Lab Exercise - Jan22nd - GLM, Multinomial

Luis Correia - Student No. 1006508566

January 15th 2020

Overview

Today we are looking at data on infant deaths (deaths in the first year of life) in the US. The dataset `infant` contains information on all deaths to the 2012 birth cohort. For today, we are interested in investigating differences in neonatal deaths (i.e. deaths in the first month of life) and cause of death.

What to hand in

As with last week, please push your Rmd and compiled document (html or pdf) to GitHub. **The questions for this week are dispersed throughout the lab.**

The dataset

Read it in and have a look to see what's in there. Variables are

- `sex`: sex of baby
- `aged`: age at death (in days)
- `race`: race of mother
- `gest`: gestation in weeks
- `ucod`: cause of death (ICD-10 code)
- `cod`: cause of death, descriptive groups
- `mom_age`: mother age in years
- `mom_age_group`: mother age group

```r
library(tidyverse)
path <- "C:/Users/LuisAlvaro/Documents/GitHub/applied-stats/data/infant.RDS"
d <- read_rds(path)
head(d)
```

```
## # A tibble: 6 x 8
##   sex    aged race   gest ucod  cod       mom_age mom_age_group
##   <chr> <dbl> <chr> <dbl> <chr> <chr>       <dbl> <fct>
## 1 F         0 NHW      27 P832  peri_oth       30 30
## 2 M         0 NHW      36 Q913  cong_mal       32 30
## 3 M         8 NHW      44 P360  peri_inf       25 25
## 4 F         0 NHB      21 P072  peri_comp      29 25
## 5 M         8 NHB      26 P220  peri_resp      23 20
## 6 M        17 NHW      39 Q249  cong_mal       34 30
```

Descriptives

Let's create some new variables that will be useful:

- `neo_death`: equals 1 if the death occurred in the first 28 days

- `preterm`: equals 1 if gestational age is less than 37 weeks
- `cod_group`: reduced number of categories of cause of death

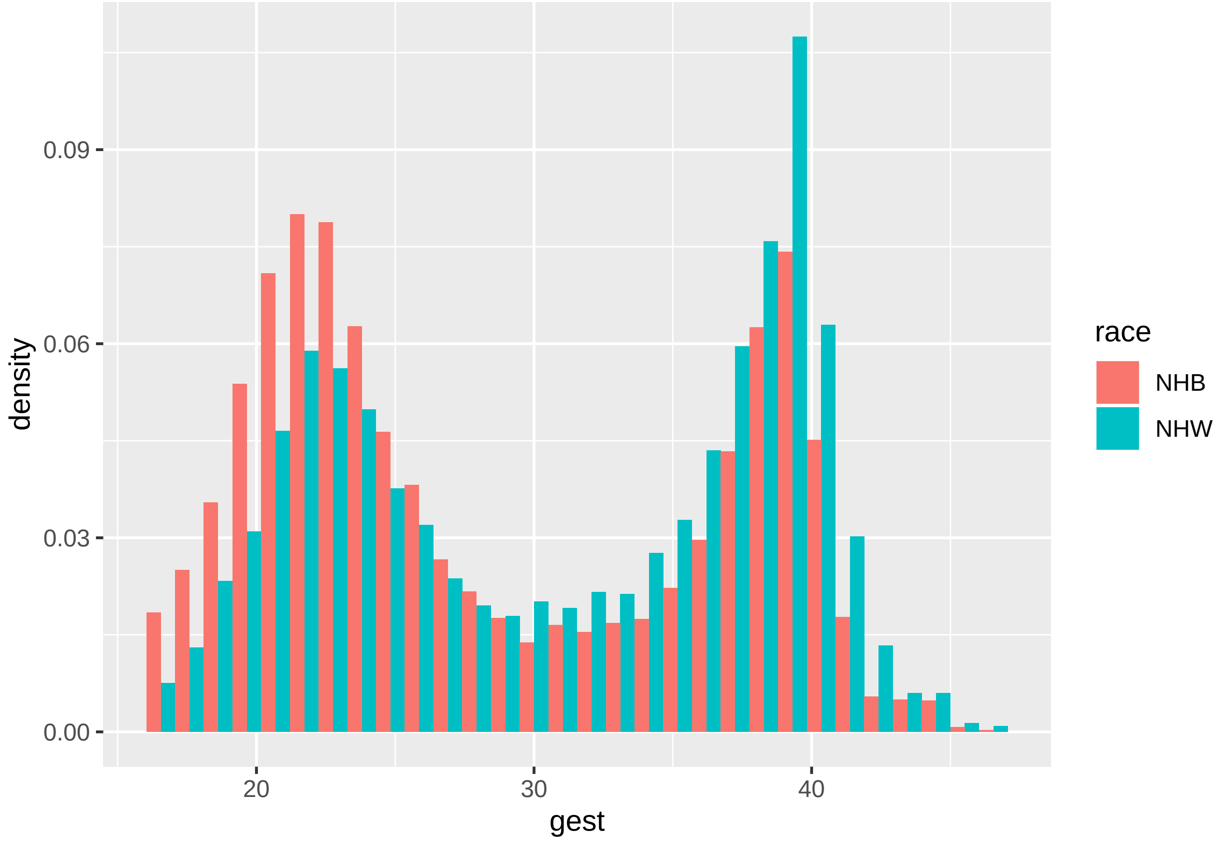
Also, removing the observations where we don't know gestational age or the mother's age.

```
d <- d %>%
  mutate(neo_death = ifelse(aged<=28, 1, 0),
         cod_group = case_when(
           str_starts(cod, "peri") ~ "perinatal",
           cod %in% c("other", "unknown") ~ "oth_unk",
           cod %in% c("sids", "maltreatment", "infection") ~ "exogenous",
           cod %in% c("resp", "heart") ~ "resp_heart",
           TRUE ~ cod
         ),
         preterm = ifelse(gest<37, 1, 0)) %>%
  filter(gest<99, !is.na(mom_age_group))
```

Distribution of gestational ages

Let's plot the distribution of gestational ages by race. It's quite bi-modal. Notice the difference in densities by race.

```
d %>%
  ggplot(aes(gest, fill = race)) + geom_histogram(position = 'dodge', aes(y = ..density..))
```



Question 1

Calculate the proportion of deaths that are neonatal by race and prematurity. Which group has the highest proportion of neonatal deaths?

```
dtmod <- d %>%
  group_by(race, preterm, neo_death) %>%
  summarise(deaths = n()) %>%
  group_by(race, preterm) %>%
  mutate(prop = deaths/sum(deaths)) %>%
  arrange(-prop)
head(dtmod)
```

```
## # A tibble: 6 x 5
## # Groups:   race, preterm [4]
##   race  preterm neo_death deaths  prop
##   <chr>   <dbl>     <dbl>  <int> <dbl>
## 1 NHW         1         1   5453 0.824
## 2 NHB         1         1   3746 0.802
## 3 NHB         0         0   1183 0.691
## 4 NHW         0         0   2464 0.618
## 5 NHW         0         1   1525 0.382
## 6 NHB         0         1    529 0.309
```

```
cat("\n\n The group with highest proportion of neonatal deaths is ",dtmod$race[1]," with ",dtmod$prop[1
```

```
##
##
##  The group with highest proportion of neonatal deaths is  NHW  with  82.4214 % proportion rate.
```

Causes of death

Let's make `cod_group` a factor with congenital malformations as the reference.

```
d <- d %>%
  mutate(cod_group = factor(cod_group, levels = c("cong_mal", "perinatal", "resp_heart", "exogenous", "
                           labels = c("cong_mal", "perinatal", "resp_heart", "exogenous", "oth_unk")))
```

The following code calculates the proportion of deaths by cause group, race, sex and prematurity

```
prop_cause <- d %>%
  group_by(race, preterm, sex, cod_group) %>%
  summarise(n = n()) %>%
  group_by(race, preterm, sex) %>%
  mutate(prop = n/sum(n)) %>%
  ungroup() %>%
  mutate(preterm = ifelse(preterm==1,"pre-term", "full-term"))
```

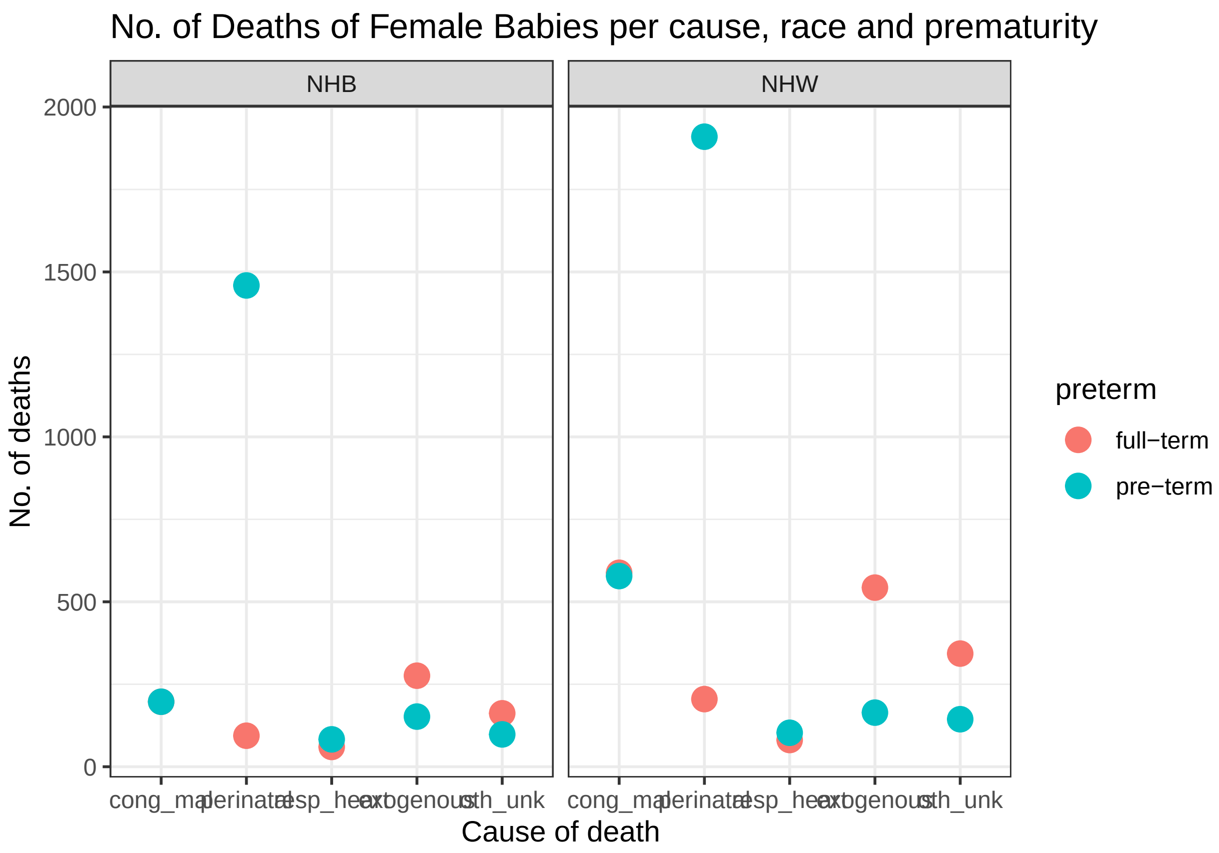
Question 2

Using the `prop_cause` above, filter to just look at female babies, and make a graph to help visualize differences in cause by race and prematurity.

```
head(prop_cause)
```

```
## # A tibble: 6 x 6
##   race  preterm   sex   cod_group      n   prop
##   <chr> <chr>     <chr> <fct>      <int>  <dbl>
## 1 NHB   full-term F     cong_mal     197 0.250
## 2 NHB   full-term F     perinatal     94 0.119
## 3 NHB   full-term F     resp_heart    60 0.0760
## 4 NHB   full-term F     exogenous    276 0.350
## 5 NHB   full-term F     oth_unk      162 0.205
## 6 NHB   full-term M     cong_mal     220 0.238
```

```
prop_cause %>%
  filter(sex == "F") %>%
  ggplot(mapping = aes(x = cod_group, y = n, color = preterm)) +
  geom_point(size = 4) +
  facet_wrap(~race) +
  theme_bw() +
  ggtitle("No. of Deaths of Female Babies per cause, race and prematurity") +
  ylab("No. of deaths") + xlab("Cause of death")
```



Logistic regression

First, let's do logistic regression to explore differences in neonatal deaths. Here's a model with prematurity, sex, race, and mom's age

```
mod <- glm(neo_death~ preterm + sex + race + race:preterm + mom_age, data = d, family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = neo_death ~ preterm + sex + race + race:preterm +
##     mom_age, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1160  -0.9126   0.6028   0.6825   1.6966
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.467493   0.095651 -15.342  < 2e-16 ***
## preterm          2.191541   0.064204  34.134  < 2e-16 ***
## sexM            -0.131361   0.036994  -3.551 0.000384 ***
## raceNHW          0.290680   0.061988   4.689 2.74e-06 ***
## mom_age          0.028682   0.002997   9.569  < 2e-16 ***
## preterm:raceNHW -0.191417   0.078973  -2.424 0.015359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21723  on 16986  degrees of freedom
## Residual deviance: 18115  on 16981  degrees of freedom
## AIC: 18127
##
## Number of Fisher Scoring iterations: 4
```

Question 3

Rerun the model above with instead of `mom_age`, include a new variable `mom_age_c` which centers mother's age around its mean.

```
d <- d %>%
  mutate(mom_age_c = mom_age-mean(mom_age))  # Centering the Mom_age arround pop. mean

mod <- glm(neo_death~ preterm + sex + race + race:preterm + mom_age_c, data = d, family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = neo_death ~ preterm + sex + race + race:preterm +
##     mom_age_c, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1160  -0.9126   0.6028   0.6825   1.6966
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.690052   0.056100 -12.300  < 2e-16 ***
## preterm          2.191541   0.064204  34.134  < 2e-16 ***
## sexM            -0.131361   0.036994  -3.551 0.000384 ***
## raceNHW          0.290680   0.061988   4.689 2.74e-06 ***
## mom_age_c        0.028682   0.002997   9.569  < 2e-16 ***
## preterm:raceNHW -0.191417   0.078973  -2.424 0.015359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21723  on 16986  degrees of freedom
## Residual deviance: 18115  on 16981  degrees of freedom
## AIC: 18127
##
## Number of Fisher Scoring iterations: 4
```

Question 4

Interpret the `preterm`, `race` and the interaction `preterm:race` coefficients.

Interpretation:- The coefficients of the new model adjusted suggest that:

- 1. preterm: babies have higher probabiliy of death than non-preterm babies (approx. 2,19 times higher);
- 2. race: NHW (Non-Hispanic White) has 29,1
- 3. for babies NHW and pre-term, the probability of death is reduced

Multinomial regression

Now let's do multinomial regression with cause of death as the outcome. We need to get the data in a different format to run the regression:

```
d$mom_age_c <- d$mom_age - mean(d$mom_age)
d_wide <- d %>%
  group_by(sex, race, cod_group, preterm,mom_age_c) %>%
  summarise(deaths = n()) %>%
  pivot_wider(names_from = cod_group, values_from = deaths) %>%
  mutate_all(.funs = funs(ifelse(is.na(.), 0, .)))

d_wide$Y <- as.matrix(d_wide[,c("cong_mal","perinatal","resp_heart", "exogenous", "oth_unk")])
```

Now run the regression

```
library(nnet)
mod2 <- multinom(Y ~ sex+race+ mom_age_c+ preterm, data = d_wide)
```

```
## # weights:  30 (20 variable)
## initial  value 27339.521819
## iter  10 value 22475.496335
## iter  20 value 19882.612578
## iter  30 value 19389.722462
## final  value 19389.720141
## converged
```

```
summary(mod2)
```

```
## Call:
## multinom(formula = Y ~ sex + race + mom_age_c + preterm, data = d_wide)
##
## Coefficients:
##             (Intercept)      sexM    raceNHW   mom_age_c    preterm
## perinatal   -0.53315841 0.0657566 -0.6249840 -0.01906239  2.4190484
## resp_heart -1.21149941 0.1350905 -0.6303309 -0.03765107  0.1683872
## exogenous   0.40732759 0.2070469 -0.5359725 -0.07602388 -1.0125410
## oth_unk    -0.09571594 0.1682889 -0.4746567 -0.04122105 -0.7189038
##
## Std. Errors:
##             (Intercept)       sexM    raceNHW    mom_age_c    preterm
## perinatal    0.06335013 0.04452244 0.04846619 0.003515557 0.05452877
## resp_heart   0.09493043 0.08128540 0.08477048 0.006591926 0.08175145
## exogenous    0.06236608 0.05376561 0.05842537 0.004532928 0.05637655
## oth_unk      0.06948526 0.06015565 0.06522548 0.004905661 0.06185463
##
## Residual Deviance: 38779.44
## AIC: 38819.44
```

Question 5

Using the `predict` function, find the predicted probabilities of each cause by race, sex and prematurity for the mothers of mean age. You can use this prediction dataframe to get all the combinations you need.

```
pred_df <- tibble(preterm = c(rep(0, 4), rep(1, 4)),
       sex = rep(c(rep("F", 2), rep("M", 2)),2),
       race = rep(c("NHB", "NHW"), 4),
       mom_age_c = 0)
```

Calculating the predicted probabilities of each cause by race, sex and prematurity:

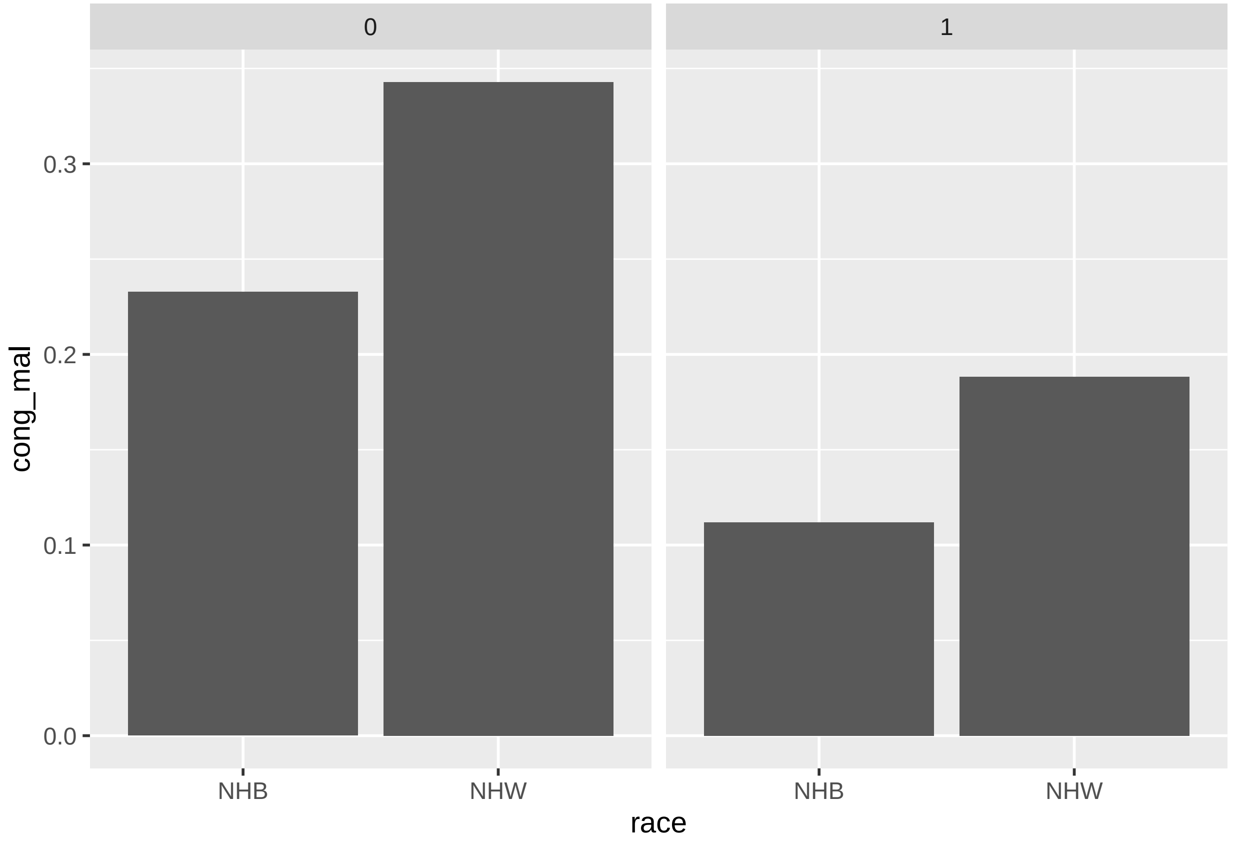
```
preds <- predict(mod2, newdata = pred_df, type = 'probs')

pred_df <- pred_df %>% cbind(preds); pred_df
```

```
##   preterm sex race mom_age_c  cong_mal perinatal resp_heart  exogenous
## 1       0   F  NHB         0 0.2327737 0.1365797 0.06930849 0.34981148
## 2       0   F  NHW         0 0.3427948 0.1076612 0.05434223 0.30141473
## 3       0   M  NHB         0 0.2044481 0.1281130 0.06967944 0.37792232
## 4       0   M  NHW         0 0.3060581 0.1026566 0.05553620 0.33101935
## 5       1   F  NHB         0 0.1119402 0.7379343 0.03944284 0.06111464
## 6       1   F  NHW         0 0.1882840 0.6643820 0.03532207 0.06014542
## 7       1   M  NHB         0 0.1037441 0.7303862 0.04184219 0.06966934
## 8       1   M  NHW         0 0.1753752 0.6608917 0.03765910 0.06890911
##      oth_unk
## 1 0.21152662
## 2 0.19378697
## 3 0.21983718
## 4 0.20472982
## 5 0.04956798
## 6 0.05186659
## 7 0.05435823
## 8 0.05716489
```

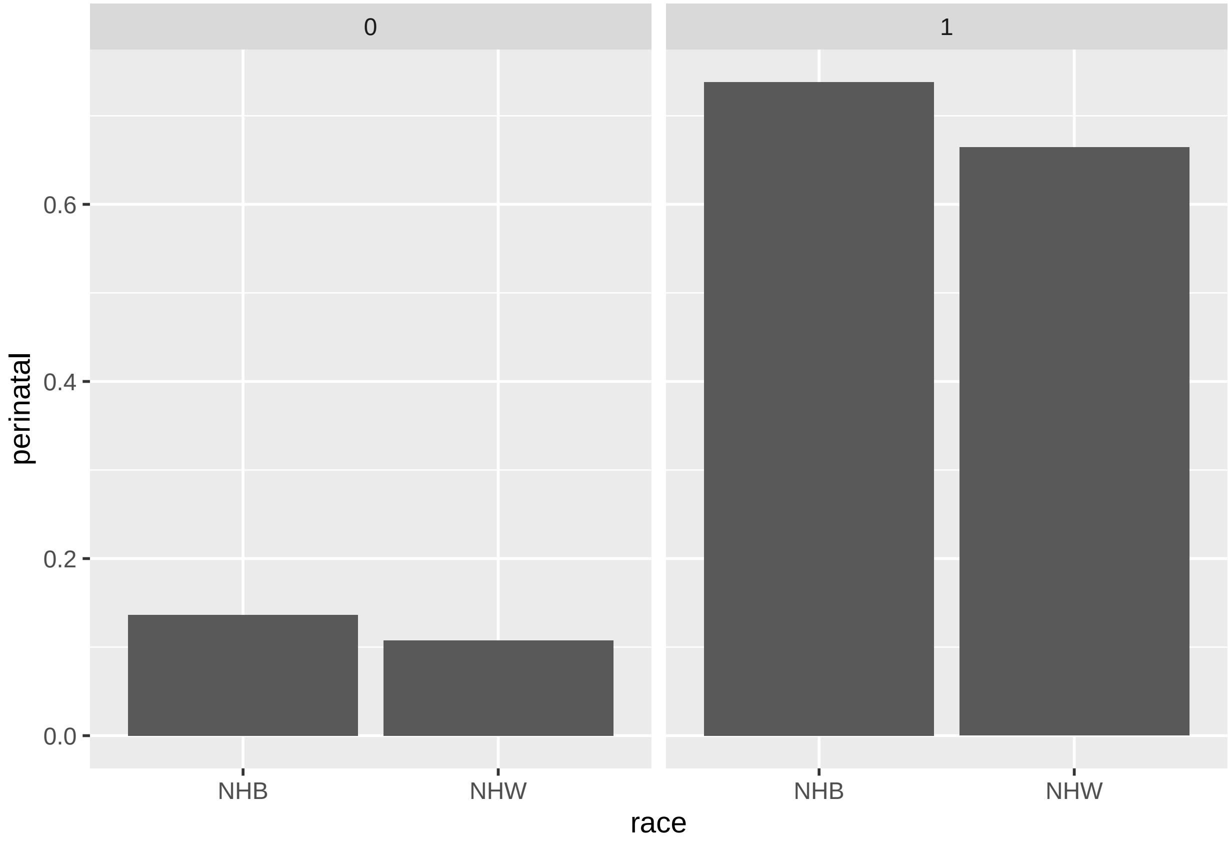
Question 6

Plot the predicted probabilities for female babies.

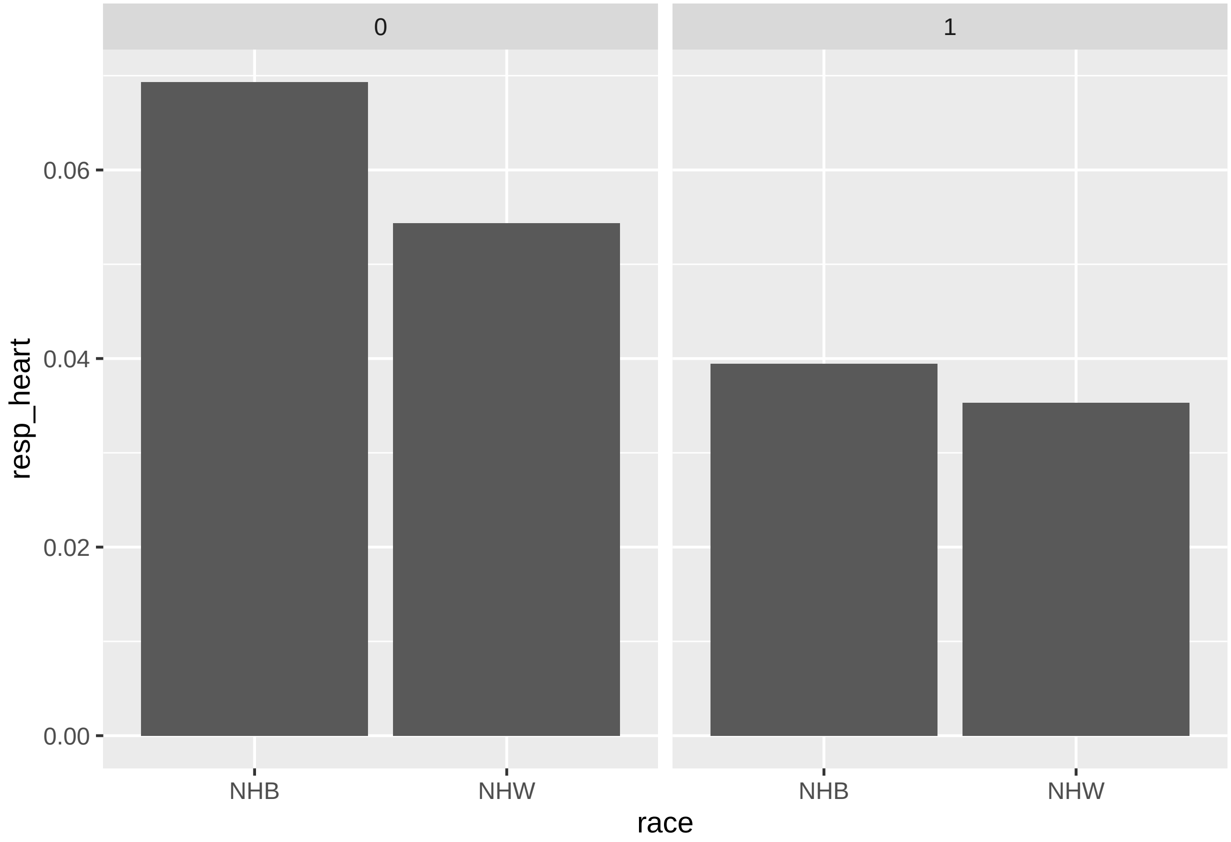
```
pred_df %>%
  filter(sex == "F") %>%
  ggplot(aes(x = race, y = cong_mal)) +
  facet_wrap(~preterm) +
  geom_col()
```



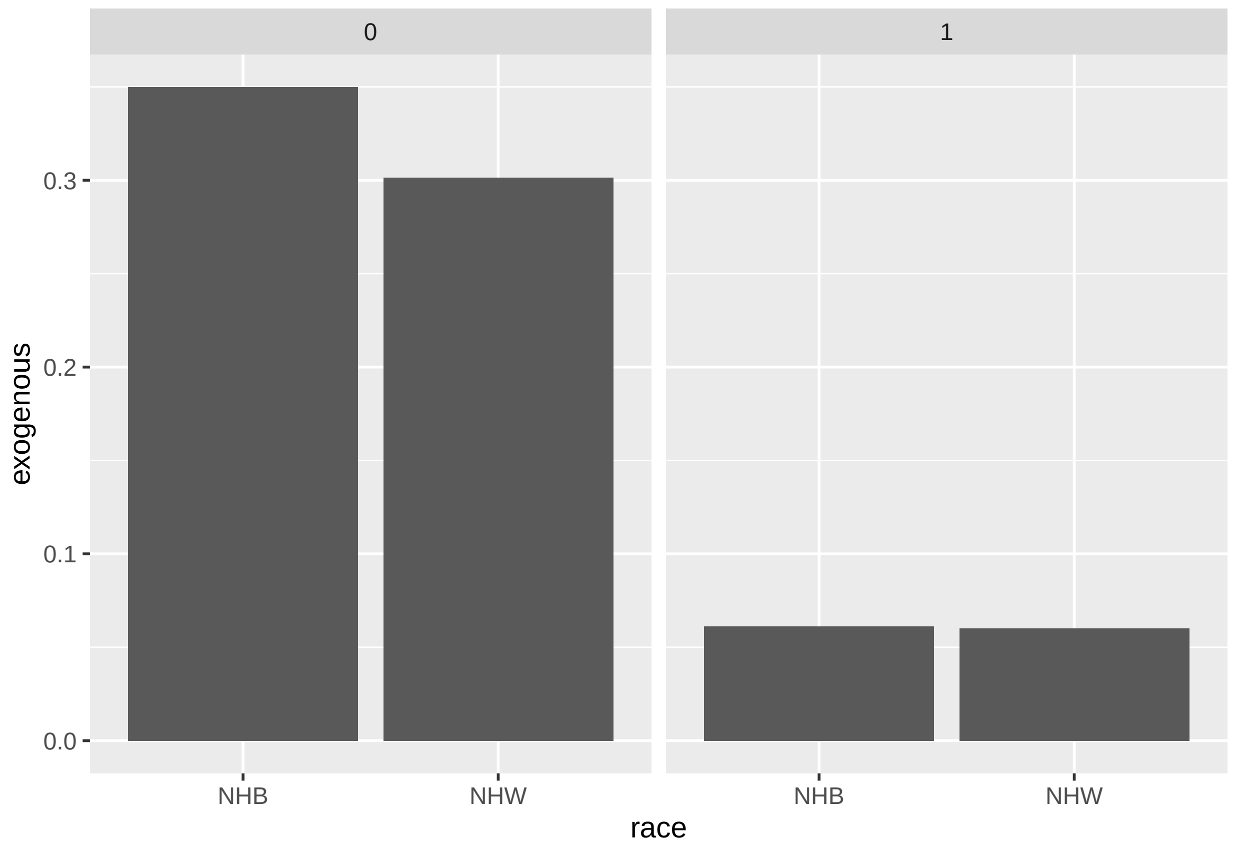
```r
pred_df %>%
  filter(sex == "F") %>%
  ggplot(aes(x = race, y = perinatal)) +
  facet_wrap(~preterm) +
  geom_col()
```



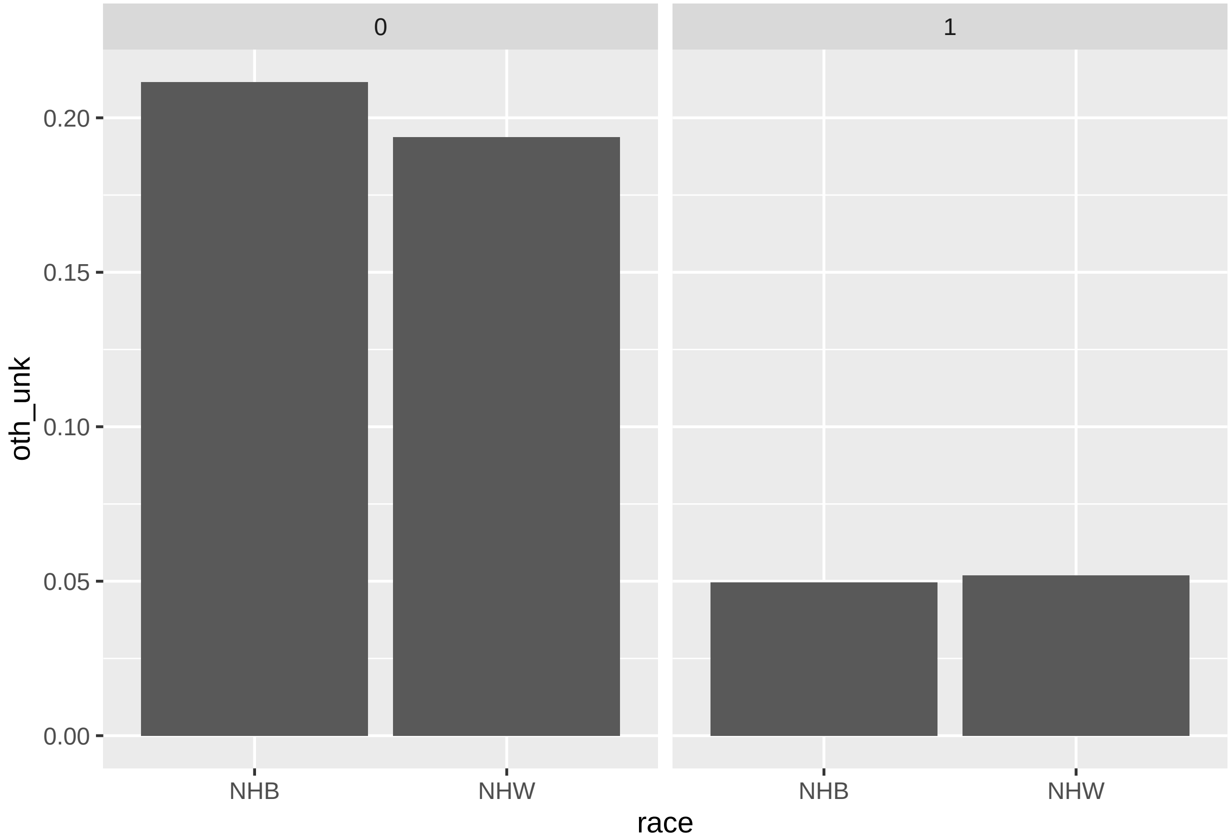
```r
pred_df %>%
  filter(sex == "F") %>%
  ggplot(aes(x = race, y = resp_heart)) +
  facet_wrap(~preterm) +
  geom_col()
```



```r
pred_df %>%
  filter(sex == "F") %>%
  ggplot(aes(x = race, y = exogenous)) +
  facet_wrap(~preterm) +
  geom_col()
```



```
pred_df %>%
  filter(sex == "F") %>%
  ggplot(aes(x = race, y = oth_unk)) +
  facet_wrap(~preterm) +
  geom_col()
```



Question 7

What race/prematurity/ cause group has the highest probability?

Response:- It is NHB (Non-Hispanic Black), Preterm, Perinatal which has 0.7379343 probability in female babies.

How does this compared to the observed proportion in the same group?

```
obsprop <- prop_cause %>%
  filter(sex == "F", race=="NHB", preterm == "pre-term", cod_group == "perinatal")
cat("\n\nThe observer proportion for this population is ",obsprop$prop)
```

```
##
##
## The observer proportion for this population is  0.7335344
```

```
cat("\n ... and the relation (predict/observed) is ", 0.7379343/obsprop$prop)
```

```
##
##  ... and the relation (predict/observed) is  1.005998
```