title: “Exam One(Redo for HW)” author: “Collin Rafferty” date: “10/22/2021” output: pdf_document

Question One:

H_0: P_A-P_B=0 H_A: P_A-P_B≠ 0

Where P_A is the proportion of people with at least some college in the Northeastern and Western regions of the U.S. that have been vaccinated, P_B is the proportion of people with a high school diploma in the Midwestern and Southern regions of the U.S. that have been vaccinated.

$$P_A\ Estimate = \frac{23{,}363}{25{,}907} = 90.18\% \quad P_B\ Estimate = \frac{3{,}478}{4{,}678} = 74.35\%$$

$$SE = \sqrt{\frac{\widehat{P_A}(1-\widehat{P_A})}{n_1} + \frac{\widehat{P_B}(1-\widehat{P_B})}{n_2}} = \sqrt{\frac{.9018(1-.9018)}{25{,}907} + \frac{.7435(1-.7435)}{4{,}678}} = .009$$

$$t - stat = \frac{\bar{X}}{SE} = \frac{.9018 - .7435}{.009} = 17.58889$$

I chose to use a .05 significance level. This means the results are significant at p < .05, and the p-value is < .00001, so the results are highly statistically significant. This means we succeeded in rejecting the null and showing sufficient evidence there is a statistically significant difference between the vaccination rates of the two samples.

$$Confidence\ Interval\text{:}\ (.9018 - .7435 \pm .009) = (.1493, .1676)$$

We are 95% confident that the average difference between the vaccination rate of sample A and sample B is between 14.93% and 16.73%

Question Two:

H_0: P_A-P_B=0 H_A: P_A-P_B≠ 0

P_A is the proportion of males with an associate degree or less who have been vaccinated, and P_B is the proportion of people who identify as transgender or other and have a bachelor’s degree or great and have been vaccinated.

$$P_A\ Estimate = \frac{8{,}486}{10{,}290} = 82.45\% \quad P_B\ Estimate = \frac{313}{389} = 80.46\%$$

$$SE = \sqrt{\frac{\widehat{P_A}(1-\widehat{P_A})}{n_1} + \frac{\widehat{P_B}(1-\widehat{P_B})}{n_2}} = \sqrt{\frac{.8245(1-.8245)}{10{,}290} + \frac{.8046(1-.8046)}{389}} = .0201$$

$$t-stat = \frac{\bar{X}}{SE} = \frac{.8245-.8046}{.0201} = .99$$

I chose to use a .10 significance level. This means the results are significant at p < .10, and the p-value is .322197, so the results are not statistically significant. This means we failed to reject the null hypothesis, which means there was not sufficient evidence to conclude the vaccination rate between these two groups was statistically significant.

$$Confidence\ Interval\colon\ (.8245-.8046 \pm .0210) = (-.0011, .0409)$$

We are 95% confident that the average difference between the vaccination rate of sample A and sample B is between -.11% and 4.09%

These were not expected results. I thought because of their educational attainment levels, sample A would have a much higher vaccination rate than sample B. One explanation for this disparity may be the difference in the sizes of the samples. Sample A is only about 4% the size of sample B

Question Three:

A)

```
load("/cloud/project/Household_Pulse_data.RData")

#Creation of Subgroup
attach(Household_Pulse_data)
use_varb <- ((RRACE=="White") |(RRACE=="Black")) & (INCOME== "HH income $75 -
99.9") & (RHISPANIC== "Not Hispanic") & ((EEDUC=="adv deg") | (EEDUC=="bach d
eg"))
sub1 <- subset(Household_Pulse_data,use_varb)
detach(Household_Pulse_data)
attach(sub1)

# Summary Statistics
summary(sub1)

##         RHISPANIC      RRACE               EEDUC              MS
##  Not Hispanic:5996   White:5684   less than hs:   0   NA       :  20
##  Hispanic    :   0   Black: 312   some hs     :   0   married  :4525
##                      Asian:   0   HS diploma  :   0   widowed  : 204
##                      Other:   0   some coll   :   0   divorced : 502
##                                   assoc deg   :   0   separated:  52
##                                   bach deg    :3005   never    : 693
##                                   adv deg     :2991
##  EGENID_BIRTH      GENID_DESCRIBE      SEXUAL_ORIENTATION
##  male  :2560   NA          :  32   NA            :  49
##  female:3436   male        :2525   gay or lesbian: 206
##                female      :3394   straight      :5501
##                transgender:   6   bisexual      : 158
##                other       :  39   something else:  46
##                                    dont know     :  36
```

```
##
##                        KIDS_LT5Y                        KIDS_5_11Y
##  NA                        :5268   NA                       :4984
##  Yes children under 5 in HH: 728   Yes children 5 - 11 in HH:1012
##
##
##
##
##
##                       KIDS_12_17Y                                ENROLLNONE
##  NA                        :5074   NA                                :5501
##  Yes children 12 - 17 in HH: 922   children not in any type of school: 495
##
##
##
##
##
##                RECVDVACC                         DOSESRV
##  NA                 :  13   NA                       : 330
##  yes got vaxx       :5682   yes got all doses        :5570
##  no did not get vaxx: 301   yes plan to get all doses:  83
##                             no will not get all doses:  13
##
##
##
##                      GETVACRV                                  KIDDOSES
##  NA                      :5687   NA                                :5078
##  definitely will get vaxx:  15   Yes kids got or will get all doses: 701
##  probably will get vaxx  :  23   no kids did not or will not       : 217
##  unsure about vaxx       :  67
##  probably not            :  82
##  definitely not          : 122
##
##                     KIDGETVAC                         HADCOVID
##  NA                      :5776   NA                       :  17
##  definitely will get vaxx:  31   yes doctor told had covid: 670
##  probably will get vaxx  :  28   no did not               :5288
##  unsure about vaxx       :  46   not sure                 :  21
##  probably not            :  44
##  definitely not          :  58
##  dont know yet           :  13
##                   WRKLOSSRV                             ANYWORK
##  NA                    :  10   NA                           :   9
##  yes recent HH job loss: 349   yes employment in last 7 days:4113
##  no recent HH job loss :5637   no employment in last 7 days :1874
##
##
##
##
##                 KINDWORK              RSNNOWRKRV
##  NA                 :1916   NA             :4144  
##  work for govt      : 943   retired        :1378  
##  work for private co:1937   other          : 176  
##  work for nonprofit : 734   caring for kids: 103  
##  self employed      : 409   did not want   :  78  
##  work in family biz :  57   laid off       :  51  
##                             (Other)        :  66  
##                                      CHLDCARE   
##  NA                                       :4806  
##  yes impacts to childcare because pandemic: 287  
##  no                                       : 903  
##                                                  
##                                                  
##                                                  
##                                                  
##                          CURFOODSUF  
##  NA                            :  13  
##  had enough food               :5615  
##  had enough but not what wanted: 344  
##  sometimes not enough food     :  22  
##  often not enough food         :   2  
##                                       
##                                       
##                                               CHILDFOOD   
##  NA                                                 :5814  
##  often kids not eating enough because couldnt afford:   3  
##  sometimes kids not eating enough                   :  22  
##  kids got enough food                               : 157  
##                                                            
##                                                            
##                                                            
##                                            ANXIOUS     
##  NA                                             :   8  
##  no anxiety over past 2 wks                     :2721  
##  several days anxiety over past 2 wks           :2088  
##  more than half the days anxiety over past 2 wks: 543  
##  nearly every day anxiety                       : 636  
##                                                        
##                                                        
##                                             WORRY      
##  NA                                             :   9  
##  no worry over past 2 wks                       :3501  
##  several days worried over past 2 wks           :1711  
##  more than half the days worried over past 2 wks: 388  
##  nearly every day worry                         : 387  
##                                                        
##                                                        
##                           TENURE    
##  NA                           :  11  
##  housing owned free and clear :1752  
##  housing owned with mortgage  :3548
##  housing rented               : 664
##  housing occupied without rent:  21
##
##
##                                 LIVQTRRV               RENTCUR
##  live in detached 1 family          :4875   NA             :5332
##  live in bldg w 5+ apts             : 470   current on rent: 654
##  live in 1 family attached to others: 444   behind on rent :  10
##  live in building with 3-4 apts     :  80
##  live in bldg w 2 apartments        :  68
##  live in mobile home                :  30
##  (Other)                            :  29
##                 MORTCUR                                            EVICT
##  NA                 :2449   NA                                        :598
6
##  current on mortgage:3475   very likely evicted in next 2 months      :
1
##  behind on mortgage :  72   somewhat likely evicted in next 2 months  :
0
##                             not very likely evicted in next 2 months  :
4
##                             not at all likely evicted in next 2 months:
5
##
##
##                                           FORCLOSE           EST_ST
##  NA                                           :5924   California: 337
##  very likely forclosed in next 2 months       :   5   Texas     : 272
##  somewhat likely forclosed in next 2 months   :   8   Washington: 233
##  not very likely forclosed in next 2 months   :  17   Florida   : 201
##  not at all forclosed evicted in next 2 months:  42   Maryland  : 201
##                                                       Utah      : 196
##                                                       (Other)   :4556
##                    PRIVHLTH                          PUBHLTH          REGION
##  has private health ins:5420   has public health ins:1779   Northeast:1013
##  no private health ins : 459   no public health ins :3852   South    :1899
##  NA                    : 117   NA                   : 365   Midwest  :1305
##                                                             West     :1779
##
##
##
##                       INCOME     Num_kids_Pub_School Num_kids_Priv_School
##  HH income $75 - 99.9    :5996   Min.   :0.000       Min.   :0.000
##  NA                      :   0   1st Qu.:1.000       1st Qu.:0.000
##  HH income less than $25k:   0   Median :2.000       Median :1.000
##  HH income $25k - $34.9k :   0   Mean   :1.748       Mean   :1.011
##  HH income $35k - 49.9   :   0   3rd Qu.:2.000       3rd Qu.:2.000
##  HH income $50k - 74.9   :   0   Max.   :4.000       Max.   :2.000
##  (Other)                 :   0   NA's   :4685        NA's   :5727
##  Num_kids_homeschool        Works_onsite           works_remote       
##  Min.   :0.00        NA           :  94   NA             : 182  
##  1st Qu.:0.00        worked onsite:3579   worked remotely:3109  
##  Median :1.00        no           :2323   no             :2705  
##  Mean   :0.72                                                   
##  3rd Qu.:1.00                                                   
##  Max.   :2.00                                                   
##  NA's   :5839                                                   
##            Shop_in_store                  eat_in_restaurant
##  NA              :  86   NA                       : 104   
##  shopped in store:5485   eat at restaurant indoors:3394   
##  no              : 425   no                       :2498   
##                                                           
##                                                           
##                                                           
##                                                           
```

```
round(prop.table(table(RRACE=="White", ANXIOUS)),2)
```

```
##        ANXIOUS
##           NA no anxiety over past 2 wks several days anxiety over past 2 w
ks
##   FALSE 0.00                       0.03                                 0.
02
##   TRUE  0.00                       0.43                                 0.
33
##        ANXIOUS
##         more than half the days anxiety over past 2 wks
##   FALSE                                            0.00
##   TRUE                                             0.09
##        ANXIOUS
##         nearly every day anxiety
##   FALSE                     0.00
##   TRUE                      0.10
```

```
round(prop.table(table(RRACE=="Black", ANXIOUS)),2)
```

```
##        ANXIOUS
##           NA no anxiety over past 2 wks several days anxiety over past 2 w
ks
##   FALSE 0.00                       0.43                                 0.
33
##   TRUE  0.00                       0.03                                 0.
02
##        ANXIOUS
##         more than half the days anxiety over past 2 wks
##   FALSE                                            0.09
##   TRUE                                             0.00
##        ANXIOUS
##         nearly every day anxiety
##   FALSE                         0.10
##   TRUE                          0.00
```

The subgroup I created is revealing because it looks at how upper-middle-class college-educated whites have coped with the pandemic compared to upper-middle-class college-educated African Americans. If you look at the proportion tables of how the anxious levels differ between the two groups, it is very eye-opening. For example, 43% of African Americans in this subgroup reported feeling no anxiety over the past two weeks compared to 43% of whites that reported feeling anxiety over the past two weeks. In the very next factor level, however, these two groups' anxiety levels flip. Much of the pandemic pernicious effects disproportionally affected the poor and underrepresented, but it is also important to remember that it has had a profound impact on every level of society. I think my subgroup helps to show that.

B) I thought it would be interesting to look how the percentage of people with private health insurance differs between the subgroup, and the larger sample.

H_0: P_A=P_B H_A: $P_A \neq P_B$

Where P_A is the proportion of people in the subgroup who have private health insurance, and P_B is the proportion of people in the greater sample that has private health insurance.

```
summary(sub1$PRIVHLTH)

## has private health ins  no private health ins                     NA
##                    5420                    459                    117

summary(Household_Pulse_data$PRIVHLTH)

## has private health ins  no private health ins                     NA
##                   46869                  11275                  10970
```

$$P_A\ Estimate = \frac{5{,}420}{5{,}996} = 90.39\% \quad P_B\ Estimate = \frac{46{,}869}{69{,}114} = 67.81\%$$

$$SE = \sqrt{\frac{\widehat{P_A}(1-\widehat{P_A})}{n_1} + \frac{\widehat{P_B}(1-\widehat{P_B})}{n_2}} = \sqrt{\frac{.9039(1-.9039)}{5{,}996} + \frac{.6781(1-.6781)}{69{,}114}} = .0042$$

$$t - stat = \frac{\bar{X}}{SE} = \frac{.9039-.6781}{.0042} = 53.76$$

I chose to use a .05 significance level. This means the results are significant at p < .05, and the p-value is < .00001, so the results are highly statistically significant. This means we succeeded in rejecting the null and showing sufficient evidence there is a statistically significant difference between the percentage of people with private health insurance in the subgroup compared to the larger sample.

$$Confidence\ Interval\colon\ (.9036 - .6781 \pm .0042) = (.2216, .23)$$

We are 95% confident that the average difference between the people with private health insurance in the subgroup and the larger sample is between 22.16% and 23%.

C)

```
##require(tidyverse)

HH1 <-
Household_Pulse_data%>%mutate(INCOME5=as.numeric(INCOME),INCOME5=case_when(IN
COME5==5~NA_integer_,TRUE~as##.integer(INCOME5)))

HH2 <-
Household_Pulse_data%>%mutate(GEN1=as.numeric(GENID_DESCRIBE),GEN1=case_when(
GEN1==1~NA_integer_,TRUE~as. integer(GEN1)))



HH3 <- Household_Pulse_data%>%mutate(VAX1=as.numeric(RECVDVACC),
##VAX1=case_when(VAX1==1~NA_integer_,TRUE~as.integer(VAX1)))



HH4 <- Household_Pulse_data%>%mutate(ANX2=as.numeric(ANXIOUS),
ANX2=case_when(ANX2==2~NA_integer_,TRUE~as.integer(ANX2)))



H5 <- Household_Pulse_data%>%mutate(EEDUC3=as.numeric(EEDUC),
##EEDUC3=case_when(EEDUC3==3~NA_integer_,TRUE~as.integer(EEDUC3)))

norm_varb <- function(X_in) {(X_in - min(X_in, na.rm = TRUE))/( max(X_in,
##na.rm = TRUE) - min(X_in, na.rm = TRUE) )}



##norm_INCOME5 <- norm_varb(HH1$INCOME5)

##norm_GEN1 <- norm_varb(HH2$GEN1)

##norm_VAX1<- norm_varb(HH3$VAX1)

##norm_ANX2 <- norm_varb(HH4$ANX2)

##norm_EEDUC3<- norm_varb(HH5$EEDUC3)



##data_use <- data.frame(norm_INCOME5,norm_GEN1,norm_VAX1,norm_ANX2,
##norm_EEDUC3)

##good_obs_data_use <- complete.cases(data_use,PUBHLTH)

##dat_use <- subset(data_use,good_obs_data_use)

##y_use <- subset(PUBHLTH,good_obs_data_use)
```

```
##set.seed(12345)
##NN_obs <- sum(good_obs_data_use == 1)
##select1 <- (runif(NN_obs) < 0.8)

##train_data <- subset(dat_use,select1)
##test_data <- subset(dat_use,(!select1))
##cl_data <- y_use[select1]
##true_data <- y_use[!select1]
```

```
##summary(cl_data)
has public health ins no public health ins                   NA
7146                 12890                 6319
prop.table(summary(cl_data))
has public health ins  no public health ins                   NA
0.2711440             0.4890913            0.2397648
summary(train_data)
norm_INCOME5      norm_GEN1        norm_VAX1        norm_ANX2
norm_EEDUC3
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.
:0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.5000   1st
Qu.:0.5000
Median :0.2857   Median :0.3333   Median :0.0000   Median :0.5000   Median
:0.8333
Mean   :0.3629   Mean   :0.2307   Mean   :0.1101   Mean   :0.5475   Mean
:0.7538
3rd Qu.:0.7143   3rd Qu.:0.3333   3rd Qu.:0.0000   3rd Qu.:0.7500   3rd
Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.
:1.0000
```

```
require(class)

for (indx in seq(1, 9, by= 2)) {

pred_PUBHLTH <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob =
FALSE, use.all = TRUE)}
```

```
num_correct_labels <- sum(pred_PUBHLTH  == true_data)

correct_rate <- num_correct_labels/length(true_data)

print(c(indx,correct_rate))

> print(c(indx,correct_rate))

[1] 9.0000000 0.7019365
```

With the predictor variables I included in the classifier, it had an accuracy of 70% of determining if a person has public health insurance or not. 70% is not terrible, but it could be improved. If I were to do another iteration, I would include the variables of ANYWORK,RSNNOWRKRV, and KINDWORK because if a person is a government employee, unemployed, or retired, they would be more likely to have public health insurance.

D)

```
attach(Household_Pulse_data)

## The following objects are masked from sub1:
## 
##     ANXIOUS, ANYWORK, CHILDFOOD, CHLDCARE, CURFOODSUF, DOSESRV,
##     eat_in_restaurant, EEDUC, EGENID_BIRTH, ENROLLNONE, EST_ST, EVICT,
##     FORCLOSE, GENID_DESCRIBE, GETVACRV, HADCOVID, INCOME, KIDDOSES,
##     KIDGETVAC, KIDS_12_17Y, KIDS_5_11Y, KIDS_LT5Y, KINDWORK, LIVQTRRV,
##     MORTCUR, MS, Num_kids_homeschool, Num_kids_Priv_School,
##     Num_kids_Pub_School, PRIVHLTH, PUBHLTH, RECVDVACC, REGION, RENTCUR,
##     RHISPANIC, RRACE, RSNNOWRKRV, SEXUAL_ORIENTATION, Shop_in_store,
##     TENURE, Works_onsite, works_remote, WORRY, WRKLOSSRV

reg1<- lm(as.numeric(Num_kids_Pub_School) ~ INCOME+EEDUC+CHLDCARE+Works_onsit
e)

require(stargazer)

## Loading required package: stargazer

## 
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary St
atistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(reg1,type = "text", title = "Table 1: Regression Results",out = "ta
ble1.txt")
```

```
## 
## Table 1: Regression Results
## ==========================================================================
=======
##                                                         Dependent variable
:
##                                                   ------------------------
-------
##                                                   as.numeric(Num_kids_Pub_
School)
## --------------------------------------------------------------------------
-------
## 25k                                                           -0.087**
##                                                                (0.037)
## 
## 34.9k                                                          -0.028
##                                                                (0.032)
## 
## 35k - 49.9                                                     -0.048*
##                                                                (0.028)
## 
## 50k - 74.9                                                     -0.031
##                                                                (0.028)
## 
## 75 - 99.9                                                       0.026
##                                                                (0.025)
## 
## 100k - 149                                                      0.021
##                                                                (0.031)
## 
## 150 - 199                                                       0.040
##                                                                (0.029)
## 
## EEDUCsome hs                                                   -0.151
##                                                                (0.101)
## 
## EEDUCHS diploma                                               -0.291***
##                                                                (0.087)
## 
## EEDUCsome coll                                                -0.365***
##                                                                (0.086)
## 
```

```
## EEDUCassoc deg                                               -0.316***
##                                                               (0.088)
##
## EEDUCbach deg                                                -0.403***
##                                                               (0.086)
##
## EEDUCadv deg                                                 -0.412***
##                                                               (0.087)
##
## CHLDCAREyes impacts to childcare because pandemic            0.101***
##                                                               (0.026)
##
## CHLDCAREno                                                   0.212***
##                                                               (0.016)
##
## Works_onsiteworked onsite                                    -0.057**
##                                                               (0.029)
##
## Works_onsiteno                                               -0.114***
##                                                               (0.030)
##
## Constant                                                     2.059***
##                                                               (0.087)
##
## --------------------------------------------------------------------------
-------
## Observations                                                  14,006
## R2                                                             0.018
## Adjusted R2                                                    0.017
## Residual Std. Error                                      0.878 (df = 13988)
## F Statistic                                          15.008*** (df = 17; 13
988)
## ==========================================================================
=======
## Note:                                                   *p<0.1; **p<0.05; **
*p<0.01
```

```
require(ggplot2)
```

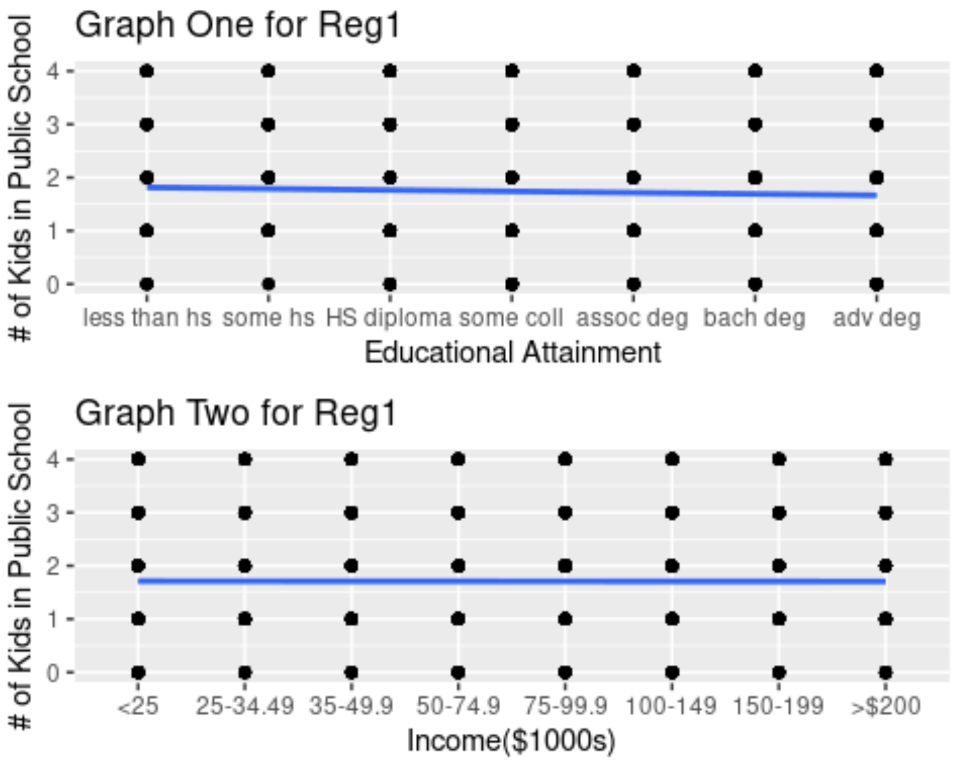
```
## Loading required package: ggplot2
```

```
require(ggthemes)
```

```
## Loading required package: ggthemes
```

```
Graph1 <-ggplot(Household_Pulse_data, aes(y=Num_kids_Pub_School, x= EEDUC, gr
oup=1))+geom_point()+geom_smooth(method=lm)+labs(x="Educational Attainment",y
="# of Kids in Public School", title = "Graph One for Reg1")
```

```
Graph2 <-ggplot(Household_Pulse_data, aes(y=Num_kids_Pub_School, x= INCOME, g
roup=1))+geom_point()+geom_smooth(method=lm)+labs(x="Income($1000s)", y="# of
Kids in Public School", title = "Graph Two for Reg1") + scale_x_discrete(labe
ls = c('<25','25-34.49','35-49.9', '50-74.9','75-99.9','100-149','150-199','>
$200' ))

gridExtra::grid.arrange(Graph1,Graph2)

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

## Warning: Removed 55108 rows containing non-finite values (stat_smooth).

## Warning: Removed 55108 rows containing missing values (geom_point).

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

## Warning: Removed 55108 rows containing non-finite values (stat_smooth).

## Warning: Removed 55108 rows containing missing values (geom_point).
```


Graph One for Reg1

Graph Two for Reg1

The regression provides some revealing insights about the data. For example, the number of children in public is negatively correlated with educational attainment level after a person graduates high school. A person with an advanced degree has -.412 fewer children in public school than someone who does not have an advanced degree. Of course, you can't have a proportion of a child, but it helps to show the picture. This result was to be expected, but what was interesting was how income affects the number of children a person has in public school. The coefficients of income are negative until 75K than positive for the rest of the income levels. Intuitively, you would think the opposite would be true: as income increases, the number of children a person has in public school decreases because these high earners have fewer children or send their children to private school.