Capstone Final

Yusei Hara

2020/11/14

Table of Contents

## ASER Pakistan 2016

In this piece of paper, a set of data obtained from Annual Status of Education Report (ASER) is explored. The raw data was downloaded from the link here. <https://palnetwork.org/aser-centre/>

### Preparation

#### Packages Used

library(tidyverse)  
library(ggplot2)  
library(ggthemes)  
library(ggrepel)  
library(gghighlight)  
library(stringr)  
library(dplyr)  
library(sf)  
library(scatterplot3d)  
library(car)  
library(ResourceSelection) # to excute Hosmer-Lemeshow test

## Warning: package 'ResourceSelection' was built under R version 4.0.3

library(equatiomatic) # to convert model to equation

## Warning: package 'equatiomatic' was built under R version 4.0.3

library(caret)  
library(patchwork) # to put some plots togather

## Warning: package 'patchwork' was built under R version 4.0.3

require(ggiraph)

## Warning: package 'ggiraph' was built under R version 4.0.3

require(ggiraphExtra)

### Data Installation

#### ASER2016

provdist <- read.csv("aser/ASER2016ProvDist.csv")  
school <- read.csv("aser/ASER2016GSchool.csv")  
child <- read.csv("aser/ASER2016Child.csv")  
pschool <- read.csv("aser/ASER2016PvtSchool.csv")  
gschool <- read.csv("aser/ASER2016GSchool.csv")  
parent <- read.csv("aser/ASER2016Parent.csv")  
house <- read.csv("aser/ASER2016HouseholdIndicators.csv")

RegionName <- c("2" = "Panjab",   
 "3" = "Sindh",   
 "4" = "Balochistan",   
 "5" = "Khyber Pakhtunkhwa",   
 "6" = "Gilgit-Baltistan",   
 "7" = "Azad Jammu and Kashmir",   
 "8" = "Islamabad - ICT",   
 "9" = "Federally Administrated Tribal Areas")  
Gender <- c("0" = "Male",  
 "-1" = "Female")

#### Spatial Data

ica <- sf::st\_read("map/pak\_ica\_categories\_areas\_geonode\_apr2017.shp")

## Reading layer `pak\_ica\_categories\_areas\_geonode\_apr2017' from data source `C:\Program Files\R\capstone\map\pak\_ica\_categories\_areas\_geonode\_apr2017.shp' using driver `ESRI Shapefile'  
## Simple feature collection with 156 features and 8 fields  
## geometry type: MULTIPOLYGON  
## dimension: XY  
## bbox: xmin: 60.8786 ymin: 23.69468 xmax: 77.83397 ymax: 37.08942  
## geographic CRS: WGS 84

### Data Wrangling

ica\_df <- ica %>%   
 mutate(centroid = st\_centroid(geometry),  
 x = st\_coordinates(centroid)[,1],  
 y = st\_coordinates(centroid)[,2]) %>%   
 as.data.frame()

## Warning: Problem with `mutate()` input `centroid`.  
## x st\_centroid does not give correct centroids for longitude/latitude data  
## i Input `centroid` is `st\_centroid(geometry)`.

## Warning in st\_centroid.sfc(geometry): st\_centroid does not give correct  
## centroids for longitude/latitude data

ica\_df <- ica\_df %>% select(Province, Districts, x, y)  
ica\_df <- ica\_df %>% summarize(Province = tolower(Province), Districts = tolower(Districts), x = x, y = y)

#### Child and ProvDist data combination

child\_dname <- child %>% left\_join(provdist[-1])

## Joining, by = "DID"

child\_dname <- child\_dname %>% mutate(dname = tolower(DNAME))

ica\_df\_3 <- ica\_df %>% filter(Province == "sindh")  
  
ica\_df\_3$Districts <- ica\_df\_3$Districts %>%   
 str\_replace("ghotki", "gotki") %>%  
 str\_replace("mirpur khas", "mirpurkhas") %>%   
 str\_replace("malir karachi", "karachi-malir-rural") %>%   
 str\_replace("naushahro feroze", "nowshero feroze") %>%   
 str\_replace("kambar shahdad kot", "qambar shahdadkot") %>%   
 str\_replace("sujawal", "sajawal") %>%   
 str\_replace("shaheed benazir abad", "shaheed benazirabad") %>%   
 str\_replace("tando allahyar", "tando allah yar") %>%   
 as.vector()  
  
child\_dname\_3 <- child\_dname %>% filter(RNAME == "Sindh") %>% left\_join(ica\_df\_3, by = c("dname" = "Districts"))  
  
child\_dname\_3 %>% group\_by(dname) %>% summarize(n = sum(x))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 25 x 2  
## dname n  
## <chr> <dbl>  
## 1 badin 132177.  
## 2 dadu 130944.  
## 3 gotki 138955.  
## 4 hyderabad 126370.  
## 5 jacobabad 113124.  
## 6 jamshoro 95520.  
## 7 karachi-malir-rural 75832.  
## 8 karachi-west-rural NA   
## 9 kashmore 132646.  
## 10 khairpur 107834.  
## # ... with 15 more rows

ica\_df\_3

## Province Districts x y  
## 1 sindh badin 68.84219 24.72277  
## 2 sindh central karachi 67.05813 24.94777  
## 3 sindh dadu 67.49687 26.86763  
## 4 sindh east karachi 67.13429 24.93583  
## 5 sindh gotki 69.65163 27.82827  
## 6 sindh hyderabad 68.45633 25.34110  
## 7 sindh jacobabad 68.47705 28.20721  
## 8 sindh jamshoro 67.79244 25.73650  
## 9 sindh qambar shahdadkot 67.71121 27.63562  
## 10 sindh kashmore 69.23065 28.27468  
## 11 sindh khairpur 69.08038 26.82907  
## 12 sindh korangi karachi 67.14997 24.84413  
## 13 sindh larkana 68.19492 27.52356  
## 14 sindh karachi-malir-rural 67.28693 25.11444  
## 15 sindh matiari 68.45326 25.76496  
## 16 sindh mirpurkhas 69.16431 25.34044  
## 17 sindh nowshero feroze 68.12521 26.87884  
## 18 sindh sanghar 69.29576 25.95919  
## 19 sindh shaheed benazirabad 68.33324 26.36101  
## 20 sindh shikarpur 68.60669 27.94883  
## 21 sindh south karachi 66.87388 24.89373  
## 22 sindh sajawal 68.15257 24.24510  
## 23 sindh sukkur 69.17648 27.50249  
## 24 sindh tando allah yar 68.77546 25.46575  
## 25 sindh tando muhammad khan 68.50212 25.00184  
## 26 sindh tharparkar 70.17772 24.78158  
## 27 sindh thatta 67.76193 24.73303  
## 28 sindh umer kot 69.77834 25.38386  
## 29 sindh west karachi 67.01128 24.99852

remove(ica\_df\_2, ica\_df\_3, ica\_df\_4,  
 ica\_df\_5, ica\_df\_6, ica\_df\_7,  
 ica\_df\_8, ica\_df\_9)

remove(child\_dname\_2, child\_dname\_3, child\_dname\_4,  
 child\_dname\_5, child\_dname\_6, child\_dname\_7,  
 child\_dname\_8, child\_dname\_9)

### Preparation for Logistic Regression Analysis

#### Making Dataframe With Dummy Variables

child\_ica\_dummy <- child\_ica %>% filter(!is.na(C002), !is.na(C003), !is.na(PR004), !is.na(PR009), !is.na(H003))  
  
child\_ica\_dummy$Age <- child\_ica\_dummy$C001  
child\_ica\_dummy$Gender <- ifelse(child\_ica\_dummy$C002 == -1, 1, 0)  
child\_ica\_dummy$School.Enrollment <- ifelse(child\_ica\_dummy$C003 == 3, 1, 0) # currently-enrolled  
child\_ica\_dummy$C003\_1\_01 <- ifelse(child\_ica\_dummy$C003 == 1, 1, 0) # never-enrolled  
child\_ica\_dummy$C003\_2\_01 <- ifelse(child\_ica\_dummy$C003 == 2, 1, 0) # drop-out  
child\_ica\_dummy$PR004\_01 <- ifelse(child\_ica\_dummy$PR004 == -1, 1, 0)  
child\_ica\_dummy$PR009\_01 <- ifelse(child\_ica\_dummy$PR009 == -1, 1, 0)  
child\_ica\_dummy$Parents.Edu1 <- ifelse(  
 child\_ica\_dummy$PR004 == -1 | child\_ica\_dummy$PR009 == -1, 1, 0)  
  
  
child\_ica\_dummy$PR004\_only\_01 <- ifelse(  
 child\_ica\_dummy$PR004 == -1 & child\_ica\_dummy$PR009 == 0, 1, 0)  
child\_ica\_dummy$PR009\_only\_01 <- ifelse(  
 child\_ica\_dummy$PR004 == 0 & child\_ica\_dummy$PR009 == -1, 1, 0)  
child\_ica\_dummy$Parents.Edu2 <- ifelse(  
 child\_ica\_dummy$PR004 == -1 & child\_ica\_dummy$PR009 == -1, 1, 0)  
  
child\_ica\_dummy$Household.Ownership <- ifelse(  
 child\_ica\_dummy$H003 == -1, 1, 0)

Num.of.Child <- child\_ica\_dummy %>%   
 group\_by(HHID) %>%   
 summarize(Num.of.Child = length(unique(CID)))

## `summarise()` ungrouping output (override with `.groups` argument)

child\_ica\_dummy <- child\_ica\_dummy %>% left\_join(Num.of.Child)

## Joining, by = "HHID"

child\_ica\_dummy$H002\_1\_01 <- ifelse(child\_ica\_dummy$H002 == 1, 1, 0)  
child\_ica\_dummy$H002\_2\_01 <- ifelse(child\_ica\_dummy$H002 == 2, 1, 0)  
child\_ica\_dummy$H002\_3\_01 <- ifelse(child\_ica\_dummy$H002 == 3, 1, 0)

#### Regional Dummy Variables

child\_ica\_dummy <- child\_ica\_dummy  
child\_ica\_dummy$Panjab <- ifelse(child\_ica\_dummy$RID == 2, 1, 0)  
child\_ica\_dummy$Sindh <- ifelse(child\_ica\_dummy$RID == 3, 1, 0)  
child\_ica\_dummy$Balochistan <- ifelse(child\_ica\_dummy$RID == 4, 1, 0)  
child\_ica\_dummy$Khyber\_Pakhtunkhwa <- ifelse(child\_ica\_dummy$RID == 5, 1, 0)  
child\_ica\_dummy$Gilgit\_Baltistan <- ifelse(child\_ica\_dummy$RID == 6, 1, 0)  
child\_ica\_dummy$Azad\_Jammu\_and\_Kashmir <- ifelse(child\_ica\_dummy$RID == 7, 1, 0)  
child\_ica\_dummy$Islamabad\_ICT <- ifelse(child\_ica\_dummy$RID == 8, 1, 0)  
child\_ica\_dummy$Federally\_Administrated\_Tribal\_Areas <- ifelse(child\_ica\_dummy$RID == 9, 1, 0)

#### Factor Variables

child\_ica\_dummy$DID <- as.factor(child\_ica\_dummy$DID)  
child\_ica\_dummy$C002 <- as.factor(child\_ica\_dummy$C002)  
child\_ica\_dummy$C003 <- as.factor(child\_ica\_dummy$C003)  
# child\_ica\_dummy$School.Enrollment <- as.factor(child\_ica\_dummy$School.Enrollment)  
child\_ica\_dummy$Household.Ownership <- as.factor(child\_ica\_dummy$Household.Ownership)  
child\_ica\_dummy$Parents.Edu1 <- as.factor(child\_ica\_dummy$Parents.Edu1)  
child\_ica\_dummy$Parents.Edu2 <- as.factor(child\_ica\_dummy$Parents.Edu2)  
child\_ica\_dummy$Gender <- as.factor(child\_ica\_dummy$Gender)

### Notes on Data Wrangling

#### Eliminated NAs

child\_ica %>%   
 summarize(  
 School.Enrollment\_NA = sum(is.na(C003)),  
 Gender\_NA = sum(is.na(C002)),  
 Parents.Edu\_NA = sum(is.na(PR004))+sum(is.na(PR009)),  
 Household.Ownership\_NA = sum(is.na(H003))  
 )

## School.Enrollment\_NA Gender\_NA Parents.Edu\_NA Household.Ownership\_NA  
## 1 0 66 10775 4500

#### Eliminated Rows in Total

data.frame(original\_rows = nrow(child\_ica),  
 eliminated\_rows = nrow(child\_ica) - nrow(child\_ica\_dummy),  
 ratio = (nrow(child\_ica)-nrow(child\_ica\_dummy))/nrow(child\_ica))

## original\_rows eliminated\_rows ratio  
## 1 255196 13218 0.05179548

child\_ica %>%  
 summarize(  
 "Variables" = c(  
 "School.Enrollment\_NA",   
 "Gender\_NA",   
 "Parents.Edu\_NA",   
 "Household.Ownership\_NA",  
 "Total"),   
 "Deleted Rows" = c(  
 sum(is.na(C003)),   
 sum(is.na(C002)),   
 sum(is.na(PR004))+sum(is.na(PR009)),   
 sum(is.na(H003)),  
 nrow(child\_ica %>% filter(is.na(C003) | is.na(C002) | is.na(PR004) | is.na(PR009) | is.na(H003)))),  
 "Ratio" = c(  
 sum(is.na(C003))/nrow(child\_ica),  
 sum(is.na(C002))/nrow(child\_ica),   
 (sum(is.na(PR004))+sum(is.na(PR009)))/nrow(child\_ica),   
 sum(is.na(H003))/nrow(child\_ica),  
 nrow(child\_ica %>% filter(is.na(C003) | is.na(C002) | is.na(PR004) | is.na(PR009) | is.na(H003)))/nrow(child\_ica))  
 )

## Variables Deleted Rows Ratio  
## 1 School.Enrollment\_NA 0 0.0000000000  
## 2 Gender\_NA 66 0.0002586247  
## 3 Parents.Edu\_NA 10775 0.0422224486  
## 4 Household.Ownership\_NA 4500 0.0176335052  
## 5 Total 13218 0.0517954827

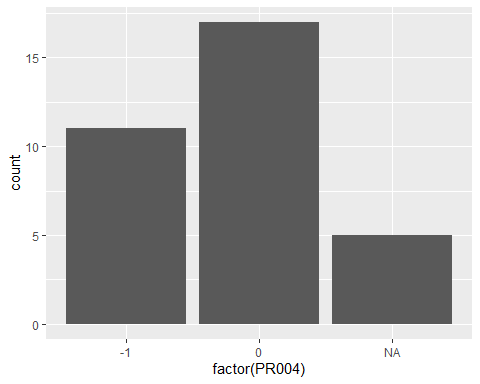
#### Eliminated NAs Hunza

child\_ica %>%   
 filter(DID == 266) %>%   
 summarize(  
 "Variables" = c(  
 "School.Enrollment\_NA",   
 "Gender\_NA",   
 "Parents.Edu\_NA",   
 "Household.Ownership\_NA",  
 "Total"),   
 "Deleted Rows" = c(  
 sum(is.na(C003)),   
 sum(is.na(C002)),   
 sum(is.na(PR004))+sum(is.na(PR009)),   
 sum(is.na(H003)),  
 nrow(child\_ica %>% filter(DID == 266, is.na(C003) | is.na(C002) | is.na(PR004) | is.na(PR009) | is.na(H003)))),  
 "Ratio" = c(  
 sum(is.na(C003))/nrow(child\_ica %>% filter(DID ==266)),  
 sum(is.na(C002))/nrow(child\_ica %>% filter(DID ==266)),   
 (sum(is.na(PR004))+sum(is.na(PR009)))/nrow(child\_ica %>% filter(DID ==266)),   
 sum(is.na(H003))/nrow(child\_ica %>% filter(DID ==266)),  
 nrow(child\_ica %>% filter(DID == 266, is.na(C003) | is.na(C002) | is.na(PR004) | is.na(PR009) | is.na(H003)))/nrow(child\_ica %>% filter(DID ==266)))  
 )

## Variables Deleted Rows Ratio  
## 1 School.Enrollment\_NA 0 0.00000000  
## 2 Gender\_NA 0 0.00000000  
## 3 Parents.Edu\_NA 33 0.02010969  
## 4 Household.Ownership\_NA 33 0.02010969  
## 5 Total 59 0.03595369

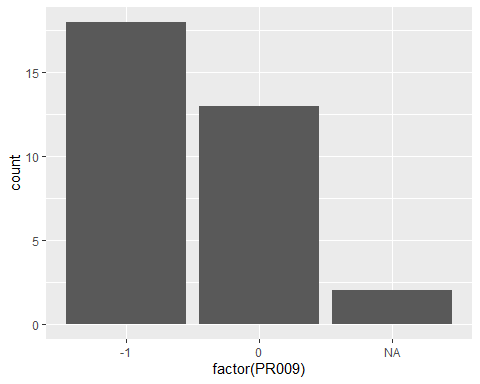
child\_ica %>%   
 filter(DID == 266, is.na(H003)) %>%   
 ggplot(aes(factor(PR004))) +  
 geom\_histogram(stat = "count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



child\_ica %>%   
 filter(DID == 266, is.na(H003)) %>%   
 ggplot(aes(factor(PR009))) +  
 geom\_histogram(stat = "count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



#### Eliminated Rows in Total Hunza

data.frame(original\_rows = nrow(child\_ica %>% filter(DID == 266)),  
 eliminated\_rows = nrow(child\_ica %>% filter(DID == 266)) - nrow(child\_ica\_dummy %>% filter(DID == 266)),  
 ratio = (nrow(child\_ica %>% filter(DID == 266) %>% filter(DID == 266))-nrow(child\_ica\_dummy %>% filter(DID == 266)))/nrow(child\_ica %>% filter(DID == 266)))

## original\_rows eliminated\_rows ratio  
## 1 1641 59 0.03595369

## Hunza. Generalized Linear Models

### Formulae

### Model 1 [NofChild + HHOwnership]

#### Hunza. GLM Age >= 5 School.Enrollment ~ Age + Gender + Parents.Edu1

glm\_child <- glm(School.Enrollment ~ Age + Gender + Parents.Edu1, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, !is.na(Household.Ownership), DID == 266))  
temp <- glm\_child %>% summary()  
temp

##   
## Call:  
## glm(formula = School.Enrollment ~ Age + Gender + Parents.Edu1,   
## family = "binomial", data = child\_ica\_dummy %>% filter(C001 >=   
## 5, !is.na(Household.Ownership), DID == 266))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8774 0.1900 0.2034 0.2199 0.4033   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.78379 0.61250 4.545 5.49e-06 \*\*\*  
## Age -0.01973 0.04690 -0.421 0.674031   
## Gender1 0.21597 0.31581 0.684 0.494052   
## Parents.Edu11 1.22242 0.31976 3.823 0.000132 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 376.54 on 1387 degrees of freedom  
## Residual deviance: 361.23 on 1384 degrees of freedom  
## AIC: 369.23  
##   
## Number of Fisher Scoring iterations: 6

temp$coefficients %>% as.data.frame() %>% pull(`Pr(>|z|)`) %>% .[3]

## [1] 0.4940516

# ggPredict(glm\_child, se = TRUE, colorAsFactor = TRUE, show.summary = TRUE, point = TRUE) +  
# theme(text = element\_text(family = "serif")) +  
# xlab("Number of Children in Each House Hold") +  
# ylab("Enrollment") +  
# labs(color = "House Hold Ownership", fill = "House Hold Ownership")

extract\_eq(glm\_child)

extract\_eq(glm\_child, use\_coefs = TRUE)

##### exponential transformation

exp(glm\_child$coefficients)

## (Intercept) Age Gender1 Parents.Edu11   
## 16.1802176 0.9804678 1.2410692 3.3954058

##### confidence interval (intercept and coefficient)

confint(glm\_child, level = 0.95)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 1.6272809 4.03584453  
## Age -0.1129899 0.07166459  
## Gender1 -0.4045039 0.84225180  
## Parents.Edu11 0.5915790 1.85347528

##### exponential transformation of confidence interval (odds ratio of intercept and coefficient)

exp(confint(glm\_child, level = 0.95))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 5.0900154 56.590692  
## Age 0.8931597 1.074295  
## Gender1 0.6673078 2.321589  
## Parents.Edu11 1.8068392 6.381960

##### Hosmer-Lemeshow

hoslem.test(x = glm\_child$y, y = fitted(glm\_child), g = 10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: glm\_child$y, fitted(glm\_child)  
## X-squared = 5.2282, df = 8, p-value = 0.7329

##### AIC

extractAIC(glm\_child)

## [1] 4.0000 369.2321

##### BIC

extractAIC(glm\_child, k = log(nrow(glm\_child$data)))

## [1] 4.0000 390.1746

##### effectiveness of explanatory variables

glm\_child\_null <- glm(School.Enrollment ~ 1, family = "binomial",   
 data = child\_ica\_dummy %>% filter(!is.na(Household.Ownership), C001 >= 5, DID == 266))  
anova(glm\_child\_null, glm\_child, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: School.Enrollment ~ 1  
## Model 2: School.Enrollment ~ Age + Gender + Parents.Edu1  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 1387 376.54   
## 2 1384 361.23 3 15.312 0.001569 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### variables selection

step(glm\_child\_null, direction = "both",  
 scope = (~ Age + Gender + Num.of.Child + Parents.Edu1 + Household.Ownership))

## Start: AIC=378.54  
## School.Enrollment ~ 1  
##   
## Df Deviance AIC  
## + Household.Ownership 1 357.40 361.40  
## + Num.of.Child 1 357.92 361.92  
## + Parents.Edu1 1 361.88 365.88  
## <none> 376.54 378.54  
## + Age 1 375.57 379.57  
## + Gender 1 376.15 380.15  
##   
## Step: AIC=361.4  
## School.Enrollment ~ Household.Ownership  
##   
## Df Deviance AIC  
## + Num.of.Child 1 342.04 348.04  
## + Parents.Edu1 1 347.54 353.54  
## <none> 357.40 361.40  
## + Age 1 356.13 362.13  
## + Gender 1 357.22 363.22  
## - Household.Ownership 1 376.54 378.54  
##   
## Step: AIC=348.04  
## School.Enrollment ~ Household.Ownership + Num.of.Child  
##   
## Df Deviance AIC  
## + Parents.Edu1 1 338.06 346.06  
## <none> 342.04 348.04  
## + Gender 1 341.38 349.38  
## + Age 1 341.47 349.47  
## - Num.of.Child 1 357.40 361.40  
## - Household.Ownership 1 357.92 361.92  
##   
## Step: AIC=346.06  
## School.Enrollment ~ Household.Ownership + Num.of.Child + Parents.Edu1  
##   
## Df Deviance AIC  
## <none> 338.06 346.06  
## + Gender 1 337.44 347.44  
## + Age 1 337.85 347.85  
## - Parents.Edu1 1 342.04 348.04  
## - Num.of.Child 1 347.54 353.54  
## - Household.Ownership 1 350.58 356.58

##   
## Call: glm(formula = School.Enrollment ~ Household.Ownership + Num.of.Child +   
## Parents.Edu1, family = "binomial", data = child\_ica\_dummy %>%   
## filter(!is.na(Household.Ownership), C001 >= 5, DID == 266))  
##   
## Coefficients:  
## (Intercept) Household.Ownership1 Num.of.Child   
## 3.1253 1.4707 -0.3562   
## Parents.Edu11   
## 0.7103   
##   
## Degrees of Freedom: 1387 Total (i.e. Null); 1384 Residual  
## Null Deviance: 376.5   
## Residual Deviance: 338.1 AIC: 346.1

### Model 2 [NofChild + HHOwnership]

#### Hunza. GLM Age >= 5 School.Enrollment ~ Age + Gender + Num.of.Child + Household.Ownership

glm\_child <- glm(School.Enrollment ~ Age + Gender + Num.of.Child + Household.Ownership, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, !is.na(Household.Ownership), DID == 266))  
glm\_child %>% summary()

##   
## Call:  
## glm(formula = School.Enrollment ~ Age + Gender + Num.of.Child +   
## Household.Ownership, family = "binomial", data = child\_ica\_dummy %>%   
## filter(C001 >= 5, !is.na(Household.Ownership), DID == 266))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0509 0.1585 0.1954 0.2496 0.9613   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.05970 0.76684 5.294 1.20e-07 \*\*\*  
## Age -0.03707 0.04695 -0.790 0.430   
## Gender1 0.27406 0.32403 0.846 0.398   
## Num.of.Child -0.44037 0.11207 -3.930 8.51e-05 \*\*\*  
## Household.Ownership1 1.61823 0.37005 4.373 1.23e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 376.54 on 1387 degrees of freedom  
## Residual deviance: 340.76 on 1383 degrees of freedom  
## AIC: 350.76  
##   
## Number of Fisher Scoring iterations: 7

# ggPredict(glm\_child, se = TRUE, colorAsFactor = TRUE, show.summary = TRUE, point = TRUE) +  
# theme(text = element\_text(family = "serif")) +  
# xlab("Number of Children in Each House Hold") +  
# ylab("Enrollment") +  
# labs(color = "House Hold Ownership", fill = "House Hold Ownership")

extract\_eq(glm\_child)

extract\_eq(glm\_child, use\_coefs = TRUE)

##### exponential transformation

exp(glm\_child$coefficients)

## (Intercept) Age Gender1   
## 57.9571028 0.9636089 1.3152993   
## Num.of.Child Household.Ownership1   
## 0.6437983 5.0441582

##### confidence interval (intercept and coefficient)

confint(glm\_child, level = 0.95)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 2.6141836 5.6269705  
## Age -0.1304561 0.0544374  
## Gender1 -0.3620688 0.9166426  
## Num.of.Child -0.6611333 -0.2203641  
## Household.Ownership1 0.8585333 2.3205211

##### exponential transformation of confidence interval (odds ratio of intercept and coefficient)

exp(confint(glm\_child, level = 0.95))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 13.6560628 277.8191819  
## Age 0.8776950 1.0559464  
## Gender1 0.6962345 2.5008798  
## Num.of.Child 0.5162659 0.8022266  
## Household.Ownership1 2.3596972 10.1809787

##### Hosmer-Lemeshow

hoslem.test(x = glm\_child$y, y = fitted(glm\_child))

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: glm\_child$y, fitted(glm\_child)  
## X-squared = 11.36, df = 8, p-value = 0.1821

##### AIC

extractAIC(glm\_child)

## [1] 5.0000 350.7572

##### BIC

extractAIC(glm\_child, k = log(nrow(glm\_child$data)))

## [1] 5.0000 376.9353

##### effectiveness of explanatory variables

glm\_child\_null <- glm(School.Enrollment ~ 1, family = "binomial",   
 data = child\_ica\_dummy %>% filter(!is.na(Household.Ownership), C001 >= 5, DID == 266))  
anova(glm\_child\_null, glm\_child, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: School.Enrollment ~ 1  
## Model 2: School.Enrollment ~ Age + Gender + Num.of.Child + Household.Ownership  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 1387 376.54   
## 2 1383 340.76 4 35.787 3.201e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

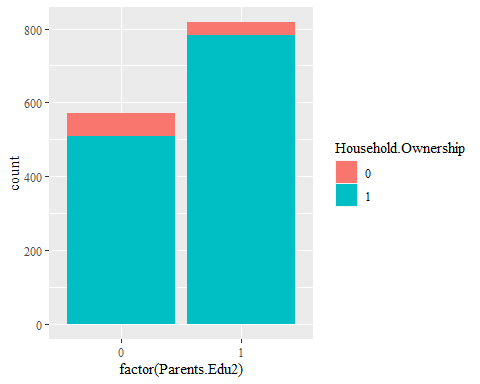
##### variables selection

step(glm\_child\_null, direction = "both",  
 scope = (~ Age + Gender + Num.of.Child + Parents.Edu1 + Household.Ownership))

## Start: AIC=378.54  
## School.Enrollment ~ 1  
##   
## Df Deviance AIC  
## + Household.Ownership 1 357.40 361.40  
## + Num.of.Child 1 357.92 361.92  
## + Parents.Edu1 1 361.88 365.88  
## <none> 376.54 378.54  
## + Age 1 375.57 379.57  
## + Gender 1 376.15 380.15  
##   
## Step: AIC=361.4  
## School.Enrollment ~ Household.Ownership  
##   
## Df Deviance AIC  
## + Num.of.Child 1 342.04 348.04  
## + Parents.Edu1 1 347.54 353.54  
## <none> 357.40 361.40  
## + Age 1 356.13 362.13  
## + Gender 1 357.22 363.22  
## - Household.Ownership 1 376.54 378.54  
##   
## Step: AIC=348.04  
## School.Enrollment ~ Household.Ownership + Num.of.Child  
##   
## Df Deviance AIC  
## + Parents.Edu1 1 338.06 346.06  
## <none> 342.04 348.04  
## + Gender 1 341.38 349.38  
## + Age 1 341.47 349.47  
## - Num.of.Child 1 357.40 361.40  
## - Household.Ownership 1 357.92 361.92  
##   
## Step: AIC=346.06  
## School.Enrollment ~ Household.Ownership + Num.of.Child + Parents.Edu1  
##   
## Df Deviance AIC  
## <none> 338.06 346.06  
## + Gender 1 337.44 347.44  
## + Age 1 337.85 347.85  
## - Parents.Edu1 1 342.04 348.04  
## - Num.of.Child 1 347.54 353.54  
## - Household.Ownership 1 350.58 356.58

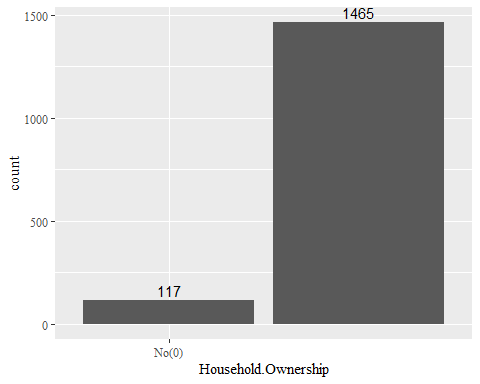
##   
## Call: glm(formula = School.Enrollment ~ Household.Ownership + Num.of.Child +   
## Parents.Edu1, family = "binomial", data = child\_ica\_dummy %>%   
## filter(!is.na(Household.Ownership), C001 >= 5, DID == 266))  
##   
## Coefficients:  
## (Intercept) Household.Ownership1 Num.of.Child   
## 3.1253 1.4707 -0.3562   
## Parents.Edu11   
## 0.7103   
##   
## Degrees of Freedom: 1387 Total (i.e. Null); 1384 Residual  
## Null Deviance: 376.5   
## Residual Deviance: 338.1 AIC: 346.1

child\_ica\_dummy %>%   
 filter(!is.na(Household.Ownership), Age >= 5, DID == 266) %>%   
 group\_by(Household.Ownership) %>%   
 ggplot(aes(factor(Parents.Edu2), fill = Household.Ownership)) +  
 geom\_bar(stat = "count") +  
 theme(text = element\_text(family = "serif"))

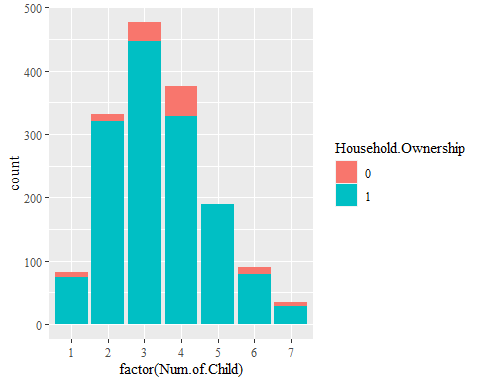


There seems multi-colinearity between Household Ownership and Both-Parents Education

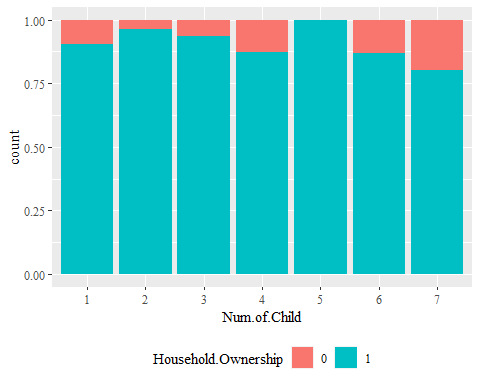
child\_ica\_dummy %>%   
 filter(DID == 266) %>%   
 ggplot(aes(Household.Ownership)) +  
 geom\_bar() +  
 geom\_text(aes(label = ..count..), stat = "count", vjust = -.3) +  
 scale\_x\_discrete(breaks = c(-1, 0, NA), labels = c("Yes(-1)", "No(0)", "NA")) +  
 theme(text = element\_text(family = "serif"))



child\_ica\_dummy %>%   
 filter(DID == 266) %>%   
 ggplot(aes(factor(Num.of.Child), fill = Household.Ownership)) +  
 geom\_bar() +  
 theme(text = element\_text(family = "serif"))



plot\_3 <- child\_ica\_dummy %>%   
 filter(DID == 266, !is.na(Household.Ownership)) %>%   
 ggplot(aes(factor(Num.of.Child), fill = Household.Ownership)) +  
 geom\_bar(stat = "count", position = "fill") +  
 theme(text = element\_text(family = "serif")) +  
 xlab("Num.of.Child") +  
 theme(legend.position = "bottom")  
plot\_3



## Pak. Generalized Linear Model

### Model 1

#### Each Dist. GLM Age >= 5 School.Enrollment ~ Age + Gender + Num.of.Child + Household.Ownership

each\_dist <- sapply(as.numeric(as.character(unique(child\_ica\_dummy$DID))), function(id){  
 child\_ica\_dummy <- child\_ica\_dummy %>% filter(DID == id)  
 glm\_child <- glm(School.Enrollment ~ Age + Gender + Num.of.Child + Household.Ownership,   
 family = "binomial", data = child\_ica\_dummy %>% filter(Age >= 5, !is.na(Household.Ownership)))  
 temp <- glm\_child %>% summary()  
 gender\_p <- temp$coefficients %>% as.data.frame() %>% pull(`Pr(>|z|)`) %>% .[3]  
 data.frame(DID = id, gender\_p = gender\_p)  
}) %>% t()

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

each\_dist <- each\_dist %>%   
 as.data.frame()   
each\_dist$DID <- each\_dist %>%   
 pull(DID) %>%   
 as.numeric()  
each\_dist$gender\_p <- each\_dist %>%   
 pull(gender\_p) %>%   
 as.numeric()  
each\_dist

## DID gender\_p  
## 1 146 1.152199e-01  
## 2 147 2.844143e-05  
## 3 148 1.838574e-01  
## 4 149 2.157592e-14  
## 5 150 7.968106e-02  
## 6 151 2.767307e-03  
## 7 152 8.921458e-06  
## 8 153 1.748082e-04  
## 9 154 2.613560e-01  
## 10 155 1.110949e-06  
## 11 156 2.124136e-01  
## 12 157 6.768599e-04  
## 13 158 9.574338e-03  
## 14 159 1.474403e-01  
## 15 160 2.011791e-02  
## 16 161 6.942851e-02  
## 17 162 2.781561e-01  
## 18 163 1.690200e-01  
## 19 164 4.323676e-01  
## 20 165 3.663836e-05  
## 21 166 1.123763e-02  
## 22 167 8.575475e-02  
## 23 169 7.563703e-14  
## 24 170 6.864079e-05  
## 25 171 9.112118e-01  
## 26 172 2.838786e-03  
## 27 173 9.006032e-01  
## 28 174 1.423419e-08  
## 29 175 5.125052e-07  
## 30 176 1.099576e-02  
## 31 177 3.845881e-08  
## 32 178 9.364775e-01  
## 33 179 5.988359e-03  
## 34 180 4.602131e-02  
## 35 181 8.405621e-02  
## 36 182 4.581134e-06  
## 37 183 5.490538e-04  
## 38 184 8.981879e-26  
## 39 185 1.193554e-01  
## 40 186 5.300262e-31  
## 41 187 1.211660e-13  
## 42 188 1.138239e-09  
## 43 189 2.683444e-03  
## 44 190 1.276993e-07  
## 45 191 6.638645e-10  
## 46 192 2.155260e-07  
## 47 193 1.597974e-01  
## 48 194 7.327253e-13  
## 49 195 8.495331e-22  
## 50 196 6.378021e-04  
## 51 197 6.611147e-03  
## 52 198 8.096264e-14  
## 53 199 1.238056e-01  
## 54 200 2.127809e-12  
## 55 202 1.191718e-24  
## 56 203 2.639115e-06  
## 57 204 7.436311e-04  
## 58 315 4.512939e-01  
## 59 316 1.250442e-02  
## 60 320 3.556029e-03  
## 61 205 1.325419e-24  
## 62 206 4.831684e-14  
## 63 207 2.483859e-14  
## 64 208 5.682013e-19  
## 65 209 1.032982e-19  
## 66 210 8.754268e-28  
## 67 211 1.318612e-04  
## 68 212 5.499238e-49  
## 69 213 2.311793e-28  
## 70 214 1.701341e-30  
## 71 215 7.205465e-08  
## 72 216 4.387150e-42  
## 73 217 3.923845e-36  
## 74 218 1.211821e-24  
## 75 219 2.006133e-12  
## 76 220 5.260703e-30  
## 77 221 1.536115e-15  
## 78 222 4.046673e-39  
## 79 223 2.565042e-04  
## 80 224 9.604156e-44  
## 81 225 1.336385e-11  
## 82 226 2.250309e-02  
## 83 227 2.565280e-17  
## 84 228 1.989928e-21  
## 85 229 4.272250e-07  
## 86 230 1.793333e-29  
## 87 231 5.110399e-39  
## 88 232 2.190714e-02  
## 89 233 1.062736e-06  
## 90 234 2.982213e-05  
## 91 318 9.591080e-05  
## 92 319 2.560452e-16  
## 93 235 4.209184e-11  
## 94 236 4.624229e-04  
## 95 237 1.460199e-05  
## 96 238 9.374841e-01  
## 97 239 3.844386e-02  
## 98 240 4.076127e-29  
## 99 241 4.190512e-07  
## 100 242 3.734747e-04  
## 101 243 8.619540e-01  
## 102 244 1.250597e-05  
## 103 245 2.375571e-01  
## 104 246 9.694425e-04  
## 105 247 2.101667e-12  
## 106 248 3.228014e-39  
## 107 249 3.337582e-02  
## 108 250 3.159246e-16  
## 109 251 2.323006e-05  
## 110 252 4.220670e-24  
## 111 253 1.031579e-04  
## 112 254 5.441949e-19  
## 113 255 5.361938e-14  
## 114 256 1.874575e-08  
## 115 257 9.500591e-01  
## 116 258 1.388522e-74  
## 117 259 8.979378e-02  
## 118 260 6.783308e-02  
## 119 261 5.433576e-102  
## 120 262 9.890232e-01  
## 121 263 9.715313e-03  
## 122 264 7.217898e-03  
## 123 265 9.273759e-01  
## 124 266 3.976686e-01  
## 125 267 3.241365e-01  
## 126 268 9.962323e-03  
## 127 269 5.077509e-01  
## 128 270 3.527810e-01  
## 129 271 1.163825e-01  
## 130 272 1.702130e-01  
## 131 273 5.552924e-01  
## 132 274 3.673593e-01  
## 133 275 2.419437e-01  
## 134 276 3.238612e-01  
## 135 277 1.760644e-01  
## 136 278 2.131649e-55  
## 137 279 1.963788e-47  
## 138 280 1.655510e-30  
## 139 281 9.822463e-21  
## 140 282 1.092476e-07  
## 141 284 1.557763e-09  
## 142 287 4.137354e-18  
## 143 289 1.183242e-64  
## 144 290 5.897799e-28

#### Dists with P value <= 0.05

DID\_gender\_P0.05 <- each\_dist %>%   
 filter(gender\_p >= 0.05) %>%   
 pull(DID)

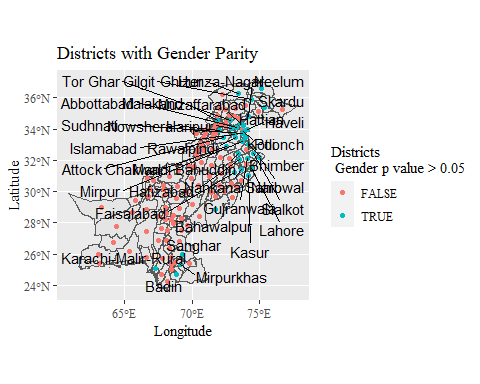
ica %>%   
 mutate(centroid = st\_centroid(geometry),  
 x = st\_coordinates(centroid)[,1],  
 y = st\_coordinates(centroid)[,2]) %>%   
 ggplot(ratio = 1) +  
 coord\_sf() +  
 geom\_sf(ratio = 1) +  
 geom\_point(data = child\_ica\_dummy %>%   
 filter(!is.na(x)) %>%   
 summarize(x = unique(x),  
 y = unique(y),  
 DID = unique(DID)),  
 aes(x, y, color = DID %in% DID\_gender\_P0.05)) +  
 geom\_text\_repel(data = child\_ica\_dummy %>%   
 filter(DID %in% DID\_gender\_P0.05, !is.na(x)) %>%   
 summarize(x = unique(x),  
 y = unique(y),  
 DID = unique(DID),  
 DNAME = unique(DNAME)),  
 aes(x, y, label = DNAME), force = 2) +  
 ggtitle("Districts with Gender Parity") +  
 labs(color = "Districts \n Gender p value > 0.05") +  
 theme(text = element\_text(family = "serif")) +  
 xlab("Longitude") +  
 ylab("Latitude")

## Warning: Problem with `mutate()` input `centroid`.  
## x st\_centroid does not give correct centroids for longitude/latitude data  
## i Input `centroid` is `st\_centroid(geometry)`.

## Warning in st\_centroid.sfc(geometry): st\_centroid does not give correct  
## centroids for longitude/latitude data

## Warning: Ignoring unknown parameters: ratio

## Coordinate system already present. Adding new coordinate system, which will replace the existing one.



ica %>%   
 mutate(centroid = st\_centroid(geometry),  
 x = st\_coordinates(centroid)[,1],  
 y = st\_coordinates(centroid)[,2]) %>%   
 ggplot(ratio = 1) +  
 coord\_sf() +  
 geom\_sf(ratio = 1, aes(fill = Province)) +  
 scale\_fill\_brewer(palette = "Pastel2", direction = -1, type = qual)+  
 geom\_point(data = child\_ica\_dummy %>%   
 filter(!is.na(x)) %>%   
 summarize(x = unique(x),  
 y = unique(y),  
 DID = unique(DID)),  
 aes(x, y, color = DID %in% DID\_gender\_P0.05)) +  
 scale\_color\_manual(values = c("darkgrey", "red")) +  
 geom\_text\_repel(data = child\_ica\_dummy %>%   
 filter(DID %in% DID\_gender\_P0.05, !is.na(x)) %>%   
 summarize(x = unique(x),  
 y = unique(y),  
 DID = unique(DID),  
 DNAME = unique(DNAME)),  
 aes(x, y, label = DNAME), family = "serif", force = 10, color = "black", label.padding = .1, box.padding = .5) +  
 ggtitle("") +  
 guides(color = FALSE) +  
 theme(text = element\_text(family = "serif"),title = element\_text(family = "serif"),aspect.ratio = 1, legend.position = "bottom" ) +  
 xlab("") +  
 ylab("")

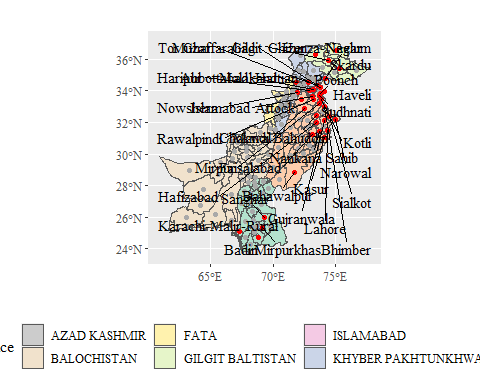
## Warning: Problem with `mutate()` input `centroid`.  
## x st\_centroid does not give correct centroids for longitude/latitude data  
## i Input `centroid` is `st\_centroid(geometry)`.

## Warning in st\_centroid.sfc(geometry): st\_centroid does not give correct  
## centroids for longitude/latitude data

## Warning: Ignoring unknown parameters: ratio

## Coordinate system already present. Adding new coordinate system, which will replace the existing one.

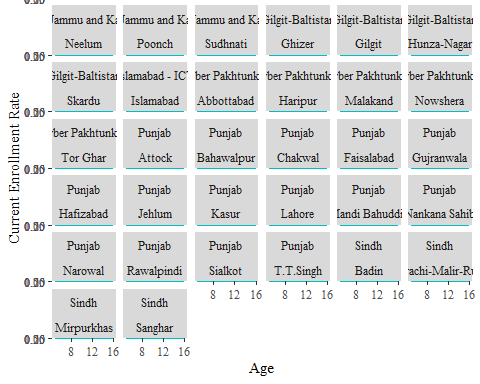
## Warning: Ignoring unknown parameters: label.padding



ggsave("map\_gender.png", unit = "cm", height = 20, dpi = 300)

## Saving 12.7 x 20 cm image

child\_ica\_dummy %>%   
 filter(Age >= 5, DID %in% DID\_gender\_P0.05, !is.na(Household.Ownership)) %>%   
 group\_by(DID, Age, Gender) %>%   
 mutate(rate = mean(School.Enrollment == 1)) %>%   
 ggplot(aes(Age, rate, color = factor(Gender))) +  
 geom\_line() +  
 facet\_wrap(RNAME~DNAME, ncol = 6) +  
 labs(color = "Gender \n 1: Female, 0: Male") +  
 xlab("Age") +  
 ylab("Current Enrollment Rate") +  
 theme(text = element\_text(family = "serif"), legend.position = "bottom")



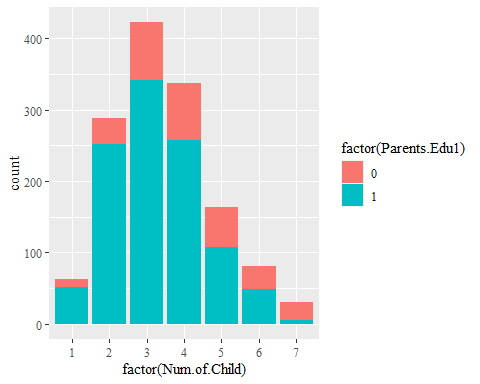
ggsave("dists\_gender.png",height = 22, unit = "cm", dpi = 300)

## Saving 12.7 x 22 cm image

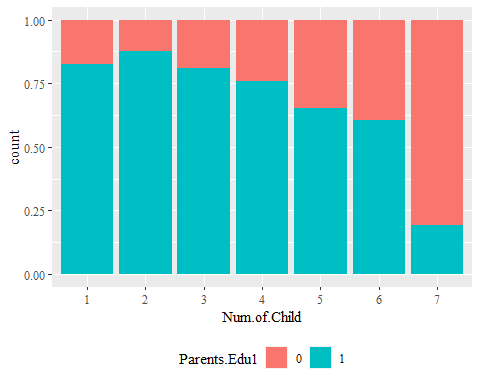
## Appendix

### Parents.Edu1 & Num.of.Child

child\_ica\_dummy %>%   
 filter(DID == 266, Age >= 5) %>%   
 ggplot(aes(factor(Num.of.Child), fill = factor(Parents.Edu1))) +  
 geom\_bar(stat = "count", position = "stack") +  
 theme(text = element\_text(family = "serif"))

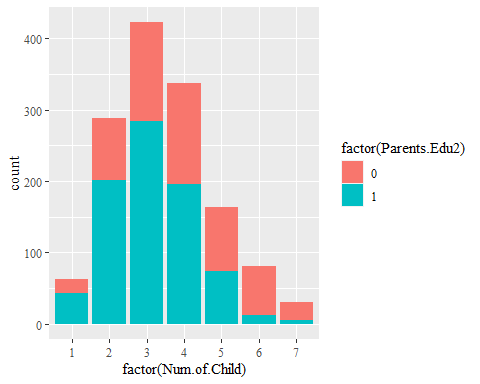


plot\_1 <- child\_ica\_dummy %>%   
 filter(DID == 266, Age >= 5) %>%   
 ggplot(aes(factor(Num.of.Child), fill = factor(Parents.Edu1))) +  
 geom\_bar(stat = "count", position = "fill") +  
 theme(text = element\_text(family = "serif"), legend.position = "bottom") +  
 labs(fill = "Parents.Edu1") +  
 xlab("Num.of.Child")   
plot\_1

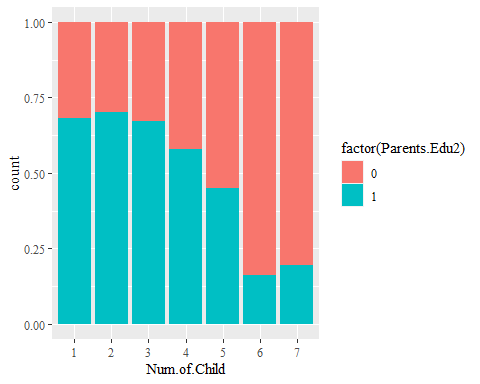


### Parents.Edu2 & Num.of.Child

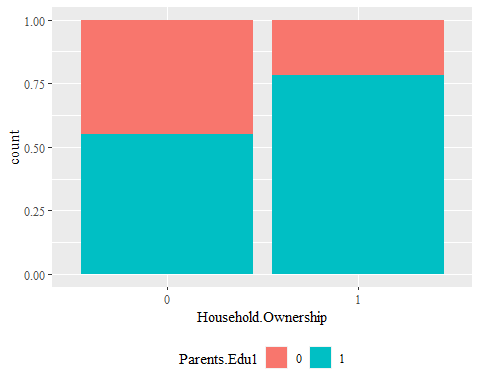
child\_ica\_dummy %>%   
 filter(DID == 266, Age >= 5) %>%   
 ggplot(aes(factor(Num.of.Child), fill = factor(Parents.Edu2))) +  
 geom\_bar(stat = "count", position = "stack") +  
 theme(text = element\_text(family = "serif"))



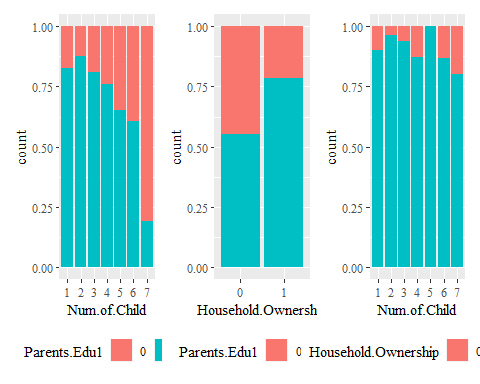
child\_ica\_dummy %>%   
 filter(DID == 266, Age >= 5) %>%   
 ggplot(aes(factor(Num.of.Child), fill = factor(Parents.Edu2))) +  
 geom\_bar(stat = "count", position = "fill") +  
 theme(text = element\_text(family = "serif")) +  
 xlab("Num.of.Child")



plot\_2 <- child\_ica\_dummy %>%   
 filter(DID == 266, Age >= 5) %>%   
 ggplot(aes(Household.Ownership, fill = factor(Parents.Edu1))) +  
 geom\_bar(stat = "count", position = position\_fill()) +  
 theme(text = element\_text(family = "serif"),legend.position = "bottom") +  
 labs(fill = "Parents.Edu1")  
plot\_2



plot\_1 + plot\_2 + plot\_3



ggsave("plot\_multico.png", unit = "cm", dpi = 300, width = 30)

## Saving 30 x 10.2 cm image

# child\_ica\_dummy %>%  
# mosaicplot( ~ School.Enrollment + Gender + Num.of.Child,  
# highlighting = "School.Enrollment", highlighting\_fill = c("lightblue", "pink"),  
# direction = c("v", "h", "h"))

child\_ica\_dummy %>%   
 filter(DID == 266) %>%   
 select(School.Enrollment, Age, Gender) %>%   
 summary()

## School.Enrollment Age Gender   
## Min. :0.0000 Min. : 3.000 0:766   
## 1st Qu.:1.0000 1st Qu.: 6.000 1:816   
## Median :1.0000 Median :10.000   
## Mean :0.9286 Mean : 9.776   
## 3rd Qu.:1.0000 3rd Qu.:13.000   
## Max. :1.0000 Max. :16.000

child\_ica\_dummy %>%   
 filter(DID == 266) %>%   
 select(Household.Ownership, Num.of.Child) %>%   
 summary()

## Household.Ownership Num.of.Child   
## 0: 117 Min. :1.000   
## 1:1465 1st Qu.:2.000   
## Median :3.000   
## Mean :3.424   
## 3rd Qu.:4.000   
## Max. :7.000