Assignment 5: Q # 6 and 7

Adam Patterson

12/17/2021

This pdf acts as a supplement to my homework as I could not figure out the wald test in python for question 6 and I could not get the ordered probit to run in question 7.

Data Management

```
setwd("C:/Users/AdamL/OneDrive/Desktop/Metrics/dta")
d<-read_dta("jcity.dta")

# Create dummy variables
# Education
d$edlt10<-ifelse(d$edcat==1, 1, 0)
d$ed10_11<-ifelse(d$edcat==2, 1, 0)
d$ed12<-ifelse(d$edcat==3, 1, 0)
d$ed13_15<-ifelse(d$edcat==4, 1, 0)
d$edgt15<-ifelse(d$edcat==5, 1, 0)

# Race
d$white<-ifelse(d$race==1, 1, 0)
d$black<-ifelse(d$race==2, 1, 0)
d$hispanic<-ifelse(d$race==3, 1, 0)
d$asianoth<-ifelse(d$race==4, 1, 0)
```

Run Model

```
# Start multinomial model
# Instruct R what group to use as base reference
d$trtmnt1<- relevel(as.factor(d$trtmnt), ref = "3")
# Run Multinomial model
multinomial_mod<- multinom(trtmnt1~age+nvrwrk+edlt10+ed10_11+ed13_15+edgt15+white+black+hispanic, data
```

```
## # weights:  33 (20 variable)
## initial  value 569.081166
## iter  10 value 537.728371
## iter  20 value 536.838728
## final  value 536.838538
## converged
```

Print Model Summary

```
summary(multinomial_mod)
```

```
## Call:
## multinom(formula = trtmnt1 ~ age + nvrwrk + edlt10 + ed10_11 +
##     ed13_15 + edgt15 + white + black + hispanic, data = d)
##
## Coefficients:
##   (Intercept)         age     nvrwrk     edlt10     ed10_11   ed13_15    edgt15
## 1   0.8978566 -0.01877284  0.4838519 -0.1790391 -0.05990046 0.6195847 0.0300192
## 2   0.3245199 -0.01948170 -0.6276590  0.2972645 -0.14041992 0.2850586 0.2580708
##       white      black   hispanic
## 1 0.2484582 -1.2305383 -0.5521050
## 2 0.5817162  0.2193995 -0.3968595
##
## Std. Errors:
##   (Intercept)        age    nvrwrk    edlt10   ed10_11   ed13_15    edgt15
## 1   0.6863775 0.01553461 0.4176293 0.4015204 0.2787961 0.3935512 0.6410782
## 2   0.7122803 0.01408594 0.5093736 0.3426199 0.2425907 0.3668645 0.5737204
##       white     black  hispanic
## 1 0.6465034 0.4963665 0.5348986
## 2 0.7261817 0.5647437 0.6200577
##
## Residual Deviance: 1073.677
## AIC: 1113.677
```

6a) Discussing the model results, we note that the characteristic of higher educated has higher influence on log odds for on the job training compared to in class training. There could be selection bias in the program ? more educated individuals may omit this program through their education alone. Of course this increase in log odds is in comparison to the other group, it is not a direct result.

We notice that individuals that have never worked are more likely to choose to the in class training rather than other. We see that individuals that have never worked are less likely to go to on the job training than other. This is intuitive as if an individual has never been to work in that industry or job, he/she/they may be timid to start training in a live setting without classroom or "other" training.

Print Likelihood and wald test

```
lrtest(multinomial_mod)
```

```
## # weights:  6 (2 variable)
## initial  value 569.081166
## final  value 558.093857
## converged
```

```
## Likelihood ratio test
##
## Model 1: trtmnt1 ~ age + nvrwrk + edlt10 + ed10_11 + ed13_15 + edgt15 +
##     white + black + hispanic
```

```
## Model 2: trtmnt1 ~ 1
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  20 -536.84
## 2   2 -558.09 -18 42.511  0.0009379 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Print p value of each individual coefficient . I also saw an F distribution test to calculate wald online?

The wald test options still give problem. code provided upon request for details and attempts/packages installed

```
# Calculate Wald Z
z<-summary(multinomial_mod)$coefficients/summary(multinomial_mod)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
##   (Intercept)       age    nvrwrk    edlt10    ed10_11    ed13_15    edgt15
## 1   0.1908363 0.2268731 0.2466324 0.6556673 0.8298811 0.1154082 0.9626518
## 2   0.6486726 0.1666465 0.2178679 0.3856013 0.5627006 0.4371508 0.6528403
##       white      black  hispanic
## 1 0.7007481 0.01317173 0.3019936
## 2 0.4230960 0.69765056 0.5221489
```

Question 7

Data Management

```
#### Question 7
setwd("C:/Users/AdamL/OneDrive/Desktop/Metrics/dta")
data<-read_dta("corpus.dta")

# Create Edcat Variables
data$EdCat<-2
data[(data$edlt10 | data$ed10_11 ==1),11]<-1
data[(data$ed13_15 | data$edgt15 ==1),11]<-3
```

```
# Try polr package found online as z
m<-polr(as.factor(EdCat)~age+black+hisp, data = data, Hess=TRUE)
summary(m)
```

```
## Call:
## polr(formula = as.factor(EdCat) ~ age + black + hisp, data = data,
##     Hess = TRUE)
##
## Coefficients:
##             Value Std. Error t value
## age   -0.002036   0.007145 -0.2850
```

```
## black -0.120167   0.199145 -0.6034
## hisp  -0.978408   0.129310 -7.5664
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.7603  0.2582    -2.9448
## 2|3  1.0036  0.2599     3.8607
##
## Residual Deviance: 2240.935
## AIC: 2250.935
```

b)

```r
# Grab mean predicted prob for each Edcat
mean_less_than_12<-mean(m$fitted.values[,1])
mean_12<-mean(m$fitted.values[,2])
mean_greater_than_12<-mean(m$fitted.values[,3])
```

```r
# Print observed percentage of each category
table1<-table(data$EdCat)
round(prop.table(table1),2)
```

```
##
##    1    2    3
## 0.48 0.35 0.17
```

```r
# Print the predicted probability for each cateogry
print('Predicted probability less than 12 yr educ:')
```

```
## [1] "Predicted probability less than 12 yr educ:"
```

```r
mean_less_than_12
```

```
## [1] 0.4801103
```

```r
print('Predicted probability of 12 yr educ:')
```

```
## [1] "Predicted probability of 12 yr educ:"
```

```r
mean_12
```

```
## [1] 0.3524768
```

```r
print('Predicted probability of greater than 12 yr educ:')
```

```
## [1] "Predicted probability of greater than 12 yr educ:"
```

```r
mean_greater_than_12
```

```
## [1] 0.1674129
```

We see that predictions are close to actual values

c. For this ordered probit model, calculate the partial effect of being Hispanic rather than being white (the omitted category) on the probability of having exactly 12 years of schooling and on the probability of having more than 12 years of schooling. These partial effects should be estimated as mean finite differences. Interpret the resulting partial effects.(Cannot get at post equivalent to run in python)

```r
# Predict probabilities using 'probs' option
m.pred <- predict(m, type="probs")
summary(m.pred)
```

```
##        1                2                3
##  Min.   :0.3284   Min.   :0.3088   Min.   :0.1099
##  1st Qu.:0.3383   1st Qu.:0.3153   1st Qu.:0.1143
##  Median :0.5663   Median :0.3177   Median :0.1160
##  Mean   :0.4801   Mean   :0.3525   Mean   :0.1674
##  3rd Qu.:0.5703   3rd Qu.:0.4106   3rd Qu.:0.2510
##  Max.   :0.5813   Max.   :0.4121   Max.   :0.2595
```

The Mean value printed above is the predicted probabilities of each category when all predictors are at their mean value.

3c) Use something similar: new_data <- data.frame(x1=rep(mean(data$age), 3), x2 = rep(mean(data$black),3),x3=rep(mean(
m1.pred <- predict(m, newdata=new_data,type="probs") summary(m1.pred)