SOCI 30005\_PS1\_Hinojosa

Cintia Hinojosa

April 26, 2019

nals <- read\_csv("C://Users/cinti/Box Sync/Booth 2017-2018/Spring 2019/Statistical Methods of Research 2/TA Sessions/nals\_synthetic.csv", col\_names = TRUE)

## Parsed with column specification:  
## cols(  
## id = col\_double(),  
## annearn = col\_double(),  
## occprest = col\_double(),  
## sei = col\_double(),  
## age = col\_double(),  
## yearsed = col\_double(),  
## gender = col\_double(),  
## ln\_earn = col\_double(),  
## literacy = col\_double(),  
## unemp = col\_double(),  
## parented = col\_double(),  
## Ethnicity = col\_double(),  
## Language = col\_double(),  
## Education = col\_double()  
## )

ls()

## [1] "nals"

names(nals)

## [1] "id" "annearn" "occprest" "sei" "age"   
## [6] "yearsed" "gender" "ln\_earn" "literacy" "unemp"   
## [11] "parented" "Ethnicity" "Language" "Education"

summary(nals)

## id annearn occprest sei   
## Min. :1.010e+10 Min. : 15 Min. :17.00 Min. :17.00   
## 1st Qu.:3.110e+10 1st Qu.: 12954 1st Qu.:33.00 1st Qu.:32.00   
## Median :5.080e+10 Median : 20800 Median :43.00 Median :42.00   
## Mean :4.791e+10 Mean : 25903 Mean :43.79 Mean :48.52   
## 3rd Qu.:6.670e+10 3rd Qu.: 33280 3rd Qu.:51.00 3rd Qu.:64.00   
## Max. :8.570e+10 Max. :416000 Max. :86.00 Max. :97.00   
## NA's :1316   
## age yearsed gender ln\_earn   
## Min. :25.00 Min. : 0.00 Min. :0.0000 Min. :1.099   
## 1st Qu.:31.00 1st Qu.:12.00 1st Qu.:0.0000 1st Qu.:5.580   
## Median :38.00 Median :13.00 Median :0.0000 Median :6.043   
## Mean :38.89 Mean :13.31 Mean :0.4902 Mean :6.018   
## 3rd Qu.:46.00 3rd Qu.:16.00 3rd Qu.:1.0000 3rd Qu.:6.479   
## Max. :59.00 Max. :18.00 Max. :1.0000 Max. :9.508   
## NA's :344   
## literacy unemp parented Ethnicity   
## Min. : 49.25 Min. :0.0000 Min. : 0.00 Min. :1.000   
## 1st Qu.:253.73 1st Qu.:0.0000 1st Qu.: 9.00 1st Qu.:2.000   
## Median :292.81 Median :0.0000 Median :12.00 Median :5.000   
## Mean :286.18 Mean :0.0879 Mean :10.93 Mean :3.982   
## 3rd Qu.:329.13 3rd Qu.:0.0000 3rd Qu.:13.00 3rd Qu.:5.000   
## Max. :441.40 Max. :1.0000 Max. :18.00 Max. :5.000   
##   
## Language Education   
## Min. :1.000 Min. :1.000   
## 1st Qu.:5.000 1st Qu.:3.000   
## Median :5.000 Median :3.000   
## Mean :4.477 Mean :3.487   
## 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :6.000   
##

## Probability

Consider the population of US adults in the labor force in 1992. We are interested in the relationship between educational attainment and the risk of unemployment. Educational attainment has six possible values: no degree, GED, high school degree, associates degree, bachelors degree, and masters degree or higher.

### Q1. Construct a theoretical contingency table with two rows (values of unemployment) and six columns (values of educational attainment) in which the entries are the joint and marginal probabilities. Use Greek letters to represent these (for example use it).

# Label values for easy interpretation  
nals$unemp <- factor(nals$unemp,  
levels = c(0,1),   
labels = c("employed", "unemployed"))  
  
nals$Education <- factor(nals$Education,  
levels = c(1,2,3,4,5,6),   
labels = c("none", "ged", "hs", "aa", "ba", "grad"))  
  
  
# Contingency table with educational attainment (x) and unemployment status(y)  
nals.tab <- prop.table(table(nals$Education,nals$unemp,   
 dnn = c("Education Attainment", "Unemployment Status")))  
  
# Convert to data frame  
nals.tab <- as.data.frame.matrix(nals.tab)  
   
# Get marginals and joint probabilities for rows  
nals.tab$marginal.education <- rowSums(nals.tab)  
   
# Add joint probabilities to get unemployment marginals  
nals.tab["marginal.employment",] <- colSums(nals.tab)  
   
nals.tab

## employed unemployed marginal.education  
## none 0.10654819 0.019132245 0.12568044  
## ged 0.03330131 0.005203330 0.03850464  
## hs 0.39225104 0.040185719 0.43243676  
## aa 0.11711495 0.010246558 0.12736151  
## ba 0.16970861 0.008885687 0.17859430  
## grad 0.09317963 0.004242715 0.09742235  
## marginal.employment 0.91210375 0.087896254 1.00000000

# Crosstabulation of unemployment status and educational attainment  
# N/T = joint probability  
# Col and row percentages = marginal probability  
crosstab <- CrossTable(nals$unemp, nals$Education,  
 expected = FALSE,   
 prop.r = TRUE,   
 prop.c = TRUE,  
 prop.t = TRUE,  
 prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 12492   
##   
##   
## | nals$Education   
## nals$unemp | none | ged | hs | aa | ba | grad | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## employed | 1331 | 416 | 4900 | 1463 | 2120 | 1164 | 11394 |   
## | 0.117 | 0.037 | 0.430 | 0.128 | 0.186 | 0.102 | 0.912 |   
## | 0.848 | 0.865 | 0.907 | 0.920 | 0.950 | 0.956 | |   
## | 0.107 | 0.033 | 0.392 | 0.117 | 0.170 | 0.093 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## unemployed | 239 | 65 | 502 | 128 | 111 | 53 | 1098 |   
## | 0.218 | 0.059 | 0.457 | 0.117 | 0.101 | 0.048 | 0.088 |   
## | 0.152 | 0.135 | 0.093 | 0.080 | 0.050 | 0.044 | |   
## | 0.019 | 0.005 | 0.040 | 0.010 | 0.009 | 0.004 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 1570 | 481 | 5402 | 1591 | 2231 | 1217 | 12492 |   
## | 0.126 | 0.039 | 0.432 | 0.127 | 0.179 | 0.097 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

$$
Pr(Y=y)=\phi^y (1-\phi)^{1-y}\text{, for }y\text{ }\epsilon \{0,1\}\\
y:\{1=\text{unemployed, }0=\text{employed\}}\ \\
Pr(Y=1)=\phi \\
Pr(Y=0)=1-\phi
$$

### Q2. Define the marginal probability of unemployment and decompose it into the sum of the relevant joint probabilities.

#### Marginal Probabilities for Unemployment Status:

The marginal probabilities tell us the probability of an event independent of other variables. Here, there is a 91.2% probability that a randomly selected individual will be employed and 8.8% probability they would be unemployed.

#### Joint Probabilities for Unemployment Status and Educational Attainment:

j.u\_nd <- .19  
j.u\_ged <- .005  
j.u\_hs <- .040  
j.u\_aa <- .01  
j.u\_ba <- .009  
j.u\_ma <- .004

The joint probabilities tell us the probability that an individual has two attributes of certain levels. For example, from observing the joint probabilities in the data, there is a 1.9% probability of an individual being unemployed and having no educational degree, and for the most part, this probability shrinks as the level of educational attainment grows. Relative to the other probability estimates, unemployed individuals with a high school degree break away from this pattern, with the highest observed joint probability with unemployment at 4%.

### Q3.For each possible level of education, define the conditional probability of unemployment.

#### Conditional Probabilities for Unemployment:

Conditional probability shows us the probability that an individual will experience an event, given another event. In this case, we are examining the probability that an individual is unemployed *given* the level of education they have attained. The data reveals a pattern in which the conditional probability of unemployment decreases as level of education attained increases, with the joint probability of being unemployed and having no degree at 15.1% while being unemployed and having a master's degree or higher is at 4.1%.

### Q4. Decompose the joint probability of having no degree and being unemployed into the relevant marginal and conditional probabilities.

$$
Pr(X=1, Y=1) = Pr(Y=1|X=1)\*Pr(X=1) \\
Pr(Y=1, X=1) = Pr(X=1|Y=1)\*Pr(Y=1)
$$

We can break down the joint probability equation to see observe the conditional and marginal probabilities of unemployment and educational attainment. The joint probability for an individual to be unemployed and have no educational degree is 1.9%, the conditional probability that they will be unemployed given that they have no degree is 15.1%, and the marginal probability of having no educational degree is 12.6%.

### Q5. Now decompose the marginal probability of being unemployed into a function of the relevant marginal and conditional probabilities.

$ $

### Q6. Using the NALS data, estimate the conditional probabilities of unemployment for each level of education. What does this seem to say about the association between education and unemployment?

# Checkout variable labels  
names(nals)

## [1] "id" "annearn" "occprest" "sei" "age"   
## [6] "yearsed" "gender" "ln\_earn" "literacy" "unemp"   
## [11] "parented" "Ethnicity" "Language" "Education"

names(nals.tab)

## [1] "employed" "unemployed" "marginal.education"

# Regression of marginal education on unemployed status  
lm\_unemp\_edu <- lm(unemployed ~ marginal.education, data=nals.tab)  
stargazer(lm\_unemp\_edu, type='text')

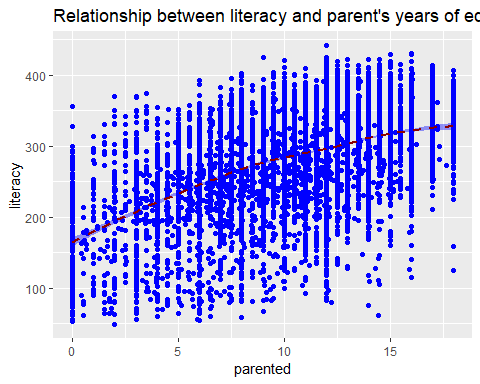
##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## unemployed   
## -----------------------------------------------  
## marginal.education 0.088\*\*\*   
## (0.006)   
##   
## Constant -0.0001   
## (0.003)   
##   
## -----------------------------------------------  
## Observations 7   
## R2 0.975   
## Adjusted R2 0.970   
## Residual Std. Error 0.005 (df = 5)   
## F Statistic 193.195\*\*\* (df = 1; 5)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

### Q7. Again using NALS, assume that unemployment and education were independent. What would then be the estimated conditional probability of unemployment given no degree? Under this scenario, how many of those with degree would we expect to be unemployed? Compare this to the number of those with no degree who were in fact unemployed and comment on how education is associated with joblessness for this group.

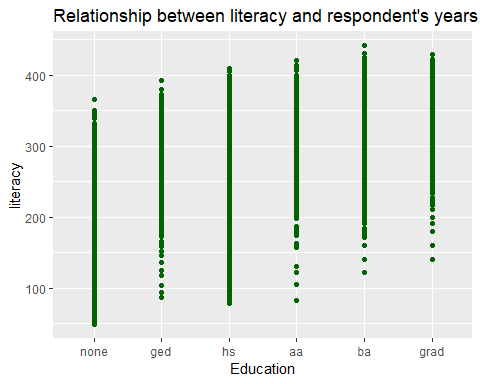
## Expectation

We are again going to work with NALS, now using three variables: parent years of education (X), respondent years of education (Z), and adult literacy (Y).

# Relationship between literacy and parental years of education  
ggplot(nals, aes(parented,literacy)) + geom\_point(color="blue") +   
 labs(title="Relationship between literacy and parent's years of education") +  
 geom\_smooth(method="loess", linetype="dashed",color="darkred", fill="blue")



# Relationship between literacy and respondent's education  
ggplot(nals, aes(Education,literacy)) + geom\_point(color="darkgreen") +   
 labs(title="Relationship between literacy and respondent's years of education") +   
 geom\_smooth(method="loess", linetype="dashed",color="darkred", fill="green")



There is a general positive linear relationships between the outcome variable, respondent's literacy level, with both of the independent variables, parent's years of education and respondent's educational attainment. There also doesn't seem to be any extreme outliners that stand out, so we can continue examining for statistically significant linear relationships through regression models.

### Q1. Write down a theoretical linear regression model in which Y is a function of X and Z. Define the terms in the model. We will assume this is the true model" of the relationship between X,Z, and Y.

$$
Y\_i=\alpha+\beta X\_i+\gamma Z\_i+\varepsilon\_i, \\
\varepsilon\_i ~ N(0,\sigma\_{y|x,z})
$$

In the above model, is the literacy level of a particular individual in the population, is the y-intercept parameter for the model's regression line, meaning that it is the average literacy level for an individual who has no educational degree () and parents who have 0 years of education (). () is the regression coefficient for the slope parameter for parental education and () is the parameter for respondent's educational attainment. is the difference between an individual's actual literacy level and the average literacy level for all individuals in the population who share the same values in regards to their educational attainment, Z, and their parent's years of education, X. is the standard deviation for a particular subset of individuals in the populations with the same values for Z and X.

### Q2. Write down two other linear regression models: a) Z is a function of X; and b) X is a function of Z;

$$
Z\_i=\alpha+\beta X\_i+\varepsilon\_i, \\
\varepsilon\_i ~ N(0,\sigma\_{z|x})
$$

$$
X\_i=\alpha+\beta Z\_i+\varepsilon\_i, \\
\varepsilon\_i ~ N(0,\sigma\_{x|z})
$$

### Q3. Now suppose someone estimated a model using only Z as a predictor. That means the person would be studying the expected value of Y given Z alone.

#### a. Using (1), find E(Y|Z).

$$\mathbb E(Y|Z)=\mathbb E(\alpha) + \mathbb E(\beta Z)$$

#### b. Using (1) and (2b), define the bias involved. Show that it has two parts and define them.

In the above model, is the naive estimate of the coefficient for the respondent's educational attainment as a sole predictor of their literacy level, . The potential bias in the model is captured by . The two conditions for bias to equal 0 is for or if there was no difference in the predicted average literacy level between individuals with varying years of education.

# linear regression of respondent's educational attainment on adult literacy  
lm\_literacy\_ed <- lm(nals$literacy ~ nals$Education)  
summary(lm\_literacy\_ed)

##   
## Call:  
## lm(formula = nals$literacy ~ nals$Education)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -219.725 -27.976 3.964 33.882 160.501   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 204.398 1.221 167.38 <2e-16 \*\*\*  
## nals$Educationged 65.202 2.522 25.86 <2e-16 \*\*\*  
## nals$Educationhs 74.697 1.387 53.84 <2e-16 \*\*\*  
## nals$Educationaa 98.976 1.721 57.50 <2e-16 \*\*\*  
## nals$Educationba 120.800 1.594 75.78 <2e-16 \*\*\*  
## nals$Educationgrad 131.235 1.848 71.01 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 48.39 on 12486 degrees of freedom  
## Multiple R-squared: 0.3777, Adjusted R-squared: 0.3775   
## F-statistic: 1516 on 5 and 12486 DF, p-value: < 2.2e-16

stargazer(lm\_literacy\_ed, type='text')

##   
## ================================================  
## Dependent variable:   
## ----------------------------  
## literacy   
## ------------------------------------------------  
## Educationged 65.202\*\*\*   
## (2.522)   
##   
## Educationhs 74.697\*\*\*   
## (1.387)   
##   
## Educationaa 98.976\*\*\*   
## (1.721)   
##   
## Educationba 120.800\*\*\*   
## (1.594)   
##   
## Educationgrad 131.235\*\*\*   
## (1.848)   
##   
## Constant 204.398\*\*\*   
## (1.221)   
##   
## ------------------------------------------------  
## Observations 12,492   
## R2 0.378   
## Adjusted R2 0.377   
## Residual Std. Error 48.388 (df = 12486)   
## F Statistic 1,515.971\*\*\* (df = 5; 12486)  
## ================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

lm\_yz.ged <- 65.202   
lm\_yz.hs <- 74.697  
lm\_yz.aa <- 98.976  
lm\_yz.ba <- 120.800  
lm\_yz.grad <- 131.235

In a linear regression of respondent educational on adult literacy, the coefficients vary by level of educational attainment. The model predicts an average literacy score increase of 65.2 for respondents with a ged, 74.7 for those with a high school degree, 99 for those with an associates degree, 120.8 for those with a bachelor's degree, and 131.2 for respondents with a master's degree or higher. Respondents with no educational degree having a predicted average literacy score of 204.4 under this model. All coefficient estimates are statistically significant (p<0.001\*\*\*) and have large magnitudes. Having a GED is predicted to have the smallest estimated average effect on literacy scores compared to the other educational degrees, with a large t-score of 25.8.

$$
y\_i = 204.4 + z\_{\text{no\_degree}} + \epsilon \\
y\_i = 204.4 + 65.2z\_{ged} + \epsilon \\
y\_i = 204.4 + 74.7z\_{hs} + \epsilon \\
y\_i = 204.4 + 98.98z\_{aa} + \epsilon \\
y\_i = 204.4 + 120.8z\_{ba} + \epsilon \\
y\_i = 204.4 + 131.24z\_{grad} + \epsilon \\
$$

### Q4. Now suppose someone estimated a model using only X as a predictor. That means this person would be studying the expected value of Y given X alone (2c).

#### a. Using (1), find E(Y|X). Define the “total effect" of X on Y.

$$\mathbb E(Y|X)=\mathbb E(\alpha) + \mathbb E(\beta X)$$

#### b. What is the direct effect of X on Y based on your theoretical model (1)?

#### c. Find the indirect effect of X on Y as it operates through Z.

#### d. Show that the total effect of X on Y is the sum of the direct and indirect effects you have defined.

$y&=\theta x+\varepsilon\_i$

### Q5. Estimate these total, direct, and indirect effects using NALS, and comment on what you have learned about how parent education and respondent education are linked to adult literacy.

# linear regression of parental education (X) on adult literacy (Y)  
lm\_yx <- lm(nals$literacy ~ nals$parented)  
summary(lm\_yx)

##   
## Call:  
## lm(formula = nals$literacy ~ nals$parented)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -254.977 -30.513 4.537 36.879 163.180   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 192.4736 1.5413 124.88 <2e-16 \*\*\*  
## nals$parented 8.5697 0.1341 63.92 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 53.24 on 12490 degrees of freedom  
## Multiple R-squared: 0.2465, Adjusted R-squared: 0.2465   
## F-statistic: 4086 on 1 and 12490 DF, p-value: < 2.2e-16

stargazer(lm\_yx, type='text')

##   
## ================================================  
## Dependent variable:   
## ----------------------------  
## literacy   
## ------------------------------------------------  
## parented 8.570\*\*\*   
## (0.134)   
##   
## Constant 192.474\*\*\*   
## (1.541)   
##   
## ------------------------------------------------  
## Observations 12,492   
## R2 0.247   
## Adjusted R2 0.246   
## Residual Std. Error 53.238 (df = 12490)   
## F Statistic 4,086.296\*\*\* (df = 1; 12490)  
## ================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

direct\_yx <- 8.5697  
direct\_yx

## [1] 8.5697

In a linear regression of parental education on adult literacy, parental education has a significant total effect on literacy with a coefficient of $b=$8.57(.134), , . This means that the model estimates an average increase of 8.57 units in a respondent's literacy level for every added 1 year of parental education, with respondents with 0 years of parental edcuation having a predicted average literacy score of 192.47.

# linear regression of parental education (X) on educational attainment (Z)  
lm\_xz <- lm(parented ~ Education, data=nals)  
summary(lm\_xz)

##   
## Call:  
## lm(formula = parented ~ Education, data = nals)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.8086 -1.8085 0.2704 1.9622 10.5471   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.45287 0.07946 93.79 <2e-16 \*\*\*  
## Educationged 2.50642 0.16408 15.28 <2e-16 \*\*\*  
## Educationhs 3.08494 0.09027 34.17 <2e-16 \*\*\*  
## Educationaa 4.55388 0.11200 40.66 <2e-16 \*\*\*  
## Educationba 5.35568 0.10372 51.64 <2e-16 \*\*\*  
## Educationgrad 5.27675 0.12025 43.88 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.149 on 12486 degrees of freedom  
## Multiple R-squared: 0.2151, Adjusted R-squared: 0.2148   
## F-statistic: 684.4 on 5 and 12486 DF, p-value: < 2.2e-16

stargazer(lm\_xz, type='text')

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## parented   
## -----------------------------------------------  
## Educationged 2.506\*\*\*   
## (0.164)   
##   
## Educationhs 3.085\*\*\*   
## (0.090)   
##   
## Educationaa 4.554\*\*\*   
## (0.112)   
##   
## Educationba 5.356\*\*\*   
## (0.104)   
##   
## Educationgrad 5.277\*\*\*   
## (0.120)   
##   
## Constant 7.453\*\*\*   
## (0.079)   
##   
## -----------------------------------------------  
## Observations 12,492   
## R2 0.215   
## Adjusted R2 0.215   
## Residual Std. Error 3.149 (df = 12486)   
## F Statistic 684.446\*\*\* (df = 5; 12486)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

ind\_xz.ged <- 2.50642  
ind\_xz.hs <- 3.08494  
ind\_xz.aa <- 4.55388  
ind\_xz.ba <- 5.35568  
ind\_xz.grad <- 5.27675

There is a statistically significant effect of years of parent's education on respondents' educational attainment that is robust across all degree levels (p<0.001\*\*\*).

Total effect = 8.5697+`r ind

There is a statistically significant effect of years of parent's education on educational attainment that is very robust across all levels of educational attainment. According to the NALS dataset, both respondent educational attainment and parental years of education are significant predictors of average literacy scores. Educational attainment is also a predictor of parental years of education and they are both moderately correlated with each other (*r*=.5). This suggests there may be an indirect or mediating relationship between these variables and literacy levels. It makes sense that academic factors like years of parental education and educational attainment would have a significant impact on an academic outcome like literacy scores. More investigation can be done on how well literacy scores may predict influential life outcomes like log annual earnings, to see if these strong associations have impact on life outcomes beyond academic advancement.