542Regression2

Group 4

After our group meeting, we decided to have 2 hypotheses for this project.

This is a document for our second hypothesis.

Step 1: Import our merged data by using the raw link, and named the data set as "fromPy"

```
link="https://raw.githubusercontent.com/Public-Policy-COVID/students_merge/main/Merged_data.csv"
fromPy=read.csv(link, header = T)
row.names(fromPy)=NULL
```

Step 2: Verifying the data structure by using the following code

This step can tell you the type of our variables. We can change their types in future clustering or regression.

```
# verifying data structure
str(fromPy,width = 50,strict.width='cut')
```

```
## 'data.frame':    133 obs. of  19 variables:
##  $ Number_of_beds       : num  3667 0 52 553 25 ..
##  $ Number_of_hospitals  : num  22 0 1 6 1 1 10 1..
##  $ Location             : chr  "Alameda_CA" "Al"..
##  $ Urban_Rural_Code     : chr  "Large central m"..
##  $ Deaths_COVID         : int  573 0 31 101 12 1..
##  $ Deaths_total         : int  10908 0 415 2313 ..
##  $ never                : num  0.019 0.025 0.045..
##  $ rarely               : num  0.008 0.085 0.013..
##  $ sometimes            : num  0.055 0.088 0.099..
##  $ frequently           : num  0.123 0.19 0.188 ..
##  $ always               : num  0.795 0.612 0.655..
##  $ mask_score           : num  3.67 3.28 3.4 3.3..
##  $ total_population     : num  1671329 1129 3975..
##  $ white_total_pct      : num  49.3 67.9 89.7 85..
##  $ black_total_pct      : num  11.03 0.35 2.68 1..
##  $ aian_total_pct       : num  1.06 25.69 2.33 2..
##  $ asian_total_pct      : num  32.33 1.59 1.67 5..
```

```
##  $ nhopi_total_pct      : num  0.94 0 0.29 0.29 ..
##  $ multiracial_total_pct: num  5.35 4.43 3.38 4...
```

Step 3: Convert integer variables to decimal variables

This step is not necessary, as integer variables are also numeric variables. It wouldn't influence our regression. I just want to keep the variable structure constant in this analysis.

```
fromPy$Deaths_COVID <- as.numeric(fromPy$Deaths_COVID)
fromPy$Deaths_total <- as.numeric(fromPy$Deaths_total)
str(fromPy,width = 50,strict.width='cut')
```

```
## 'data.frame':    133 obs. of  19 variables:
##  $ Number_of_beds       : num  3667 0 52 553 25 ..
##  $ Number_of_hospitals  : num  22 0 1 6 1 1 10 1..
##  $ Location             : chr  "Alameda_CA" "Al"..
##  $ Urban_Rural_Code     : chr  "Large central m"..
##  $ Deaths_COVID         : num  573 0 31 101 12 1..
##  $ Deaths_total         : num  10908 0 415 2313 ..
##  $ never                : num  0.019 0.025 0.045..
##  $ rarely               : num  0.008 0.085 0.013..
##  $ sometimes            : num  0.055 0.088 0.099..
##  $ frequently           : num  0.123 0.19 0.188 ..
##  $ always               : num  0.795 0.612 0.655..
##  $ mask_score           : num  3.67 3.28 3.4 3.3..
##  $ total_population     : num  1671329 1129 3975..
##  $ white_total_pct      : num  49.3 67.9 89.7 85..
##  $ black_total_pct      : num  11.03 0.35 2.68 1..
##  $ aian_total_pct       : num  1.06 25.69 2.33 2..
##  $ asian_total_pct      : num  32.33 1.59 1.67 5..
##  $ nhopi_total_pct      : num  0.94 0 0.29 0.29 ..
##  $ multiracial_total_pct: num  5.35 4.43 3.38 4...
```

Step 4: Summary of the data set

This step is for understanding the basic information of each variable. Such as the minimum, maximum, median, mean, etc.

```
summary(fromPy)
```

```
##  Number_of_beds    Number_of_hospitals   Location         Urban_Rural_Code
##  Min.   :    0.0   Min.   :  0         Length:133         Length:133
##  1st Qu.:   25.0   1st Qu.:  1         Class :character   Class :character
##  Median :  131.0   Median :  2         Mode  :character   Mode  :character
```

```
##  Mean   :  885.4   Mean   :  5
##  3rd Qu.:  553.0   3rd Qu.:  4
##  Max.   :26672.0   Max.   :112
##   Deaths_COVID   Deaths_total       never             rarely
##  Min.   :   0   Min.   :    0   Min.   :0.00100   Min.   :0.00000
##  1st Qu.:   0   1st Qu.:    0   1st Qu.:0.01600   1st Qu.:0.01400
##  Median :  22   Median :  637   Median :0.02600   Median :0.02800
##  Mean   : 206   Mean   : 2896   Mean   :0.03513   Mean   :0.03806
##  3rd Qu.: 128   3rd Qu.: 2537   3rd Qu.:0.04500   3rd Qu.:0.05600
##  Max.   :8034   Max.   :75463   Max.   :0.14000   Max.   :0.20600
##    sometimes         frequently         always         mask_score
##  Min.   :0.00400   Min.   :0.0580   Min.   :0.3050   Min.   :2.470
##  1st Qu.:0.04800   1st Qu.:0.1410   1st Qu.:0.6160   1st Qu.:3.301
##  Median :0.06900   Median :0.1680   Median :0.6810   Median :3.464
##  Mean   :0.07167   Mean   :0.1736   Mean   :0.6814   Mean   :3.428
##  3rd Qu.:0.09100   3rd Qu.:0.2040   3rd Qu.:0.7540   3rd Qu.:3.591
##  Max.   :0.21300   Max.   :0.3320   Max.   :0.8890   Max.   :3.822
##  total_population   white_total_pct black_total_pct  aian_total_pct
##  Min.   :    1129   Min.   :49.28   Min.   : 0.000   Min.   : 0.590
##  1st Qu.:   24658   1st Qu.:82.16   1st Qu.: 0.770   1st Qu.: 1.430
##  Median :   79481   Median :88.64   Median : 1.260   Median : 2.010
##  Mean   :  385537   Mean   :85.50   Mean   : 2.318   Mean   : 2.985
##  3rd Qu.:  283111   3rd Qu.:91.84   3rd Qu.: 2.620   3rd Qu.: 3.070
##  Max.   :10039107   Max.   :96.13   Max.   :14.770   Max.   :25.690
##  asian_total_pct  nhopi_total_pct  multiracial_total_pct
##  Min.   : 0.500   Min.   :0.0000   Min.   :1.200
##  1st Qu.: 1.210   1st Qu.:0.2100   1st Qu.:3.160
##  Median : 1.870   Median :0.2800   Median :3.720
##  Mean   : 4.961   Mean   :0.3838   Mean   :3.856
##  3rd Qu.: 5.840   3rd Qu.:0.4500   3rd Qu.:4.440
##  Max.   :39.020   Max.   :1.7100   Max.   :7.800
```

Step 5: State the second hypothesis, and name it "hypo2"

1. hypo2 = hypothesis 2: state with higher Deaths_COVID number has more Number_of_beds in hospitals.

2. Besides that, we think the hospital beds would be correlated with the total population, suggesting county with more population would have more beds.

3. What's more, we also want to know if race variables are significant in this analysis, thereby we added all race variables in this regression to check their relationship with the number of beds.

```
hypo2=formula(Number_of_beds~ Deaths_COVID+total_population+black_total_pct+a
ian_total_pct+asian_total_pct+nhopi_total_pct+multiracial_total_pct)
```

Step 6: Using (glm) code to compute the regression model glm stands for 'Generalized Linear Models'

```
gauss2=glm(hypo2,
           data = fromPy,
           family = 'gaussian')
```

Step 7: See the result of our regression

By using code (summary), we are able to check the result of our regression.

```
summary(gauss2)
```

```
##
## Call:
## glm(formula = hypo2, family = "gaussian", data = fromPy)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2149.50    -94.91     -2.91     70.70   2337.16
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.155e+02  1.400e+02   0.825 0.411063
## Deaths_COVID           1.313e+00  2.778e-01   4.727 6.04e-06 ***
## total_population       1.551e-03  2.199e-04   7.054 1.05e-10 ***
## black_total_pct       -2.731e+01  2.062e+01  -1.324 0.187765
## aian_total_pct         4.344e+00  1.191e+01   0.365 0.716016
## asian_total_pct        3.176e+01  8.762e+00   3.625 0.000419 ***
## nhopi_total_pct       -1.723e+02  1.567e+02  -1.100 0.273487
## multiracial_total_pct -3.622e+01  4.310e+01  -0.841 0.402221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 169887.2)
##
##     Null deviance: 930593834  on 132  degrees of freedom
## Residual deviance:  21235896  on 125  degrees of freedom
## AIC: 1988.9
##
## Number of Fisher Scoring iterations: 2
```

RESULTS: Based on the results of this regression, we can tell that variable 'Deaths_COVID', 'total_population', and 'asian_total_pct' is statistically significantly correlated with our dependent variable 'Number_of_beds' at 99% confidence interval. This suggests:

1. For each county, with a 1 case increase in COVID death, there will be a 1.31 increase in the number of hospital beds.

2. For each county, with 1 person increase in the total population, there will be a 1.55 increase in the number of beds in the number of hospital beds.

3. For each county, with a 1 percent increase in the proportion of the Asian population, there will be a 3.18 increase in the number of hospital beds.

Step 8: Get the R square of this regression

R square can tell us the percentage of the response variable variation that is explained by our model. This step is to check whether this regression is an effective model. Normally, the higher the R-squared, the better the model fits our data.

```
library(rsq)
rsq(gauss2, adj = T)
```

```
## [1] 0.9759024
```

Step 9: To have some summary plots for our analysis

9-1 This plot is for the coefficient estimates. We need to load the required package 'ggplot2' for it.

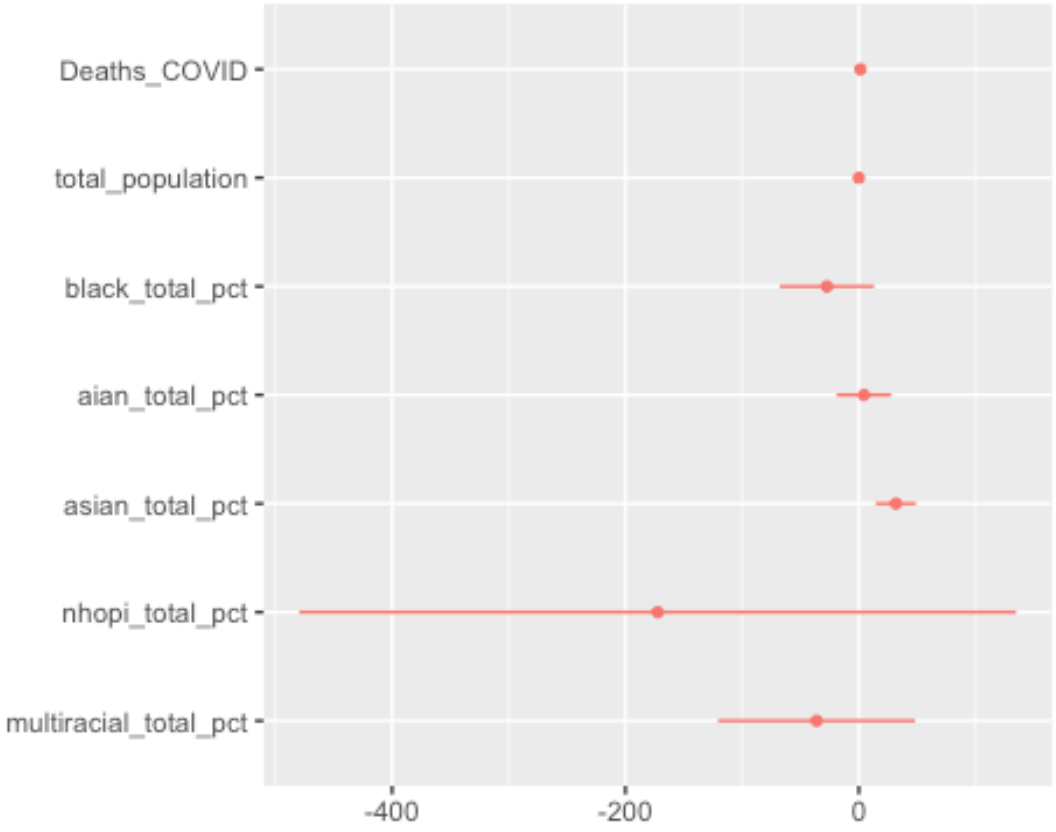
```
library(dotwhisker)
```

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

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency dete
cted.
## TMB was built with Matrix version 1.3.2
## Current Matrix version is 1.2.18
## Please re-install 'TMB' from source using install.packages('TMB', type = '
source') or ask CRAN for a binary version of 'TMB' matching CRAN's 'Matrix' p
ackage
```

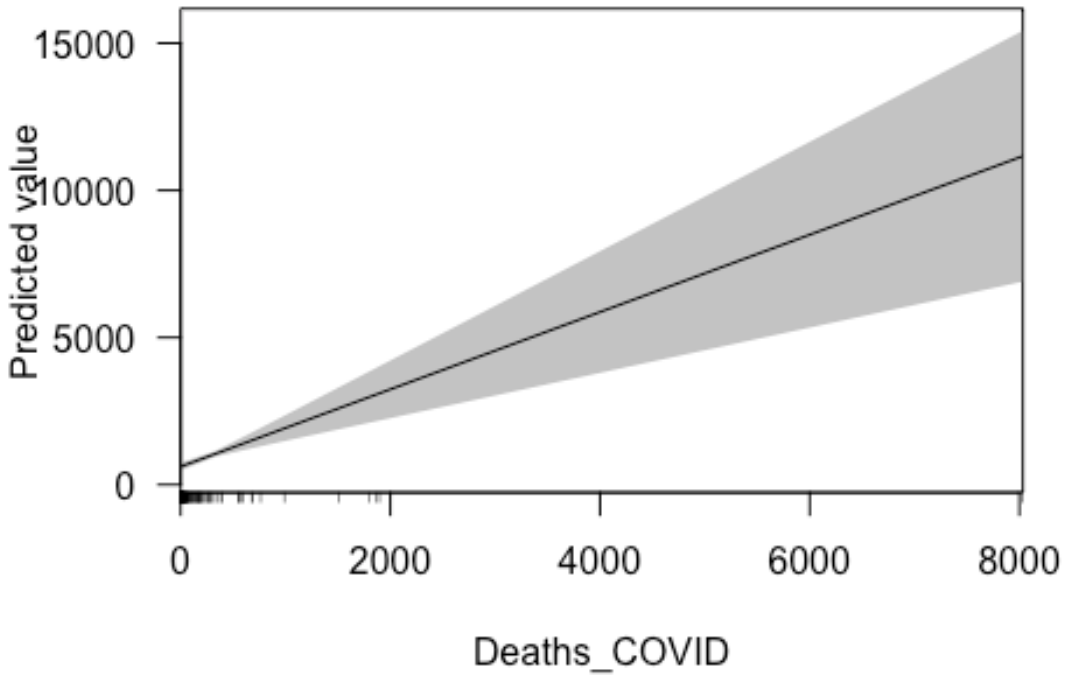
```
## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```

```
dwplot(gauss2,by_2sd = F)
```

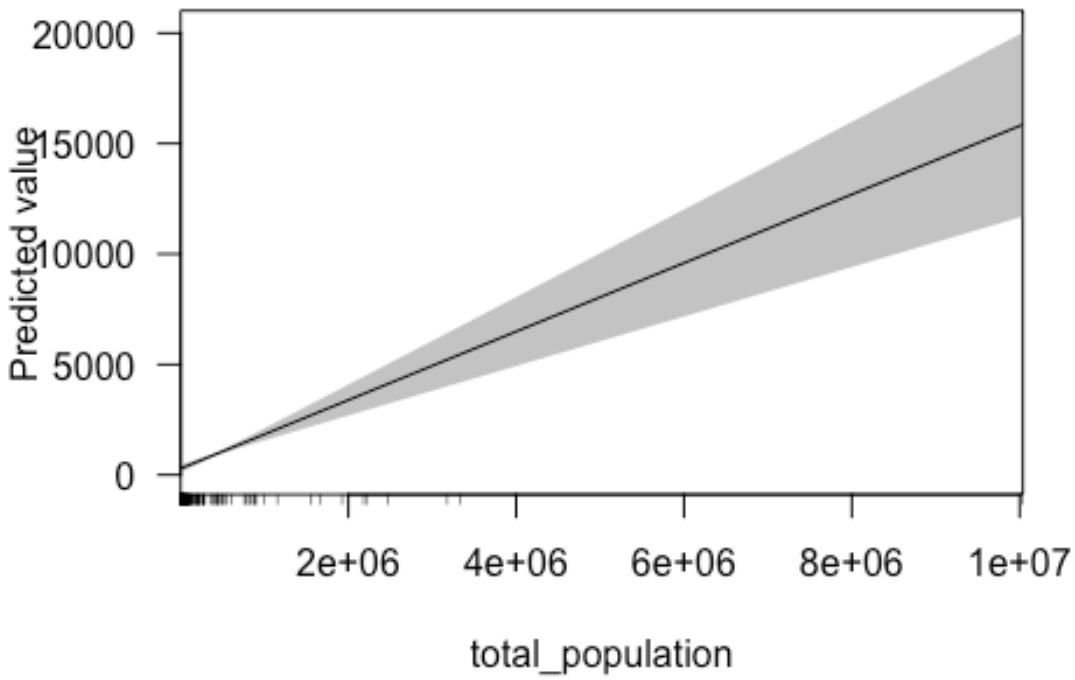


9-2 The following plots are for the margins of each independent variable. We need to use the margins library package.

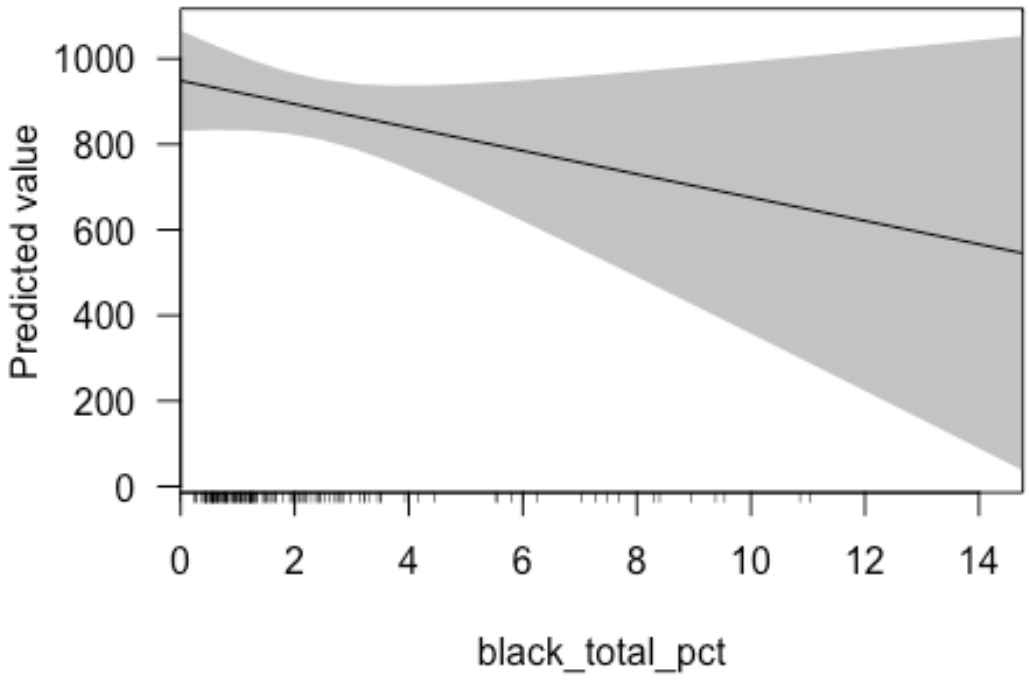
```
library(margins)

cplot(gauss2,'Deaths_COVID')
```



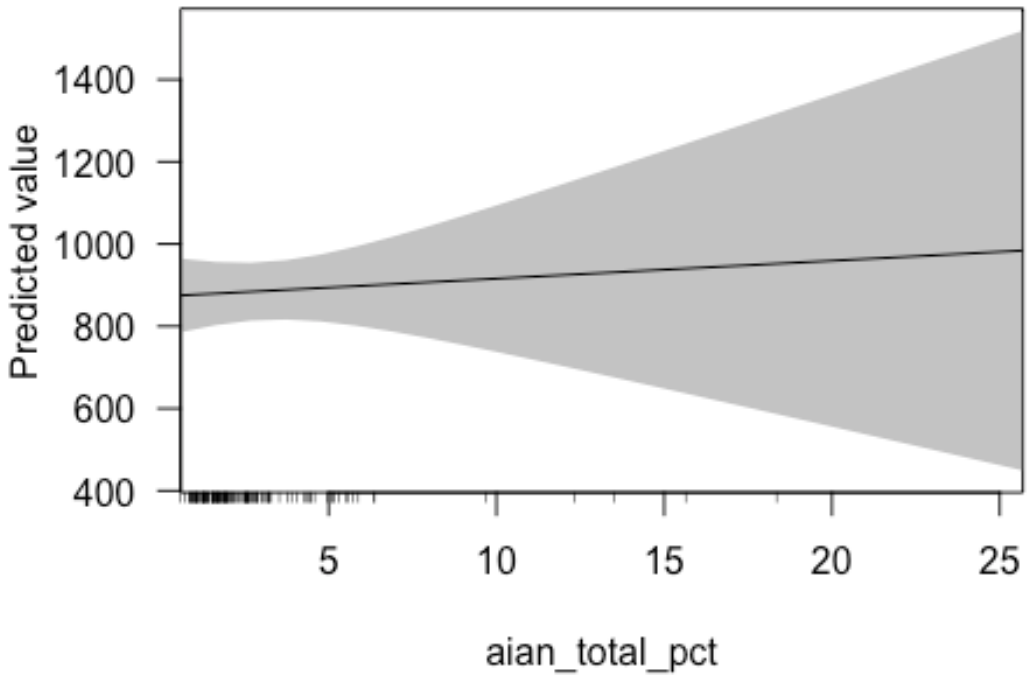

```
cplot(gauss2,'total_population')
```



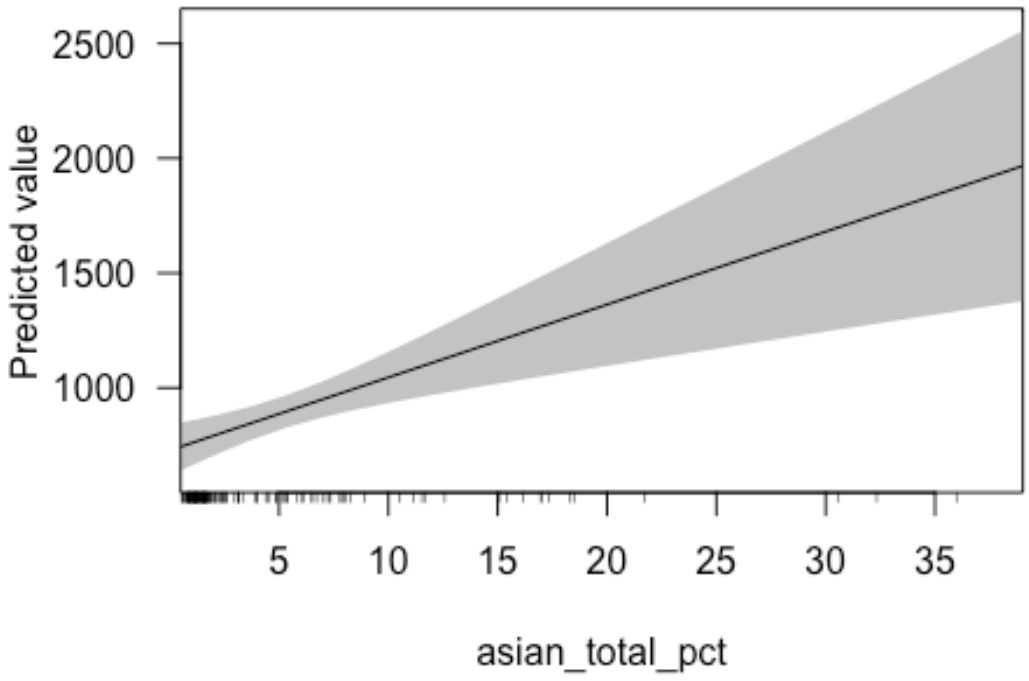

```
cplot(gauss2,'black_total_pct')
```



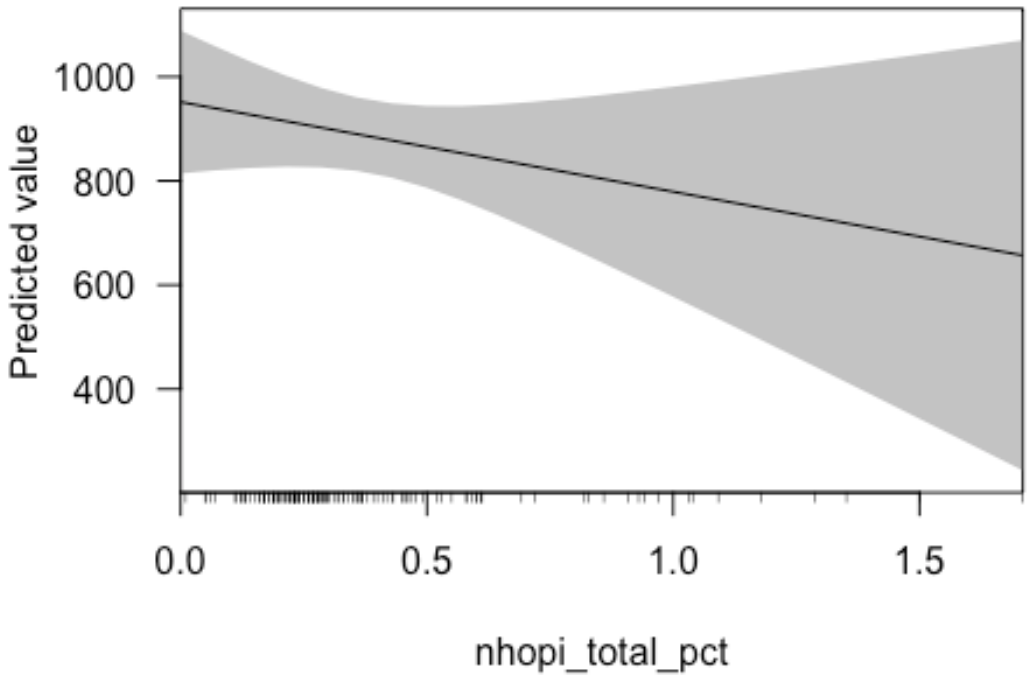

```
cplot(gauss2,'aian_total_pct')
```



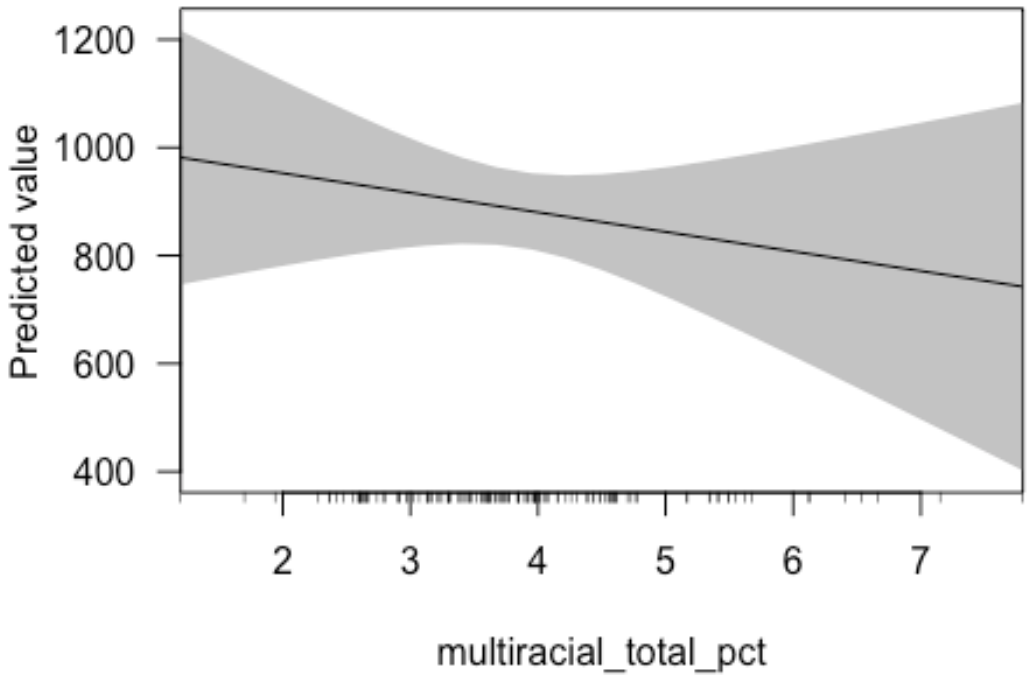

```
cplot(gauss2,'asian_total_pct')
```




```
cplot(gauss2,'nhopi_total_pct')
```

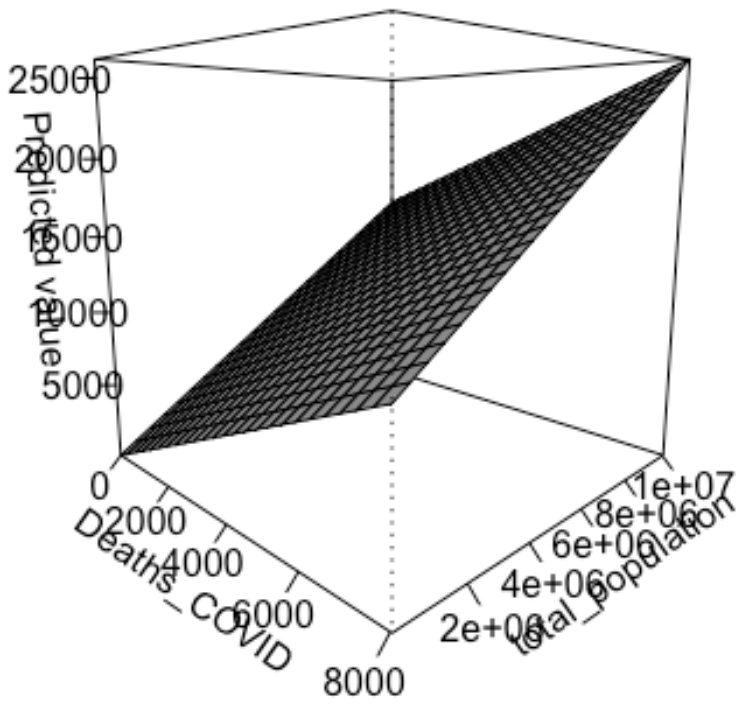



```
cplot(gauss2,'multiracial_total_pct')
```



9-3 To plot the ineraction between our independent variables.

```
persp(gauss2)
```


25000
20000
15000
10000
5000
Predicted value
0
2000
4000
6000
8000
Deaths_COVID
2e+06
4e+06
6e+06
8e+06
1e+07
Total_population