ECON 607 Project 2

Jingda Li

2025-12-16

Project 2

Q1.

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```
#Import Project 2 Dataset into R
Project2_Data <- read_excel("Project 2 Data.xlsx")
```

```
## New names:
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
```

```
dat <- Project2_Data[, colSums(!is.na(Project2_Data)) > 0]

#Construct the necessary indicator (or dummy) variables
dat$Gender <- factor(dat$gender)
dat$Minority <- factor(dat$'minority flag')
dat$PayGrade <- factor(dat$grade)
dat$Rating <- factor(dat$rating)
dat$Location <- factor(dat$location)

#Choose the reference levels
dat$Gender<-relevel(dat$Gender, ref = "m")
dat$PayGrade <- relevel(dat$PayGrade, ref = 1)
dat$Rating <- relevel(dat$Rating, ref = 1)
dat$Location <- relevel(dat$Location, ref = "Canada")

#Print the first 6 rows and last 5 columns in "dat"
head(dat[, (ncol(dat)-4):ncol(dat)], 6)
```

```
## # A tibble: 6 x 5
##   Gender Minority PayGrade Rating Location
##   <fct>  <fct>    <fct>    <fct>  <fct>
## 1 f      0        4        3      Canada
## 2 f      0        4        3      Canada
## 3 f      0        4        3      Canada
## 4 f      0        4        3      Canada
## 5 f      0        4        3      Canada
## 6 f      0        4        3      Canada
```

```
#Fit the linear regression model
model1 <- lm(salary~Gender + Minority + PayGrade +
               Rating + Location,
             data = dat, na.action = na.omit)
summary(model1)
```

```
##
## Call:
## lm(formula = salary ~ Gender + Minority + PayGrade + Rating +
##     Location, data = dat, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27743.9  -5698.6    -51.9   6109.2  22204.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44166.3     1043.5  42.326  < 2e-16 ***
## Genderf      -3096.3      588.2  -5.264 1.63e-07 ***
## Minority1      475.0      622.7   0.763    0.446
## PayGrade3     5631.4     1006.3   5.596 2.64e-08 ***
## PayGrade4    12406.0     1023.4  12.122  < 2e-16 ***
## PayGrade5    28890.8     1082.3  26.693  < 2e-16 ***
## PayGrade6    41718.2     1077.5  38.716  < 2e-16 ***
## PayGrade7    57937.6     1185.0  48.891  < 2e-16 ***
## PayGrade8    83616.0     1351.8  61.854  < 2e-16 ***
## Rating4       2639.2      475.8   5.546 3.49e-08 ***
## LocationUSA   2943.7      473.0   6.223 6.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8394 on 1381 degrees of freedom
##   (25 observations deleted due to missingness)
## Multiple R-squared:  0.8925, Adjusted R-squared:  0.8917
## F-statistic:  1147 on 10 and 1381 DF,  p-value: < 2.2e-16
```

The explanatory variables that are included in this model are:

- Gender
  - Levels in the data set: m, f
  - Dummy variable(s): Genderf

  - where: Genderf $= I[\text{Gender Female}] = \begin{cases} 1, & \text{if Gender} = f \\ 0, & \text{if Gender} = m \end{cases}$

- Minority
  - Levels in the data set: 0, 1
  - Dummy variable(s): Minority1
  - where: Minority1 $= I[\text{Minority 1}] = \begin{cases} 1, & \text{if Minority} = 1 \\ 0, & \text{if Minority} = 0 \end{cases}$

- Pay Grade
  - Levels in the data set: 2, 3, 4, 5, 6, 7, 8
  - Dummy variable(s): PayGrade3, PayGrade4,. . . ,PayGrade8
  - where: PayGradeK $= I[\text{PayGrade K}] = \begin{cases} 1, & \text{if PayGrade} = K \\ 0, & \text{otherwise} \end{cases}, K = 3, 4, ..., 8$

- Rating
  - Levels in the data set: 3, 4
  - Dummy variable(s): Rating4
  - where: Rating4 $= I[\text{Rating 4}] = \begin{cases} 1, & \text{if Rating} = 4 \\ 0, & \text{otherwise} \end{cases}$

- Location
  - Levels in the data set: Canada, USA
  - Dummy variable(s): LocationUSA
  - where: LocationUSA $= I[\text{Location USA}] = \begin{cases} 1, & \text{if Location} = \text{USA} \\ 0, & \text{if Location} = \text{Canada} \end{cases}$

Note: I treated each explanatory variable as a categorical variable with multiple levels. As a result, I created dummy variables for each categorical variable in the model.

Interpretation of coefficients:

- Coefficient on "Genderf": The expected salary of an employee that's female is about \$3096.30 lower than the expected salary of a male employee, controlling for all other explanatory variables.
- Coefficient on "Minority1": The expected salary of an employee who is a member of a minority group is about \$475 higher than the expected salary of an employee who is not in a minority group, controlling for all other explanatory variables.
- Coefficient on "PayGrade3", . . . , "PayGrade8": The expected salary of an employee in Pay Grade K, for K = 3,. . . ,8 is about:
  - \$5631.40
  - \$12406
  - \$28890.80
  - \$41718.20
  - \$57937.60

– $83616

higher, respectively, than the expected salary of an employee in Pay Grade 2, controlling for all other explanatory variables.

- Coefficient on "Rating4": The expected salary of an employee with a performance rating of 4 is about $2639.20 higher than the expected salary of an employee with a performance rating of 3, controlling for all other explanatory variables.
- Coefficient on "LocationUSA": The expected salary of an employee in the USA is about $2943.70 higher than the expected salary of an employee in Canada, controlling for all other explanatory variables.

Statistical Significance:

- The coefficient on "Genderf" is negative and is statistically significant at the 1% level, which indicates that there is very strong evidence that the expected salary of female and male employees differ, controlling for all other explanatory variables, and female employees tend to earn less on average than male employees, controlling for all other explanatory variables. This suggests that there is a material risk of disparate impact discrimination based on gender, looking only at the results of this model.
- The coefficient on "Minority1" is positive and statistically insignificant, which indicates that there is very weak or almost no evidence that the expected salary of minority and non-minority employees differ, controlling for all other explanatory variables. This suggests that there is little to no risk of disparate impact discrimination based on minority status.

Q2.

```
#Creating a new dataset containing a sub-sample with only employees in their first year
dat$hiredate <- as.Date(dat$hiredate)
dat$hire_year <- as.integer(format(dat$hiredate, "%Y"))

dat2 <- subset(dat, year == hire_year)

#Fit the linear regression model for  SalaryAtHire.
model2 <- lm(salary~Gender + Minority + PayGrade +
               Rating + Location,
             data = dat2, na.action = na.omit)
summary(model2)
```

```
##
## Call:
## lm(formula = salary ~ Gender + Minority + PayGrade + Rating +
##     Location, data = dat2, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13456.0  -4462.1   -911.8   3895.4  17871.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39737.91    2708.47  14.672  < 2e-16 ***
## Genderf     -4767.28    1405.19  -3.393 0.000909 ***
```

```
## Minority1    1323.88    1654.27   0.800 0.424953
## PayGrade3    8782.27    2746.17   3.198 0.001724 **
## PayGrade4   15154.54    2705.14   5.602 1.14e-07 ***
## PayGrade5   29715.00    2853.85  10.412  < 2e-16 ***
## PayGrade6   40350.21    2918.14  13.827  < 2e-16 ***
## PayGrade7   59238.67    3026.74  19.572  < 2e-16 ***
## PayGrade8   80252.61    3625.88  22.133  < 2e-16 ***
## Rating4      2158.83    1181.44   1.827 0.069865 .
## LocationUSA   -20.56    1211.58  -0.017 0.986483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6750 on 135 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.9173, Adjusted R-squared:  0.9111
## F-statistic: 149.7 on 10 and 135 DF,  p-value: < 2.2e-16
```

Interpretation of coefficients:

- Coefficient on "Genderf": The expected salary at hire of an employee that's female is about $4767.28 lower than the expected salary at hire of a male employee, controlling for all other explanatory variables.
- Coefficient on "Minority1": The expected salary at hire of an employee who is a member of a minority group is about $1323.88 higher than the expected salary at hire of an employee who is not in a minority group, controlling for all other explanatory variables.
- Coefficient on "PayGrade3", ..., "PayGrade8": The expected salary at hire of an employee in Pay Grade K, for K = 3,...,8 is about:
  - $8782.27
  - $15154.54
  - $29715
  - $40350.21
  - $59238.67
  - $80252.61

  higher, respectively, than the expected salary at hire of an employee in Pay Grade 2, controlling for all other explanatory variables.
- Coefficient on "Rating4": The expected salary at hire of an employee with a performance rating of 4 is about $2158.83 higher than the expected salary at hire of an employee with a performance rating of 3, controlling for all other explanatory variables.
- Coefficient on "LocationUSA": The expected salary at hire of an employee in the USA is about $20.56 lower than the expected salary at hire of an employee in Canada, controlling for all other explanatory variables.

Statistical Significance:

- The coefficient on "Genderf" is negative and is statistically significant at the 1% level, which indicates that there is very strong evidence that the expected salary at hire of female and male employees differ, controlling for all other explanatory variables, and female employees tend to earn less upon hire on average than male employees, controlling for all other explanatory variables. This suggests that there is a material risk of disparate impact discrimination based on gender, looking only at the results of this model.

- The coefficient on "Minority1" is positive and statistically insignificant, which indicates that there is very weak or almost no evidence that the expected salary at hire of minority and non-minority employees differ, controlling for all other explanatory variables. This suggests that there is little to no risk of disparate impact discrimination based on minority status.

Q3.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## 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
```

```
#Collapsing the original dataset to the individual level
dat3 <- ungroup(
  summarise(
    group_by(dat, id),
    gender = first(Gender),
    minority = first(Minority),
    hiredate = first(hiredate),
    location = first(Location),
    min_grade = min(grade, na.rm = TRUE),
    max_grade = max(grade, na.rm = TRUE),
    promoted = as.integer(max_grade > min_grade)
  )
)

dat3$min_grade <- factor(dat3$min_grade)
dat3$max_grade <- factor(dat3$max_grade)
dat3$hire_year <- factor(as.integer(format(dat3$hiredate, "%Y")))
```

```
#Fit the linear regression model for Promoted to the collapsed dataset

model3 <- lm(promoted ~ gender + minority +
               location + min_grade + max_grade + hire_year,
             data = dat3, na.action = na.omit)
summary(model3)
```

```
##
## Call:
## lm(formula = promoted ~ gender + minority + location + min_grade +
```

```
##     max_grade + hire_year, data = dat3, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80809 -0.05641 -0.02596  0.02577  0.42828
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.030339   0.058606  -0.518   0.6053
## genderf        0.024413   0.028845   0.846   0.3984
## minority1     -0.021437   0.031248  -0.686   0.4935
## locationUSA    0.048465   0.027702   1.750   0.0818 .
## min_grade3     0.291643   0.118899   2.453   0.0151 *
## min_grade4    -0.547354   0.098881  -5.536 1.01e-07 ***
## min_grade5    -1.205201   0.106843 -11.280  < 2e-16 ***
## min_grade6    -1.823166   0.109087 -16.713  < 2e-16 ***
## min_grade7    -2.397017   0.124683 -19.225  < 2e-16 ***
## min_grade8    -3.303878   0.144575 -22.852  < 2e-16 ***
## max_grade3    -0.303764   0.130312  -2.331   0.0208 *
## max_grade4     0.577649   0.107285   5.384 2.12e-07 ***
## max_grade5     1.274331   0.116064  10.980  < 2e-16 ***
## max_grade6     1.859864   0.116093  16.020  < 2e-16 ***
## max_grade7     2.458289   0.127714  19.248  < 2e-16 ***
## max_grade8     3.282658   0.141131  23.260  < 2e-16 ***
## hire_year2002  0.002206   0.097463   0.023   0.9820
## hire_year2003 -0.023696   0.039011  -0.607   0.5443
## hire_year2004  0.052222   0.044855   1.164   0.2458
## hire_year2005  0.043882   0.060909   0.720   0.4721
## hire_year2006  0.001591   0.036998   0.043   0.9658
## hire_year2007 -0.045204   0.079284  -0.570   0.5692
## hire_year2008 -0.039602   0.052852  -0.749   0.4546
## hire_year2009 -0.042879   0.118552  -0.362   0.7180
## hire_year2010  0.005586   0.068865   0.081   0.9354
## hire_year2011 -0.077157   0.099760  -0.773   0.4402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.159 on 191 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8493
## F-statistic: 49.69 on 25 and 191 DF,  p-value: < 2.2e-16
```

The explanatory variables that are included in this model are:

- gender
  - Levels in the data set: m, f
  - Dummy variable(s): genderf
- minority
  - Levels in the data set: 0, 1
  - Dummy variable(s): minority1
- location

  - Levels in the data set: Canada, USA
  - Dummy variable(s): locationUSA

- min_grade
  - Levels in the data set: 2, 3, 4, 5, 6, 7, 8
  - Dummy variable(s): min_grade3,. . . ,min_grade8

- max_grade
  - Levels in the data set: 2, 3, 4, 5, 6, 7, 8
  - Dummy variable(s): max_grade3, . . . , max_grade8

- hire_year
  - Levels in the data set: 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011
  - Dummy variable(s): hire_year2002,. . . , hire_year2011

Interpretation of Coefficients:

- Coefficient on "genderf": The expected probability of being promoted for a female employee is about 0.0244 or 2.44% points higher than the expected probability of being promoted for a male employee, controlling for all other explanatory variables.
- Coefficient on "minority1": The expected probability of promotion for employees in a visible minority group is about 0.02144 or 2.144% points lower than the expected probability of promotion for non-minority employees, controlling for all other explanatory variables.
- Coefficient on "locationUSA": The expected probability of promotion for employees in the USA is about 0.0485 or 4.85% points higher than the expected probability of promotion for employees in Canada, controlling for all other explanatory variables.
- Coefficient on "min_grade3",. . . ,"min_grade8": The expected probability of promotion for employees with Grade K as their starting Pay Grade, for K = 3,. . . ,8, is about:
  - 0.2916 or 29.16% points higher
  - 0.5474 or 54.74% points lower
  - 1.2052 or 120.52% points lower
  - 1.8232 or 182.32% points lower
  - 2.3970 or 239.7% points lower
  - 3.3039 or 330.39% points lower

  respectively, than the expected probability of promotion for employees with Grade 2 as their starting Pay Grade. controlling for all other explanatory variables.
- Coefficient on "max_grade3",. . . ,"max_grade8": The expected probability of promotion for employees with Grade K as their highest reported Pay Grade, for K = 3,. . . ,8, is about:
  - 0.3038 or 30.38% points lower
  - 0.5776 or 57.76% points higher
  - 1.2743 or 127.43% points higher
  - 1.8599 or 185.99% points higher
  - 2.4583 or 245.83% points higher

  - 3.2827 or 328.27% points higher

  respectively, than the expected probability of promotion for employees with Grade 2 as their highest reported Pay Grade, controlling for all other explanatory variables.

- Coefficient on "hire_year2002",...,"hire_year2011": The expected probability of promotion for employees that were hired in year t, for t = 2002,...,2011 is about:

  - 0.0022 or 0.22% points higher
  - 0.0237 or 2.37% points lower
  - 0.0522 or 5.22% points higher
  - 0.0439 or 4.39% points higher
  - 0.0016 or 0.16% points higher
  - 0.0452 or 4.52% points lower
  - 0.0396 or 3.96% points lower
  - 0.0429 or 4.29% points lower
  - 0.0056 or 0.56% points higher
  - 0.0772 or 7.72% points lower

  respectively, than the expected probability of promotion for employees that were hired in 2001, controlling for all other explanatory variables.

Statistical Significance:

- The coefficient on "genderf" is positive and is statistically insignificant, which indicates that there is very weak or almost no evidence that the expected probability of promotion for female and male employees differ, controlling for all other explanatory variables. This suggests that there is little to no risk of disparate impact discrimination based on gender, under this Promotion model.
- The coefficient on "minority1" is negative and statistically insignificant, which indicates that there is very weak or almost no evidence that the expected probability of promotion for minority and non-minority employees differ, controlling for all other explanatory variables. This suggests that there is little to no risk of disparate impact discrimination based on minority status, under this Promotion model.

Drawbacks of the Promotions model:

- There are many omitted variables, such as prior experience, education, performance rating, and department, that could also affect the probability of promotion for an employee, which means that Gender and Minority status are not properly controlled for. This would make it harder to confidently determine whether there are any human rights disparate impact risks using this model.
- Because the promotions model is a linear probability model and was fitted using OLS just like a regular linear regression model, which is for continuous outcomes, the predicted probabilities of this model can fall outside of the closed interval [0, 1]. This makes the risk assessment using this model less reliable since this model could underestimate or overestimate the magnitude of the coefficients, which could result in predicted probabilities that are below 0 or above 1, and thus underestimate or overestimate the risk of disparate impact discrimination.