Lab 14 - Bivariate Regression & Interpretation

Your name here

November 28, 2017

Complete the following exercises below and include all code used to find the answers. Knit together the PDF document and commit both the Lab 14 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Select the main focal relationship you're interested in exploring for your poster project.

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

The population number per tract is directly and strongly correlated with the proportion of total population.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
AsianAlone <- read.csv("DEC_10_SF2_PCT1_with_ann.csv")

Totalpopulation <- read_csv("../Datasets/TotalPopulation/ACS_10_5YR_B01003_with_ann.csv",skip=0)
```

```
## Parsed with column specification:
## cols(
##   GEO.id = col_character(),
##   GEO.id2 = col_character(),
##   `GEO.display-label` = col_character(),
##   HD01_VD01 = col_character(),
##   HD02_VD01 = col_character()
## )
```

```r
data_subset_total <- Totalpopulation %>%
  filter(GEO.id != "Id") %>%
  mutate(TotalCount = as.numeric(as.character(HD01_VD01)))

#transform from factor to numeric
Data_Asian_Alone <- AsianAlone %>%
  filter(GEO.id != "Id") %>%
  filter(POPGROUP.id != "012" & POPGROUP.id != "031") %>% #removing the Asian Alone and Asian in combin
  mutate(Count = as.numeric(as.character(D001))) #creating a new column for the population number as nu

#barchart
mergedata <- data_subset_total %>%
  left_join(Data_Asian_Alone, by="GEO.id2") %>%
  select(-GEO.id.y, -ends_with("label.y")) %>%
  mutate(totalprop = Count/ TotalCount)
```

```
## Warning: Column `GEO.id2` joining character vector and factor, coercing
## into character vector
```

```r
#total number of tracts that doesn't have Asian Alone or in combination
sum(is.na(mergedata$Count))
```

```
## [1] 117
```

```r
bivariatepopsimp <- glm(Count~TotalCount,  data = mergedata, family = poisson)
summary(bivariatepopsimp)
```

```
##
## Call:
## glm(formula = Count ~ TotalCount, family = poisson, data = mergedata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -15.523   -7.943   -4.527    2.893   48.765
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 5.133e+00  9.567e-03  536.47   <2e-16 ***
## TotalCount  7.176e-05  1.716e-06   41.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 61663  on 594  degrees of freedom
## Residual deviance: 59922  on 593  degrees of freedom
##   (117 observations deleted due to missingness)
## AIC: 64189
##
## Number of Fisher Scoring iterations: 5
```

```r
#this is the simple bivariate model, taking into account the Total population per tract and the Total a

#regression:
#AsianN = Bcamb + .00008893 * tractN +sum(alfa_group*group_dummy)
#Asian prop = AsianN/tractN = Bcamb + .00008893/tractN
```

```
#Intercept =  4.485e+00 (Cambodian as reference)

#1-(bivariatepopsimp$deviance/bivariatepopsimp$null.deviance)^2
#1-(bivariatepopsimp$deviance/bivariatepop$null.deviance)^2
#amount of variance explained when taking in consideration the tract and the subgroup
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

It is a positive relationship, the slope of the total count is very small (8.893e-05) describing the magnitude, and the p-value is very small as well, showing that there is a strong evidence agaist the Ho of no relationship.

d. What is the meaning of the model intercept? The value of the intercept is 5.133e+00, demosntrating that for a Census tract at 0 population, there is 5.1 Asian representation.

e. How well does the bivariate model fit the data? How is this information calculated?

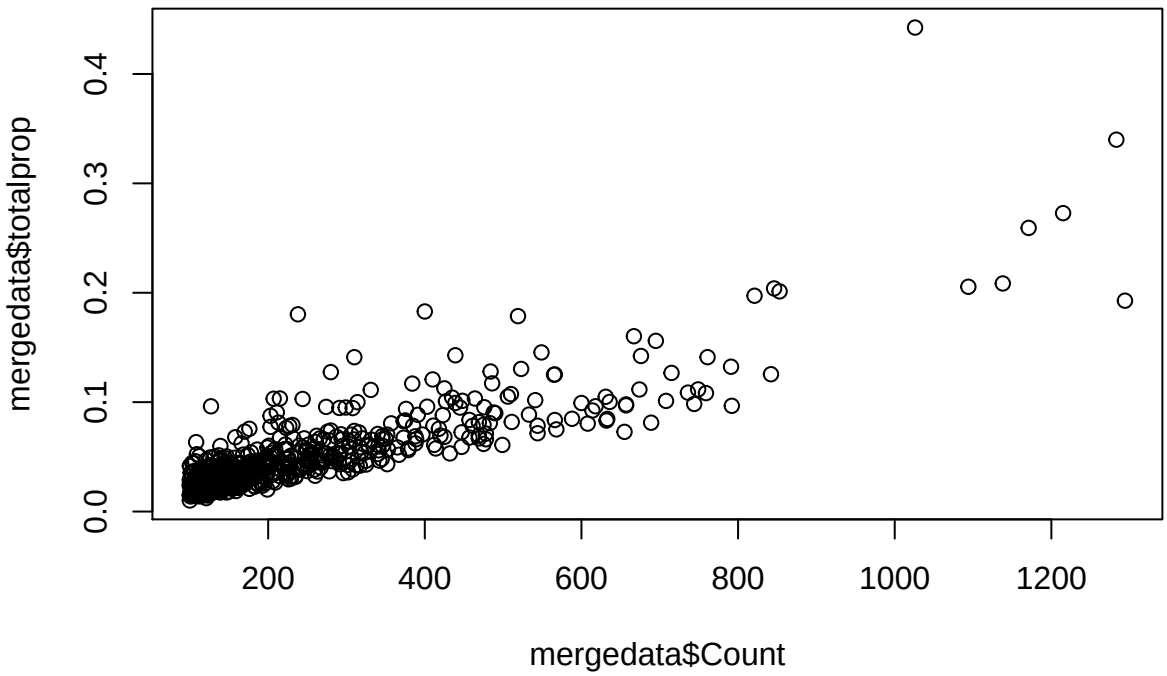
AsianN = Bcamb + .0008893 * tractN +sum(alfa_group*group_dummy) The R^2 is 0.2863632, meaning that 28.6% of the variance of the Asian subgroup population in any given tract is explained by the population of the tract and subgroup identity.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

It is not consistent with the hypothesis as it as assumed in the beginning that there would be a positive association of the Asian subgroup propoetion and the total population in the tract.

2. Select a different focal relationship related to your project. This could be:

```
library(ggplot2)
plot(x=mergedata$Count, y=mergedata$totalprop)
```



- **A different response and a different explanatory variable**
- **A different response and the same explanatory variable**
- **The same response and a different explanatory variable**

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

There is a positive relationship between the two variables, as the Asian population count increases, the Asian population proportion also increases.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

```
simplebivariate <- lm(totalprop~Count, data = mergedata)
summary(simplebivariate)
```

```
##
## Call:
## lm(formula = totalprop ~ Count, data = mergedata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.056760 -0.009223 -0.003286  0.006247  0.244896
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.033e-03  1.459e-03   2.764  0.00589 **
## Count       1.886e-04  4.732e-06  39.856  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02122 on 593 degrees of freedom
##   (117 observations deleted due to missingness)
## Multiple R-squared:  0.7282, Adjusted R-squared:  0.7277
## F-statistic:  1588 on 1 and 593 DF,  p-value: < 2.2e-16
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

d. What is the meaning of the model intercept?

THe intercept y=b if population count is zero

e. How well does the bivariate model fit the data? How is this information calculated?

It fits pretty well, 72% of the variation in the outcome is explained by the variability of the explanatory variable (x-axis) The information is calcularet through R2.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

Yes, it is consistent as it demonstrate positive association expected between the Total count of the population and the population proportion.