FinalReport_RCode

Shengyuan Wang

12/10/2021

Load Packages

```
library(dplyr)
library(readr)
library(ggplot2)
library(broom)
library(ggmosaic)
```

Load Data

```
library(NHANES)
data(NHANES)
```

Clean Data

Filter Missing Values and Other Filters

```
# filter the Age ranging from 15 to 70
NHANES_1 <- NHANES %>%
  filter(Age >= 15 & Age <= 70)
# filter out the cases with no Physical Activity, Diabetes Status and BMI rank information
# filter out ouliter cases with weight over 190kg and combined systolic blood pressure higher than 200mmHg
NHANES_2 <- subset(NHANES_1, PhysActive != "NA" & Diabetes != "NA" & BMI_WHO != "NA") %>%
  filter(Weight <= 190) %>%
  filter(BPSysAve < 200)
# show the cases left other each filtering process
nrow(NHANES)
```

```
## [1] 10000
```

```
nrow(NHANES_1)
```

```
## [1] 7101
```

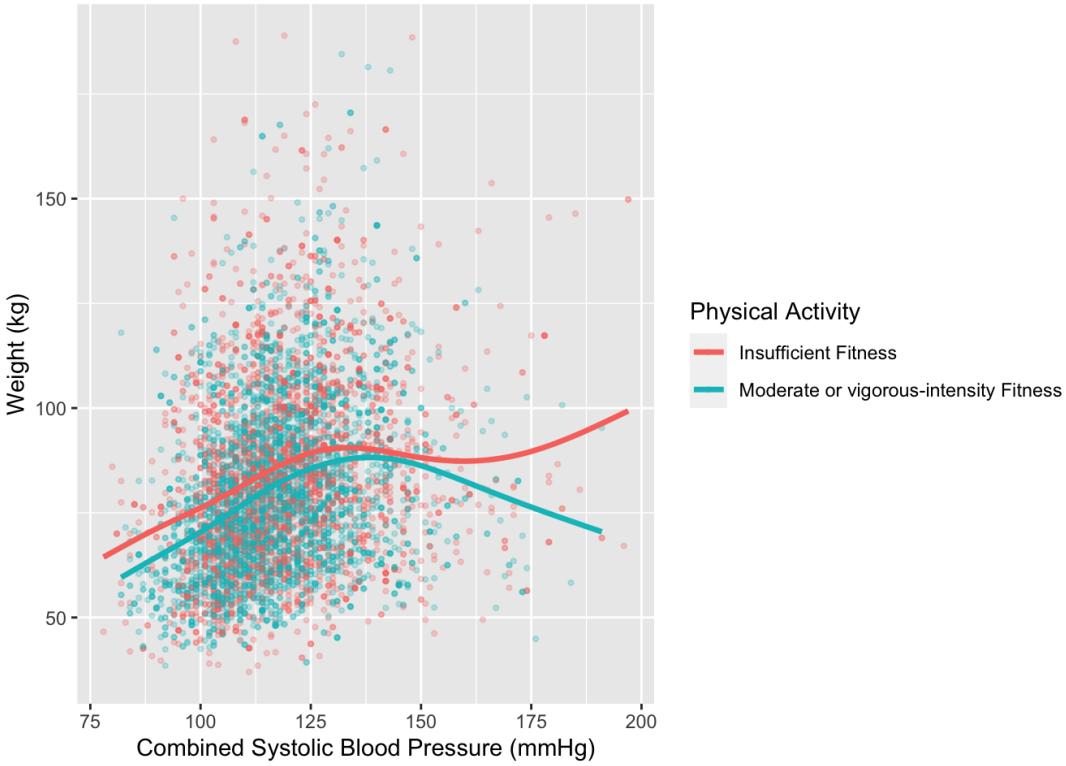
```
nrow(NHANES_2)
```

```
## [1] 6735
```

Research Question #1

Visualization

```
# draw scatter plot with combined systolic blood pressure, weight and physical activity
NHANES_2 %>%
  ggplot(aes(x = BPSysAve, y = Weight, color = PhysActive)) +
  geom_point(alpha = 0.25, size = 0.8) +
  xlab("Combined Systolic Blood Pressure (mmHg)") +
  ylab("Weight (kg)") +
  scale_color_discrete(name ="Physical Activity", label = c("Insufficient Fitness", "Moderate or vigorous-i
ntensity Fitness")) +
  geom_smooth(se = FALSE, size = 1.2)
```



Create Models

The original big model

```
Lmod1 <- NHANES_2 %>% #Original big model
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ Height + PhysActive + BPSysAve + Gender + Race3))
```

Model without Race3

```
Lmod2 <- NHANES_2 %>% #Model without Race3
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ Height + PhysActive + BPSysAve + Gender))
```

Model without Height

```
Lmod3 <- NHANES_2 %>% #Model without Height
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ PhysActive + BPSysAve + Gender + Race3))
```

Model without Combined systolic blood pressure

```
Lmod4 <- NHANES_2 %>% #Model without BPSysAve
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ Height + PhysActive + Gender + Race3))
```

Model without Physical activity

```
Lmod5 <- NHANES_2 %>% #Model without PhysActive
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ Height + BPSysAve + Gender + Race3))
```

Model without Sleep hours per night (Final Model)

```
Lmod6 <- NHANES_2 %>% #Model without SleepHrsNight
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ Height + BPSysAve + PhysActive + Race3))
```

Model Selection

Original Big Model

```
glance(Lmod1)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.258         0.256  18.2      129. 1.48e-209     9 -14503. 29027. 29095.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
summary(Lmod1)
```

```
##
## Call:
## lm(formula = Weight ~ Height + PhysActive + BPSysAve + Gender +
##     Race3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.876 -12.320  -2.927   9.861 106.732
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -104.31762    7.81493 -13.348  < 2e-16 ***
## Height           0.87820    0.04490  19.559  < 2e-16 ***
## PhysActiveYes   -4.63461    0.64617  -7.172 9.03e-13 ***
## BPSysAve         0.26542    0.02163  12.271  < 2e-16 ***
## Gendermale       0.32056    0.88832   0.361    0.718
## Race3Black      11.69149    1.62064   7.214 6.68e-13 ***
## Race3Hispanic    8.75462    1.79433   4.879 1.12e-06 ***
## Race3Mexican     9.75633    1.68873   5.777 8.29e-09 ***
## Race3White       7.96812    1.38374   5.758 9.26e-09 ***
## Race3Other      10.91955    2.30856   4.730 2.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.25 on 3346 degrees of freedom
## Multiple R-squared:  0.2582, Adjusted R-squared:  0.2562
## F-statistic: 129.4 on 9 and 3346 DF,  p-value: < 2.2e-16
```

Model without Race

```
glance(Lmod2)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.245         0.244  18.4      272. 5.93e-203     4 -14532. 29075. 29112.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
anova(Lmod1, Lmod2)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Height + PhysActive + BPSysAve + Gender + Race3
## Model 2: Weight ~ Height + PhysActive + BPSysAve + Gender
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   3346 1114041
## 2   3351 1133303 -5    -19262 11.571 4.213e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model without Height

```
glance(Lmod3)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.173         0.171  19.3      87.7 1.74e-132     8 -14684. 29389. 29450.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Model without Combined systolic blood pressure

```
glance(Lmod4)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.225         0.223  18.7      121. 7.70e-179     8 -14577. 29173. 29234.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Model without Physical activity

```
glance(Lmod5)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.247         0.245  18.4      137. 1.28e-199     8 -14528. 29077. 29138.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
anova(Lmod1, Lmod5)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Height + PhysActive + BPSysAve + Gender + Race3
## Model 2: Weight ~ Height + BPSysAve + Gender + Race3
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   3346 1114041
## 2   3347 1131170 -1    -17128 51.444 9.026e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model without Sleep hours per night (Final Model)

```
glance(Lmod6)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.258         0.256  18.2      146. 1.27e-210     8 -14503. 29026. 29087.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

From the p-values above, we see that the p-value of height, combined systolic blood pressure, physical activity and race($p < 0.0001$) are lower than 0.05 threshold. If we exclude race, combined blood pressure, physical activity and height separately from the model, we see the adjusted R-squared will decrease. It means that the model including race, combined blood pressure, physical activity and height are better for predictions.

Fit Model

The model:

$$E[Weight|Height, BPSysAve, PhysActive, Race3] = \beta_0 + \beta_1 * Height + \beta_2 * BPSysAve + \beta_3 * PhysActiveYes + \beta_4 * Race3Black$$
$$+\beta_5 * Race3Hispanic + \beta_6 * Race3Mexican + \beta_7 * Race3White + \beta_8 * Race3Other$$

```
Lmod <- NHANES_2 %>% #Model without SleepHrsNight
  select(Weight, Height, PhysActive, BPSysAve, Gender, Race3) %>%
  na.omit() %>%
  with(lm(Weight ~ Height + BPSysAve + PhysActive + Race3))
```

Inference

```
summary(Lmod)
```

```
##
## Call:
## lm(formula = Weight ~ Height + BPSysAve + PhysActive + Race3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.783 -12.308  -2.967   9.910 106.648
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -106.16583    5.90196 -17.988  < 2e-16 ***
## Height           0.88944    0.03235  27.497  < 2e-16 ***
## BPSysAve         0.26679    0.02129  12.529  < 2e-16 ***
## PhysActiveYes   -4.63882    0.64598  -7.181 8.48e-13 ***
## Race3Black      11.62467    1.60982   7.221 6.35e-13 ***
## Race3Hispanic    8.75248    1.79409   4.879 1.12e-06 ***
## Race3Mexican     9.77492    1.68772   5.792 7.61e-09 ***
## Race3White       7.90039    1.37077   5.763 8.99e-09 ***
## Race3Other      10.87896    2.30552   4.719 2.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.24 on 3347 degrees of freedom
## Multiple R-squared:  0.2582, Adjusted R-squared:  0.2564
## F-statistic: 145.6 on 8 and 3347 DF,  p-value: < 2.2e-16
```

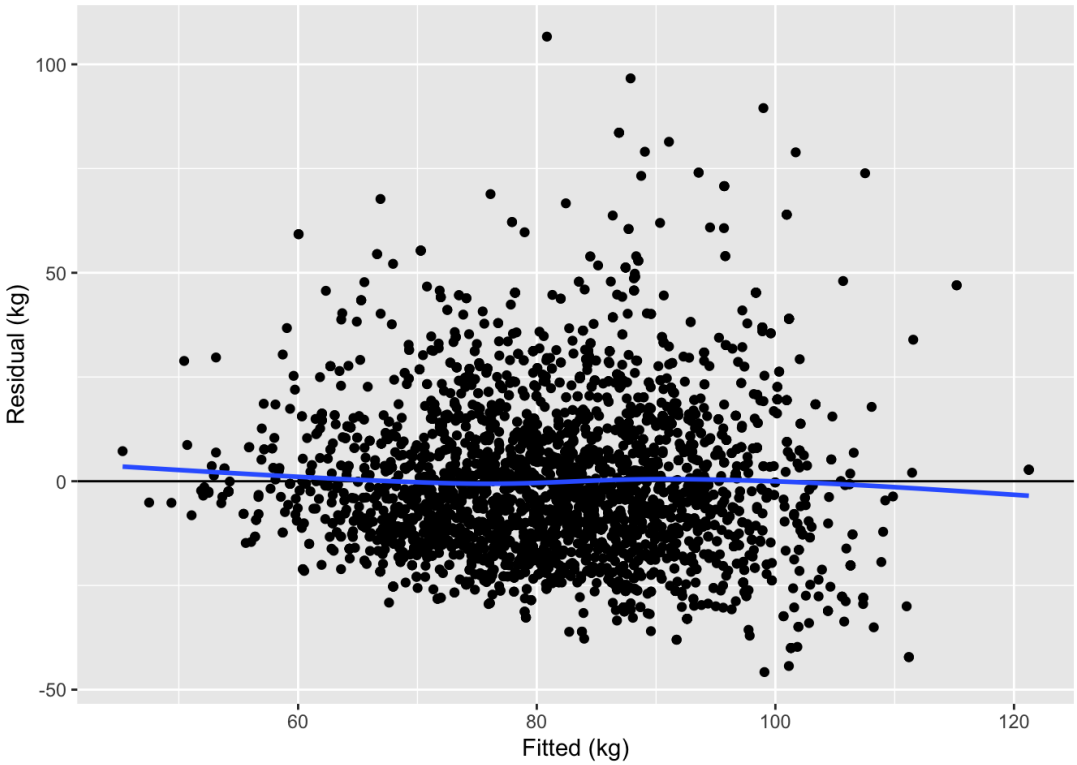
```
Lmod %>%
  confint()
```

```
##                      2.5 %      97.5 %
## (Intercept)   -117.7376389 -94.5940262
## Height           0.8260176   0.9528598
## BPSysAve         0.2250394   0.3085410
## PhysActiveYes   -5.9053682  -3.3722715
## Race3Black       8.4683462  14.7810000
## Race3Hispanic    5.2348592  12.2701097
## Race3Mexican     6.4658480  13.0839978
## Race3White       5.2127550  10.5880331
## Race3Other       6.3585881  15.3993342
```

Evaluation

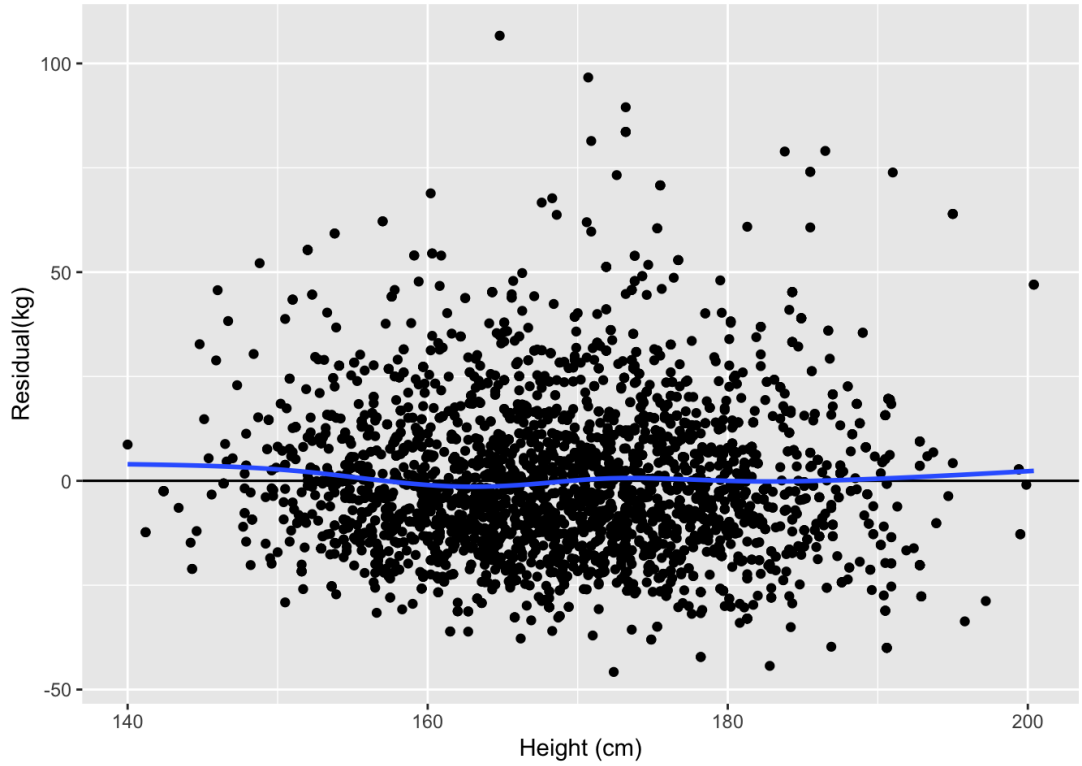
Residual vs. Fitted plot

```
# residuals vs fitted
augment(Lmod) %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  xlab('Fitted (kg)') +
  ylab('Residual (kg)') +
  geom_hline(yintercept = 0)+
  geom_smooth(se=FALSE)
```



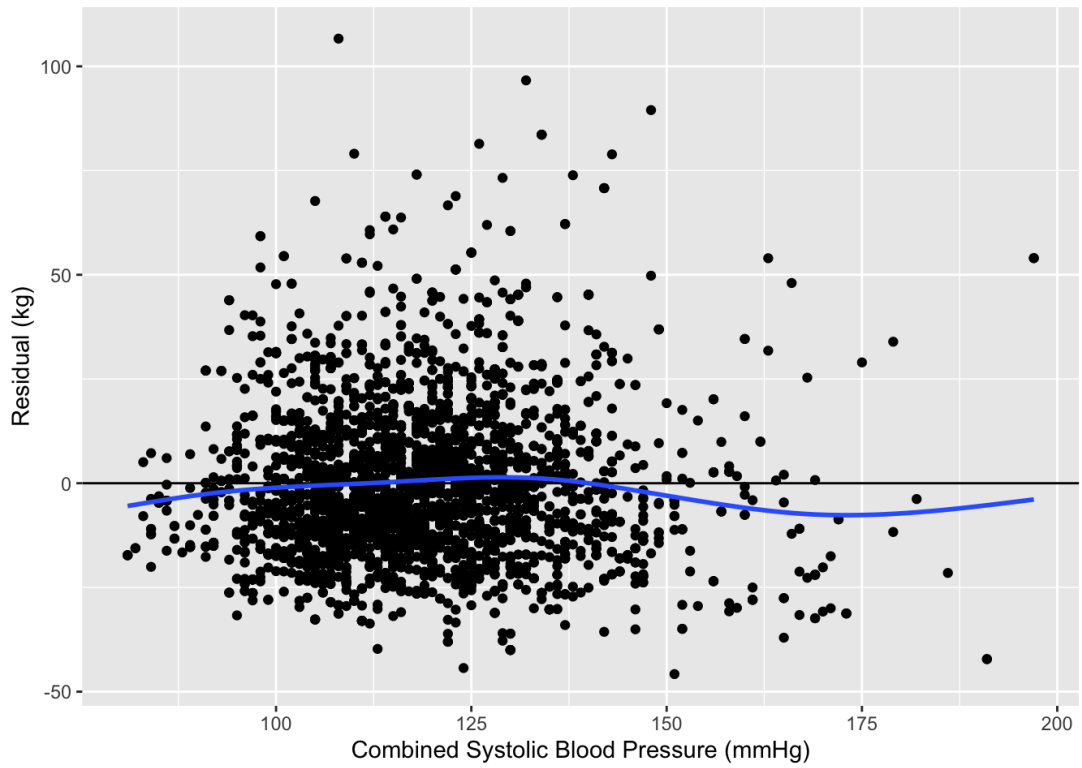
Residual vs. Height plot

```
# residuals vs Height
augment(Lmod) %>%
  ggplot(aes(x = Height, y = .resid)) +
  geom_point() +
  xlab('Height (cm)') +
  ylab('Residual(kg)') +
  geom_hline(yintercept = 0)+
  geom_hline(yintercept = 0)+
  geom_smooth(se=FALSE)
```



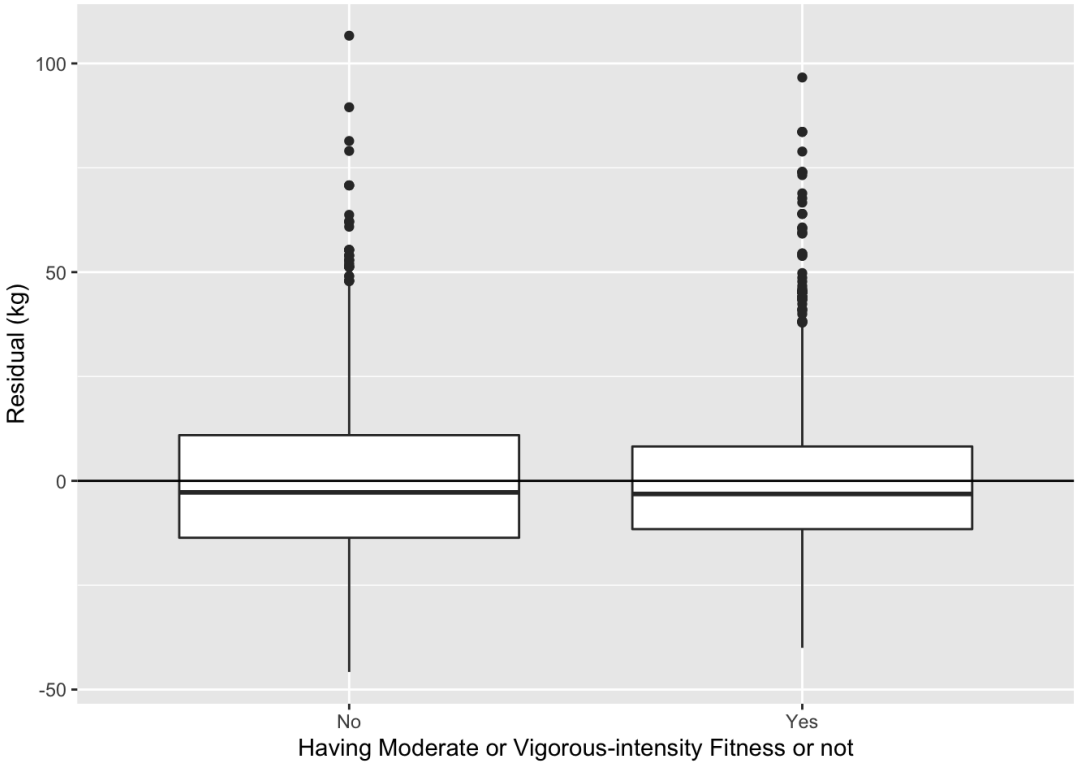
Residual vs. Combined Systolic blood pressure plot

```
# residuals vs BPSysAve
augment(Lmod) %>%
  ggplot(aes(x = BPSysAve, y = .resid)) +
  geom_point() +
  xlab('Combined Systolic Blood Pressure (mmHg)') +
  ylab('Residual (kg)') +
  geom_hline(yintercept = 0)+
  geom_smooth(se=FALSE)
```



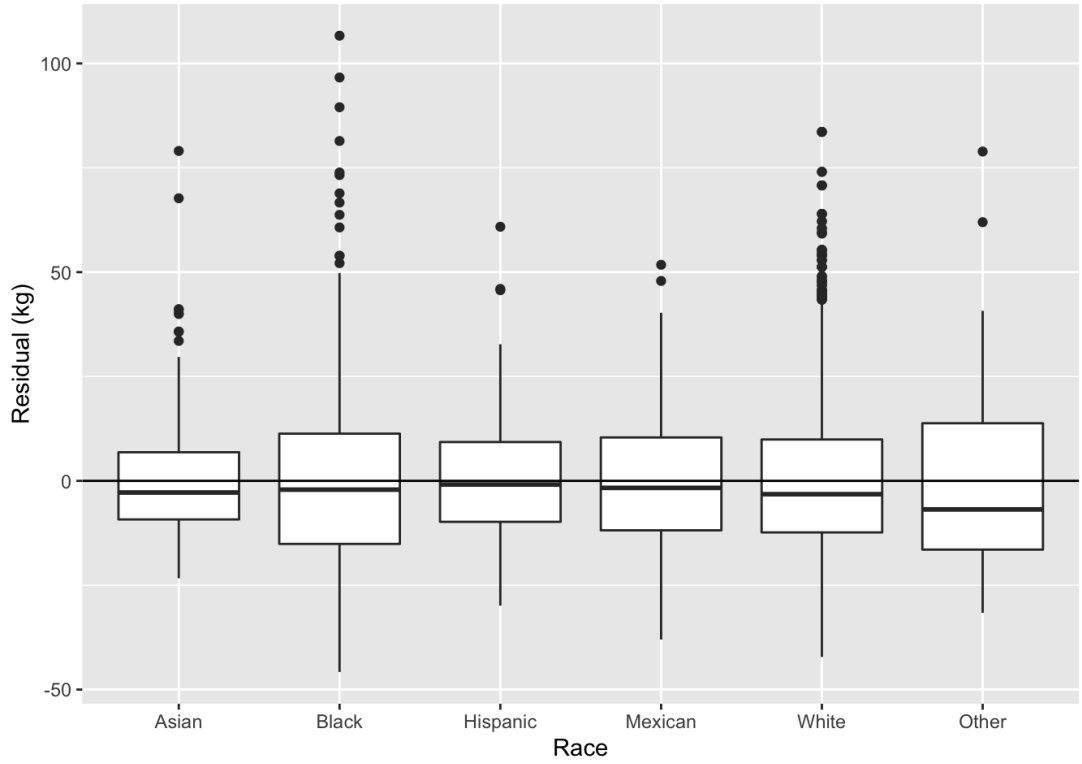
Residual vs. Physical Activity plot

```
# residuals vs PhysActive
augment(Lmod) %>%
  ggplot(aes(x = PhysActive, y = .resid)) +
  geom_boxplot() +
  xlab('Having Moderate or Vigorous-intensity Fitness or not') +
  ylab('Residual (kg)') +
  geom_hline(yintercept = 0)
```



Residual vs. Race plot

```
# residuals vs Race3
augment(Lmod) %>%
  ggplot(aes(x = Race3, y = .resid)) +
  geom_boxplot() +
  xlab('Race') +
  ylab('Residual (kg)') +
  geom_hline(yintercept = 0)
```




```
summary(Lmod)
```

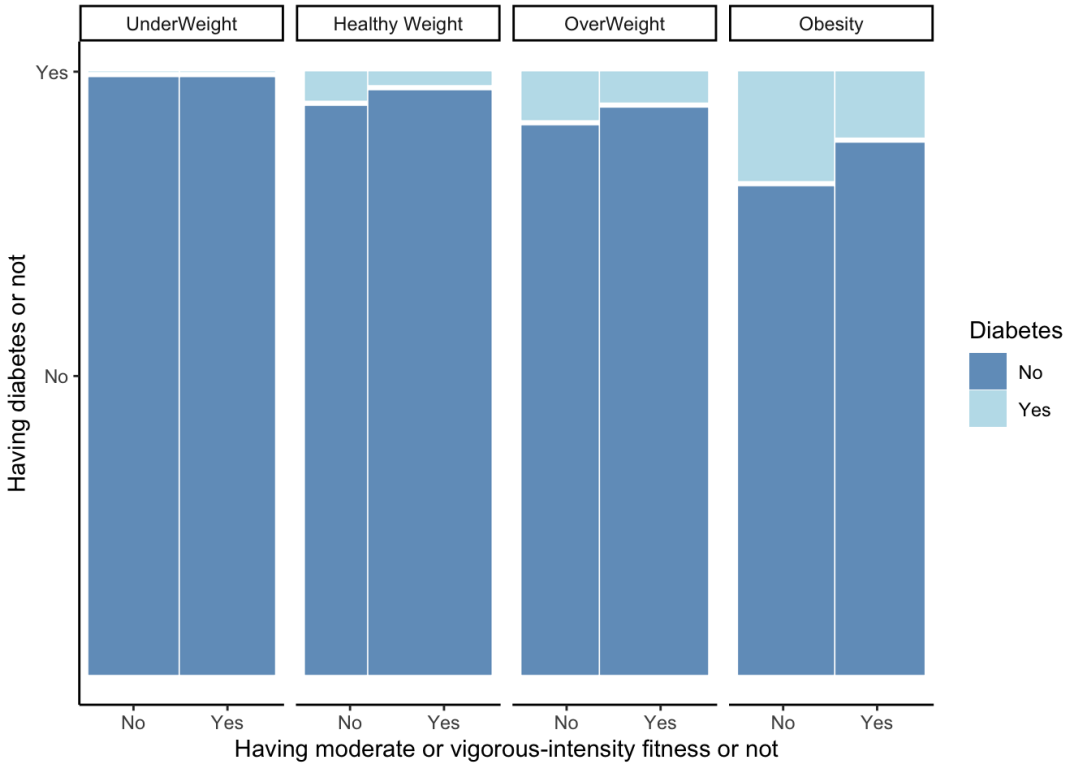
```
##
## Call:
## lm(formula = Weight ~ Height + BPSysAve + PhysActive + Race3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.783 -12.308  -2.967   9.910 106.648
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -106.16583    5.90196 -17.988  < 2e-16 ***
## Height           0.88944    0.03235  27.497  < 2e-16 ***
## BPSysAve         0.26679    0.02129  12.529  < 2e-16 ***
## PhysActiveYes   -4.63882    0.64598  -7.181 8.48e-13 ***
## Race3Black      11.62467    1.60982   7.221 6.35e-13 ***
## Race3Hispanic    8.75248    1.79409   4.879 1.12e-06 ***
## Race3Mexican     9.77492    1.68772   5.792 7.61e-09 ***
## Race3White       7.90039    1.37077   5.763 8.99e-09 ***
## Race3Other      10.87896    2.30552   4.719 2.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.24 on 3347 degrees of freedom
## Multiple R-squared:  0.2582, Adjusted R-squared:  0.2564
## F-statistic: 145.6 on 8 and 3347 DF,  p-value: < 2.2e-16
```

Research Question #2

Visualization

```
# change labels in BMI_WHO from BMI ranges(i.e. "12.0_18.5") into names(i.e."UnderWeight")
labels <- c("12.0_18.5" = "UnderWeight", "18.5_to_24.9" = "Healthy Weight", "25.0_to_29.9" = "OverWeight",
"30.0_plus" = "Obesity")

# draw mosaic plot with Diabtes status, physical activity and BMI ranges
NHANES_2 %>%
  ggplot()+
  geom_mosaic(aes(x = product(Diabetes,PhysActive), fill = Diabetes)) +
  facet_grid(. ~ BMI_WHO, labeller = labeller(BMI_WHO = labels)) +
  scale_fill_manual(values = c("steelblue", "lightblue")) +
  xlab("Having moderate or vigorous-intensity fitness or not") +
  ylab("Having diabetes or not") +
  theme_classic()
```



Creat Models

Original Big Model

```
Rmod1 <- NHANES_2 %>% # Original big model
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3, family = binomial))
summary(Rmod1)
```

```
##
## Call:
## glm(formula = Diabetes ~ BMI + Gender + BPSysAve + PhysActive +
##     Race3, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3450  -0.4200  -0.3066  -0.2170   2.8488
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.152064   0.608719 -13.392  < 2e-16 ***
## BMI            0.094487   0.008752  10.797  < 2e-16 ***
## Gendermale     0.140245   0.136012   1.031  0.30248
## BPSysAve       0.031013   0.004190   7.402 1.34e-13 ***
## PhysActiveYes -0.613150   0.135666  -4.520 6.20e-06 ***
## Race3Black    -0.393265   0.302107  -1.302  0.19300
## Race3Hispanic -0.626516   0.353047  -1.775  0.07596 .
## Race3Mexican  -0.737843   0.337499  -2.186  0.02880 *
## Race3White    -0.867844   0.269408  -3.221  0.00128 **
## Race3Other    -0.671919   0.458575  -1.465  0.14286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922  on 3355  degrees of freedom
## Residual deviance: 1653  on 3346  degrees of freedom
## AIC: 1673
##
## Number of Fisher Scoring iterations: 6
```

Model without Race

```
Rmod2 <- NHANES_2 %>% #model without Race3
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + Gender + BPSysAve + PhysActive, family = binomial))
```

Model without Physcial Activity

```
Rmod3 <- NHANES_2 %>% #model without PhysActive
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + Gender + BPSysAve + Race3, family = binomial))
```

Model without Combined systolic blood pressure

```
Rmod4 <- NHANES_2 %>% #Model without BPSysAve
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + Gender + PhysActive + Race3, family = binomial))
```

Model without Gender

```
Rmod5 <- NHANES_2 %>% #Model without Gender
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + Gender + PhysActive + Race3, family = binomial))
```

Model without BMI

```
Rmod6 <- NHANES_2 %>% # Model without BMI
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~  Gender + BPSysAve + PhysActive + Race3, family = binomial))
```

Model without Gender and Race(Final Model)

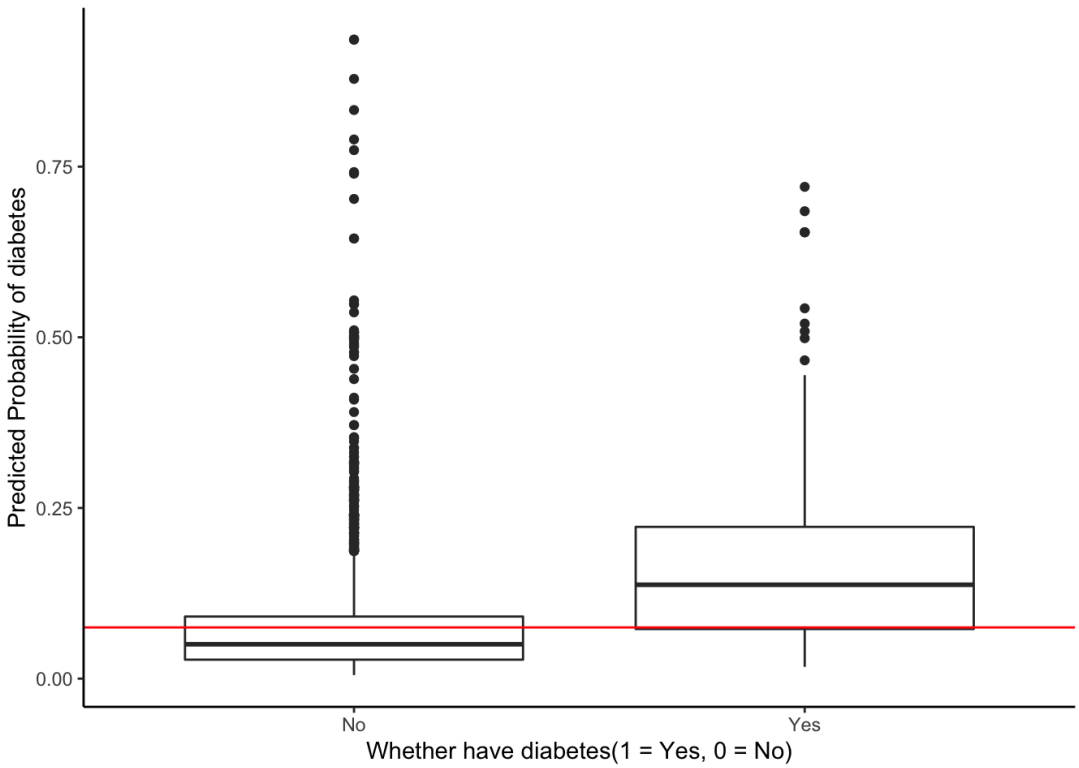
```
Rmod7 <- NHANES_2 %>% #Model without Gender and Race3
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + BPSysAve + PhysActive, family = binomial))
```

Model Selection

Original Big Model

```
#set the threshold
threshold <- 0.075

Rmod1 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Whether have diabetes(1 = Yes, 0 = No)') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod1 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

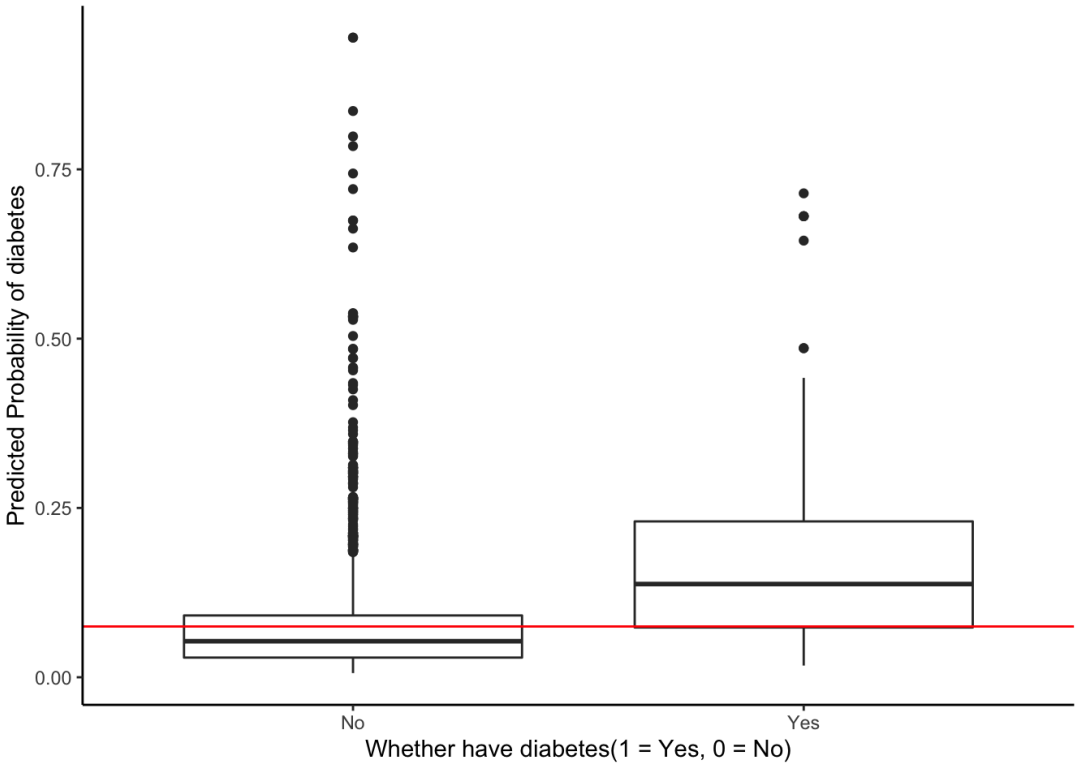
```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2090    0.679
## 2 No       TRUE              987    0.321
## 3 Yes      FALSE              73    0.262
## 4 Yes      TRUE              206    0.738
```

Model without Race

```
anova(Rmod1, Rmod2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3
## Model 2: Diabetes ~ BMI + Gender + BPSysAve + PhysActive
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3346       1653
## 2      3351       1667 -5   -13.95  0.01593 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rmod2 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Whether have diabetes(1 = Yes, 0 = No)') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod2 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

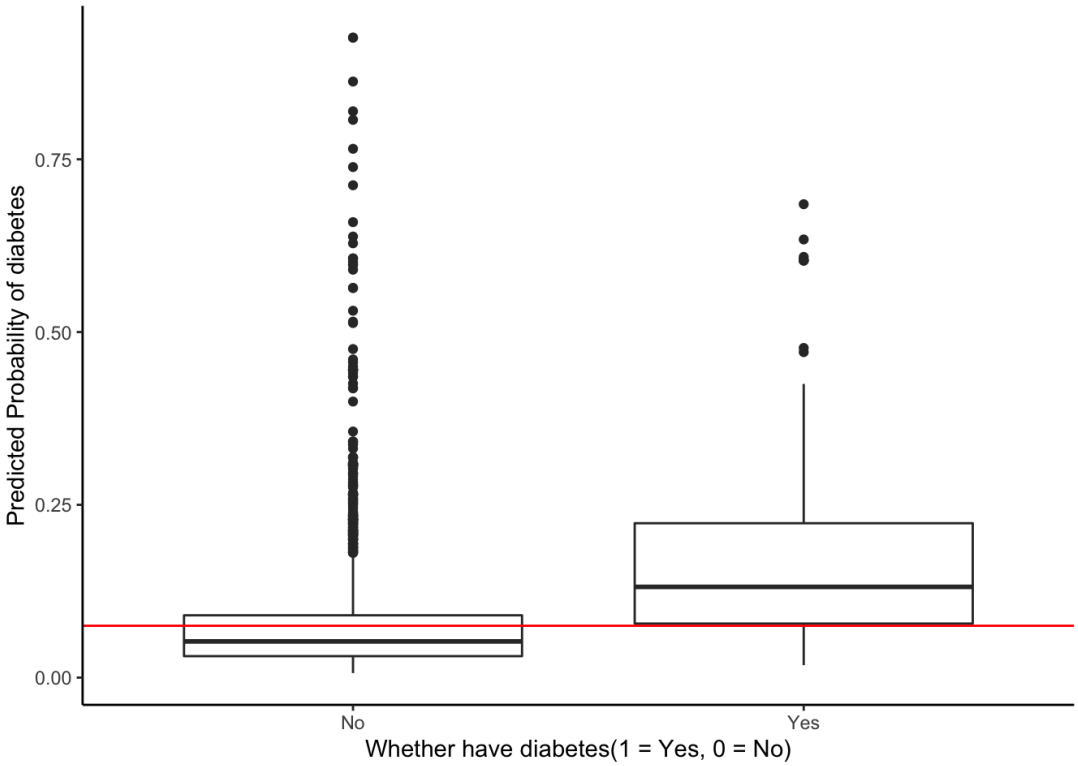
```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2092    0.680
## 2 No       TRUE              985    0.320
## 3 Yes      FALSE              71    0.254
## 4 Yes      TRUE              208    0.746
```

Model without Physcial Activity

```
anova(Rmod1, Rmod3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3
## Model 2: Diabetes ~ BMI + Gender + BPSysAve + Race3
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3346     1653.0
## 2      3347     1673.7 -1  -20.742 5.254e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rmod3 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Whether have diabetes(1 = Yes, 0 = No)') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod3 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

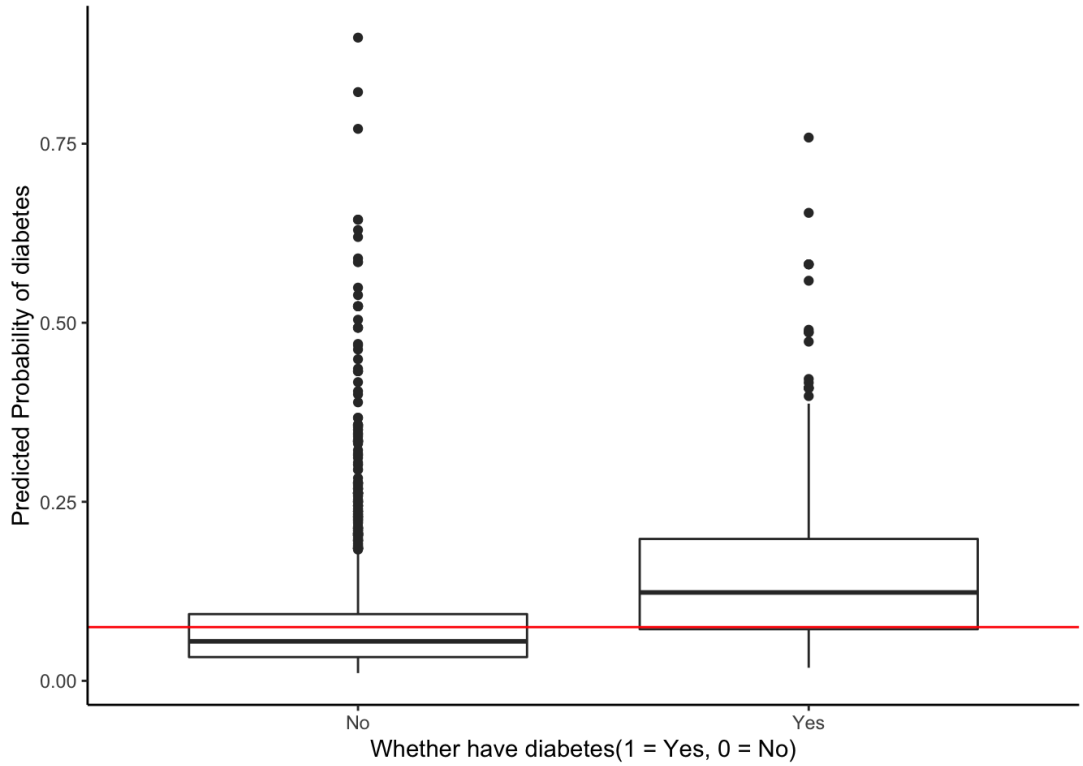
```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2075    0.674
## 2 No       TRUE             1002    0.326
## 3 Yes      FALSE              66    0.237
## 4 Yes      TRUE              213    0.763
```

Model without Combined systolic blood pressure

```
anova(Rmod1, Rmod4, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3
## Model 2: Diabetes ~ BMI + Gender + PhysActive + Race3
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3346     1653.0
## 2      3347     1706.4 -1   -53.38 2.749e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rmod4 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Whether have diabetes(1 = Yes, 0 = No)') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod4 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

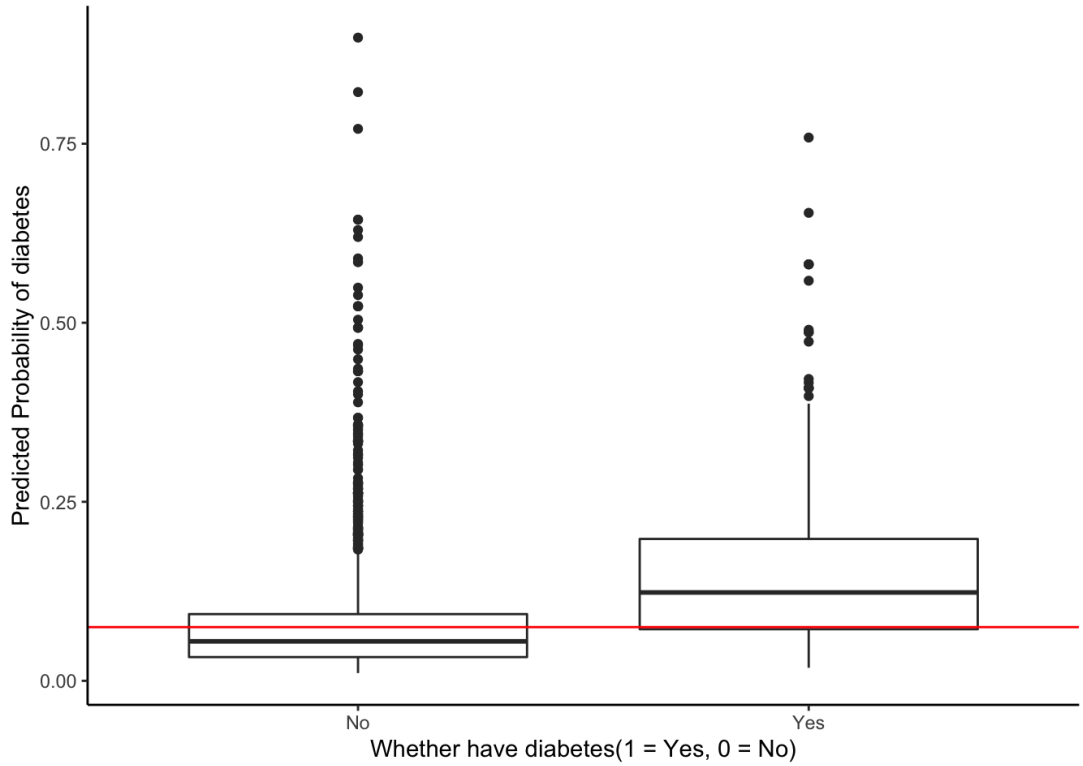
```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2012    0.654
## 2 No       TRUE             1065    0.346
## 3 Yes      FALSE              75    0.269
## 4 Yes      TRUE              204    0.731
```

Model without Gender

```
anova(Rmod1, Rmod5, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3
## Model 2: Diabetes ~ BMI + Gender + PhysActive + Race3
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3346     1653.0
## 2      3347     1706.4 -1   -53.38 2.749e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rmod5 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Whether have diabetes(1 = Yes, 0 = No)') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod5 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

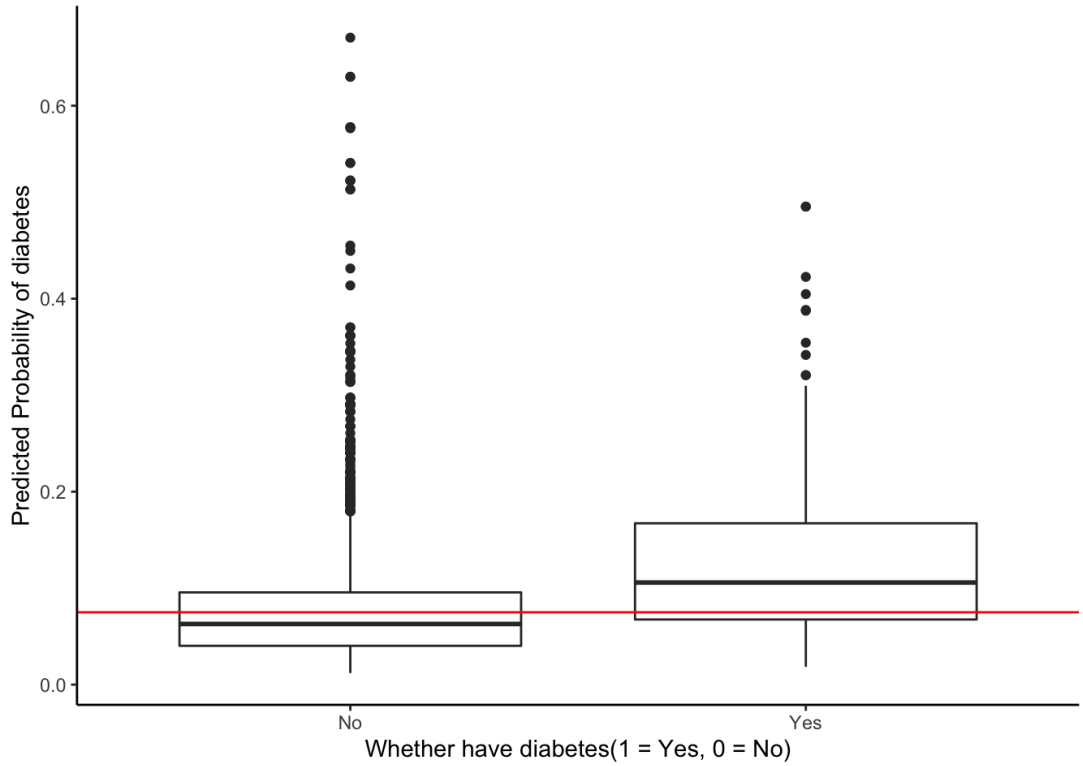
```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2012    0.654
## 2 No       TRUE             1065    0.346
## 3 Yes      FALSE              75    0.269
## 4 Yes      TRUE              204    0.731
```

Model without BMI

```
anova(Rmod1, Rmod6, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3
## Model 2: Diabetes ~ Gender + BPSysAve + PhysActive + Race3
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3346     1653.0
## 2      3347     1771.8 -1  -118.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rmod6 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Whether have diabetes(1 = Yes, 0 = No)') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod6 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

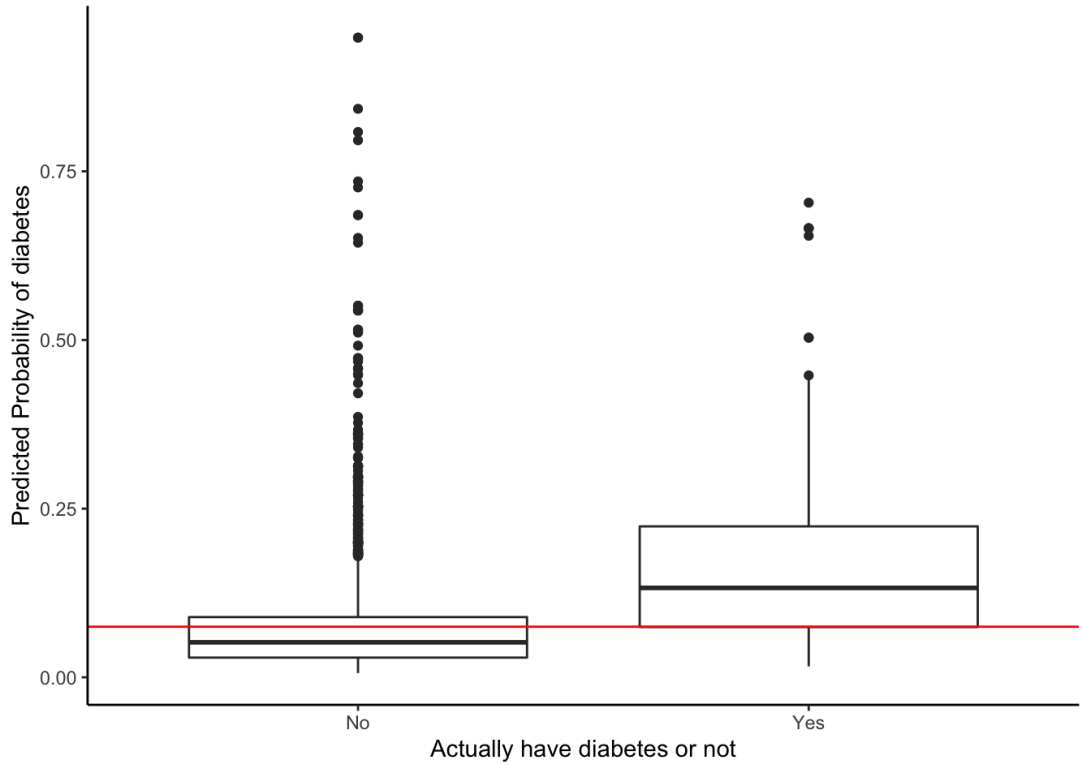
```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            1861    0.605
## 2 No       TRUE             1216    0.395
## 3 Yes      FALSE              80    0.287
## 4 Yes      TRUE              199    0.713
```

Model without Gender and Race(Final Model)

```
anova(Rmod1, Rmod7, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Diabetes ~ BMI + Gender + BPSysAve + PhysActive + Race3
## Model 2: Diabetes ~ BMI + BPSysAve + PhysActive
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3346     1653.0
## 2      3352     1667.8 -6  -14.781  0.02203 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rmod7 %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Actually have diabetes or not') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```




```
Rmod7 %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2102    0.683
## 2 No       TRUE              975    0.317
## 3 Yes      FALSE              71    0.254
## 4 Yes      TRUE              208    0.746
```

In this context, we focus on maximizing the overall accuracy, so I set the threshold as 0.075 here. When we exclude gender and race from the model, at the threshold of 0.075, both the specificity, sensitivity, and accuracy increases. Plus, the low p-value(< 0.0001) shows up in the nested hypothesis test of physical activity, BMI, and combined systolic blood pressure.Thus, the low p-values indicate that we should choose the model with physical activity, BMI, and combined systolic blood pressure for prediction.

Final Model:

$$log(Odds[Diabetes|BMI, BPSysAve, PhysActive]) = \beta_0 + \beta_1 * BMI + \beta_2 * BPSysAve + \beta_3 * PhysActiveYes$$

Fit Model

```
Rmod <- NHANES_2 %>% #Model without Gender and Race3
  select(Diabetes, BMI, Gender, BPSysAve, PhysActive, Race3) %>%
  na.omit() %>%
  with(glm(Diabetes ~ BMI + BPSysAve + PhysActive, family = binomial))
```

Inference

```
Rmod %>%
  coef() %>%
  exp()
```

```
##   (Intercept)           BMI      BPSysAve PhysActiveYes
##  0.0001519416  1.0961304593  1.0323045071  0.5371895357
```

```
Rmod %>%
  confint() %>%
  exp()
```

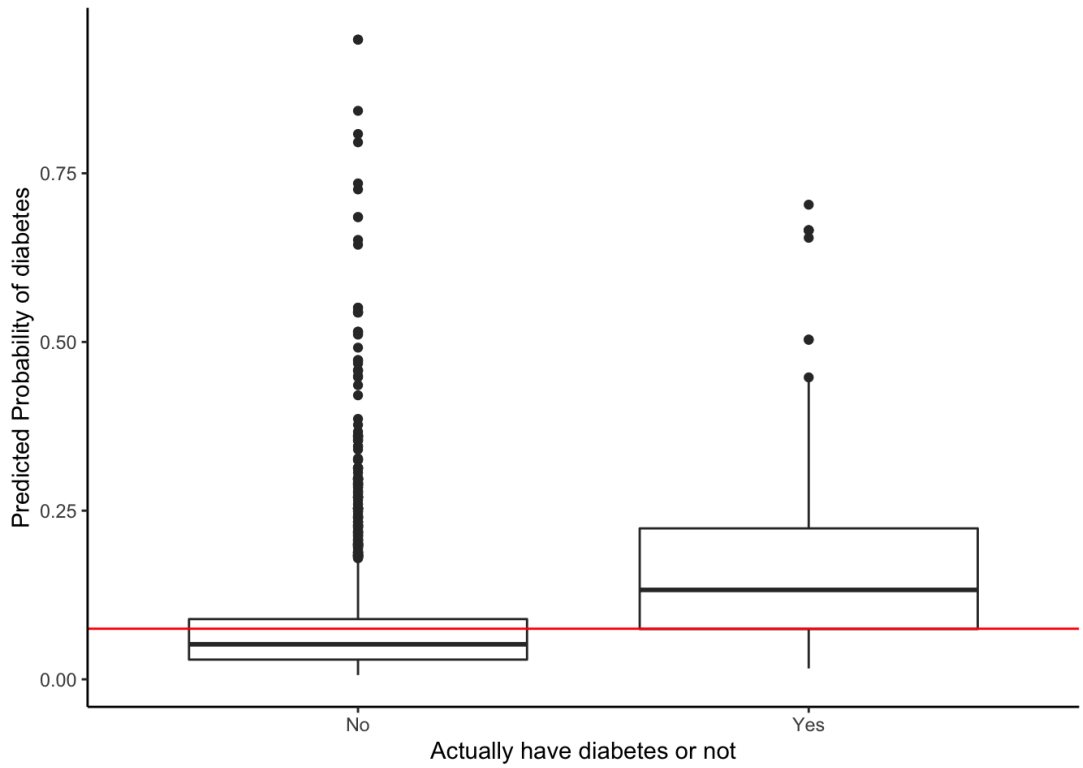
```
##                      2.5 %       97.5 %
## (Intercept)   4.821544e-05 0.0004654146
## BMI           1.078088e+00 1.1147053075
## BPSysAve      1.023963e+00 1.0407002054
## PhysActiveYes 4.120372e-01 0.6983065489
```

```
summary(Rmod)
```

```
##
## Call:
## glm(formula = Diabetes ~ BMI + BPSysAve + PhysActive, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4324  -0.4189  -0.3125  -0.2253   2.8724
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.792015   0.578061 -15.209  < 2e-16 ***
## BMI            0.091786   0.008512  10.783  < 2e-16 ***
## BPSysAve       0.031794   0.004133   7.693 1.43e-14 ***
## PhysActiveYes -0.621404   0.134444  -4.622 3.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.0  on 3355  degrees of freedom
## Residual deviance: 1667.8  on 3352  degrees of freedom
## AIC: 1675.8
##
## Number of Fisher Scoring iterations: 6
```

Evaluation

```
Rmod %>%
  augment(type.predict = "response") %>%
  ggplot(aes(y = .fitted, x = Diabetes)) +
  geom_boxplot() +
  ylab('Predicted Probability of diabetes') +
  xlab('Actually have diabetes or not') +
  geom_hline(yintercept = threshold, color = 'red') +
  theme_classic()
```



```
Rmod %>%
  augment(type.predict = 'response') %>%
  mutate(predictDiabetes = .fitted >= threshold) %>%
  count(Diabetes, predictDiabetes) %>%
  group_by(Diabetes) %>%
  mutate(condprop = n / sum(n))
```

```
## # A tibble: 4 × 4
## # Groups:   Diabetes [2]
##   Diabetes predictDiabetes     n condprop
##   <fct>    <lgl>           <int>    <dbl>
## 1 No       FALSE            2102    0.683
## 2 No       TRUE              975    0.317
## 3 Yes      FALSE              71    0.254
## 4 Yes      TRUE              208    0.746
```

```
(accuracy = (2101+208)/(2102+975+71+208))
```

```
## [1] 0.6880215
```