Data Analysis Project

BIOST 536

In order to make the household density variable that we are interested in in order to perform a grouped linear adjustment with ordinal variables, we must mutate our data by dividing family size by number of rooms.

```
tbdata = tbdata %>% mutate(famSize = case_when(
  famSize == "<4" ~ 1,
  famSize == "4-6" ~ 2,
  famSize == ">6" ~ 3
))


tbdata = tbdata %>% mutate(roomNo = case_when(
  roomNo == "1" ~ 1,
  roomNo == "2" ~ 2,
  roomNo == "3+" ~ 3
))

tbdata = tbdata %>% mutate(householdDensity = (famSize / roomNo))
```

Now, we can fit a logistic regression model to estimate the risk of TB infection associated with the number of windows in a household, adjusting for sex and our household density variable.

```
tb_mod_glm = glm(case ~ windowNo + male + householdDensity + income,
                 family=binomial,
                 data=tbdata)

summary(tb_mod_glm)
```

```
##
## Call:
## glm(formula = case ~ windowNo + male + householdDensity + income,
##     family = binomial, data = tbdata)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.856  -1.019  -0.133   1.110   1.502
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.5578     0.3147    1.77  0.07637 .
## windowNo1          -0.9966     0.2574   -3.87  0.00011 ***
## windowNo2-3        -1.3068     0.2863   -4.57    5e-06 ***
## windowNo4+         -0.5615     0.3923   -1.43  0.15239
## male                0.0216     0.1890    0.11  0.90920
## householdDensity    0.0336     0.1462    0.23  0.81847
## incomelow           0.8798     0.2511    3.50  0.00046 ***
```

```
## incomemid          0.1667     0.2265    0.74  0.46173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.15  on 499  degrees of freedom
## Residual deviance: 649.72  on 492  degrees of freedom
## AIC: 665.7
##
## Number of Fisher Scoring iterations: 4
```

```
tb_mod_glm_null = glm(case ~ male + householdDensity + income,
                      family=binomial,
                      data=tbdata)
```

In this case, the model estimates that the odds of TB infection for a subpopulation of individuals with 1 window in their house, with the same sex and household density, is 0.369 times the odds of TB infection for a subpopulation of individuals without windows in their house with a 95% CI using model based SEs of [0.221, 0.607].

Comparing individuals with 2-3 windows vs no windows, the OR is now 0.271 with a 95% CI using model based SEs of [0.153, 0.471].

Comparing individuals with 4 or more windows vs no windows, the OR is now 0.57 with a 95% CI using model based SEs of [0.264, 1.233].

Since we have full and null models at this point, we can also run an LRtest in order to assess the significance of the number of windows in a house being a risk factor for TB infection. We'll use the significance level $\alpha = 0.05$

```
lrtest(tb_mod_glm, tb_mod_glm_null)
```

```
## Likelihood ratio test
##
## Model 1: case ~ windowNo + male + householdDensity + income
## Model 2: case ~ male + householdDensity + income
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   8   -325
## 2   5   -338 -3  25.4    1.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At the aforementioned significance level, we can reject the null hypothesis that the number of windows in a house does not have an effect on the risk of TB infection for this population of individuals.