Ramirez Stats SYE .csv

# Setup

rm(list = ls())  
library(here)  
library(tidyverse)  
library(ggfortify)  
library(emmeans)  
library(betareg)  
library(multcomp)  
  
covid<-read.csv(here("data","CovidbyRaceandState.csv"))  
covid<-covid|>dplyr::select(c(2:11))  
glimpse(covid)

fixing data types

covid$Date<-ymd(covid$Date)  
covid$State<-as.factor(covid$State)  
covid$Race<-as.factor(covid$Race)  
covid$Region<-as.factor(covid$Region)  
covid$Gov\_Control<-as.factor(covid$Gov\_Control)  
glimpse(covid)

Rows: 408  
Columns: 10  
$ Date <date> 2021-03-07, 2021-03-07, 2021-03-07, 2021-03-07, 20…  
$ State <fct> Alaska, Alaska, Alaska, Alaska, Alaska, Alaska, Ala…  
$ Race <fct> Total, White, Black, Asian, AIAN, NHPI, Multiracial…  
$ Cases <int> 59332, 18300, 1499, 2447, 12238, 1508, 4453, 7130, …  
$ Deaths <int> 305, 127, 9, 29, 111, 17, 4, 1, 10148, 4730, 2223, …  
$ Hosp <int> 1293, 423, 45, 116, 336, 126, 67, 54, NA, NA, NA, N…  
$ Tests <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ Region <fct> Pacific, Pacific, Pacific, Pacific, Pacific, Pacifi…  
$ Gov\_Control <fct> Red, Red, Red, Red, Red, Red, Red, Red, Red, Red, R…  
$ Population\_By\_Race <int> 733391, 450472, 23395, 47464, 104957, 11209, 82727,…

# Multiple Regression

## Setting up for our model

# Extract total population from rows where Race == "Total"  
state\_totals <- covid|>  
 filter(Race == "Total")|>  
 dplyr::select(State, Population\_By\_Race)|>  
 dplyr::rename(Total\_Population = Population\_By\_Race)  
# Remove "Total" rows from the main data  
covid\_filtered <- covid|>  
 filter(Race != "Total")|>  
 left\_join(state\_totals, by = "State")  
# Create Race\_Proportion variable  
covid1 <- covid\_filtered|>  
 mutate(Race\_Proportion = Population\_By\_Race / Total\_Population)  
summary(covid1$Race\_Proportion)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
0.0001945 0.0084955 0.0394199 0.1427775 0.1036212 0.9258088

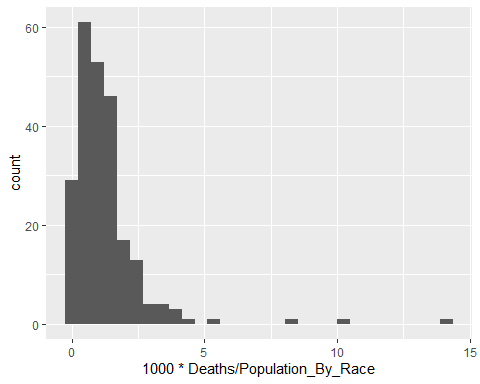
Taking out super low proportions due to skewing caused by such low numbers, also taking zeros out.

covidna<-covid1|>  
 filter(!is.na(Deaths))|>  
 filter(Deaths>0)  
covidna<-covidna|>  
 mutate(death\_rate\_prop = Deaths / Population\_By\_Race)  
covidna<-covidna|>  
 filter(Race\_Proportion>=0.005)

## Visualize distribution of death rate

ggplot(covidna,aes(1000\*Deaths/Population\_By\_Race))+  
 geom\_histogram()

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Fitting and analyzing a model

must set the baseline levels to compare to

covidna <- covidna|>  
 mutate(Race = relevel(factor(Race), ref = "White"))  
covidna <- covidna|>  
 mutate(Gov\_Control = relevel(factor(Gov\_Control), ref = "Red"))  
covidna <- covidna|>  
 mutate(Region = relevel(factor(Region), ref = "Northeast"))

model\_beta <- betareg(death\_rate\_prop ~ Race+Gov\_Control+Region, data = covidna)  
summary(model\_beta)

Call:  
betareg(formula = death\_rate\_prop ~ Race + Gov\_Control + Region, data = covidna)  
  
Quantile residuals:  
 Min 1Q Median 3Q Max   
-3.1370 -0.5002 -0.0806 0.3629 5.4567   
  
Coefficients (mean model with logit link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -6.555597 0.170129 -38.533 < 2e-16 \*\*\*  
RaceAIAN 0.293952 0.153727 1.912 0.055854 .   
RaceAsian -0.478085 0.136091 -3.513 0.000443 \*\*\*  
RaceBlack -0.030232 0.123184 -0.245 0.806129   
RaceMultiracial -1.296711 0.250529 -5.176 2.27e-07 \*\*\*  
RaceNHPI 0.532099 0.318884 1.669 0.095192 .   
RaceOther -0.438776 0.135841 -3.230 0.001238 \*\*   
Gov\_ControlBlue -0.008681 0.127522 -0.068 0.945729   
Gov\_ControlDivided -0.026301 0.117851 -0.223 0.823402   
RegionMidatlantic 0.056811 0.179548 0.316 0.751693   
RegionMidwest 0.243234 0.163726 1.486 0.137382   
RegionPacific -0.440964 0.183123 -2.408 0.016039 \*   
RegionSouth 0.207406 0.172673 1.201 0.229694   
RegionWest 0.016221 0.162292 0.100 0.920386   
  
Phi coefficients (precision model with identity link):  
 Estimate Std. Error z value Pr(>|z|)   
(phi) 1369.4 136.5 10.03 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
  
Type of estimator: ML (maximum likelihood)  
Log-likelihood: 1377 on 15 Df  
Pseudo R-squared: 0.3409  
Number of iterations: 63 (BFGS) + 5 (Fisher scoring)

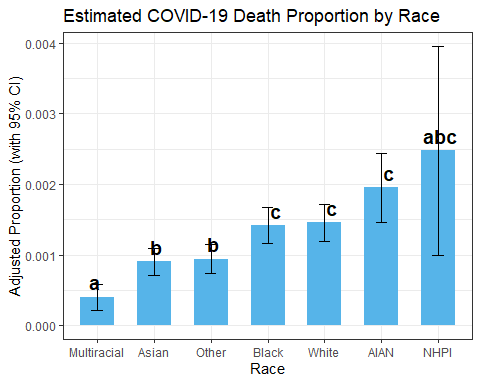
exp(coef(model\_beta))

(Intercept) RaceAIAN RaceAsian RaceBlack   
 0.001422133 1.341719616 0.619969301 0.970220296   
 RaceMultiracial RaceNHPI RaceOther Gov\_ControlBlue   
 0.273429693 1.702501321 0.644825174 0.991356966   
Gov\_ControlDivided RegionMidatlantic RegionMidwest RegionPacific   
 0.974041904 1.058455304 1.275366648 0.643415560   
 RegionSouth RegionWest (phi)   
 1.230482136 1.016352981 Inf

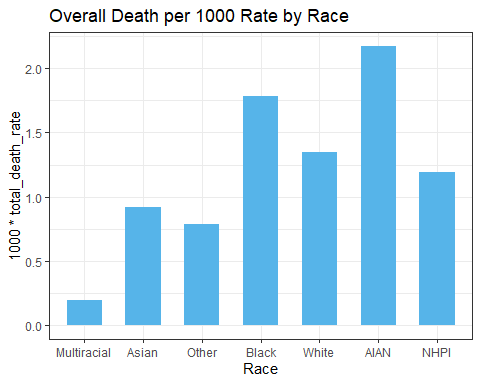
# Get marginal means (still on response scale)  
race\_emm <- emmeans(model\_beta, ~ Race)  
# Get confidence intervals on response scale  
race\_ci <- confint(race\_emm, type = "response")|>  
 as.data.frame()  
# Check to verify:  
names(race\_ci)

[1] "Race" "emmean" "SE" "df" "asymp.LCL" "asymp.UCL"

# Get compact letter display (still from response-scale estimates)  
race\_cld <- cld(race\_emm, Letters = letters, type = "response")  
  
# Merge group letters into CI data  
race\_df <- race\_ci %>%  
 left\_join(dplyr::select(as.data.frame(race\_cld), Race, .group), by = "Race")  
  
# Plot  
ggplot(race\_df, aes(x = reorder(Race, emmean), y = emmean)) +  
 geom\_col(fill = "#56B4E9", width = 0.6) +  
 geom\_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL), width = 0.2) +  
 geom\_text(aes(label = .group), vjust = -0.5, size = 5, fontface = "bold") +  
 labs(  
 title = "Estimated COVID-19 Death Proportion by Race",  
 x = "Race",  
 y = "Adjusted Proportion (with 95% CI)"  
 ) +  
 theme\_bw()

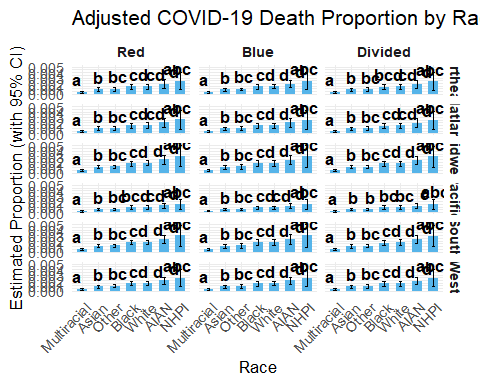


#ggsave("EstDeathRate.jpg",width = 10,height = 6)  
  
covidnasum1<-covidna|>  
 group\_by(Race)|>  
 summarize(total\_death\_rate=sum(Deaths)/sum(Population\_By\_Race))  
ggplot(covidnasum1,aes(factor(Race,levels = c("Multiracial","Asian","Other","Black","White","AIAN","NHPI")),1000\*total\_death\_rate))+  
 geom\_col(fill = "#56B4E9", width = 0.6)+  
 labs(x="Race",title = "Overall Death per 1000 Rate by Race")+theme\_bw()



#ggsave("RealDeathProp.jpg",width = 10,height = 6)

# Get marginal means by Race × Gov\_Control × Region  
emm\_race\_full <- emmeans(model\_beta, ~ Race | Gov\_Control \* Region)  
# Convert to data frame with CIs  
race\_ci <- confint(emm\_race\_full, type = "response")|>  
 as.data.frame()  
# Significance groups per panel (Gov\_Control × Region)  
race\_cld <- cld(emm\_race\_full, Letters = letters, type = "response")  
# Merge group letters into CI data  
race\_df <- race\_ci %>%  
 left\_join(dplyr::select(as.data.frame(race\_cld), Race, Gov\_Control, Region, .group),  
 by = c("Race", "Gov\_Control", "Region"))  
ggplot(race\_df, aes(x = reorder(Race, emmean), y = emmean)) +  
 geom\_col(fill = "#56B4E9", width = 0.6) +  
 geom\_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL), width = 0.2) +  
 geom\_text(aes(label = .group), vjust = -0.5, size = 4.5, fontface = "bold") +  
 facet\_grid(Region ~ Gov\_Control) + # Facet by Region (rows) and Gov\_Control (columns)  
 labs(  
 title = "Adjusted COVID-19 Death Proportion by Race, Region, and Government Control",  
 x = "Race",  
 y = "Estimated Proportion (with 95% CI)"  
 ) +  
 theme\_minimal(base\_size = 13) +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 strip.text = element\_text(face = "bold")  
 )



## Setting up for our model

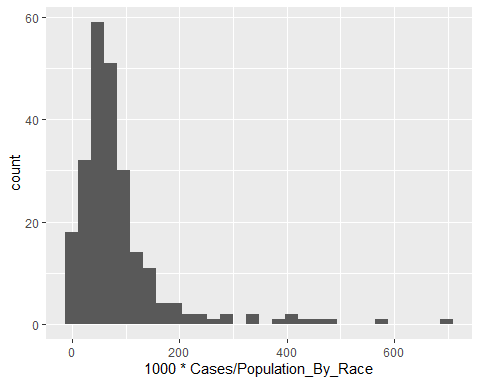
Taking out super low proportions due to skewing caused by such low numbers, also taking zeros out.

covidnacase<-covid1|>  
 filter(!is.na(Cases))|>  
 filter(Cases>0)  
covidnacase<-covidnacase|>  
 mutate(case\_rate\_prop = Cases / Population\_By\_Race)  
covidnacase<-covidnacase|>  
 filter(Race\_Proportion>=0.005)

## Visualize distribution of death rate

ggplot(covidnacase,aes(1000\*Cases/Population\_By\_Race))+  
 geom\_histogram()

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Fitting and analyzing a model

must set the baseline levels to compare to

covidnacase <- covidnacase|>  
 mutate(Race = relevel(factor(Race), ref = "White"))  
covidnacase <- covidnacase|>  
 mutate(Gov\_Control = relevel(factor(Gov\_Control), ref = "Red"))  
covidnacase <- covidnacase|>  
 mutate(Region = relevel(factor(Region), ref = "Northeast"))

model\_betacase <- betareg(case\_rate\_prop ~ Race+Gov\_Control+Region, data = covidnacase)  
summary(model\_betacase)

Call:  
betareg(formula = case\_rate\_prop ~ Race + Gov\_Control + Region, data = covidnacase)  
  
Quantile residuals:  
 Min 1Q Median 3Q Max   
-5.0452 -0.3937 -0.0818 0.3009 4.7887   
  
Coefficients (mean model with logit link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -2.39794 0.19050 -12.588 < 2e-16 \*\*\*  
RaceAIAN 0.45697 0.18058 2.531 0.01139 \*   
RaceAsian -0.12718 0.15485 -0.821 0.41147   
RaceBlack 0.20941 0.14672 1.427 0.15350   
RaceMultiracial -0.71540 0.24308 -2.943 0.00325 \*\*   
RaceNHPI 0.92654 0.34262 2.704 0.00684 \*\*   
RaceOther 0.65742 0.14016 4.690 2.73e-06 \*\*\*  
Gov\_ControlBlue -0.41119 0.13526 -3.040 0.00237 \*\*   
Gov\_ControlDivided -0.03444 0.12552 -0.274 0.78378   
RegionMidatlantic 0.14957 0.19346 0.773 0.43944   
RegionMidwest 0.06751 0.17689 0.382 0.70270   
RegionPacific -0.18543 0.19597 -0.946 0.34403   
RegionSouth -0.10043 0.18934 -0.530 0.59582   
RegionWest 0.11553 0.17648 0.655 0.51271   
  
Phi coefficients (precision model with identity link):  
 Estimate Std. Error z value Pr(>|z|)   
(phi) 17.822 1.698 10.5 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
  
Type of estimator: ML (maximum likelihood)  
Log-likelihood: 377.7 on 15 Df  
Pseudo R-squared: 0.2992  
Number of iterations: 23 (BFGS) + 1 (Fisher scoring)

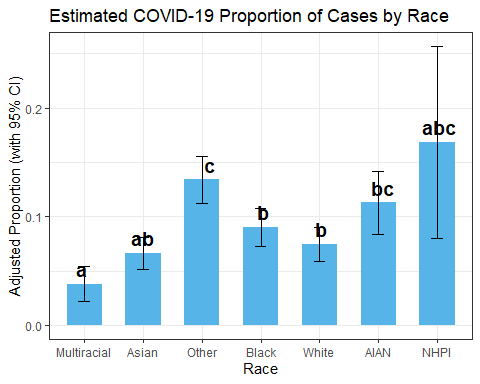
exp(coef(model\_betacase))

(Intercept) RaceAIAN RaceAsian RaceBlack   
 9.090493e-02 1.579283e+00 8.805773e-01 1.232952e+00   
 RaceMultiracial RaceNHPI RaceOther Gov\_ControlBlue   
 4.889974e-01 2.525753e+00 1.929808e+00 6.628623e-01   
Gov\_ControlDivided RegionMidatlantic RegionMidwest RegionPacific   
 9.661434e-01 1.161336e+00 1.069844e+00 8.307448e-01   
 RegionSouth RegionWest (phi)   
 9.044484e-01 1.122470e+00 5.496396e+07

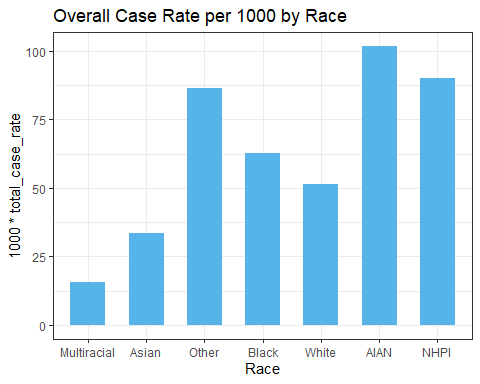
# Get marginal means (still on response scale)  
race\_emmcase <- emmeans(model\_betacase, ~ Race)  
# Get confidence intervals on response scale  
race\_cicase <- confint(race\_emmcase, type = "response")|>  
 as.data.frame()  
# Check to verify:  
names(race\_cicase)

[1] "Race" "emmean" "SE" "df" "asymp.LCL" "asymp.UCL"

# Get compact letter display (still from response-scale estimates)  
race\_cldcase <- cld(race\_emmcase, Letters = letters, type = "response")  
  
# Merge group letters into CI data  
race\_dfcase <- race\_cicase %>%  
 left\_join(dplyr::select(as.data.frame(race\_cldcase), Race, .group), by = "Race")  
  
# Plot  
ggplot(race\_dfcase, aes(x = factor(Race,levels = c("Multiracial","Asian","Other","Black","White","AIAN","NHPI")), y = emmean)) +  
 geom\_col(fill = "#56B4E9", width = 0.6) +  
 geom\_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL), width = 0.2) +  
 geom\_text(aes(label = .group), vjust = -0.5, size = 5, fontface = "bold") +  
 labs(  
 title = "Estimated COVID-19 Proportion of Cases by Race",  
 x = "Race",  
 y = "Adjusted Proportion (with 95% CI)"  
 ) +  
 theme\_bw()



# ggsave("EstCaseRate.jpg",width = 10,height = 6)  
covidnasumcase<-covidnacase|>  
 group\_by(Race)|>  
 summarize(total\_case\_rate=sum(Cases)/sum(Population\_By\_Race))  
ggplot(covidnasumcase,aes(factor(Race,levels = c("Multiracial","Asian","Other","Black","White","AIAN","NHPI")),1000\*total\_case\_rate))+  
 geom\_col(fill = "#56B4E9", width = 0.6)+  
 labs(x="Race",title = "Overall Case Rate per 1000 by Race")+theme\_bw()



#ggsave("RealCaseProp.jpg",width = 10,height = 6)

emm\_race\_fullcase <- emmeans(model\_betacase, ~ Race | Gov\_Control \* Region)  
race\_cicase <- confint(emm\_race\_fullcase, type = "response")|>  
 as.data.frame()  
race\_cldcase <- cld(emm\_race\_fullcase, Letters = letters, type = "response")  
race\_dfcase <- race\_cicase %>%  
 left\_join(dplyr::select(as.data.frame(race\_cldcase), Race, Gov\_Control, Region, .group),  
 by = c("Race", "Gov\_Control", "Region"))  
ggplot(race\_dfcase, aes(x = reorder(Race, emmean), y = emmean)) +  
 geom\_col(fill = "#56B4E9", width = 0.6) +  
 geom\_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL), width = 0.2) +  
 geom\_text(aes(label = .group), vjust = -0.5, size = 4.5, fontface = "bold") +  
 facet\_grid(Region ~ Gov\_Control) +   
 labs(  
 title = "Adjusted COVID-19 Case Proportion by Race, Region, and Government Control",  
 x = "Race",  
 y = "Estimated Proportion (with 95% CI)"  
 ) +  
 theme\_minimal(base\_size = 13) +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 strip.text = element\_text(face = "bold")  
 )

