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)  
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))  
child\_ica\_dummy$C002\_01 <- ifelse(child\_ica\_dummy$C002 == -1, 1, 0)  
child\_ica\_dummy$C003\_01 <- ifelse(child\_ica\_dummy$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$PR004ORPR009\_01 <- ifelse(  
 child\_ica\_dummy$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$PR004ANDPR009\_01 <- ifelse(  
 child\_ica\_dummy$PR004 == -1 & child\_ica\_dummy$PR009 == -1, 1, 0)

NCH <- child\_ica\_dummy %>%   
 group\_by(HHID) %>%   
 summarize(NCH = length(unique(CID)))

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

child\_ica\_dummy <- child\_ica\_dummy %>% left\_join(NCH)

## 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$H003 <- as.factor(child\_ica\_dummy$H003)

### Notes on Data Wrangling

#### Eliminated NAs

child\_ica %>%   
 summarize(C002\_na = sum(is.na(C002)),  
 C003\_na = sum(is.na(C003)),  
 PR004\_na = sum(is.na(PR004)),  
 PR009\_na = sum(is.na(PR009)))

## C002\_na C003\_na PR004\_na PR009\_na  
## 1 66 0 6709 4066

#### 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 9449 0.03702644

#### Eliminated NAs Hunza

child\_ica %>%   
 filter(DID == 266) %>%   
 summarize(C002\_na = sum(is.na(C002)),  
 C003\_na = sum(is.na(C003)),  
 PR004\_na = sum(is.na(PR004)),  
 PR009\_na = sum(is.na(PR009)))

## C002\_na C003\_na PR004\_na PR009\_na  
## 1 0 0 15 18

#### 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 31 0.01889092

## Hunza. Generalized Linear Models

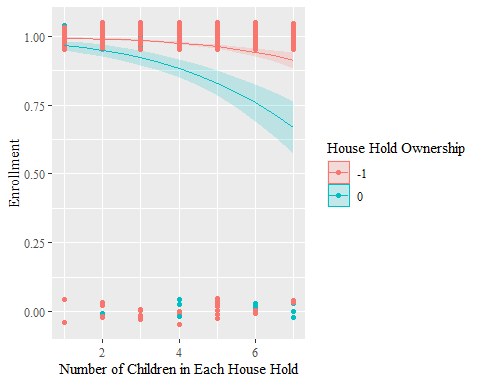
### Model 1 [NofChild + HHOwnership]

#### Hunza. GLM Age >= 5 C003\_01 ~ NCH + H003

glm\_child <- glm(C003\_01 ~ NCH + H003, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, !is.na(H003), DID == 266))  
  
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")

##   
## Call:  
## glm(formula = C003\_01 ~ NCH + H003, family = "binomial", data = child\_ica\_dummy %>%   
## filter(C001 >= 5, !is.na(H003), DID == 266))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1543 0.1464 0.1819 0.2259 0.8965   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.4051 0.5016 10.775 < 2e-16 \*\*\*  
## NCH -0.4373 0.1105 -3.958 7.54e-05 \*\*\*  
## H0030 -1.6398 0.3680 -4.456 8.36e-06 \*\*\*  
## ---  
## 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: 342.04 on 1385 degrees of freedom  
## AIC: 348.04  
##   
## Number of Fisher Scoring iterations: 7

## Warning in eval(family$initialize): non-integer #successes in a binomial glm!  
  
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!



extract\_eq(glm\_child)

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

##### exponential transformation

exp(glm\_child$coefficients)

## (Intercept) NCH H0030   
## 222.5420713 0.6457498 0.1940219

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

confint(glm\_child, level = 0.95)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 4.4624151 6.4342992  
## NCH -0.6549398 -0.2204113  
## H0030 -2.3381830 -0.8840013

##### 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) 86.69663780 622.8459563  
## NCH 0.51947333 0.8021888  
## H0030 0.09650283 0.4131266

##### 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 = 13.475, df = 8, p-value = 0.09653

##### AIC

extractAIC(glm\_child)

## [1] 3.0000 348.0375

##### BIC

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

## [1] 3.0000 363.7443

##### effectiveness of explanatory variables

glm\_child\_null <- glm(C003\_01 ~ 1, family = "binomial",   
 data = child\_ica\_dummy %>% filter(!is.na(H003), C001 >= 5, DID == 266))  
anova(glm\_child\_null, glm\_child, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: C003\_01 ~ 1  
## Model 2: C003\_01 ~ NCH + H003  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 1387 376.54   
## 2 1385 342.04 2 34.506 3.214e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

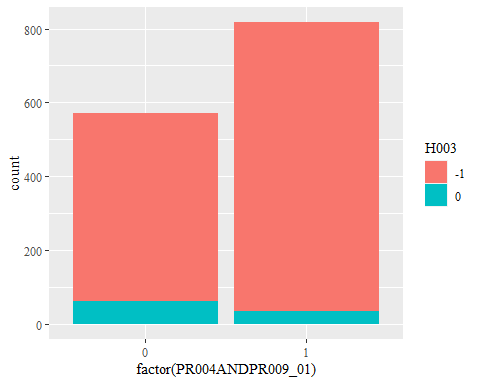
##### variables selection

step(glm\_child\_null, direction = "both",  
 scope = (~ C001 + C002\_01 + NCH + H003))

## Start: AIC=378.54  
## C003\_01 ~ 1  
##   
## Df Deviance AIC  
## + H003 1 357.40 361.40  
## + NCH 1 357.92 361.92  
## <none> 376.54 378.54  
## + C001 1 375.57 379.57  
## + C002\_01 1 376.15 380.15  
##   
## Step: AIC=361.4  
## C003\_01 ~ H003  
##   
## Df Deviance AIC  
## + NCH 1 342.04 348.04  
## <none> 357.40 361.40  
## + C001 1 356.13 362.13  
## + C002\_01 1 357.22 363.22  
## - H003 1 376.54 378.54  
##   
## Step: AIC=348.04  
## C003\_01 ~ H003 + NCH  
##   
## Df Deviance AIC  
## <none> 342.04 348.04  
## + C002\_01 1 341.38 349.38  
## + C001 1 341.47 349.47  
## - NCH 1 357.40 361.40  
## - H003 1 357.92 361.92

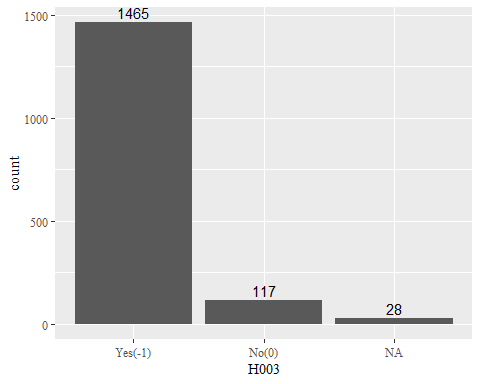
##   
## Call: glm(formula = C003\_01 ~ H003 + NCH, family = "binomial", data = child\_ica\_dummy %>%   
## filter(!is.na(H003), C001 >= 5, DID == 266))  
##   
## Coefficients:  
## (Intercept) H0030 NCH   
## 5.4051 -1.6398 -0.4373   
##   
## Degrees of Freedom: 1387 Total (i.e. Null); 1385 Residual  
## Null Deviance: 376.5   
## Residual Deviance: 342 AIC: 348

child\_ica\_dummy %>%   
 filter(!is.na(H003), C001 >= 5, DID == 266) %>%   
 group\_by(H003) %>%   
 ggplot(aes(factor(PR004ANDPR009\_01), fill = H003)) +  
 geom\_bar(stat = "count") +  
 theme(text = element\_text(family = "serif"))

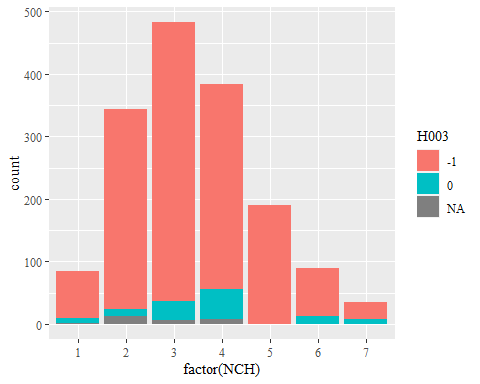


There seems multi-colinearity between Household Ownership and Both-Parents Education

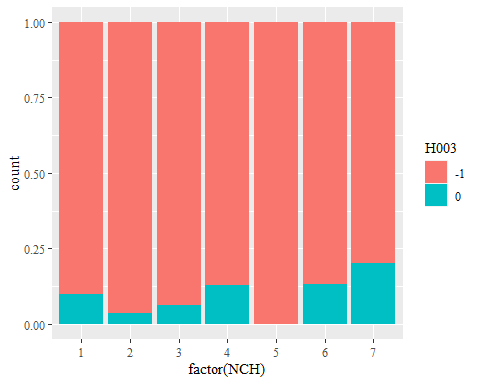
child\_ica\_dummy %>%   
 filter(DID == 266) %>%   
 ggplot(aes(H003)) +  
 geom\_bar() +  
 geom\_text(aes(label = ..count..), stat = "count", vjust = -.3) +  
 scale\_x\_discrete(breaks = c(-1, 0, NA), labels = c("Yes(-1)", "No(0)", "NA")) +  
 theme(text = element\_text(family = "serif"))



child\_ica\_dummy %>%   
 filter(DID == 266) %>%   
 ggplot(aes(factor(NCH), fill = H003)) +  
 geom\_bar() +  
 theme(text = element\_text(family = "serif"))



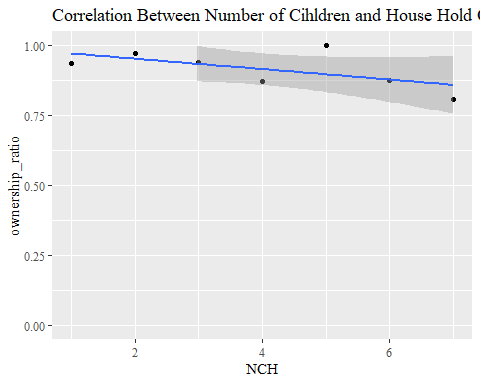
child\_ica\_dummy %>%   
 filter(DID == 266, !is.na(H003)) %>%   
 ggplot(aes(factor(NCH), fill = H003)) +  
 geom\_bar(stat = "count", position = "fill") +  
 theme(text = element\_text(family = "serif"))



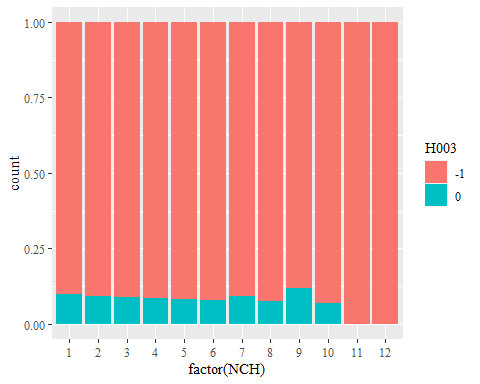
child\_ica\_dummy %>%   
 filter(C001 >= 5, DID == 266, !is.na(H003)) %>%   
 group\_by(NCH) %>%   
 summarize(ownership\_ratio = mean(H003 == -1)) %>%   
 ggplot(aes(NCH, ownership\_ratio)) +  
 geom\_point() +  
 geom\_smooth(method = "lm")+  
 ylim(c(0,1)) +  
 labs(title = "Correlation Between Number of Cihldren and House Hold Ownership?") +  
 theme(text = element\_text(family = "serif"))

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

## `geom\_smooth()` using formula 'y ~ x'



child\_ica\_dummy %>%   
 filter(!is.na(H003)) %>%   
 ggplot(aes(factor(NCH), fill = H003)) +  
 geom\_bar(stat = "count", position = "fill") +  
 theme(text = element\_text(family = "serif"))



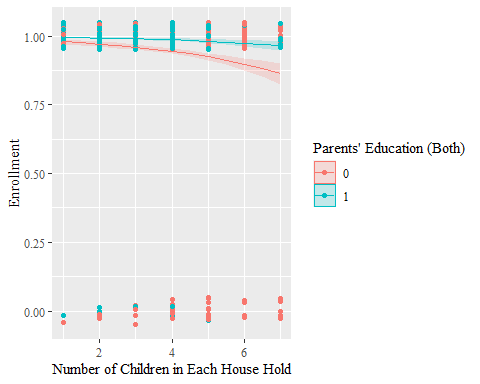
### Model 2 [NofChild + ParentsBothEdu]

#### Hunza. GLM Age >= 5 C003\_01 ~ NCH + PR004ANDPR009\_01

glm\_child <- glm(C003\_01 ~ NCH + PR004ANDPR009\_01, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, DID == 266))  
  
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 = "Parents' Education (Both)", fill = "Parents' Education (Both)")

##   
## Call:  
## glm(formula = C003\_01 ~ NCH + PR004ANDPR009\_01, family = "binomial",   
## data = child\_ica\_dummy %>% filter(C001 >= 5, DID == 266))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2554 0.1395 0.1646 0.2860 0.5418   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.1748 0.5255 7.944 1.96e-15 \*\*\*  
## NCH -0.3329 0.1099 -3.028 0.002461 \*\*   
## PR004ANDPR009\_01 1.4519 0.3926 3.698 0.000217 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 384.83 on 1409 degrees of freedom  
## Residual deviance: 349.37 on 1407 degrees of freedom  
## AIC: 355.37  
##   
## Number of Fisher Scoring iterations: 7

## Warning in eval(family$initialize): non-integer #successes in a binomial glm!  
  
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!



extract\_eq(glm\_child)

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

##### exponential transformation

exp(glm\_child$coefficients)

## (Intercept) NCH PR004ANDPR009\_01   
## 65.0249160 0.7168352 4.2711686

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

confint(glm\_child, level = 0.95)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 3.1894615 5.254292  
## NCH -0.5504811 -0.118222  
## PR004ANDPR009\_01 0.7205429 2.277683

##### 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) 24.2753514 191.3858514  
## NCH 0.5766723 0.8884988  
## PR004ANDPR009\_01 2.0555489 9.7540560

##### 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 = 5.41, df = 8, p-value = 0.713

##### AIC

extractAIC(glm\_child)

## [1] 3.0000 355.3652

##### BIC

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

## [1] 3.0000 371.1192

##### effectiveness of explanatory variables

glm\_child\_null <- glm(C003\_01~1, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, DID == 266))  
anova(glm\_child\_null, glm\_child, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: C003\_01 ~ 1  
## Model 2: C003\_01 ~ NCH + PR004ANDPR009\_01  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 1409 384.83   
## 2 1407 349.37 2 35.462 1.993e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### variables selection

step(glm\_child\_null, direction = "both",   
 scope = (~ C001 + C002\_01 + PR004ANDPR009\_01 + PR004ORPR009\_01 + NCH))

## Start: AIC=386.83  
## C003\_01 ~ 1  
##   
## Df Deviance AIC  
## + PR004ANDPR009\_01 1 358.59 362.59  
## + NCH 1 365.61 369.61  
## + PR004ORPR009\_01 1 371.38 375.38  
## <none> 384.83 386.83  
## + C001 1 384.22 388.22  
## + C002\_01 1 384.61 388.61  
##   
## Step: AIC=362.59  
## C003\_01 ~ PR004ANDPR009\_01  
##   
## Df Deviance AIC  
## + NCH 1 349.37 355.37  
## <none> 358.59 362.59  
## + C002\_01 1 358.16 364.16  
## + PR004ORPR009\_01 1 358.23 364.23  
## + C001 1 358.55 364.55  
## - PR004ANDPR009\_01 1 384.83 386.83  
##   
## Step: AIC=355.37  
## C003\_01 ~ PR004ANDPR009\_01 + NCH  
##   
## Df Deviance AIC  
## <none> 349.37 355.37  
## + C002\_01 1 348.46 356.46  
## + PR004ORPR009\_01 1 349.30 357.30  
## + C001 1 349.34 357.34  
## - NCH 1 358.59 362.59  
## - PR004ANDPR009\_01 1 365.61 369.61

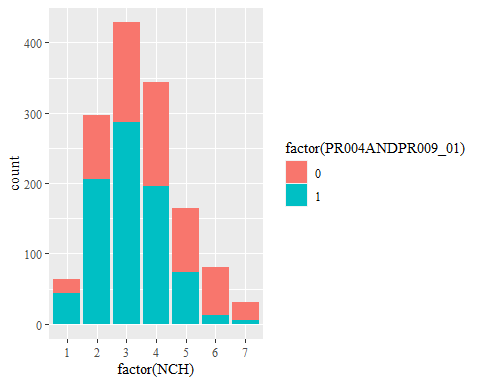
##   
## Call: glm(formula = C003\_01 ~ PR004ANDPR009\_01 + NCH, family = "binomial",   
## data = child\_ica\_dummy %>% filter(C001 >= 5, DID == 266))  
##   
## Coefficients:  
## (Intercept) PR004ANDPR009\_01 NCH   
## 4.1748 1.4519 -0.3329   
##   
## Degrees of Freedom: 1409 Total (i.e. Null); 1407 Residual  
## Null Deviance: 384.8   
## Residual Deviance: 349.4 AIC: 355.4

##### multicolinearity

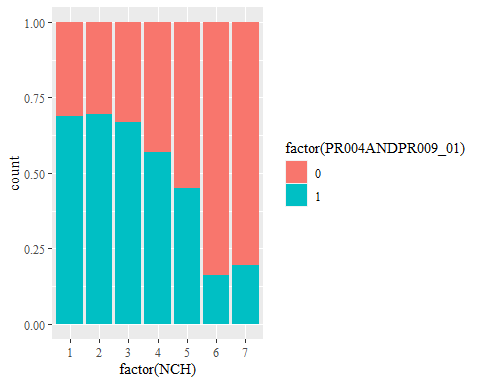
vif(glm\_child)

## NCH PR004ANDPR009\_01   
## 1.068705 1.068705

child\_ica\_dummy %>%   
 filter(DID == 266, C001 >= 5) %>%   
 ggplot(aes(factor(NCH), fill = factor(PR004ANDPR009\_01))) +  
 geom\_bar(stat = "count", position = "stack") +  
 theme(text = element\_text(family = "serif"))



child\_ica\_dummy %>%   
 filter(DID == 266, C001 >= 5) %>%   
 ggplot(aes(factor(NCH), fill = factor(PR004ANDPR009\_01))) +  
 geom\_bar(stat = "count", position = "fill") +  
 theme(text = element\_text(family = "serif"))



child\_ica\_dummy %>%   
 filter(C001 >= 5, DID == 266) %>%   
 summarize()

## data frame with 0 columns and 1 row

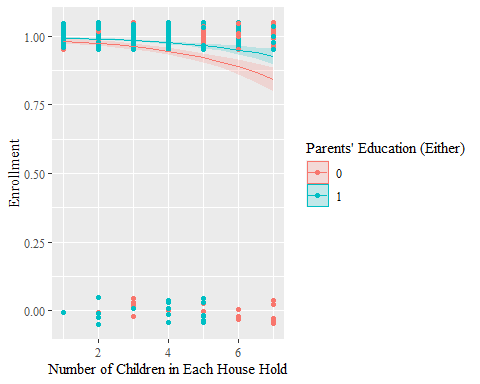
### Model 3 [NofChild + ParentsEitherEdu]

#### Hunza. GLM Age >= 5 C003\_01 ~ NCH + PR004ORPR009\_01

glm\_child <- glm(C003\_01 ~ NCH + PR004ORPR009\_01, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, DID == 266))  
  
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 = "Parents' Education (Either)", fill = "Parents' Education (Either)")

##   
## Call:  
## glm(formula = C003\_01 ~ NCH + PR004ORPR009\_01, family = "binomial",   
## data = child\_ica\_dummy %>% filter(C001 >= 5, DID == 266))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1213 0.1825 0.2212 0.2678 0.5841   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.3995 0.5767 7.629 2.37e-14 \*\*\*  
## NCH -0.3882 0.1110 -3.496 0.000472 \*\*\*  
## PR004ORPR009\_01 0.8522 0.3299 2.583 0.009791 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 384.83 on 1409 degrees of freedom  
## Residual deviance: 359.15 on 1407 degrees of freedom  
## AIC: 365.15  
##   
## Number of Fisher Scoring iterations: 6

## Warning in eval(family$initialize): non-integer #successes in a binomial glm!  
  
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!



extract\_eq(glm\_child)

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

##### exponential transformation

exp(glm\_child$coefficients)

## (Intercept) NCH PR004ORPR009\_01   
## 81.4083635 0.6782539 2.3447616

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

confint(glm\_child, level = 0.95)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 3.3155909 5.5810264  
## NCH -0.6077054 -0.1710846  
## PR004ORPR009\_01 0.1984172 1.4991924

##### 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) 27.5386612 265.3438228  
## NCH 0.5445991 0.8427502  
## PR004ORPR009\_01 1.2194710 4.4780713

##### 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 = 16.438, df = 8, p-value = 0.03652

##### AIC

extractAIC(glm\_child)

## [1] 3.0000 365.1525

##### BIC

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

## [1] 3.0000 380.9066

## Pak. Generalized Linear Model

### Model 1 [NofChild + HHOwnership]

#### Each Dist. GLM Age >= 5 C003\_01 ~ NCH + H003

each\_dist <- sapply(as.numeric(as.character(unique(child\_ica\_dummy$DID))), function(id){  
 child\_ica\_dummy <- child\_ica\_dummy %>% filter(DID == id)  
 glm\_child <- glm(C003\_01 ~ NCH + H003,   
 family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, !is.na(H003)))  
 temp\_hoslem <- hoslem.test(x = glm\_child$y, y = fitted(glm\_child))  
 data.frame(DID = id, hoslem\_p\_value = temp\_hoslem$p.value)  
}) %>% t()

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

## DID hoslem\_p\_value  
## 1 146 3.702152e-01  
## 2 147 8.402235e-01  
## 3 148 5.747210e-01  
## 4 149 9.583907e-01  
## 5 150 8.932666e-01  
## 6 151 8.977396e-01  
## 7 152 3.931301e-01  
## 8 153 1.868806e-01  
## 9 154 2.252387e-01  
## 10 155 9.988024e-01  
## 11 156 4.570744e-01  
## 12 157 1.274284e-01  
## 13 158 7.766738e-01  
## 14 159 9.786181e-01  
## 15 160 NaN  
## 16 161 6.622064e-01  
## 17 162 9.541911e-01  
## 18 163 3.094742e-01  
## 19 164 7.114949e-01  
## 20 165 1.554080e-02  
## 21 166 5.996638e-01  
## 22 167 8.584151e-01  
## 23 169 9.561827e-01  
## 24 170 2.845733e-01  
## 25 171 NaN  
## 26 172 7.646206e-01  
## 27 173 1.801413e-01  
## 28 174 4.621954e-01  
## 29 175 9.488777e-01  
## 30 176 6.698873e-01  
## 31 177 3.693581e-01  
## 32 178 6.375828e-01  
## 33 179 2.404270e-01  
## 34 180 7.712604e-02  
## 35 181 9.756719e-01  
## 36 182 8.916927e-01  
## 37 183 1.778700e-01  
## 38 184 NaN  
## 39 185 NaN  
## 40 186 7.997788e-01  
## 41 187 3.649090e-01  
## 42 188 3.434400e-01  
## 43 189 4.202915e-01  
## 44 190 8.283956e-01  
## 45 191 8.599984e-01  
## 46 192 8.128284e-01  
## 47 193 2.314288e-06  
## 48 194 3.126762e-01  
## 49 195 1.378695e-01  
## 50 196 4.316092e-01  
## 51 197 8.306028e-01  
## 52 198 2.599524e-01  
## 53 199 1.082096e-02  
## 54 200 1.518975e-02  
## 55 202 3.956771e-01  
## 56 203 1.349658e-06  
## 57 204 6.204874e-01  
## 58 315 8.202915e-01  
## 59 316 1.389904e-01  
## 60 320 5.912602e-05  
## 61 205 9.999470e-01  
## 62 206 9.963441e-01  
## 63 207 7.404630e-01  
## 64 208 9.846597e-01  
## 65 209 3.866598e-01  
## 66 210 1.507208e-01  
## 67 211 8.883225e-01  
## 68 212 9.725283e-01  
## 69 213 8.938643e-01  
## 70 214 8.987225e-01  
## 71 215 6.623776e-01  
## 72 216 9.976142e-01  
## 73 217 5.119611e-01  
## 74 218 2.221756e-01  
## 75 219 8.848497e-01  
## 76 220 3.592986e-02  
## 77 221 7.504679e-01  
## 78 222 9.552782e-02  
## 79 223 9.962632e-01  
## 80 224 9.536877e-01  
## 81 225 9.999157e-01  
## 82 226 3.045858e-01  
## 83 227 4.912747e-01  
## 84 228 9.983706e-01  
## 85 229 9.995215e-01  
## 86 230 9.992232e-01  
## 87 231 9.209521e-01  
## 88 232 3.822277e-01  
## 89 233 6.763309e-01  
## 90 234 1.920139e-01  
## 91 318 3.606647e-01  
## 92 319 5.949079e-01  
## 93 235 6.125810e-01  
## 94 236 4.774151e-01  
## 95 237 1.786874e-02  
## 96 238 6.661981e-01  
## 97 239 9.394117e-01  
## 98 240 2.511028e-01  
## 99 241 2.053414e-03  
## 100 242 7.527516e-01  
## 101 243 5.712028e-03  
## 102 244 5.000419e-01  
## 103 245 9.999208e-01  
## 104 246 5.653004e-01  
## 105 247 7.135440e-01  
## 106 248 9.079060e-03  
## 107 249 5.100203e-01  
## 108 250 8.443304e-02  
## 109 251 9.194270e-01  
## 110 252 9.796447e-01  
## 111 253 2.954387e-01  
## 112 254 2.235738e-01  
## 113 255 1.950090e-01  
## 114 256 4.500482e-01  
## 115 257 7.300007e-01  
## 116 258 4.399887e-04  
## 117 259 8.222252e-01  
## 118 260 5.329579e-01  
## 119 261 7.541846e-01  
## 120 262 5.458027e-01  
## 121 263 9.993226e-01  
## 122 264 3.744874e-01  
## 123 265 9.957911e-01  
## 124 266 9.652839e-02  
## 125 267 6.100159e-01  
## 126 268 6.908767e-01  
## 127 269 9.977134e-01  
## 128 270 4.451470e-01  
## 129 271 7.444678e-04  
## 130 272 8.059775e-01  
## 131 273 5.713590e-01  
## 132 274 8.981888e-01  
## 133 275 6.771278e-01  
## 134 276 9.998250e-01  
## 135 277 7.339147e-01  
## 136 278 1.568815e-02  
## 137 279 7.700977e-01  
## 138 280 5.990123e-01  
## 139 281 9.261617e-01  
## 140 282 9.572941e-01  
## 141 284 1.741378e-01  
## 142 287 4.787838e-01  
## 143 289 1.259827e-01  
## 144 290 1.823365e-01

#### Dists with P value <= 0.05

DID\_HosLem0.05\_model\_1 <- each\_dist %>%   
 filter(hoslem\_p\_value <= 0.05) %>%   
 pull(DID)

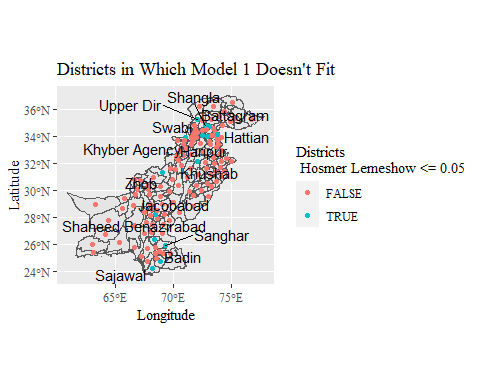
ica %>%   
 mutate(centroid = st\_centroid(geometry),  
 x = st\_coordinates(centroid)[,1],  
 y = st\_coordinates(centroid)[,2]) %>%   
 ggplot(ratio = 1) +  
 coord\_sf() +  
 geom\_sf(ratio = 1) +  
 geom\_point(data = child\_ica\_dummy %>%   
 filter(!is.na(x)) %>%   
 summarize(x = unique(x),  
 y = unique(y),  
 DID = unique(DID)),  
 aes(x, y, color = DID %in% DID\_HosLem0.05\_model\_1)) +  
 geom\_text\_repel(data = child\_ica\_dummy %>%   
 filter(DID %in% DID\_HosLem0.05\_model\_1, !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 in Which Model 1 Doesn't Fit") +  
 labs(color = "Districts \n Hosmer Lemeshow <= 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)`.

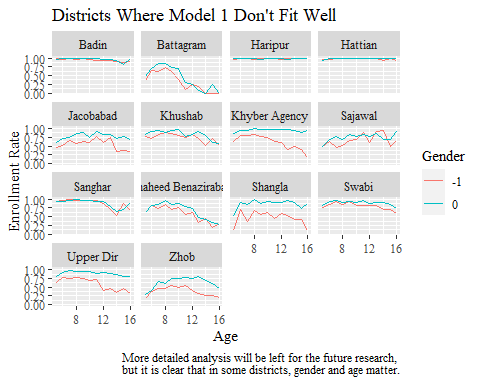
## Warning in st\_centroid.sfc(geometry): st\_centroid does not give correct  
## centroids for longitude/latitude data

## Warning: Ignoring unknown parameters: ratio

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



child\_ica\_dummy %>%   
 filter(C001 >= 5, DID %in% DID\_HosLem0.05\_model\_1) %>%   
 group\_by(DID, C001, C002) %>%   
 mutate(rate = mean(C003 == 3)) %>%   
 ggplot(aes(C001, rate, color = C002)) +  
 geom\_line() +  
 facet\_wrap(.~DNAME) +  
 xlab("Age") +  
 ylab("Enrollment Rate") +  
 labs(title = "Districts Where Model 1 Don't Fit Well",   
 caption = "More detailed analysis will be left for the future research, \n but it is clear that in some districts, gender and age matter.",  
 color = "Gender") +  
 theme(text = element\_text(family = "serif"))



### Model 2 [NofChild + ParentsBothEdu]

#### Whole Pak. GLM Age >= 5 C003\_01 ~ NCH + PR004ANDPR009\_01 + relevel(DID, ref = “266”)

glm\_child <- glm(C003\_01 ~ NCH + PR004ANDPR009\_01 + relevel(DID, ref = "266"), family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5))  
  
glm\_child %>% summary()

##   
## Call:  
## glm(formula = C003\_01 ~ NCH + PR004ANDPR009\_01 + relevel(DID,   
## ref = "266"), family = "binomial", data = child\_ica\_dummy %>%   
## filter(C001 >= 5))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2002 0.1836 0.4668 0.6971 1.5731   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.509591 0.156181 22.471 < 2e-16 \*\*\*  
## NCH -0.107577 0.003969 -27.104 < 2e-16 \*\*\*  
## PR004ANDPR009\_01 0.673143 0.017386 38.717 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")146 -0.875895 0.194633 -4.500 6.79e-06 \*\*\*  
## relevel(DID, ref = "266")147 -1.482768 0.172762 -8.583 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")148 -0.915625 0.179277 -5.107 3.27e-07 \*\*\*  
## relevel(DID, ref = "266")149 -1.643550 0.168400 -9.760 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")150 -1.607819 0.168098 -9.565 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")151 -0.807095 0.185542 -4.350 1.36e-05 \*\*\*  
## relevel(DID, ref = "266")152 -0.717054 0.196612 -3.647 0.000265 \*\*\*  
## relevel(DID, ref = "266")153 -1.736759 0.170212 -10.203 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")154 -1.553572 0.172629 -8.999 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")155 -1.670105 0.167193 -9.989 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")156 -2.140333 0.181540 -11.790 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")157 -1.103420 0.181688 -6.073 1.25e-09 \*\*\*  
## relevel(DID, ref = "266")158 -1.156450 0.177290 -6.523 6.89e-11 \*\*\*  
## relevel(DID, ref = "266")159 -2.184450 0.172128 -12.691 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")160 -1.709068 0.169995 -10.054 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")161 -2.304469 0.163081 -14.131 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")162 -0.630815 0.203630 -3.098 0.001949 \*\*   
## relevel(DID, ref = "266")163 -1.012210 0.196169 -5.160 2.47e-07 \*\*\*  
## relevel(DID, ref = "266")164 -1.659793 0.180095 -9.216 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")165 -1.903771 0.169885 -11.206 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")166 -1.684691 0.171039 -9.850 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")167 -0.396569 0.202148 -1.962 0.049789 \*   
## relevel(DID, ref = "266")169 -2.712608 0.163731 -16.568 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")170 -1.222926 0.177931 -6.873 6.28e-12 \*\*\*  
## relevel(DID, ref = "266")171 -1.205940 0.180436 -6.683 2.33e-11 \*\*\*  
## relevel(DID, ref = "266")172 -1.385540 0.175339 -7.902 2.74e-15 \*\*\*  
## relevel(DID, ref = "266")173 -1.117779 0.177247 -6.306 2.86e-10 \*\*\*  
## relevel(DID, ref = "266")174 -1.723520 0.172349 -10.000 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")175 -2.293485 0.166458 -13.778 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")176 -0.389622 0.228612 -1.704 0.088327 .   
## relevel(DID, ref = "266")177 -1.775120 0.168013 -10.565 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")178 -0.088396 0.210848 -0.419 0.675040   
## relevel(DID, ref = "266")179 -1.150526 0.179781 -6.400 1.56e-10 \*\*\*  
## relevel(DID, ref = "266")180 -1.451699 0.173220 -8.381 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")181 -1.148265 0.182264 -6.300 2.98e-10 \*\*\*  
## relevel(DID, ref = "266")182 -2.301266 0.165160 -13.934 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")183 -1.559035 0.172473 -9.039 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")184 -2.161783 0.165073 -13.096 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")185 -0.255138 0.189408 -1.347 0.177970   
## relevel(DID, ref = "266")186 -2.386320 0.164891 -14.472 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")187 -2.162162 0.166537 -12.983 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")188 -1.932412 0.167238 -11.555 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")189 -1.114197 0.175976 -6.332 2.43e-10 \*\*\*  
## relevel(DID, ref = "266")190 -1.830149 0.174173 -10.508 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")191 -1.648106 0.174426 -9.449 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")192 -1.337355 0.176637 -7.571 3.70e-14 \*\*\*  
## relevel(DID, ref = "266")193 -0.664444 0.179964 -3.692 0.000222 \*\*\*  
## relevel(DID, ref = "266")194 -2.631858 0.164251 -16.023 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")195 -2.789213 0.165049 -16.899 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")196 -1.025130 0.192010 -5.339 9.35e-08 \*\*\*  
## relevel(DID, ref = "266")197 -2.305263 0.165346 -13.942 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")198 -2.639440 0.164133 -16.081 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")199 0.401975 0.214627 1.873 0.061083 .   
## relevel(DID, ref = "266")200 -2.379658 0.167313 -14.223 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")202 -2.257160 0.166358 -13.568 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")203 -2.441492 0.163484 -14.934 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")204 -1.682783 0.169939 -9.902 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")205 -2.500618 0.163197 -15.323 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")206 -2.476645 0.163259 -15.170 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")207 -3.060126 0.163533 -18.713 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")208 -2.343127 0.164555 -14.239 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")209 -1.662136 0.175557 -9.468 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")210 -2.666322 0.162637 -16.394 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")211 -1.806394 0.173698 -10.400 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")212 -2.424800 0.165135 -14.684 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")213 -2.517405 0.163698 -15.378 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")214 -2.856169 0.163880 -17.428 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")215 -1.894926 0.163571 -11.585 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")216 -2.667465 0.165232 -16.144 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")217 -2.621078 0.164739 -15.910 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")218 -2.373729 0.162463 -14.611 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")219 -2.573537 0.165147 -15.583 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")220 -2.929650 0.161316 -18.161 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")221 -3.388760 0.164526 -20.597 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")222 -2.641607 0.162435 -16.263 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")223 -2.335626 0.171366 -13.629 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")224 -2.947502 0.163771 -17.998 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")225 -2.765260 0.164207 -16.840 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")226 -2.445937 0.166589 -14.682 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")227 -2.505052 0.163056 -15.363 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")228 -3.436061 0.165374 -20.778 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")229 -2.788615 0.166051 -16.794 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")230 -2.951117 0.163918 -18.004 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")231 -2.540898 0.163221 -15.567 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")232 -1.303620 0.172806 -7.544 4.56e-14 \*\*\*  
## relevel(DID, ref = "266")233 -3.337002 0.166821 -20.004 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")234 -2.488387 0.169736 -14.660 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")235 -1.659018 0.168694 -9.834 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")236 -1.870621 0.169750 -11.020 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")237 -1.475577 0.173747 -8.493 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")238 1.232565 0.295481 4.171 3.03e-05 \*\*\*  
## relevel(DID, ref = "266")239 -1.145610 0.177968 -6.437 1.22e-10 \*\*\*  
## relevel(DID, ref = "266")240 -1.659761 0.170571 -9.731 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")241 -2.841397 0.164413 -17.282 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")242 -1.867780 0.170169 -10.976 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")243 0.320346 0.303443 1.056 0.291104   
## relevel(DID, ref = "266")244 -1.201486 0.181179 -6.632 3.32e-11 \*\*\*  
## relevel(DID, ref = "266")245 -0.297386 0.202708 -1.467 0.142358   
## relevel(DID, ref = "266")246 -2.358940 0.166469 -14.170 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")247 -2.172820 0.171802 -12.647 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")248 -1.678299 0.165842 -10.120 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")249 -0.647963 0.196336 -3.300 0.000966 \*\*\*  
## relevel(DID, ref = "266")250 -1.467809 0.170954 -8.586 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")251 -1.356769 0.175621 -7.726 1.11e-14 \*\*\*  
## relevel(DID, ref = "266")252 -1.572819 0.170335 -9.234 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")253 -1.325726 0.178398 -7.431 1.08e-13 \*\*\*  
## relevel(DID, ref = "266")254 -2.523487 0.180001 -14.019 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")255 -1.601880 0.175767 -9.114 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")256 -0.530739 0.181964 -2.917 0.003537 \*\*   
## relevel(DID, ref = "266")257 -0.491253 0.215678 -2.278 0.022743 \*   
## relevel(DID, ref = "266")258 -2.397226 0.164104 -14.608 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")259 -0.120495 0.218934 -0.550 0.582065   
## relevel(DID, ref = "266")260 -0.579001 0.183002 -3.164 0.001557 \*\*   
## relevel(DID, ref = "266")261 -3.072381 0.162677 -18.886 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")262 -1.110709 0.174535 -6.364 1.97e-10 \*\*\*  
## relevel(DID, ref = "266")263 -0.618894 0.185937 -3.329 0.000873 \*\*\*  
## relevel(DID, ref = "266")264 -0.817633 0.179256 -4.561 5.09e-06 \*\*\*  
## relevel(DID, ref = "266")265 -0.190128 0.204767 -0.929 0.353142   
## relevel(DID, ref = "266")267 -0.203894 0.200998 -1.014 0.310388   
## relevel(DID, ref = "266")268 -1.662288 0.179179 -9.277 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")269 0.958226 0.302515 3.168 0.001537 \*\*   
## relevel(DID, ref = "266")270 -0.206918 0.198848 -1.041 0.298067   
## relevel(DID, ref = "266")271 0.511318 0.257880 1.983 0.047393 \*   
## relevel(DID, ref = "266")272 0.933463 0.295507 3.159 0.001584 \*\*   
## relevel(DID, ref = "266")273 -0.224661 0.210888 -1.065 0.286736   
## relevel(DID, ref = "266")274 0.181126 0.269851 0.671 0.502089   
## relevel(DID, ref = "266")275 -1.877979 0.170434 -11.019 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")276 1.254633 0.295322 4.248 2.15e-05 \*\*\*  
## relevel(DID, ref = "266")277 -1.036492 0.226583 -4.574 4.77e-06 \*\*\*  
## relevel(DID, ref = "266")278 -1.035538 0.167092 -6.197 5.74e-10 \*\*\*  
## relevel(DID, ref = "266")279 -2.582915 0.164997 -15.654 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")280 -1.305004 0.172797 -7.552 4.28e-14 \*\*\*  
## relevel(DID, ref = "266")281 -0.702178 0.180060 -3.900 9.63e-05 \*\*\*  
## relevel(DID, ref = "266")282 -1.272273 0.181308 -7.017 2.26e-12 \*\*\*  
## relevel(DID, ref = "266")284 -1.538773 0.170887 -9.005 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")287 -1.485674 0.175398 -8.470 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")289 -1.862308 0.166027 -11.217 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")290 -1.680815 0.170166 -9.878 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")315 -1.335461 0.190690 -7.003 2.50e-12 \*\*\*  
## relevel(DID, ref = "266")316 -2.145252 0.167764 -12.787 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")318 -2.754422 0.168236 -16.372 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")319 -3.219120 0.163022 -19.747 < 2e-16 \*\*\*  
## relevel(DID, ref = "266")320 -2.641120 0.168331 -15.690 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 212206 on 208249 degrees of freedom  
## Residual deviance: 183508 on 208104 degrees of freedom  
## AIC: 183800  
##   
## Number of Fisher Scoring iterations: 7

glm\_child %>%   
 summary() %>%   
 .$coefficient %>%   
 as.data.frame() %>%   
 select(`Pr(>|z|)`) %>%   
 filter(`Pr(>|z|)` > 0.05) %>%   
 row.names() %>%   
 str\_split("\\)") %>%   
 as.data.frame() %>%   
 gather() %>%   
 select(value) %>%   
 .[seq(2,24,2),] %>%   
 as.numeric()

## [1] 176 178 185 199 243 245 259 265 267 270 273 274

#### Each Dist. GLM Age >= 5 C003\_01 ~ NCH + PR004ANDPR009\_01

each\_dist <- sapply(as.numeric(as.character(unique(child\_ica\_dummy$DID))), function(id){  
 child\_ica\_dummy <- child\_ica\_dummy %>% filter(DID == id)  
 glm\_child <- glm(C003\_01 ~ NCH + PR004ANDPR009\_01,   
 family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5))  
 temp\_hoslem <- hoslem.test(x = glm\_child$y, y = fitted(glm\_child))  
 data.frame(DID = id, hoslem\_p\_value = temp\_hoslem$p.value)  
}) %>% t()

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

## DID hoslem\_p\_value  
## 1 146 4.833365e-01  
## 2 147 9.696368e-01  
## 3 148 2.340054e-01  
## 4 149 2.630450e-01  
## 5 150 4.052944e-01  
## 6 151 6.107987e-01  
## 7 152 5.295676e-01  
## 8 153 2.108079e-01  
## 9 154 8.521452e-01  
## 10 155 9.999996e-01  
## 11 156 4.243112e-01  
## 12 157 1.996798e-01  
## 13 158 2.698620e-01  
## 14 159 9.576960e-01  
## 15 160 1.873412e-01  
## 16 161 9.862571e-02  
## 17 162 9.700350e-01  
## 18 163 7.888207e-02  
## 19 164 8.745456e-01  
## 20 165 9.088681e-03  
## 21 166 6.417484e-01  
## 22 167 8.832939e-01  
## 23 169 9.760668e-01  
## 24 170 3.695271e-01  
## 25 171 1.588784e-01  
## 26 172 7.933903e-02  
## 27 173 9.982887e-03  
## 28 174 1.314211e-01  
## 29 175 9.932377e-01  
## 30 176 9.237029e-01  
## 31 177 5.996293e-01  
## 32 178 6.349025e-01  
## 33 179 2.552938e-01  
## 34 180 5.420001e-01  
## 35 181 7.728695e-01  
## 36 182 9.210615e-01  
## 37 183 7.729950e-03  
## 38 184 6.966427e-01  
## 39 185 1.115057e-01  
## 40 186 6.361407e-01  
## 41 187 NaN  
## 42 188 6.543812e-01  
## 43 189 5.089121e-01  
## 44 190 4.263431e-01  
## 45 191 9.611336e-01  
## 46 192 2.009149e-01  
## 47 193 6.445687e-06  
## 48 194 6.159970e-01  
## 49 195 2.881515e-01  
## 50 196 3.757288e-03  
## 51 197 9.592320e-01  
## 52 198 3.930063e-01  
## 53 199 3.020499e-01  
## 54 200 2.915802e-01  
## 55 202 8.220956e-01  
## 56 203 3.982822e-07  
## 57 204 7.540585e-01  
## 58 315 4.577361e-01  
## 59 316 1.377099e-02  
## 60 320 3.400093e-02  
## 61 205 9.858706e-01  
## 62 206 9.781201e-01  
## 63 207 9.324696e-01  
## 64 208 9.349555e-01  
## 65 209 8.148914e-01  
## 66 210 1.928247e-01  
## 67 211 9.737445e-01  
## 68 212 9.604219e-01  
## 69 213 8.909924e-01  
## 70 214 8.231951e-01  
## 71 215 6.468684e-01  
## 72 216 9.274780e-01  
## 73 217 8.091547e-01  
## 74 218 6.925247e-01  
## 75 219 8.513320e-01  
## 76 220 1.065343e-01  
## 77 221 9.281304e-01  
## 78 222 2.759677e-01  
## 79 223 8.395059e-01  
## 80 224 8.469389e-01  
## 81 225 9.953924e-01  
## 82 226 1.391444e-01  
## 83 227 4.715612e-01  
## 84 228 9.517910e-01  
## 85 229 1.000000e+00  
## 86 230 9.506553e-01  
## 87 231 9.992710e-01  
## 88 232 2.364582e-03  
## 89 233 8.167503e-01  
## 90 234 2.029032e-01  
## 91 318 8.704222e-01  
## 92 319 2.196002e-01  
## 93 235 2.113758e-01  
## 94 236 2.945804e-02  
## 95 237 2.517542e-01  
## 96 238 6.276721e-01  
## 97 239 2.157016e-01  
## 98 240 1.282194e-01  
## 99 241 5.197638e-05  
## 100 242 6.327375e-01  
## 101 243 9.503456e-01  
## 102 244 2.637431e-01  
## 103 245 9.922232e-01  
## 104 246 5.520648e-03  
## 105 247 2.720873e-01  
## 106 248 8.590541e-03  
## 107 249 4.451278e-01  
## 108 250 3.860093e-01  
## 109 251 9.431389e-01  
## 110 252 3.632820e-01  
## 111 253 9.562219e-01  
## 112 254 5.528683e-01  
## 113 255 2.014395e-01  
## 114 256 6.554420e-01  
## 115 257 6.877678e-03  
## 116 258 6.264664e-04  
## 117 259 7.785989e-01  
## 118 260 7.112161e-02  
## 119 261 8.481883e-01  
## 120 262 4.452464e-01  
## 121 263 9.315994e-01  
## 122 264 8.451285e-01  
## 123 265 9.716815e-01  
## 124 266 7.129842e-01  
## 125 267 1.644173e-01  
## 126 268 3.273428e-02  
## 127 269 9.902921e-01  
## 128 270 9.496613e-01  
## 129 271 9.579085e-01  
## 130 272 8.863188e-01  
## 131 273 4.208476e-01  
## 132 274 7.171719e-01  
## 133 275 1.355154e-02  
## 134 276 9.920801e-01  
## 135 277 6.343009e-02  
## 136 278 1.073304e-02  
## 137 279 6.630176e-01  
## 138 280 4.948066e-01  
## 139 281 9.708145e-01  
## 140 282 8.897523e-01  
## 141 284 7.440368e-01  
## 142 287 3.178372e-01  
## 143 289 5.144863e-02  
## 144 290 1.967437e-01

#### Dists with P value <= 0.05

each\_dist %>%   
 filter(hoslem\_p\_value <= 0.05)

## DID hoslem\_p\_value  
## 1 165 9.088681e-03  
## 2 173 9.982887e-03  
## 3 183 7.729950e-03  
## 4 193 6.445687e-06  
## 5 196 3.757288e-03  
## 6 203 3.982822e-07  
## 7 316 1.377099e-02  
## 8 320 3.400093e-02  
## 9 232 2.364582e-03  
## 10 236 2.945804e-02  
## 11 241 5.197638e-05  
## 12 246 5.520648e-03  
## 13 248 8.590541e-03  
## 14 257 6.877678e-03  
## 15 258 6.264664e-04  
## 16 268 3.273428e-02  
## 17 275 1.355154e-02  
## 18 278 1.073304e-02

dists\_not\_fit <- each\_dist %>%   
 filter(hoslem\_p\_value <= 0.05) %>%   
 pull(DID)

#### Not dists\_not\_fit. GLM Age >= 5 C003\_01 ~ NCH + PR004ANDPR009\_01

glm\_child <- glm(C003\_01 ~ NCH + PR004ANDPR009\_01, family = "binomial", data = child\_ica\_dummy %>% filter(C001 >= 5, !DID %in% dists\_not\_fit))  
  
glm\_child %>% summary()

##   
## Call:  
## glm(formula = C003\_01 ~ NCH + PR004ANDPR009\_01, family = "binomial",   
## data = child\_ica\_dummy %>% filter(C001 >= 5, !DID %in% dists\_not\_fit))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3662 0.3771 0.6920 0.7748 1.1695   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.564935 0.016627 94.12 <2e-16 \*\*\*  
## NCH -0.128849 0.003823 -33.70 <2e-16 \*\*\*  
## PR004ANDPR009\_01 1.300495 0.017287 75.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 185614 on 181260 degrees of freedom  
## Residual deviance: 176655 on 181258 degrees of freedom  
## AIC: 176661  
##   
## Number of Fisher Scoring iterations: 5

##### 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 = 44.192, df = 8, p-value = 5.235e-07