ECN620 - Assignment 3

2023-04-17

ECN620 - Applied Economic Analysis

Professor Andrey Stoyanov

Name: Ha Doan

Student ID: 501026801

Question 1

(a) Estimate three "wage" equations with three different dependent variables: log of income, log of wage, and log of welfare payments. Use the following variables as your explanatory variables: log density, log metpop10, and age.

```
df$log_inctot <- log(df$inctot)
```

```
## Warning in log(df$inctot): NaNs produced
```

```
df$log_incwage <- log(df$incwage)
df$log_incwelfr <- log(df$incwelfr)
df[is.na(df) | df=="-Inf"] = NA # remove negative and zero Value

model <- list(
  "Income" = lm(log_inctot ~ log(density) + log(metpop10) + log(age), data = df),
  "Wage" = lm(log_incwage ~ log(density) + log(metpop10) + log(age), data = df),
  "Welfare Payment" = lm(log_incwelfr ~ log(density) + log(metpop10) + log(age), data = df))

modelsummary(model, stars = T)
```

Based on the summary table, the first model which estimating *income* is the most effective because it has the highest R-squared and all explanatory variables are statistically significant at a 1% significance level. Number of observations for three models are different because negative and zero value of income are removed. To estimate $welfare$ in the third model, *age* is the only statistically significant variable.

(b) Pick up one of the three specifications from 1A and *sex* dummy variable.

```
df <- df %>%
  mutate(male = ifelse(sex == "male", 1, 0))
# transform sex variable into dummy variable

q1b <- lm(log_inctot ~ male, data = df)
summary(q1b)
```

	Income	Wage	Welfare Payment
(Intercept)	15.631***	12.777***	12.660***
	(0.274)	(0.301)	(0.339)
log(density)	−0.262***	−0.344***	0.005
	(0.025)	(0.027)	(0.033)
log(metpop10)	0.263***	0.334***	−0.015
	(0.028)	(0.031)	(0.036)
log(age)	−1.860***	−1.253***	−0.635***
	(0.025)	(0.026)	(0.032)
Num.Obs.	8330	5409	1246
R2	0.413	0.317	0.238
R2 Adj.	0.412	0.316	0.236
AIC	34 665.5	21 134.2	3444.9
BIC	34 700.6	21 167.1	3470.5
Log.Lik.	−17 327.747	−10 562.083	−1717.429
RMSE	1.94	1.71	0.96

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
##
## Call:
## lm(formula = log_inctot ~ male, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6261 -1.3660 -0.5005  0.3670  5.6147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.50343    0.03923  267.72   <2e-16 ***
## male         0.50900    0.05569    9.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.552 on 8397 degrees of freedom
##   (1601 observations deleted due to missingness)
## Multiple R-squared:  0.009852,   Adjusted R-squared:  0.009734
## F-statistic: 83.55 on 1 and 8397 DF,  p-value: < 2.2e-16
```

For the chosen model which estimates *wage* based on age, population, and average 2010 metropolitan area, when adding sex variable, at a 1% significance level, being a male will increase 50.9% of an individual's total income compared to being a female. The model still needs to be improved as it R-squared only explains 0.99% of the variation.

(c) Add *education* in your analysis.

For 1C, not including grade 12 into the model is to avoid perfect multicollinearity.

```
table(df$educ) # frequency
```

```
##
## 1 year o  2 years  4 years 5+ years grade 10 grade 11 grade 12 grade 5,
##     1421      553     1013      628      371      344     3622      676
```

```
##  grade 9 n/a or n  nursery
##      311      502      559
```

```r
df <- df %>%
  mutate(no_school = ifelse(educ == "n/a or n", 1, 0),
         nursery = ifelse(educ == "nursery", 1, 0),
         middle = ifelse(educ == "grade 5,", 1, 0),
         grade_9 = ifelse(educ == "grade 9", 1, 0),
         grade_10 = ifelse(educ == "grade 10", 1, 0),
         grade_11 = ifelse(educ == "grade 11", 1, 0),
         college_1 = ifelse(educ == "1 year o", 1, 0),
         college_2 = ifelse(educ == "2 years", 1, 0),
         college_4 = ifelse(educ == "4 years", 1, 0),
         college_5 = ifelse(educ == "5+ years", 1, 0))

q1c <- lm(log_inctot ~ male + no_school + nursery + middle + grade_9 + grade_10 + grade_11 + college_1 +
summary(q1c)
```

```
##
## Call:
## lm(formula = log_inctot ~ male + no_school + nursery + middle +
##     grade_9 + grade_10 + grade_11 + college_1 + college_2 + college_4 +
##     college_5, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7231  -0.5063   0.3154   0.9211   6.4589
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.45026    0.03587 263.459  < 2e-16 ***
## male         0.37749    0.03701  10.199  < 2e-16 ***
## no_school    4.96175    0.08288  59.863  < 2e-16 ***
## nursery      5.90457    0.07834  75.371  < 2e-16 ***
## middle       3.75991    0.07773  48.371  < 2e-16 ***
## grade_9      0.20890    0.12751   1.638    0.101
## grade_10    -0.55744    0.11895  -4.686 2.82e-06 ***
## grade_11    -0.55002    0.11245  -4.891 1.02e-06 ***
## college_1   -0.06237    0.05915  -1.054    0.292
## college_2    0.49948    0.08122   6.149 8.13e-10 ***
## college_4    0.95318    0.06341  15.031  < 2e-16 ***
## college_5    1.21962    0.07524  16.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.691 on 8387 degrees of freedom
##   (1601 observations deleted due to missingness)
## Multiple R-squared:  0.5656, Adjusted R-squared:  0.565
## F-statistic: 992.8 on 11 and 8387 DF,  p-value: < 2.2e-16
```

It is surprising that those education lower than high school (or elementary education level) earn from 376% to 590% more than those who with higher education level. It might be because they have more working experience.

Only those with 2+ years (3, 4, 5+ years) in post secondary education are statistically higher, while those who finished grade 10, grade 11, or 1 years of college have total income less than others. The model is improved and significant because Adjusted R-squared are 56.5%.

Another model to see if post secondary education is significant in estimating total income. Set grade 12 as a base and indicate that the individual has finished high school. Education will divide into three levels: Lower than high school, Finish high school (grade 12), and higher than high school (post secondary).

```
df <- df %>%
  mutate(lower_hs = ifelse(educ=="n/a or n"|educ=="nursery"|
                         educ=="grade 5+"|educ=="grade 9"|
                         educ=="grade 10"|educ=="grade 11", 1, 0),
         higher_hs = ifelse(educ=="1 year o"|educ=="2 years"|
                          educ=="4 years"|educ=="5+ years", 1, 0))

q1c_a <- lm(log_inctot ~ male + lower_hs + higher_hs, data = df)
summary(q1c_a)
```

```
##
## Call:
## lm(formula = log_inctot ~ male + lower_hs + higher_hs, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.387 -1.089 -0.005  0.902  6.083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.03515    0.04651 215.771  < 2e-16 ***
## male         0.40937    0.05074   8.068 8.15e-16 ***
## lower_hs     2.63136    0.06877  38.266  < 2e-16 ***
## higher_hs   -0.02850    0.05677  -0.502    0.616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.322 on 8395 degrees of freedom
##   (1601 observations deleted due to missingness)
## Multiple R-squared:  0.1803, Adjusted R-squared:   0.18
## F-statistic: 615.5 on 3 and 8395 DF,  p-value: < 2.2e-16
```

The effect of *male* decreases when including educational level, lower educational level is statistically significant at 1% significant value and those with lower educational level have their total income higher by 263%.

(d) Construct hourly wage rate from your data and estimate the return on education using the same specification as in 1C.

```
# compute hourly wage rate
df <- df %>%
  mutate(hour_wage = incwage/(wkswork1*as.numeric(uhrswork)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `hour_wage = incwage/(wkswork1 * as.numeric(uhrswork))`.
## Caused by warning:
## ! NAs introduced by coercion
```

```r
# construct model
q1d <- lm(hour_wage ~ male + no_school + nursery + middle + grade_9 + grade_10 + grade_11 + college_1 +
summary(q1d)
```

```
##
## Call:
## lm(formula = hour_wage ~ male + no_school + nursery + middle +
##     grade_9 + grade_10 + grade_11 + college_1 + college_2 + college_4 +
##     college_5, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -45.09  -12.97   -5.96    2.14 2250.18
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.248      1.717   7.716 1.49e-14 ***
## male           9.727      1.783   5.456 5.13e-08 ***
## no_school     -9.731     11.704  -0.831   0.4058
## nursery       -6.468     26.001  -0.249   0.8036
## middle        -2.349      7.899  -0.297   0.7662
## grade_9       -8.847      8.260  -1.071   0.2842
## grade_10      -1.785      6.008  -0.297   0.7664
## grade_11      -7.660      5.101  -1.502   0.1332
## college_1      1.377      2.484   0.554   0.5794
## college_2      7.826      3.350   2.336   0.0195 *
## college_4     15.733      2.646   5.946 2.96e-09 ***
## college_5     22.118      3.156   7.008 2.80e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.04 on 4303 degrees of freedom
##   (5685 observations deleted due to missingness)
## Multiple R-squared:  0.02501,    Adjusted R-squared:  0.02252
## F-statistic: 10.04 on 11 and 4303 DF,  p-value: < 2.2e-16
```

```r
q1d_a <- lm(hour_wage ~ male + lower_hs + higher_hs, data = df)
summary(q1d_a)
```

```
##
## Call:
## lm(formula = hour_wage ~ male + lower_hs + higher_hs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -33.24  -14.01   -6.31    2.16 2255.65
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.454      1.711   7.865 4.64e-15 ***
## male           9.200      1.785   5.153 2.67e-07 ***
## lower_hs      -6.033      3.555  -1.697   0.0898 .
```

```
## higher_hs     10.585      1.869   5.664 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.29 on 4311 degrees of freedom
##   (5685 observations deleted due to missingness)
## Multiple R-squared:  0.01493,    Adjusted R-squared:  0.01424
## F-statistic: 21.78 on 3 and 4311 DF,  p-value: 5.389e-14
```

When compute hourly wage, we discover that those whose education are lower than high school will earn \$6.03/*hour* less than who finish high school, and whose education are higher than high school will earn \$10.59/*hour* more than who only have high school education. Only those whose education are higher is statistically significant at a 1% significance level.

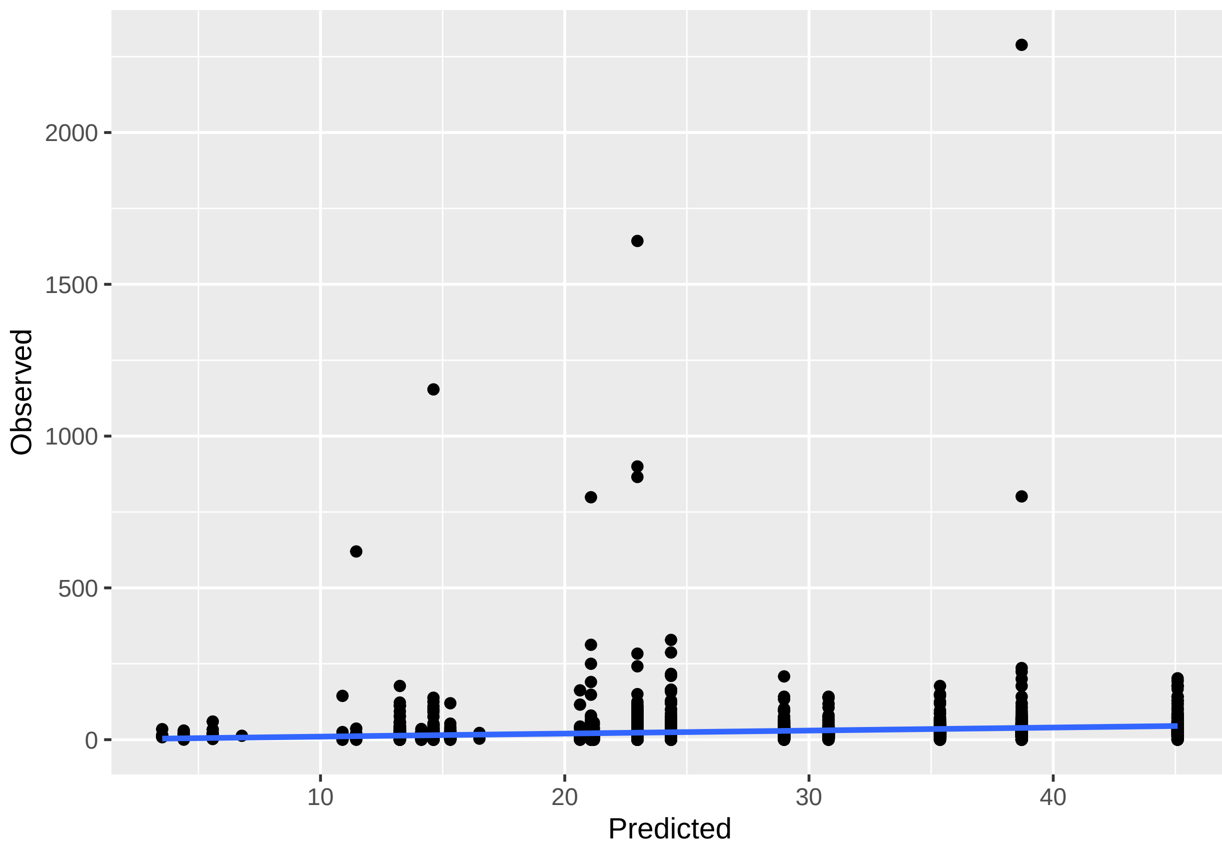
(e) Construct a scatter plot of the predicted and actual wealth status.

```r
# extract non-missing value
new_df <-subset(df, df$hour_wage != "NA")

# create data for ggplot
data_mod <- data.frame(Predicted = predict(q1d),
                       Observed = new_df$hour_wage)

# plot
ggplot(data_mod,
       aes(x = Predicted,
           y = Observed)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(f) Using hourly wage rate as the dependent variable, estimate the return to post secondary education (1 year of college and up) separately for male and female.

```
q1f <- lm(hour_wage ~ higher_hs + male + higher_hs*male, data = df)
summary(q1f)
```

```
##
## Call:
## lm(formula = hour_wage ~ higher_hs + male + higher_hs * male,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -33.52  -13.88   -6.41    2.04 2255.37
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13.078      1.956   6.688 2.56e-11 ***
## higher_hs        10.706      2.582   4.146 3.45e-05 ***
## male              8.159      2.614   3.121  0.00181 **
## higher_hs:male    1.574      3.573   0.440  0.65965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.31 on 4311 degrees of freedom
##   (5685 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.01432,    Adjusted R-squared:  0.01363
## F-statistic: 20.87 on 3 and 4311 DF,  p-value: 2.021e-13
```

higherhs is the variable that represent whether the individual return to post secondary education. Among all those with return to post secondary education (*higherhs*), being a male increase $\$1.57 + \$8.15 = \$9.72/hour$ in hourly wage. At 5% significant level, the difference in hourly wage of a male and a female that both has higher educational level is not significant.

(g) Estimate the return on education for female workers with and without children.

```r
df <- df %>%
  mutate(without_children = ifelse(nchild == "0 childr", 1, 0),
         female = ifelse(sex == "female", 1, 0))
# transform dummy variables

q1g <- lm(higher_hs ~  female + without_children + female*without_children, data = df)
summary(q1g)
```

```
##
## Call:
## lm(formula = higher_hs ~ female + without_children + female *
##     without_children, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5152 -0.3457 -0.3016  0.6543  0.6984
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.508413   0.016432  30.940   <2e-16 ***
## female                   0.006813   0.021329   0.319    0.749
## without_children        -0.206803   0.018000 -11.489   <2e-16 ***
## female:without_children  0.037305   0.023836   1.565    0.118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.474 on 9996 degrees of freedom
## Multiple R-squared:  0.02709,    Adjusted R-squared:  0.0268
## F-statistic: 92.79 on 3 and 9996 DF,  p-value: < 2.2e-16
```

The probability of return on education for female workers without children will lower than female workers with children by $3.73\% - 20.68\% = 16.95\%$. The difference in the probability of return on education for a femal worker with or without children is not statistically significant at a 5% signigicance level.