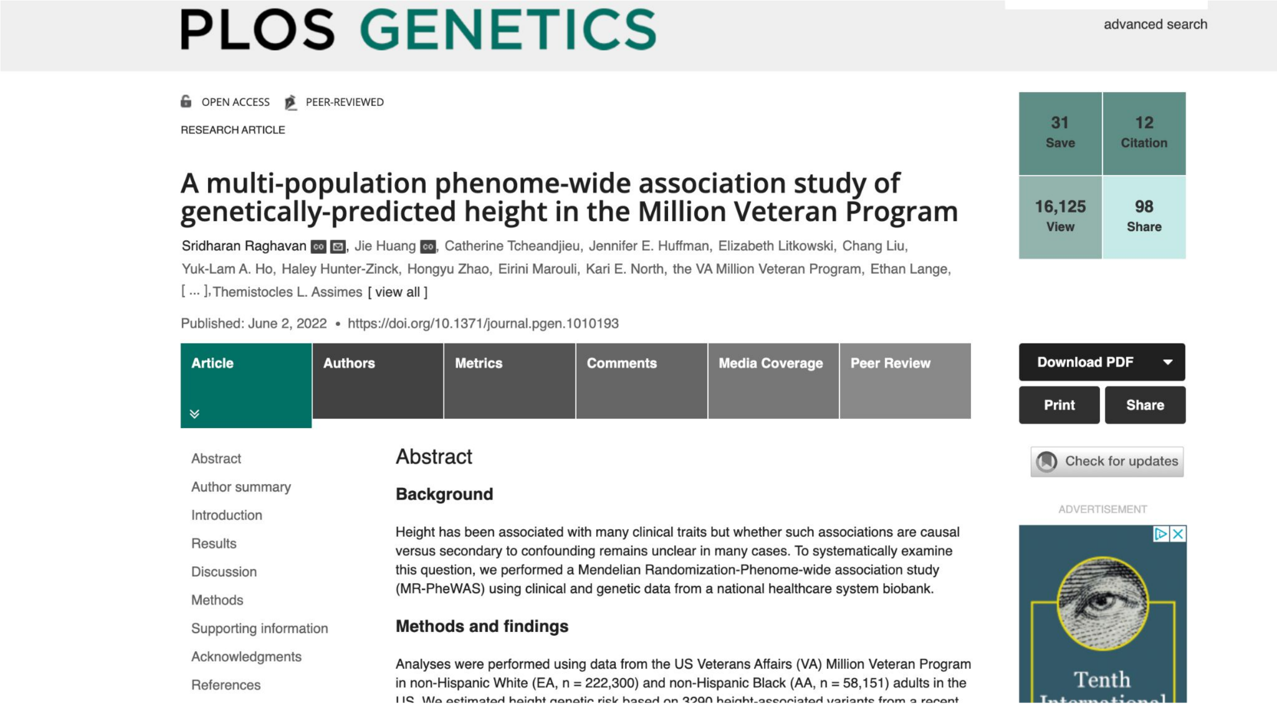
Logistic regression with data akin to the classroom simulation

Illustrative study of height as a risk or protective factor for different clinical traits, such as peripheral neuropathy

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010193



A multi-population phenome-wide association study of genetically-predicted height in the Million Veteran Program

Sridharan Raghavan, Jie Huang, Catherine Tcheandjieu, Jennifer E. Huffman, Elizabeth Litkowski, Chang Liu, Yuk-Lam A. Ho, Haley Hunter-Zinck, Hongyu Zhao, Eirini Marouli, Kari E. North, the VA Million Veteran Program, Ethan Lange, [...], Themistocles L. Assimes [view all]

Published: June 2, 2022 • https://doi.org/10.1371/journal.pgen.1010193

Abstract

Background

Height has been associated with many clinical traits but whether such associations are causal versus secondary to confounding remains unclear in many cases. To systematically examine this question, we performed a Mendelian Randomization-Phenome-wide association study (MR-PheWAS) using clinical and genetic data from a national healthcare system biobank.

Methods and findings

Analyses were performed using data from the US Veterans Affairs (VA) Million Veteran Program in non-Hispanic White (EA, n = 222,300) and non-Hispanic Black (AA, n = 58,151) adults in the US. We estimated height genetic risk based on 3290 height-associated variants from a recent

Example simulation similar to the classroom experimental design

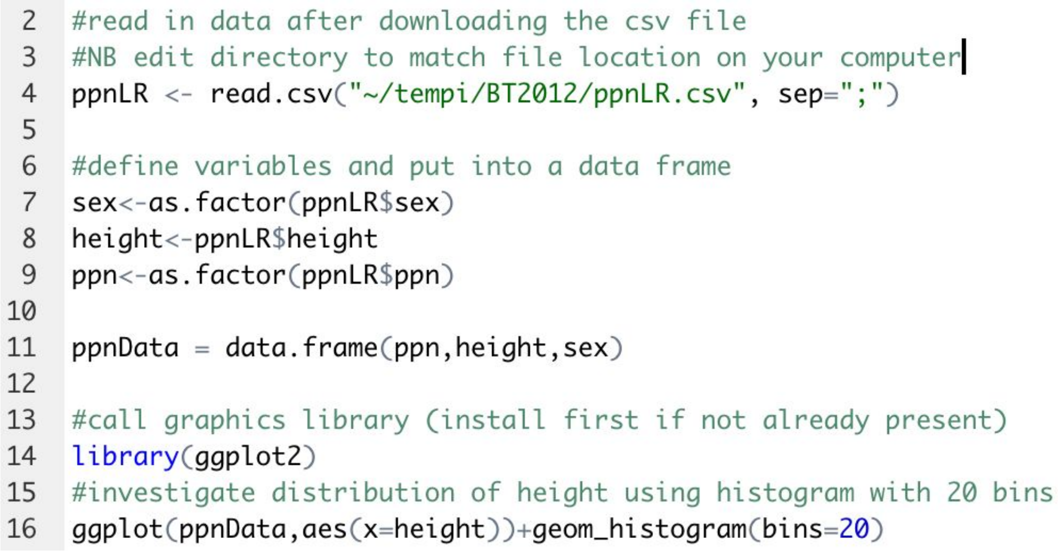
Example data ppnLR.csv downloadable from the course folder

Contains 162 observations on 3 variables: *sex* (F/M), *height* (in cm) and *peripheral neuropathy* (*ppn*, coded as 0: no *ppn*, 1: diagnosed *ppn*)

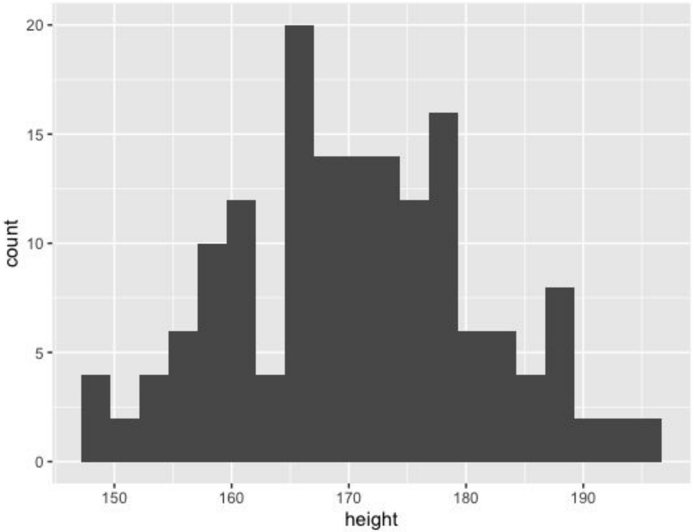
We will examine association between *ppn* and *height* using EDA and logistic regression and consider how *sex* must be taken into account

R code (height demo.R) can be downloaded from the same folder

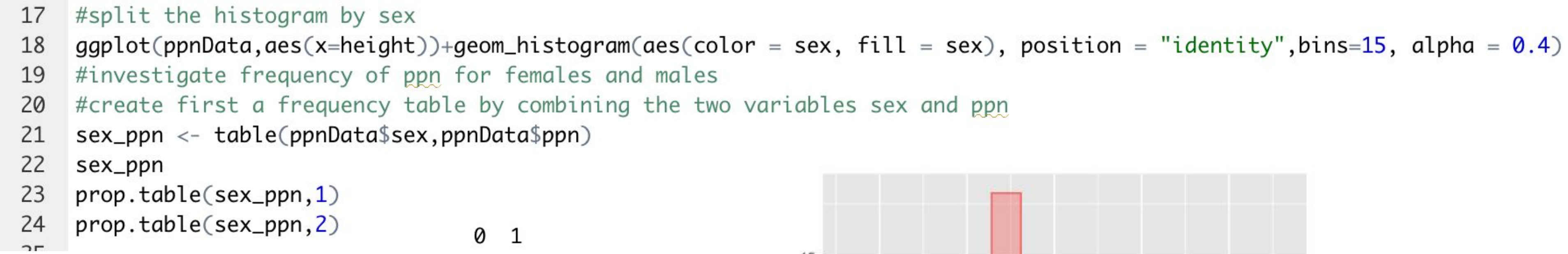
Start data investigation



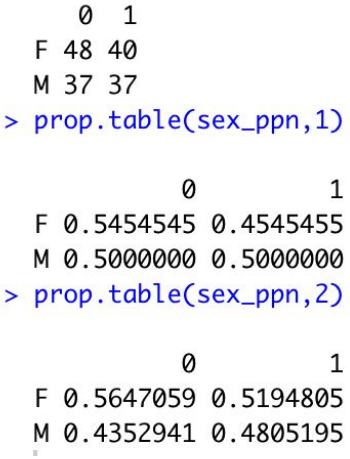
```r
#read in data after downloading the csv file
#NB edit directory to match file location on your computer
ppnLR <- read.csv("~/tempi/BT2012/ppnLR.csv", sep=";")

#define variables and put into a data frame
sex<-as.factor(ppnLR$sex)
height<-ppnLR$height
ppn<-as.factor(ppnLR$ppn)

ppnData = data.frame(ppn,height,sex)

#call graphics library (install first if not already present)
library(ggplot2)
#investigate distribution of height using histogram with 20 bins
ggplot(ppnData,aes(x=height))+geom_histogram(bins=20)
```



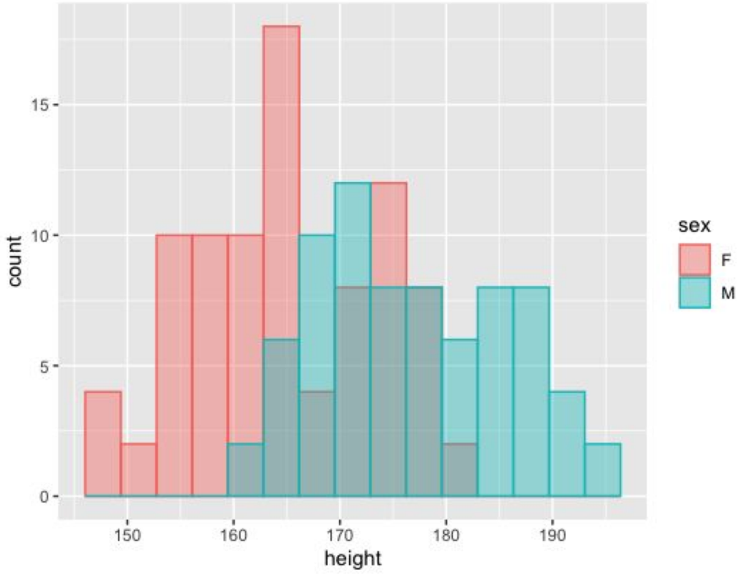
Data investigation continued



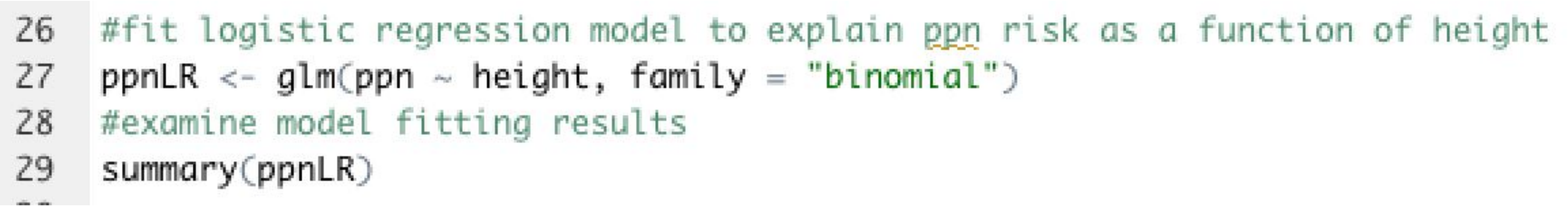
```r
#split the histogram by sex
ggplot(ppnData,aes(x=height))+geom_histogram(aes(color = sex, fill = sex), position = "identity",bins=15, alpha = 0.4)
#investigate frequency of ppn for females and males
#create first a frequency table by combining the two variables sex and ppn
sex_ppn <- table(ppnData$sex,ppnData$ppn)
sex_ppn
prop.table(sex_ppn,1)
prop.table(sex_ppn,2)
```



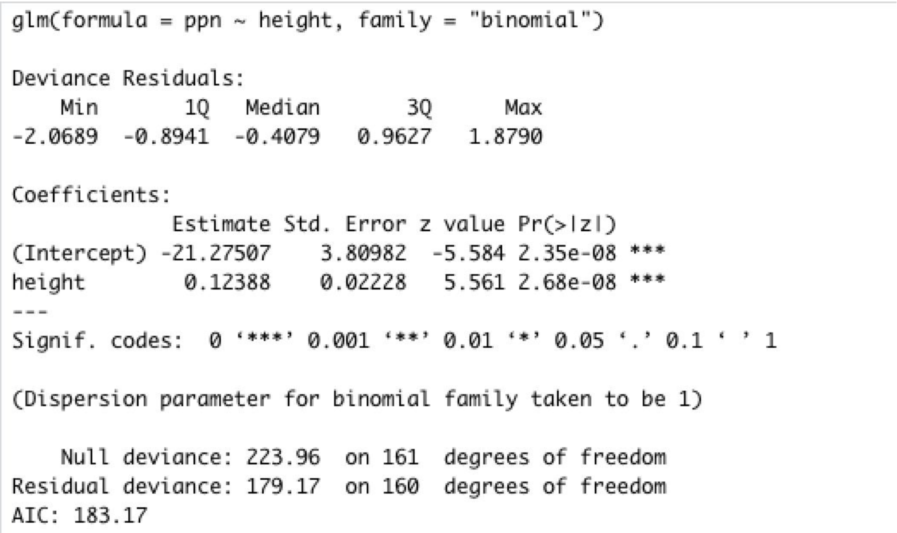
```
   0  1
F 48 40
M 37 37
> prop.table(sex_ppn,1)

            0         1
F 0.5454545 0.4545455
M 0.5000000 0.5000000
> prop.table(sex_ppn,2)

            0         1
F 0.5647059 0.5194805
M 0.4352941 0.4805195
```



Logistic regression for *ppn* as a function of *height*

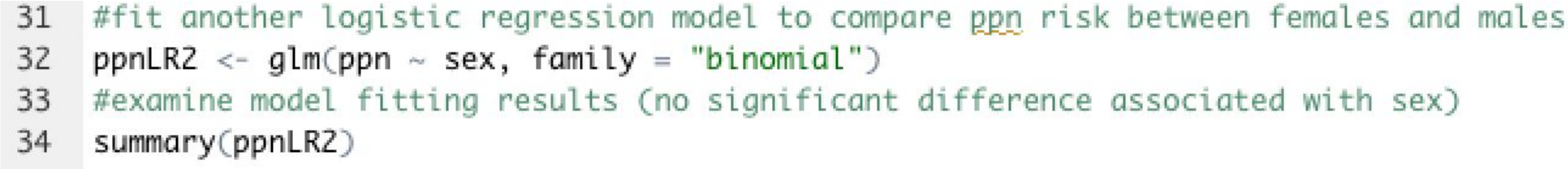


```
#fit logistic regression model to explain ppn risk as a function of height
ppnLR <- glm(ppn ~ height, family = "binomial")
#examine model fitting results
summary(ppnLR)
```

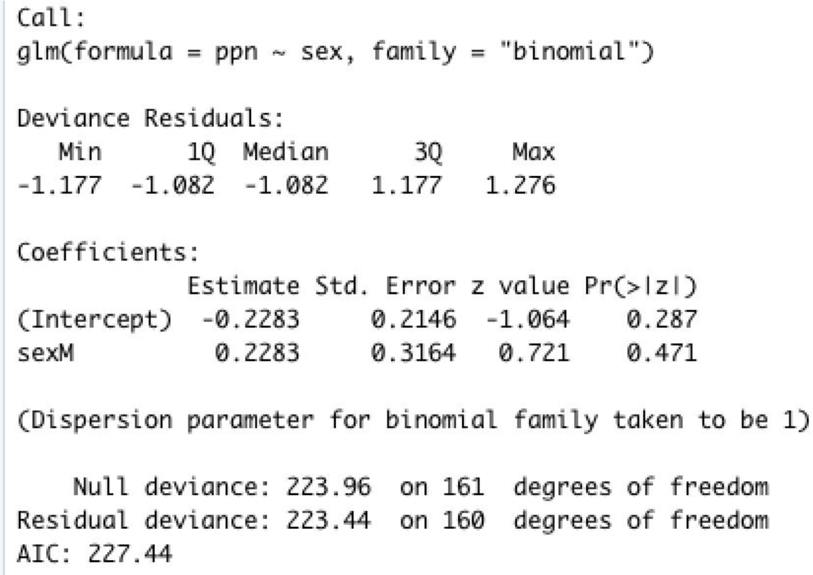


```
glm(formula = ppn ~ height, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0689  -0.8941  -0.4079   0.9627   1.8790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.27507    3.80982  -5.584 2.35e-08 ***
height        0.12388    0.02228   5.561 2.68e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 223.96  on 161  degrees of freedom
Residual deviance: 179.17  on 160  degrees of freedom
AIC: 183.17
```

Logistic regression for *ppn* as a function of *sex*

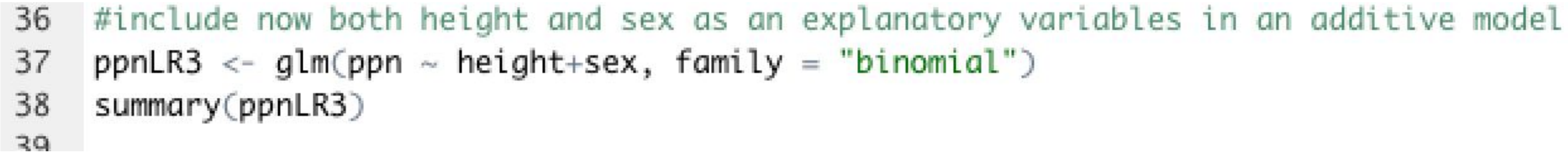


```r
#fit another logistic regression model to compare ppn risk between females and males
ppnLR2 <- glm(ppn ~ sex, family = "binomial")
#examine model fitting results (no significant difference associated with sex)
summary(ppnLR2)
```

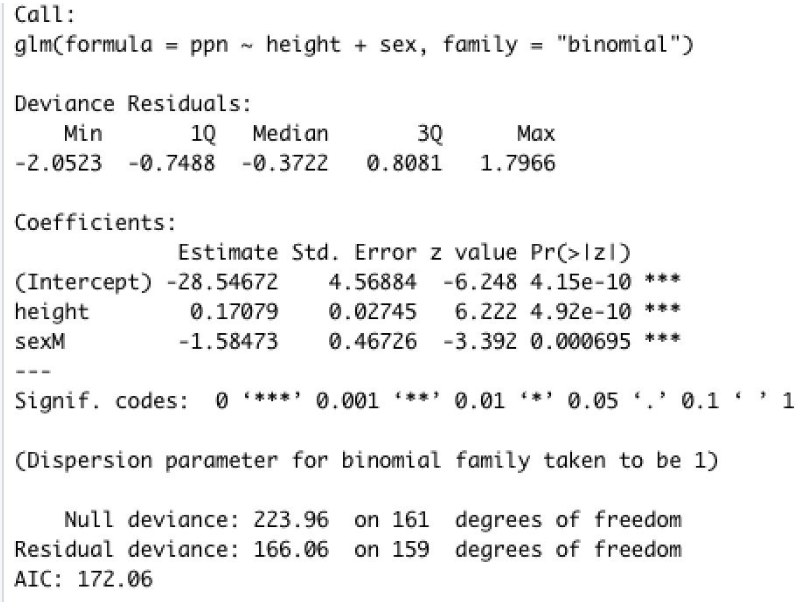


```
Call:
glm(formula = ppn ~ sex, family = "binomial")

Deviance Residuals:
   Min       1Q   Median       3Q      Max  
-1.177   -1.082   -1.082    1.177    1.276  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2283     0.2146  -1.064    0.287
sexM          0.2283     0.3164   0.721    0.471

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 223.96  on 161  degrees of freedom
Residual deviance: 223.44  on 160  degrees of freedom
AIC: 227.44
```

Logistic regression for *ppn* as a function of *height* and *sex*

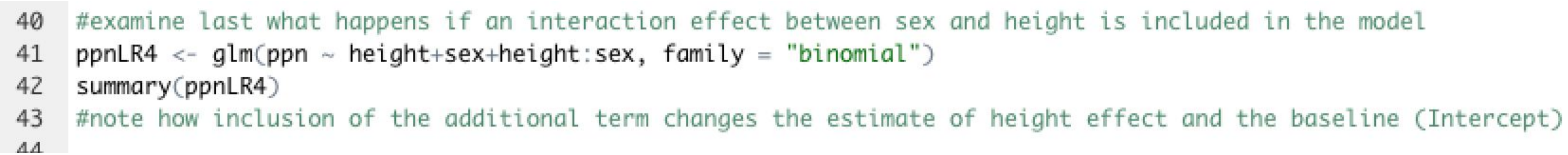


```
#include now both height and sex as an explanatory variables in an additive model
ppnLR3 <- glm(ppn ~ height+sex, family = "binomial")
summary(ppnLR3)
```

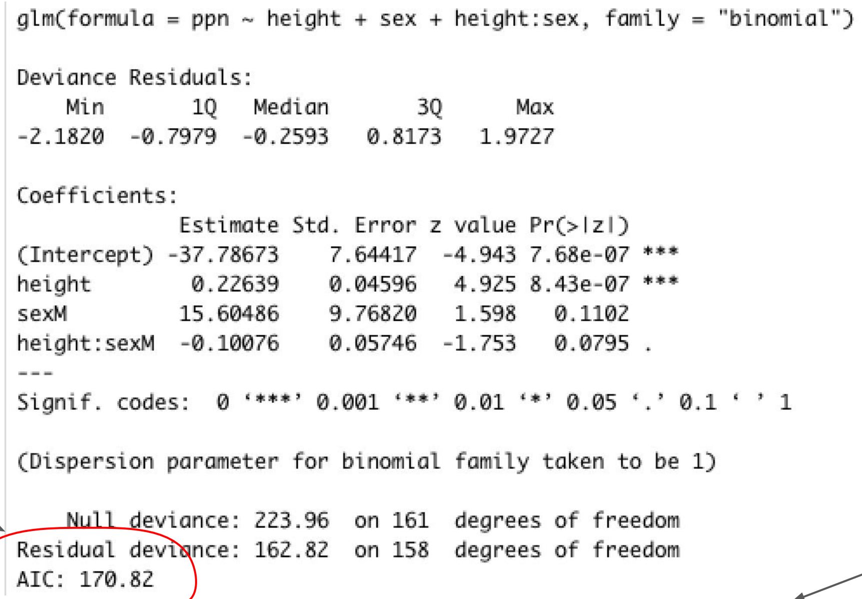


```
Call:
glm(formula = ppn ~ height + sex, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0523  -0.7488  -0.3722   0.8081   1.7966

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -28.54672    4.56884  -6.248 4.15e-10 ***
height        0.17079    0.02745   6.222 4.92e-10 ***
sexM         -1.58473    0.46726  -3.392 0.000695 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 223.96  on 161  degrees of freedom
Residual deviance: 166.06  on 159  degrees of freedom
AIC: 172.06
```

LR for *ppn* as a function of *height* and *sex* with *interaction* term



```
#examine last what happens if an interaction effect between sex and height is included in the model
ppnLR4 <- glm(ppn ~ height+sex+height:sex, family = "binomial")
summary(ppnLR4)
#note how inclusion of the additional term changes the estimate of height effect and the baseline (Intercept)
```



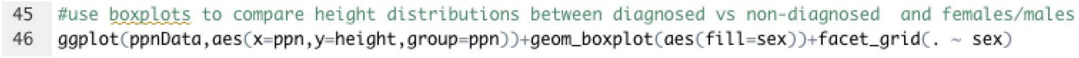
```
glm(formula = ppn ~ height + sex + height:sex, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1820  -0.7979  -0.2593   0.8173   1.9727

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -37.78673    7.64417  -4.943 7.68e-07 ***
height        0.22639    0.04596   4.925 8.43e-07 ***
sexM         15.60486    9.76820   1.598   0.1102
height:sexM  -0.10076    0.05746  -1.753   0.0795 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 223.96  on 161  degrees of freedom
Residual deviance: 162.82  on 158  degrees of freedom
AIC: 170.82
```

Compare with height+sex model AIC (172.06)

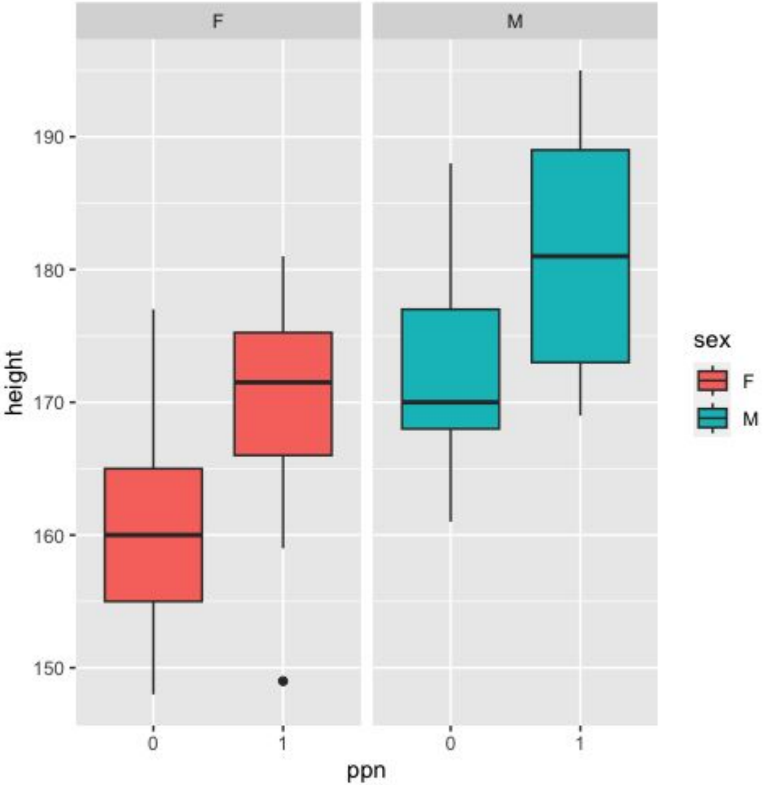
Review of model selection criteria, including AIC

https://www.sciencedirect.com/science/article/abs/pii/S0304380007005558

Dig deeper into the data using a structured box-plot



```
#use boxplots to compare height distributions between diagnosed vs non-diagnosed  and females/males
ggplot(ppnData,aes(x=ppn,y=height,group=ppn))+geom_boxplot(aes(fill=sex))+facet_grid(. ~ sex)
```



The data are not conclusive about the relative effect of *sex* on *ppn*, given *height*, which in turn has a clear effect (increases risk)

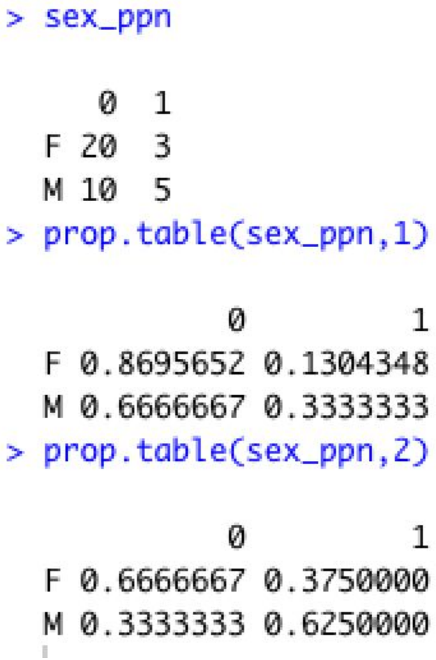
Perform the same kind of analyses for the classroom data

Data are coded in classroom.csv, downloadable from the course folder

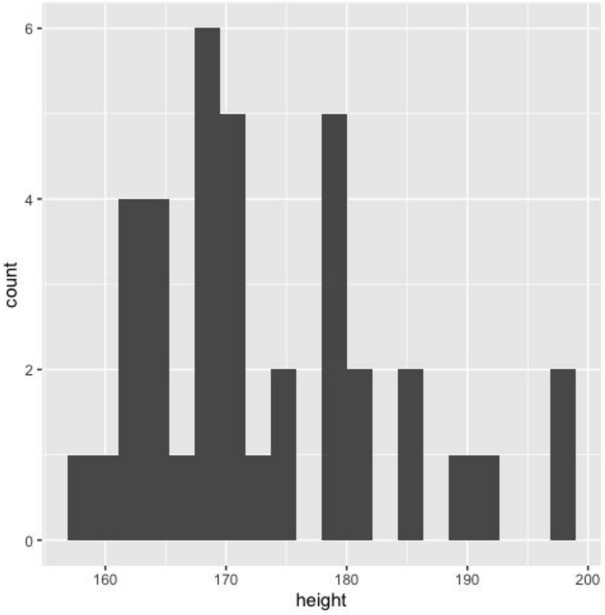
Contains 38 observations on 3 variables that are coded exactly as in the previous example

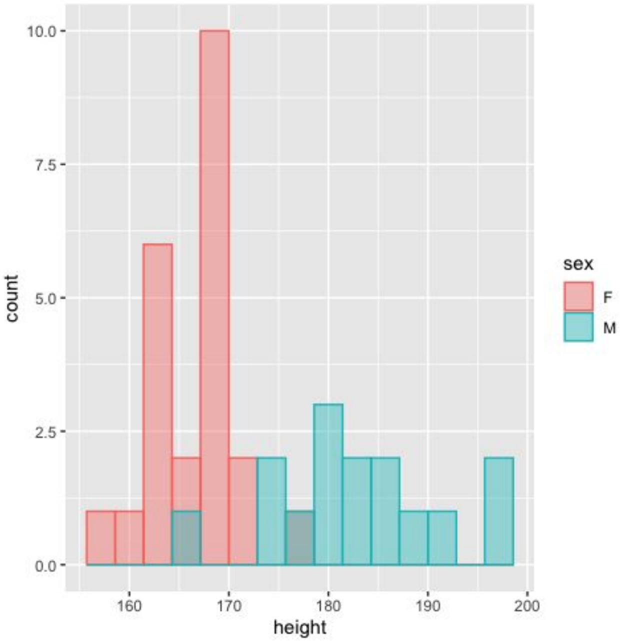
Same R code (height demo.R) can be used

Examine the data distribution



```
> sex_ppn

     0  1
  F 20  3
  M 10  5
> prop.table(sex_ppn,1)

            0         1
  F 0.8695652 0.1304348
  M 0.6666667 0.3333333
> prop.table(sex_ppn,2)

            0         1
  F 0.6666667 0.3750000
  M 0.3333333 0.6250000
```





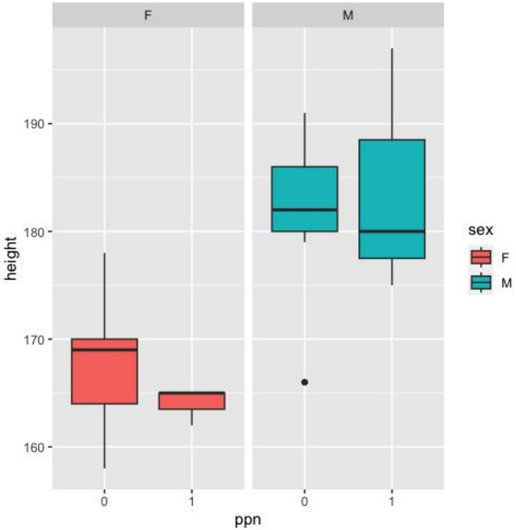
Fitting the logistic regression models

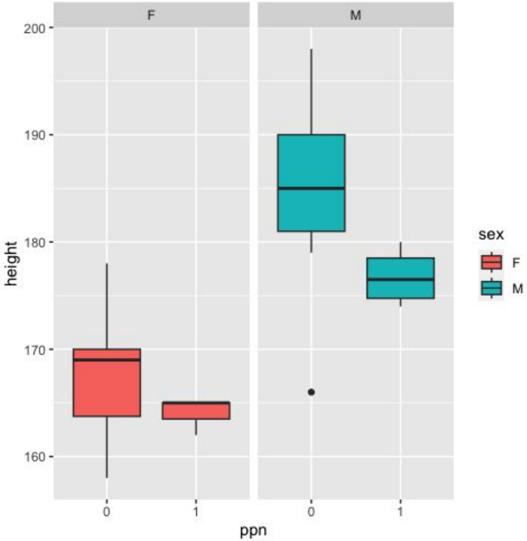
Neither *sex* nor *height* turn out as significant predictors of *ppn*

Plot on the left shows the 3-way distribution of data

Plot on the right shows the same data after changing *one male* observation to *no ppn*

Now height appears as protective factor for males and there is even an impression of the same tendency for females as their median height is also smaller for *ppn* group


F
M
height
ppn
sex
F
M


F
M
height
ppn
sex
F
M