HW5\_pnegandh

Prashil Negandhi

30 October 2019

library(NHANES)

## Warning: package 'NHANES' was built under R version 3.5.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##   
## 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(tidyr)

## Warning: package 'tidyr' was built under R version 3.5.3

nhanes = NHANES  
head(nhanes)

## # A tibble: 6 x 76  
## ID SurveyYr Gender Age AgeDecade AgeMonths Race1 Race3 Education  
## <int> <fct> <fct> <int> <fct> <int> <fct> <fct> <fct>   
## 1 51624 2009\_10 male 34 " 30-39" 409 White <NA> High Sch~  
## 2 51624 2009\_10 male 34 " 30-39" 409 White <NA> High Sch~  
## 3 51624 2009\_10 male 34 " 30-39" 409 White <NA> High Sch~  
## 4 51625 2009\_10 male 4 " 0-9" 49 Other <NA> <NA>   
## 5 51630 2009\_10 female 49 " 40-49" 596 White <NA> Some Col~  
## 6 51638 2009\_10 male 9 " 0-9" 115 White <NA> <NA>   
## # ... with 67 more variables: MaritalStatus <fct>, HHIncome <fct>,  
## # HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>,  
## # Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>, Height <dbl>,  
## # BMI <dbl>, BMICatUnder20yrs <fct>, BMI\_WHO <fct>, Pulse <int>,  
## # BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>,  
## # BPSys2 <int>, BPDia2 <int>, BPSys3 <int>, BPDia3 <int>,  
## # Testosterone <dbl>, DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>,  
## # UrineFlow1 <dbl>, UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>,  
## # DiabetesAge <int>, HealthGen <fct>, DaysPhysHlthBad <int>,  
## # DaysMentHlthBad <int>, LittleInterest <fct>, Depressed <fct>,  
## # nPregnancies <int>, nBabies <int>, Age1stBaby <int>,  
## # SleepHrsNight <int>, SleepTrouble <fct>, PhysActive <fct>,  
## # PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,  
## # TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,  
## # AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,  
## # Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,  
## # RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,  
## # SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>,  
## # SameSex <fct>, SexOrientation <fct>, PregnantNow <fct>

nhanes\_raw = NHANESraw  
head(nhanes\_raw)

## # A tibble: 6 x 78  
## ID SurveyYr Gender Age AgeMonths Race1 Race3 Education MaritalStatus  
## <int> <fct> <fct> <int> <int> <fct> <fct> <fct> <fct>   
## 1 51624 2009\_10 male 34 409 White <NA> High Sch~ Married   
## 2 51625 2009\_10 male 4 49 Other <NA> <NA> <NA>   
## 3 51626 2009\_10 male 16 202 Black <NA> <NA> <NA>   
## 4 51627 2009\_10 male 10 131 Black <NA> <NA> <NA>   
## 5 51628 2009\_10 female 60 722 Black <NA> High Sch~ Widowed   
## 6 51629 2009\_10 male 26 313 Mexi~ <NA> 9 - 11th~ Married   
## # ... with 69 more variables: HHIncome <fct>, HHIncomeMid <int>,  
## # Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>, Work <fct>,  
## # Weight <dbl>, Length <dbl>, HeadCirc <dbl>, Height <dbl>, BMI <dbl>,  
## # BMICatUnder20yrs <fct>, BMI\_WHO <fct>, Pulse <int>, BPSysAve <int>,  
## # BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>, BPSys2 <int>,  
## # BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,  
## # DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,  
## # UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>,  
## # HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,  
## # LittleInterest <fct>, Depressed <fct>, nPregnancies <int>,  
## # nBabies <int>, Age1stBaby <int>, SleepHrsNight <int>,  
## # SleepTrouble <fct>, PhysActive <fct>, PhysActiveDays <int>,  
## # TVHrsDay <fct>, CompHrsDay <fct>, TVHrsDayChild <int>,  
## # CompHrsDayChild <int>, Alcohol12PlusYr <fct>, AlcoholDay <int>,  
## # AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>, SmokeAge <int>,  
## # Marijuana <fct>, AgeFirstMarij <int>, RegularMarij <fct>,  
## # AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>, SexAge <int>,  
## # SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,  
## # SexOrientation <fct>, WTINT2YR <dbl>, WTMEC2YR <dbl>, SDMVPSU <int>,  
## # SDMVSTRA <int>, PregnantNow <fct>

nhanes = nhanes %>% select(ID, Diabetes, Age) %>% drop\_na() %>% mutate(Diabetes = ifelse(Diabetes == "Yes", 1, 0))  
head(nhanes)

## # A tibble: 6 x 3  
## ID Diabetes Age  
## <int> <dbl> <int>  
## 1 51624 0 34  
## 2 51624 0 34  
## 3 51624 0 34  
## 4 51625 0 4  
## 5 51630 0 49  
## 6 51638 0 9

nhanes\_raw = nhanes\_raw %>% select(ID, Diabetes, Age, WTINT2YR) %>% drop\_na() %>% mutate(Diabetes = ifelse(Diabetes == "Yes", 1, 0))  
head(nhanes\_raw)

## # A tibble: 6 x 4  
## ID Diabetes Age WTINT2YR  
## <int> <dbl> <int> <dbl>  
## 1 51624 0 34 80101.  
## 2 51625 0 4 53901.  
## 3 51626 0 16 13953.  
## 4 51627 0 10 11665.  
## 5 51628 1 60 20090.  
## 6 51629 0 26 22538.

# Q1.a)

with\_d = sum(nhanes$Diabetes == 1)  
total = nrow(nhanes)  
  
with\_d/total

## [1] 0.07709475

As we can see 7.7% of the people have diabetes.

# Q1.b)

with\_d\_2 = sum(nhanes\_raw$Diabetes == 1)  
total\_2 = nrow(nhanes\_raw)  
  
with\_d\_2/total\_2

## [1] 0.08766701

As we can see 8.7% of the people have diabetes without using weights.

# Q1.c)

with\_d\_3 = sum(nhanes\_raw$Diabetes \* nhanes\_raw$WTINT2YR)  
total\_3 = sum(nhanes\_raw$WTINT2YR)  
  
with\_d\_3/total\_3

## [1] 0.07983828

As we can see 7.9% of the people have diabetes if we use weights.

The number for 1.a) and 1.c) are really close. This is because using weights makes the proportion almost same. However the proportion for 1.b) is 1% away.

# Q2.a)

mod\_1 <- glm(Diabetes ~ Age, data=nhanes, family='binomial')  
summary(mod\_1)

##   
## Call:  
## glm(formula = Diabetes ~ Age, family = "binomial", data = nhanes)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9037 -0.4218 -0.2506 -0.1517 3.0274   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.19218 0.13240 -39.22 <2e-16 \*\*\*  
## Age 0.05634 0.00221 25.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5355.2 on 9857 degrees of freedom  
## Residual deviance: 4495.1 on 9856 degrees of freedom  
## AIC: 4499.1  
##   
## Number of Fisher Scoring iterations: 6

# Q2.b)

mod\_2 <- glm(Diabetes ~ Age, data = nhanes\_raw, family = 'quasibinomial', weights = WTINT2YR)  
summary(mod\_2)

##   
## Call:  
## glm(formula = Diabetes ~ Age, family = "quasibinomial", data = nhanes\_raw,   
## weights = WTINT2YR)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## 0 0 0 0 3768   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.979e+15 1.797e+13 -277.07 <2e-16 \*\*\*  
## Age 3.192e+13 4.137e+11 77.16 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasibinomial family taken to be 1.109088e+19)  
##   
## Null deviance: 334158316 on 19459 degrees of freedom  
## Residual deviance: 3454325322 on 19458 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 4

As we can see the deviance residuals are much lower for the second model. The intercepts are higher for the second model. Moreover looking at the t value and the error we can say that the second model doesn’t perform that well.

nhanes\_raw["Weighted"] = nhanes\_raw$Diabetes \* nhanes\_raw$WTINT2YR

head(nhanes\_raw)

## # A tibble: 6 x 5  
## ID Diabetes Age WTINT2YR Weighted  
## <int> <dbl> <int> <dbl> <dbl>  
## 1 51624 0 34 80101. 0   
## 2 51625 0 4 53901. 0   
## 3 51626 0 16 13953. 0   
## 4 51627 0 10 11665. 0   
## 5 51628 1 60 20090. 20090.  
## 6 51629 0 26 22538. 0

library(plyr)

## Warning: package 'plyr' was built under R version 3.5.3

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

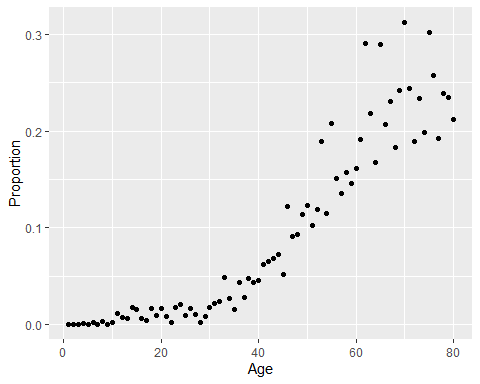
plot\_data = ddply(nhanes\_raw, .(Age), summarize, Proportion = sum(Weighted)/sum(WTINT2YR))  
head(plot\_data)

## Age Proportion  
## 1 1 0.000000000  
## 2 2 0.000000000  
## 3 3 0.000000000  
## 4 4 0.001116786  
## 5 5 0.000000000  
## 6 6 0.001513930

sum(plot\_data$Weighted\_Mean)

## [1] 0

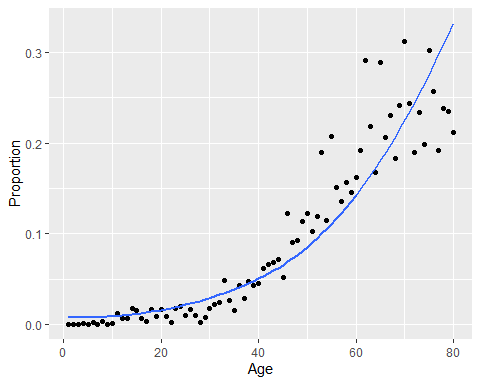
ggplot(plot\_data, aes(Age, Proportion)) + geom\_point()



newdata = data.frame("Age" = seq(1:80))  
plot\_data["Predicted"] = predict(mod\_1, newdata, type = "response")

ggplot(plot\_data, aes(Age)) + geom\_point(aes(y = Proportion)) + geom\_smooth(aes(y = Predicted))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Looking at the graphs we can say that the model performs much better for predicting diabetes proportion for lower age groups. However, as we go to higher ages the model starts performing worse. This might be because we don’t have enough people in those age groups.