parameters (df = 1342)
## Probit selection equation:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.74890    0.46928   1.596 0.110760
## HLEV2G2      0.37611    0.18938   1.986 0.047236 *
## HLEV2G3      0.53538    0.19337   2.769 0.005706 **
## HLEV2G4      0.51376    0.18022   2.851 0.004427 **
## agegp06      0.49454    0.29668   1.667 0.095768 .
## agegp07     -0.13496    0.30219  -0.447 0.655218
## agegp08      0.26617    0.30880   0.862 0.388867
## agegp09      0.07657    0.31957   0.240 0.810674
## agegp10     -0.19707    0.29106  -0.677 0.498482
## agegp11     -0.20179    0.30293  -0.666 0.505441
## agegp12     -0.38997    0.31137  -1.252 0.210640
## agegp13     -0.88208    0.31804  -2.773 0.005623 **
## marstp02     0.07917    0.29102   0.272 0.785634
## marstp03    -0.71391    0.31166  -2.291 0.022138 *
## marstp04    -0.58488    0.24004  -2.437 0.014957 *
## cfsize       0.06441    0.07418   0.868 0.385404
## cfcomp02    -0.46900    0.29749  -1.577 0.115144
## cfcomp03    -0.08194    0.27915  -0.294 0.769147
## cfcomp04    -0.06533    0.33606  -0.194 0.845892
## cfcomp05    -0.29340    0.30311  -0.968 0.333236
## cfcomp06    -0.24139    0.44383  -0.544 0.586610
## cfcomp07     0.01646    0.37773   0.044 0.965251
## cfcomp08     0.71881    0.48896   1.470 0.141773
## cfcomp09    -0.46759    0.59763  -0.782 0.434109
## yrimmg2      0.32736    0.13823   2.368 0.018013 *
## yrimmg3      0.50599    0.15926   3.177 0.001521 **
## yrimmg4      0.79959    0.20859   3.833 0.000132 ***
## yrimmg5      0.37606    0.18803   2.000 0.045704 *
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.181613   0.328261  34.063   <2e-16 ***
## HLEV2G2     -0.011268   0.286694  -0.039   0.9687
## HLEV2G3     -0.008215   0.290931  -0.028   0.9775
## HLEV2G4      0.258832   0.278697   0.929   0.3532
## expr         0.008806   0.005147   1.711   0.0873 .
## Multiple R-Squared:0.0288,   Adjusted R-Squared:0.0248
##    Error terms:
##                Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  -2.1831     0.4854  -4.497 7.47e-06 ***
## sigma           2.1890         NA      NA       NA
## rho            -0.9973         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```r
#---------------------------------------------
fit_prob_f <- predict(probit_men)
```

```r
Reg_heck_men <- lm(log10(cfearng+0.01) ~  HLEV2G +expr+fit_prob_f+fit_prob_f^2+fit_prob_f^3+fit_prob_f^4

summary(Reg_heck_men)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr + fit_prob_f +
##     fit_prob_f^2 + fit_prob_f^3 + fit_prob_f^4, data = mydata_2018_men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6012 -0.0189  0.3106  0.5964  1.8761
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.4766335  0.2091885  16.620   <2e-16 ***
## HLEV2G2     -0.0014738  0.1799047  -0.008    0.993
## HLEV2G3      0.0202823  0.1841048   0.110    0.912
## HLEV2G4      0.0928108  0.1749927   0.530    0.596
## expr         0.0001807  0.0033141   0.055    0.957
## fit_prob_f   0.7486263  0.0898615   8.331   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.454 on 1371 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.06202,    Adjusted R-squared:  0.0586
## F-statistic: 18.13 on 5 and 1371 DF,  p-value: < 2.2e-16
```

# Heckman two-step method for Women to correct the non-random selected samples

```r
#selection model for women

heckit_model_women <- heckit( fworked ~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg,
   log(cfearng+0.01) ~  HLEV2G +expr,data=mydata_2018_women, method = '2step' )

summary(heckit_model_women)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 1569 observations (413 censored and 1156 observed)
## 34 free parameters (df = 1536)
## Probit selection equation:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06232    0.38132  -0.163 0.870208
```

```
## HLEV2G2      0.48097    0.16035   3.000 0.002747 **
## HLEV2G3      0.81882    0.16083   5.091 3.99e-07 ***
## HLEV2G4      0.93834    0.15641   5.999 2.47e-09 ***
## agegp06     -0.19763    0.24265  -0.814 0.415517
## agegp07     -0.20614    0.22896  -0.900 0.368089
## agegp08     -0.33193    0.23782  -1.396 0.163012
## agegp09     -0.10500    0.23791  -0.441 0.659038
## agegp10     -0.04715    0.24427  -0.193 0.846980
## agegp11     -0.11051    0.24419  -0.453 0.650925
## agegp12     -0.45856    0.25679  -1.786 0.074335 .
## agegp13     -0.87344    0.25976  -3.363 0.000791 ***
## marstp02     0.52770    0.25082   2.104 0.035551 *
## marstp03     0.17669    0.19480   0.907 0.364527
## marstp04     0.19600    0.18609   1.053 0.292406
## cfsize      -0.06877    0.05071  -1.356 0.175244
## cfcomp02    -0.14192    0.26519  -0.535 0.592617
## cfcomp03     0.10694    0.22572   0.474 0.635730
## cfcomp04     0.08430    0.25623   0.329 0.742212
## cfcomp05    -0.08035    0.25474  -0.315 0.752481
## cfcomp06    -0.25707    0.21572  -1.192 0.233565
## cfcomp07     0.18889    0.24722   0.764 0.444966
## yrimmg2      0.42209    0.10513   4.015 6.23e-05 ***
## yrimmg3      0.67063    0.11980   5.598 2.56e-08 ***
## yrimmg4      0.45832    0.14743   3.109 0.001913 **
## yrimmg5      0.40389    0.14507   2.784 0.005433 **
## Outcome equation:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.013992   0.444888  27.005   <2e-16 ***
## HLEV2G2     -0.530468   0.340373  -1.558    0.119
## HLEV2G3     -0.449063   0.362680  -1.238    0.216
## HLEV2G4     -0.336946   0.361137  -0.933    0.351
## expr         0.002058   0.004722   0.436    0.663
## Multiple R-Squared:0.0221,   Adjusted R-Squared:0.0179
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  -1.3057     0.3645  -3.582 0.000352 ***
## sigma           1.9021         NA      NA       NA
## rho            -0.6864         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```r
fit_prob_f <- predict(probit_women)


Reg_heck_women <- lm(log10(cfearng+0.01) ~  HLEV2G +expr+  fit_prob_f+fit_prob_f^2+fit_prob_f^3+fit_prol

summary(Reg_heck_women)
```

```
##
## Call:
## lm(formula = log10(cfearng + 0.01) ~ HLEV2G + expr + fit_prob_f +
##     fit_prob_f^2 + fit_prob_f^3 + fit_prob_f^4, data = mydata_2018_women)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4063  0.0282  0.4285  0.7765  2.3650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.542427   0.236192  14.998  < 2e-16 ***
## HLEV2G2      0.200862   0.229123   0.877    0.381
## HLEV2G3      0.079740   0.247701   0.322    0.748
## HLEV2G4      0.312607   0.246319   1.269    0.205
## expr        -0.002301   0.003847  -0.598    0.550
## fit_prob_f   0.926747   0.132616   6.988 4.11e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.768 on 1559 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.06562,    Adjusted R-squared:  0.06262
## F-statistic:  21.9 on 5 and 1559 DF,  p-value: < 2.2e-16
```

## Likelihood Function for Tobit model-Censored random variable

$$L = \prod_{i=1}^{N} [1 - F_{SN}(-x_i\beta - \rho\frac{1}{\sigma_1}(y_i - x_i\gamma)) * \frac{1}{\sigma} f_{SN}(\frac{y_i - x_i\gamma}{\sigma})]^{d_i}[F(-x_i\beta]^{1-d_i}$$

## Optimize the Likelihood Function

```
Reg_men <- lm(cfearng ~  HLEV2G + expr , data = mydata_2018_men )

probit_men <- glm(fworked~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg , data = mydata_2018_men, famil

x1  <-  model.matrix(Reg_men)
x2<-  model.matrix(probit_men)

y <- mydata_2018_men$cfearng


beta <- coef(Reg_men)
gama <- coef(probit_men)


init1 <- c(beta, log_sigma = log(summary(Reg_men)$sigma))
init2 <- gama
init <- c(init1, init2)
init <- rep(0, length(init))

tobit_ll <- function(par,y, x1,x2,beta_length) {
```

```r
  sigma <- exp(par[beta_length + 1])
  beta <- par[1:beta_length]
  gama <- par[(beta_length + 2):length(par)]
  ro <- 0.5



 # create indicator depending on chosen limit

    indicator = ifelse(mydata_2018_men$fworked==1,1,0)



  lp2 <- x2 %*% gama
  lp <- x1%*% beta

 ll = sum(indicator * log(1-dnorm(-lp-ro*(1/sigma)*(y-lp2))*(1/sigma)*pnorm(y-lp2)/sigma ) +
    sum((1-indicator) * log(pnorm(-lp))))

 return(-ll)

}

fit_tobit = optim(
  par = init,
  tobit_ll,
  y  =y,
  x1  = x1,
  x2=x2,
  beta_length = length(beta),
  method  = 'BFGS'

)
fit_tobit
```

```
## $par
##  [1] -1.038899e+01 -2.897610e+00 -2.049529e+00 -3.957712e+00 -2.913340e+02
##  [6]  4.465207e-04  1.781357e-04  3.817194e-05  5.089592e-05  7.634388e-05
## [11]  1.272398e-05  2.544796e-05  1.272398e-05  2.544796e-05  1.272398e-05
## [16]  2.544796e-05  2.544796e-05  3.817194e-05  0.000000e+00  0.000000e+00
## [21]  5.089592e-05  4.835087e-04  0.000000e+00  2.544796e-05  7.634388e-05
## [26]  0.000000e+00  0.000000e+00  0.000000e+00  1.272398e-05  0.000000e+00
## [31]  3.817194e-05  3.817194e-05  2.544796e-05  2.544796e-05
##
## $value
## [1] 0
##
## $counts
## function gradient
##        8        2
##
## $convergence
## [1] 0
```

```
##
## $message
## NULL
```

```
devtools::source_url("https://raw.githubusercontent.com/MatthieuStigler/Misconometrics/master/Gelbach_d
```

```
## i SHA-1 hash of file is "409ebbbc3b2320c7e2019ad8056532467a0f90c2"
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

## Decomposition of Earnings Differences using Gelbach method

```
Reg_men<- lm(log10(cfearng+0.01)~  HLEV2G + expr , data=mydata_2018_men)
coef_men <- coef(Reg_men)

reg_women <- lm(log10(cfearng+0.01) ~  HLEV2G + expr, data = mydata_2018_women)
coef_women <- coef(Reg_women)


#difference in mean log earnings
#diff_mean <- mean(predict(reg_men) - predict(reg_women))
#diff_mean

#Decomposition part


meanexp_men <-mean(as.numeric(mydata_2018$expr[mydata_2018$sex == 1]))
meanedu2_men <- mean(as.numeric(mydata_2018_men$HLEV2G=="2"))
meanedu3_men<- mean(as.numeric(mydata_2018_men$HLEV2G=="3"))
meanedu4_men<- mean(as.numeric(mydata_2018_men$HLEV2G=="4"))

meanexp_women <-mean(as.numeric(mydata_2018$expr[mydata_2018$sex == 2]))
meanedu2_women <- mean(as.numeric(mydata_2018_women$HLEV2G=="2"))
meanedu3_women<- mean(as.numeric(mydata_2018_women$HLEV2G=="3"))
meanedu4_women<- mean(as.numeric(mydata_2018_women$HLEV2G=="4"))

#Compute the fitted log earnings for men and women

fitted_men <-  (meanexp_men * coef_men['expr'])+(meanedu2_men*coef_men['HLEV2G2'])+
  (meanedu3_men*coef_men['HLEV2G3'])+(meanedu4_men*coef_men['HLEV2G4'])



fitted_women <-  (meanexp_women *coef_women["expr"])+(meanedu2_women*coef_women["HLEV2G2"])+
```

```
  (meanedu3_women*coef_women["HLEV2G3"])+(meanedu4_women*coef_women["HLEV2G4"])

#difference in mean log earnings
Dif_earning <- fitted_women-fitted_men
Dif_earning
```

```
##      expr
## 0.4562888
```

```
#unexplained

unexplained <- (meanexp_women*(coef_women["expr"]-coef_men["expr"]))+
 (meanedu2_women*(coef_women["HLEV2G2"]-coef_men["HLEV2G2"]))+
   (meanedu3_women*(coef_women["HLEV2G3"]-coef_men["HLEV2G3"]))+
   (meanedu4_women*(coef_women["HLEV2G4"]-coef_men["HLEV2G4"]))

unexplained
```

```
##     expr
## 0.454952
```

```
explained <- ((meanexp_women-meanexp_men)*coef_men["expr"])+((meanedu2_women-meanedu2_men)*coef_men["HL
+((meanedu3_women-meanedu3_men)*coef_men["HLEV2G3"])+(((meanedu4_women-meanedu4_men)*coef_men["HLEV2G4"]
```

```
##      HLEV2G3
## -0.002348203
```

```
explained
```

```
##        expr
## 0.003685008
```

Based on the results of the decomposition analysis, we can see that the difference in log earnings between men and women is approximately 0.456. However, the majority of this difference, around 0.455, remains unexplained, while the portion of the difference that can be attributed to observable factors is very small. This suggests that there is likely discrimination in employment practices between men and women.

## Counterfactual Mean Earnings for Women Using Male Coefficients

```
coeff_men <- coef(Reg_heck_men) # coefficients for men

coeff_women <- coef(Reg_heck_women) # coefficients for women

#mean values of education and experience separately for men and women

meanexp_men <-mean(as.numeric(mydata_2018$expr[mydata_2018$sex == 1]))
meanedu2_men <- mean(as.numeric(mydata_2018_men$HLEV2G=="2"))
meanedu3_men<- mean(as.numeric(mydata_2018_men$HLEV2G=="3"))
```

```
meanedu4_men<- mean(as.numeric(mydata_2018_men$HLEV2G=="4"))

meanexp_women <-mean(as.numeric(mydata_2018$expr[mydata_2018$sex == 2]))
meanedu2_women <- mean(as.numeric(mydata_2018_women$HLEV2G=="2"))
meanedu3_women<- mean(as.numeric(mydata_2018_women$HLEV2G=="3"))
meanedu4_women<- mean(as.numeric(mydata_2018_women$HLEV2G=="4"))

#Compute the fitted log earnings for men and women

fitted_men <-  (meanexp_men * coeff_men['expr'])+(meanedu2_men*coeff_men['HLEV2G2'])+
  (meanedu3_men*coeff_men['HLEV2G3'])+(meanedu4_men*coeff_men['HLEV2G4'])



fitted_women <-  (meanexp_women *coeff_women["expr"])+(meanedu2_women*coeff_women["HLEV2G2"])+
  (meanedu3_women*coeff_women["HLEV2G3"])+(meanedu4_women*coeff_women["HLEV2G4"])

# log earning difference between men and women
diff_log <- fitted_women-fitted_men
diff_log
```

```
##       expr
## 0.09904954
```

```
# Counterfactual mean log earnings


counter_women <- (meanexp_women *coeff_men["expr"])+(meanedu2_women*coeff_men["HLEV2G2"])+
  (meanedu3_women["HLEV2G3"]*coeff_men["HLEV2G3"])+(meanedu4_women*coeff_men["HLEV2G4"])


counter_men <- (meanexp_men * coeff_women["expr"])+(meanedu2_men*coeff_women["HLEV2G2"])+
  (meanedu3_men*coeff_women[""])+(meanedu4_men*coeff_women["HLEV2G4"])



#Explained and Unexplained

unexplained <- (meanexp_women*(coeff_women["expr"]-coeff_men["expr"]))+
 (meanedu2_women*(coeff_women["HLEV2G2"]-coeff_men["HLEV2G2"]))+
   (meanedu3_women*(coeff_women["HLEV2G3"]-coeff_men["HLEV2G3"]))+
   (meanedu4_women*(coeff_women["HLEV2G4"]-coeff_men["HLEV2G4"]))

unexplained
```

```
##      expr
## 0.1001341
```

```
explained <- ((meanexp_women-meanexp_men)*coeff_men["expr"])+((meanedu2_women-meanedu2_men)*coeff_men["
+((meanedu3_women-meanedu3_men)*coeff_men["HLEV2G3"])+(((meanedu4_women-meanedu4_men)*coeff_men["HLEV2G
```

```
##     HLEV2G3
## -0.00106284
```

```
explained
```

```
##          expr
## -2.170876e-05
```

The application of the Heckman two-step method reveals that the fitted log earning values for men and women differ by approximately 0.1. This difference is entirely attributed to the unexplained part, which represents discrimination in employment and is also referred to as the structural component. It reflects the disparities in the wage structure and payment practices. Despite the lower log earning difference obtained through the Heckman two-step method, it is consistent with the previous results in indicating the presence of earning discrimination against women.

# Decomposition of Earnings Differences using Blinder-Oaxaca including IMR

```
library(oaxaca)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). oaxaca: Blinder-Oaxaca Decomposition in R.
```

```
##  R package version 0.1.5. https://CRAN.R-project.org/package=oaxaca
```

```
probit_gender<- glm(fworked~  HLEV2G + agegp + marstp+cfsize+cfcomp+yrimmg+sex, data = mydata_2018, fami
summary(probit_gender)
```

```
##
## Call:
## glm(formula = fworked ~ HLEV2G + agegp + marstp + cfsize + cfcomp +
##     yrimmg + sex, family = binomial(link = "probit"), data = mydata_2018)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.511410   0.287855   1.777 0.075630 .
## HLEV2G2      0.403183   0.117732   3.425 0.000616 ***
## HLEV2G3      0.689414   0.118973   5.795 6.84e-09 ***
## HLEV2G4      0.754385   0.114051   6.614 3.73e-11 ***
## agegp06      0.068815   0.182464   0.377 0.706067
## agegp07     -0.162826   0.176562  -0.922 0.356424
## agegp08     -0.128190   0.180167  -0.712 0.476771
## agegp09     -0.006278   0.183391  -0.034 0.972693
## agegp10     -0.045567   0.182608  -0.250 0.802946
## agegp11     -0.071139   0.186026  -0.382 0.702155
## agegp12     -0.344103   0.193430  -1.779 0.075247 .
## agegp13     -0.779003   0.196317  -3.968 7.25e-05 ***
## marstp02     0.337166   0.187840   1.795 0.072659 .
## marstp03     0.043222   0.158038   0.273 0.784477
```

```
## marstp04    -0.061969   0.141771  -0.437 0.662036    
## cfsize      -0.033172   0.040641  -0.816 0.414384    
## cfcomp02    -0.172338   0.193136  -0.892 0.372226    
## cfcomp03     0.170685   0.168596   1.012 0.311351    
## cfcomp04     0.197273   0.195133   1.011 0.312034    
## cfcomp05    -0.033460   0.189424  -0.177 0.859788    
## cfcomp06    -0.079862   0.176389  -0.453 0.650720    
## cfcomp07     0.266963   0.200896   1.329 0.183893    
## cfcomp08     0.407532   0.435586   0.936 0.349482    
## cfcomp09    -0.580497   0.555236  -1.045 0.295793    
## yrimmg2      0.363749   0.081447   4.466 7.97e-06 ***
## yrimmg3      0.592414   0.093614   6.328 2.48e-10 ***
## yrimmg4      0.523567   0.114743   4.563 5.04e-06 ***
## yrimmg5      0.356362   0.112652   3.163 0.001559 ** 
## sex2        -0.642746   0.060954 -10.545  < 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: 2869.4  on 2954  degrees of freedom
## Residual deviance: 2542.3  on 2926  degrees of freedom
## AIC: 2600.3
## 
## Number of Fisher Scoring iterations: 5
```

```r
# Compute the inverse Mills ratio
IMR <- invMillsRatio(probit_gender)

# Add the inverse Mills ratio to the dataset
mydata_2018$IMR <- IMR

mydata_2018$dummay_female <- ifelse(mydata_2018$sex==2,1,0)
dec <- oaxaca(log10(cfearng+0.01)~  HLEV2G + expr+IMR$IMR1|dummay_female, data=mydata_2018)
```

```
## oaxaca: oaxaca() performing analysis. Please wait.
```

```
## 
## Bootstrapping standard errors:
```

```
## 1 / 100 (1%)
```

```
## 10 / 100 (10%)
```

```
## 20 / 100 (20%)
```

```
## 30 / 100 (30%)
```

```
## 40 / 100 (40%)
```

```
## 50 / 100 (50%)
```

```
## 60 / 100 (60%)
```

```
## 70 / 100 (70%)
```

```
## 80 / 100 (80%)
```

```
## 90 / 100 (90%)
```

```
## 100 / 100 (100%)
```

```r
dec$y
```

```
## $y.A
## [1] 4.567068
##
## $y.B
## [1] 4.334669
##
## $y.diff
## [1] 0.2323991
```

```r
dec$twofold$overall
```

```
##      group.weight coef(explained) se(explained) coef(unexplained)
## [1,]    0.0000000       0.8026510    0.08193165        -0.5702519
## [2,]    1.0000000       0.6289277    0.12800669        -0.3965286
## [3,]    0.5000000       0.7157894    0.07811290        -0.4833903
## [4,]    0.4680489       0.7213400    0.08020556        -0.4889409
## [5,]   -1.0000000       0.5017957    0.05331636        -0.2693966
## [6,]   -2.0000000       0.7567522    0.07029817        -0.5243531
##      se(unexplained) coef(unexplained A) se(unexplained A) coef(unexplained B)
## [1,]      0.08247382       -5.702519e-01      8.247382e-02           0.0000000
## [2,]      0.15403135        0.000000e+00      0.000000e+00          -0.3965286
## [3,]      0.09907657       -2.851260e-01      4.123691e-02          -0.1982643
## [4,]      0.10187807       -3.033461e-01      3.860179e-02          -0.1855948
## [5,]      0.04115654       -1.433058e-01      2.216695e-02          -0.1260908
## [6,]      0.07880649        2.109424e-15      1.687319e-14          -0.5243531
##      se(unexplained B)
## [1,]        0.00000000
## [2,]        0.15403135
## [3,]        0.07701567
## [4,]        0.08193714
## [5,]        0.01929427
## [6,]        0.07880649
```

```r
dec$threefold$overall
```

```
##   coef(endowments)     se(endowments) coef(coefficients)   se(coefficients)
##         0.80265103         0.08193165        -0.39652860         0.15403135
##  coef(interaction)    se(interaction)
##        -0.17372334         0.14761609
```

I used the Blinder-Oaxaca decomposition to analyze the difference between the log earnings of men and women including IMR term. The results indicate that there is a difference of 0.23 between the log earnings of women and men, which can be broken down into explained and unexplained components based on all the variables used in the analysis. The decomposition reveals that out of the total difference of 0.23, around 0.8 can be attributed to differences in endowments (such as age, experience, and education) between the two groups. A difference of -0.39 can be attributed to differences in the coefficients used in the analysis, while the remaining -0.17 can be explained by the interaction between the two groups. The results of both analyses indicate that there is a difference between the log earning values, and the proportion of explained and unexplained components differs between the two analyses. However, both analyses confirm that the unexplained component accounts for the majority of the log earning difference between men and women.