**Category**

* [Regression Models](https://datascienceplus.com/category/regression-models/)

**Tags**

* [AUC](https://datascienceplus.com/tag/auc/)
* [Logistic Regression](https://datascienceplus.com/tag/logistic-regression/)
* [R Programming](https://datascienceplus.com/tag/rstats/)

In my previous post, I showed how to run a [linear regression model with medical data](https://datascienceplus.com/linear-regression-with-healthcare-data-for-beginners-in-r/). In this post, I will show how to conduct a logistic regression model. The major difference between linear and logistic regression is that the latter needs a dichotomous (0/1) dependent (outcome) variable, whereas the first, work with a continuous outcome. I will run a logistic regression to evaluate the effect of calcium and vitD on the osteoporosis.

**Let's start loading the packages:**

library(tidyverse)

library(RNHANES)

library(ggplot2)

library(pROC)

**Prepare the dataset**

Variables included for this analysis are:

* age (years)
* sex (women, men)
* serum levels of vitamin D (mg/ml)
* serum levels of calcium (mg/ml)
* osteoporosis (yes/no, 1/0).

All variables are assessed from [NHANES](https://www.cdc.gov/nchs/nhanes/index.htm) in the cycles 2007-2008 and 2009-2010.

d07 = nhanes\_load\_data("DEMO\_E", "2007-2008") %>%

select(SEQN, cycle, RIAGENDR, RIDAGEYR) %>%

transmute(SEQN=SEQN, wave=cycle, RIAGENDR, RIDAGEYR) %>%

left\_join(nhanes\_load\_data("VID\_E", "2007-2008"), by="SEQN") %>%

select(SEQN, wave, RIAGENDR, RIDAGEYR, LBXVIDMS) %>%

transmute(SEQN, wave, RIAGENDR, RIDAGEYR, vitD=LBXVIDMS) %>%

left\_join(nhanes\_load\_data("BIOPRO\_E", "2007-2008"), by="SEQN") %>%

select(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, LBXSCA) %>%

transmute(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, Calcium = LBXSCA) %>%

left\_join(nhanes\_load\_data("OSQ\_E", "2007-2008"), by="SEQN") %>%

select(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, Calcium, OSQ060) %>%

transmute(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, Calcium, Osteop = OSQ060)

d09 = nhanes\_load\_data("DEMO\_F", "2009-2010") %>%

select(SEQN, cycle, RIAGENDR, RIDAGEYR) %>%

transmute(SEQN=SEQN, wave=cycle, RIAGENDR, RIDAGEYR) %>%

left\_join(nhanes\_load\_data("VID\_F", "2009-2010"), by="SEQN") %>%

select(SEQN, wave, RIAGENDR, RIDAGEYR, LBXVIDMS) %>%

transmute(SEQN, wave, RIAGENDR, RIDAGEYR, vitD=LBXVIDMS) %>%

left\_join(nhanes\_load\_data("BIOPRO\_F", "2009-2010"), by="SEQN") %>%

select(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, LBXSCA) %>%

transmute(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, Calcium = LBXSCA) %>%

left\_join(nhanes\_load\_data("OSQ\_F", "2009-2010"), by="SEQN") %>%

select(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, Calcium, OSQ060) %>%

transmute(SEQN, wave, RIAGENDR, RIDAGEYR, vitD, Calcium, Osteop = OSQ060)

dat = bind\_rows(d07, d09) %>% as.data.frame()

**Create categories of Vitamin D**

Institute of Medicine cutoffs for Vitamin D

* Vitamin D deficiency: Serum 25OHD less than 30 nmol/L (12 ng/mL)
* Vitamin D inadequacy: Serum 25OHD 30-49 nmol/L (12-19 ng/mL)
* Vitamin D sufficiency: Serum 25OHD 50-125 nmol/L (20-50 ng/mL)

dat1 = dat %>%

mutate(

vitD\_group = case\_when(

vitD < 30 ~ "Deficiency",

vitD >= 30 & vitD < 50 ~ "Inadequacy",

vitD >= 50 & vitD <= 125 ~ "Sufficiency"))

**Exclude missings**

dat2 = dat1 %>%

filter(!is.na(vitD\_group), !is.na(Calcium), !is.na(Osteop), Osteop!=9) %>%

mutate(Gender = recode\_factor(RIAGENDR,

`1` = "Men",

`2` = "Women"),

Osteop = recode\_factor(Osteop,

`1` = 1,

`2` = 0))

head(dat2)

*## SEQN wave RIAGENDR RIDAGEYR vitD Calcium Osteop vitD\_group Gender*

*## 1 41475 2007-2008 2 62 58.8 9.5 0 Sufficiency Women*

*## 2 41477 2007-2008 1 71 81.8 10.0 0 Sufficiency Men*

*## 3 41479 2007-2008 1 52 78.4 9.0 0 Sufficiency Men*

*## 4 41482 2007-2008 1 64 61.9 9.1 0 Sufficiency Men*

*## 5 41483 2007-2008 1 66 53.3 8.9 0 Sufficiency Men*

*## 6 41485 2007-2008 2 30 39.1 9.3 0 Inadequacy Women*

**Logit regression model**

I will use the glm() function to run the logistic regression and then summary() command to get the results.

fit <- glm(Osteop ~ vitD\_group + Calcium + Gender + RIDAGEYR,

data = dat2,

family = "binomial")

summary(fit)

*##*

*## Call:*

*## glm(formula = Osteop ~ vitD\_group + Calcium + Gender + RIDAGEYR,*

*## family = "binomial", data = dat2)*

*##*

*## Deviance Residuals:*

*## Min 1Q Median 3Q Max*

*## -3.4265 0.1009 0.1894 0.3315 1.0305*

*##*

*## Coefficients:*

*## Estimate Std. Error z value Pr(>|z|)*

*## (Intercept) 7.81969 1.08054 7.237 4.59e-13 \*\*\**

*## vitD\_groupInadequacy -0.17444 0.20124 -0.867 0.38603*

*## vitD\_groupSufficiency -0.53068 0.18159 -2.922 0.00347 \*\**

*## Calcium 0.10330 0.11404 0.906 0.36506*

*## GenderWomen -2.08873 0.12298 -16.984 < 2e-16 \*\*\**

*## RIDAGEYR -0.07127 0.00330 -21.599 < 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: 4591.4 on 10064 degrees of freedom*

*## Residual deviance: 3553.1 on 10059 degrees of freedom*

*## AIC: 3565.1*

*##*

*## Number of Fisher Scoring iterations: 7*

**Transforms beta's to the odds ratio**

The output of summary() does not provide the odds ratio which are often presented in research papers. The exp of beta's give the odds.

round(exp(coef(fit)), 2)

*## (Intercept) vitD\_groupInadequacy vitD\_groupSufficiency*

*## 2489.14 0.84 0.59*

*## Calcium GenderWomen RIDAGEYR*

*## 1.11 0.12 0.93*

**Interpreting results**

From the output, I see that there is a significant association between vitamin D and osteoporosis. Compared to individuals with deficiency levels of vitamin D, those with sufficient levels of vitamin D in the blood have 41% (odds ratio: 0.59) lower risk of having osteoporosis. Inadequacy of vitamin D is not significantly (p=0.38) associated with osteoporosis.

To get the 95% confidence interval, I use confit() for confidence intervals of each variable.

round(exp(confint(fit)), 2)

*## 2.5 % 97.5 %*

*## (Intercept) 298.55 20650.10*

*## vitD\_groupInadequacy 0.56 1.24*

*## vitD\_groupSufficiency 0.41 0.83*

*## Calcium 0.89 1.39*

*## GenderWomen 0.10 0.16*

*## RIDAGEYR 0.93 0.94*

**Assessing discrimination of the model with ROC curve**

When studying a new biomarker, it is essential to illustrate the discrimination ability of the model, in addition to the association with the outcome, osteoporosis in our example. Levels of vitamin D in the blood are known to be related to osteoporosis, but here will show how much discrimination adds to the model.

First, I will run a model without vitamin D and assess the discrimination and after adding vitamin D in the model and see the differences in the ROC curve.

# model without vitamin D

fit1 <- glm(Osteop ~ Calcium + Gender + RIDAGEYR,

data = dat2,

family = "binomial")

# model with vitamin D

fit2 <- glm(Osteop ~ vitD\_group + Calcium + Gender + RIDAGEYR,

data = dat2,

family = "binomial")

dat2$prob1=predict(fit1,type=c("response"))

dat2$prob2=predict(fit2,type=c("response"))

roc(Osteop ~ prob1, data = dat2)

*##*

*## Call:*

*## roc.formula(formula = Osteop ~ prob1, data = dat2)*

*##*

*## Data: prob1 in 608 controls (Osteop 1) < 9457 cases (Osteop 0).*

*## Area under the curve: 0.8496*

roc(Osteop ~ prob2, data = dat2)

*##*

*## Call:*

*## roc.formula(formula = Osteop ~ prob2, data = dat2)*

*##*

*## Data: prob2 in 608 controls (Osteop 1) < 9457 cases (Osteop 0).*

*## Area under the curve: 0.8508*

There is a slight improvement in discrimination with including vitamin D in the model from 0.8496 to 0.8508.