Stat355, Final Project

Daniel Belkin

2021-04-10

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
setwd("C:/Users/tae0933/Desktop")
Heart <- read.csv('Heart.csv')
head(Heart)
```

	ï..age <int>	sex <int>	cp <int>	trestbps <int>	chol <int>	fbs <int>	restecg <int>	thalach <int>	exang <int>
1	63	1	3	145	233	1	0	150	0
2	37	1	2	130	250	0	1	187	0
3	41	0	1	130	204	0	0	172	0
4	56	1	1	120	236	0	1	178	0
5	57	0	0	120	354	0	1	163	1
6	57	1	0	140	192	0	1	148	0

6 rows | 1-10 of 15 columns

```
sum(is.na(Heart))
```

```
## [1] 0
```

- comment on distribution, outliers

```
Heart <- within(Heart,
     {SEXf <- factor(sex, levels=c(0,1), labels=c('female','male'))
     CPf <- factor(cp, levels=c(0,1,2,3), labels=c('asymptomatic','atypical angina','pain - no angina','typical angine'))
     FBSf <- factor(fbs, levels=c(0,1), labels=c('sugar <120mg','sugar>120mg'))
     RESTECGf <- factor(restecg, levels=c(0,1,2), labels=c('left ventricular hypertrophy','normal','abnormal'))
     EXANGf <- factor(exang, levels=c(0,1), labels=c('no angina during exercise','yes angina during exercise'))
     SLOPEf <- factor(slope, levels=c(0,1,2), labels=c('descending','flat','ascending'))
     CAf <- factor(ca)
     thalf <- factor(thal, levels=c(0,1,2,3), labels=c('null','fixed defect','normal','reversible defect'))
     Disease<-factor(target, levels = c(0,1),labels = c("yes","no"))
  })


summary(Heart)
```

```
##      ï..age           sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##       chol            fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak         slope            ca
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##       thal           target       Disease                thalf     CAf
##  Min.   :0.000   Min.   :0.0000   yes:138   null             :  2   0:175
##  1st Qu.:2.000   1st Qu.:0.0000   no :165   fixed defect     : 18   1: 65
##  Median :2.000   Median :1.0000             normal           :166   2: 38
##  Mean   :2.314   Mean   :0.5446             reversible defect:117   3: 20
##  3rd Qu.:3.000   3rd Qu.:1.0000                                     4:  5
##  Max.   :3.000   Max.   :1.0000
##         SLOPEf                          EXANGf
##  descending: 21   no angina during exercise :204
##  flat      :140   yes angina during exercise: 99
##  ascending :142
##
##
##
##                          RESTECGf             FBSf                   CPf
##  left ventricular hypertrophy:147   sugar <120mg:258   asymptomatic    :143
##  normal                       :152   sugar>120mg : 45   atypical angina : 50
##  abnormal                     :  4                      pain - no angina: 87
##                                                         typical angine  : 23
##
##
##      SEXf
##  female: 96
##  male  :207
##
##
##
##
```

age distributed normally (mean 54.37, range 29-77) sex: mostly males (207 compared to 96 females) cp: almost half are asymptomatic trestbps: normally distributed (mean 131.6, range 94-200) chol:slight negative skew (mean 246.3, range 126-564) fbs: mostly normal sugar level <120 mg (258, compared with 45 >120mg) RESTECGF: only 4 abnormal subject in the study thalach: positive skew, mean lower than median , range : 71-202 exang: mostly no angina during exercise 204, compared with 99 yes angina during exercise oldpeak: negative skew, mean 1.04, range 0- 6.2 slope: only 21 subjects with descending slope ca: mostly 0 narrow blood vessels (175), 5 null values thalf: only 18 subject with a fixed defect

Logistic regression using only one categorical variable

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

```
chisq.test(Heart$target,Heart$sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Heart$target and Heart$sex
## X-squared = 22.717, df = 1, p-value = 1.877e-06
```

```
sqrt(chisq.test(Heart$target,Heart$sex)$statistic /303)
```

```
## X-squared
## 0.2738144
```

```
chisq.test(Heart$target,Heart$cp)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Heart$target and Heart$cp
## X-squared = 81.686, df = 3, p-value < 2.2e-16
```

```
sqrt(chisq.test(Heart$target,Heart$cp)$statistic /303)
```

```
## X-squared
## 0.5192227
```

```
chisq.test(Heart$target,Heart$fbs)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Heart$target and Heart$fbs
## X-squared = 0.10627, df = 1, p-value = 0.7444
```

```
sqrt(chisq.test(Heart$target,Heart$fbs)$statistic /303)
```

```
##  X-squared
## 0.01872793
```

```
chisq.test(Heart$target,Heart$restecg)
```

```
## Warning in chisq.test(Heart$target, Heart$restecg): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Heart$target and Heart$restecg
## X-squared = 10.023, df = 2, p-value = 0.006661
```

```
sqrt(chisq.test(Heart$target,Heart$restecg)$statistic /303)
```

```
## Warning in chisq.test(Heart$target, Heart$restecg): Chi-squared approximation
## may be incorrect
```

```
## X-squared
## 0.1818777
```

```
chisq.test(Heart$target,Heart$exang)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Heart$target and Heart$exang
## X-squared = 55.945, df = 1, p-value = 7.454e-14
```

```
sqrt(chisq.test(Heart$target,Heart$exang)$statistic /303)
```

```
## X-squared
## 0.4296923
```

```
chisq.test(Heart$target,Heart$slope)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Heart$target and Heart$slope
## X-squared = 47.507, df = 2, p-value = 4.831e-11
```

```
sqrt(chisq.test(Heart$target,Heart$slope)$statistic /303)
```

```
## X-squared
## 0.3959652
```

```
chisq.test(Heart$target,Heart$ca)
```

```
## Warning in chisq.test(Heart$target, Heart$ca): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Heart$target and Heart$ca
## X-squared = 74.367, df = 4, p-value = 2.712e-15
```

```
sqrt(chisq.test(Heart$target,Heart$ca)$statistic /303)
```

```
## Warning in chisq.test(Heart$target, Heart$ca): Chi-squared approximation may be
## incorrect
```

```
## X-squared
## 0.4954134
```

```
chisq.test(Heart$target,Heart$thal)
```

```
## Warning in chisq.test(Heart$target, Heart$thal): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Heart$target and Heart$thal
## X-squared = 85.304, df = 3, p-value < 2.2e-16
```

```
sqrt(chisq.test(Heart$target,Heart$thal)$statistic /303)
```

```
## Warning in chisq.test(Heart$target, Heart$thal): Chi-squared approximation may
## be incorrect
```

```
## X-squared
## 0.5305945
```

```
summary(glm(target ~ thal, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ thal, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5462  -0.9079   0.9284   0.9284   1.4733
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.2016     0.5456   5.869 4.40e-09 ***
## thal         -1.2916     0.2249  -5.743 9.29e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 379.14  on 301  degrees of freedom
## AIC: 383.14
##
## Number of Fisher Scoring iterations: 4
```

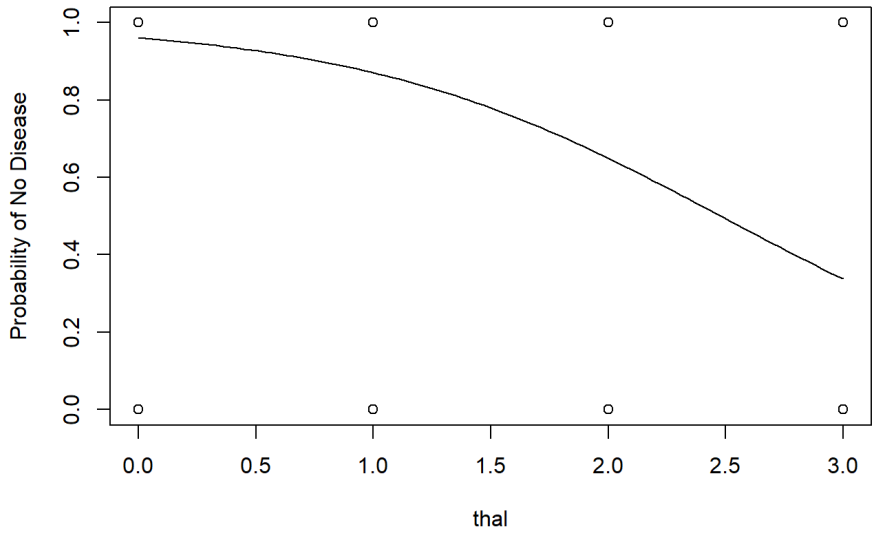
Best single statistically significant association between target and a categorical variables is for thal. with a p-value of 2.2e-16 which is statistically significant at the 99.9 confidence interval, and the correlation has a strength of 0.54.

When we use thal as our predictor variable we end up with the regression model p(y) = 3.2016-1.2916x

where p(y) is the probability of NOT developing heart disease, 3.2016 is the y-intercept, and -12916 is the slope.

It's important to note that the higher the probability the more likely the person is NOT to have heart disease as the target is codded as 1= no disease, and 0 = disease. This is reflected in the first graph as the probability of NOT developing disease drops as we move towards having a non-fixed reversible defect.

```
plot(Heart$thal,Heart$target,xlab="thal",ylab="Probability of No Disease")
g=(glm(target ~ thal, data = Heart, family = binomial))

curve(predict(g,data.frame(thal=x),type ='resp'),add=TRUE) # draws a curve based on prediction from logistic regression mode
l
```




```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

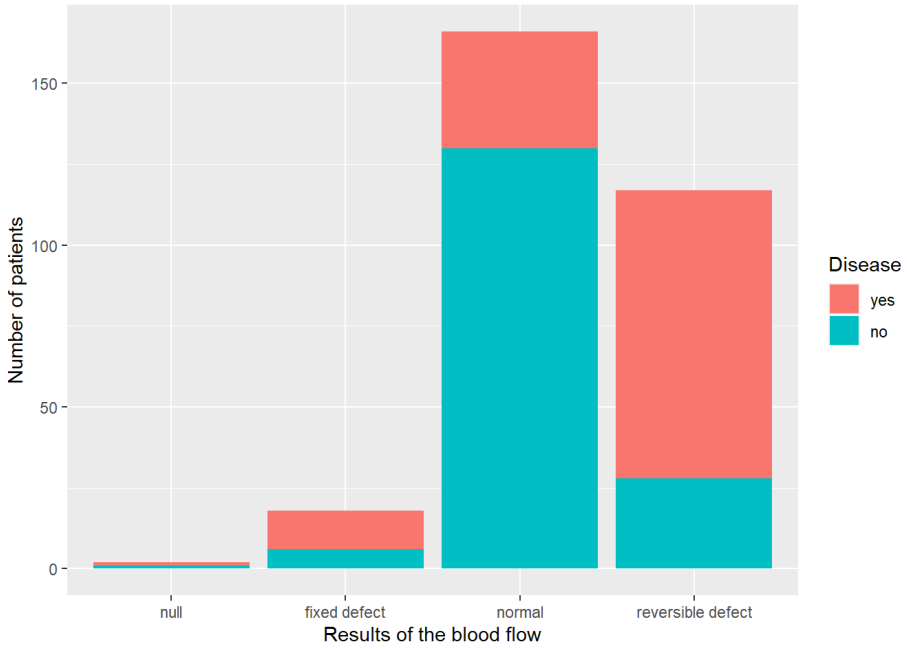
```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.4
```

```
## Warning: package 'forcats' was built under R version 4.0.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
ggplot(Heart, aes(thalf, fill= Disease)) +
  geom_bar() +
  labs( x="Results of the blood flow", y="Number of patients")
```



Logistic regression using multiple categorical variables

First using all the categorical variables available and then using only those that are statistically significant and finally a model that takes into account the interaction between thal and ca.

We see that the model using all categorical variables has 2 variables that are not statistically significant for our model. When we drop these two variables our model contains only significant variables.

The final model with the interaction between thal and ca is:

p(y) = 2.5881-1.3045(thal)-2.1776(ca)+1.3240(slope)-1.2897(exang)+0.8574(cp)-1.4253(sex)+0.5665(that)(ca)

This is the best model that only uses categorical variables as it contains all statistically significant variables as well as account for interaction between factors.

(Intercept) 2.5881 0.8877 2.915 0.003551 ** thal -1.3045 0.3426 -3.808 0.000140 ***ca -2.1776 0.7321 -2.975 0.002934*** *slope 1.3240 0.2910 4.550 5.37e-06* **exang -1.2897 0.3739 -3.450 0.000561** *cp 0.8574 0.1730 4.955 7.23e-07* **sex -1.4253 0.3998 -3.565 0.000364** * thal:ca 0.5665 0.2991 1.894 0.058207 .

```
summary(glm(target ~ thal+ca+slope+exang+cp+restecg+fbs+sex, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + restecg +
##     fbs + sex, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5758  -0.5409   0.2025   0.5965   2.6685
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.61280    0.79137   2.038 0.041550 *
## thal        -0.92557    0.26775  -3.457 0.000547 ***
## ca          -0.87546    0.17636  -4.964 6.91e-07 ***
## slope        1.23564    0.28266   4.371 1.23e-05 ***
## exang       -1.33081    0.37231  -3.574 0.000351 ***
## cp           0.82219    0.17172   4.788 1.68e-06 ***
## restecg      0.49516    0.31572   1.568 0.116796
## fbs          0.03596    0.49056   0.073 0.941565
## sex         -1.41897    0.39569  -3.586 0.000336 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 233.77  on 294  degrees of freedom
## AIC: 251.77
##
## Number of Fisher Scoring iterations: 5
```

```
summary(glm(target ~ thal+ca+slope+exang+cp+sex, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + sex,
##     family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6690  -0.5073   0.1768   0.5761   2.7570
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8279     0.7669   2.384 0.017144 *
## thal         -0.9156     0.2623  -3.490 0.000483 ***
## ca           -0.8756     0.1755  -4.990 6.03e-07 ***
## slope         1.2614     0.2818   4.477 7.59e-06 ***
## exang        -1.2993     0.3682  -3.529 0.000417 ***
## cp            0.8156     0.1688   4.831 1.36e-06 ***
## sex          -1.4335     0.3939  -3.639 0.000273 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 236.24  on 296  degrees of freedom
## AIC: 250.24
##
## Number of Fisher Scoring iterations: 5
```

```
summary(glm(target ~ thal+ca+slope+exang+cp+sex+thal*ca, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + sex +
##     thal * ca, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7557  -0.5268   0.1608   0.6011   2.5435
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.5881     0.8877   2.915 0.003551 **
## thal         -1.3045     0.3426  -3.808 0.000140 ***
## ca           -2.1776     0.7321  -2.975 0.002934 **
## slope         1.3240     0.2910   4.550 5.37e-06 ***
## exang        -1.2897     0.3739  -3.450 0.000561 ***
## cp            0.8574     0.1730   4.955 7.23e-07 ***
## sex          -1.4253     0.3998  -3.565 0.000364 ***
## thal:ca       0.5665     0.2991   1.894 0.058207 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 232.56  on 295  degrees of freedom
## AIC: 248.56
##
## Number of Fisher Scoring iterations: 5
```

Logistic Regression using one continous variables- we see that the statistically significant variables for predicting heart disease are: age,thalach, and oldpeak - with the single best predictor being oldpeak (p-value of 4.09e-15 ***)

We then combine the three variables into a multiple logistic regression test taking into account the interaction between age and thalach.

The resulting model is : p(y) = 0.1716(thalach)+0.3627(age)-0.7531(oldpeak)-0.00249(thalach)(age)-24.31

```
summary(aov(Heart$target ~ Heart$ï..age))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Heart$ï..age   1   3.82   3.819   16.12 7.52e-05 ***
## Residuals    301  71.33   0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Heart$target ~ Heart$trestbps))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Heart$trestbps   1   1.58  1.5785   6.458 0.0115 *
## Residuals      301  73.57  0.2444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Heart$target ~ Heart$chol))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Heart$chol    1   0.55  0.5460   2.203  0.139
## Residuals   301  74.60  0.2478
```

```
summary(aov(Heart$target ~ Heart$thalach))
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## Heart$thalach   1  13.37  13.366   65.12 1.7e-14 ***
## Residuals     301  61.78   0.205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Heart$target ~ Heart$oldpeak))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Heart$oldpeak   1  13.94  13.940   68.55 4.09e-15 ***
## Residuals     301  61.21   0.203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(glm(target ~ oldpeak, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ oldpeak, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6737  -1.0186   0.7522   0.8656   2.4025
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1177     0.1810   6.177 6.55e-10 ***
## oldpeak      -0.9396     0.1386  -6.779 1.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 355.00  on 301  degrees of freedom
## AIC: 359
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(target ~ oldpeak+thalach+ï..age+thalach*ï..age, data = Heart, family = binomial))
```

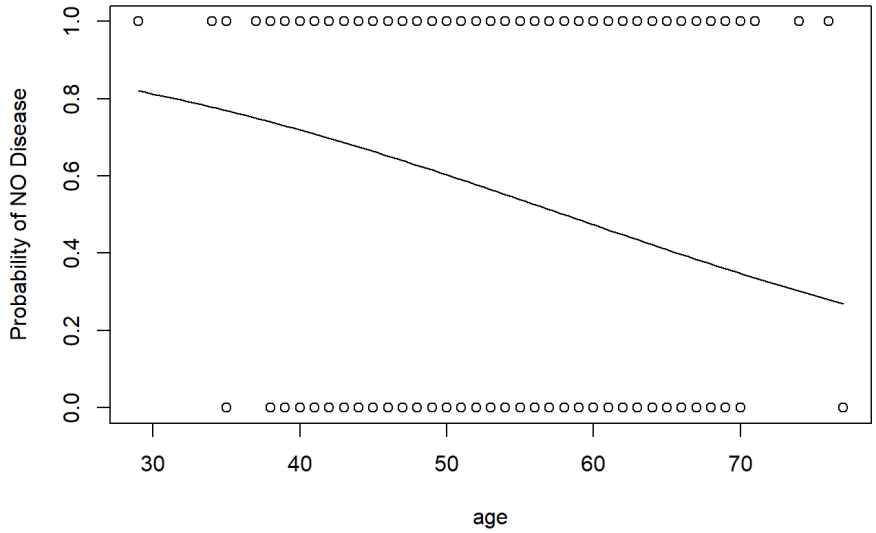
```
##
## Call:
## glm(formula = target ~ oldpeak + thalach + ï..age + thalach *
##     ï..age, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5531  -0.8206   0.3469   0.8801   2.2721
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.431e+01  7.028e+00  -3.460 0.000541 ***
## oldpeak        -7.531e-01  1.457e-01  -5.169 2.36e-07 ***
## thalach         1.716e-01  4.594e-02   3.735 0.000188 ***
## ï..age          3.627e-01  1.201e-01   3.021 0.002523 **
## thalach:ï..age -2.490e-03  7.925e-04  -3.142 0.001679 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 314.55  on 298  degrees of freedom
## AIC: 324.55
##
## Number of Fisher Scoring iterations: 5
```

```
summary(glm(target ~ oldpeak+thalach+ï..age+thalach*ï..age, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ oldpeak + thalach + ï..age + thalach *
##     ï..age, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5531  -0.8206   0.3469   0.8801   2.2721
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.431e+01  7.028e+00  -3.460 0.000541 ***
## oldpeak        -7.531e-01  1.457e-01  -5.169 2.36e-07 ***
## thalach         1.716e-01  4.594e-02   3.735 0.000188 ***
## ï..age          3.627e-01  1.201e-01   3.021 0.002523 **
## thalach:ï..age -2.490e-03  7.925e-04  -3.142 0.001679 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 314.55  on 298  degrees of freedom
## AIC: 324.55
##
## Number of Fisher Scoring iterations: 5
```

Probability of NO disease decreases as age increase

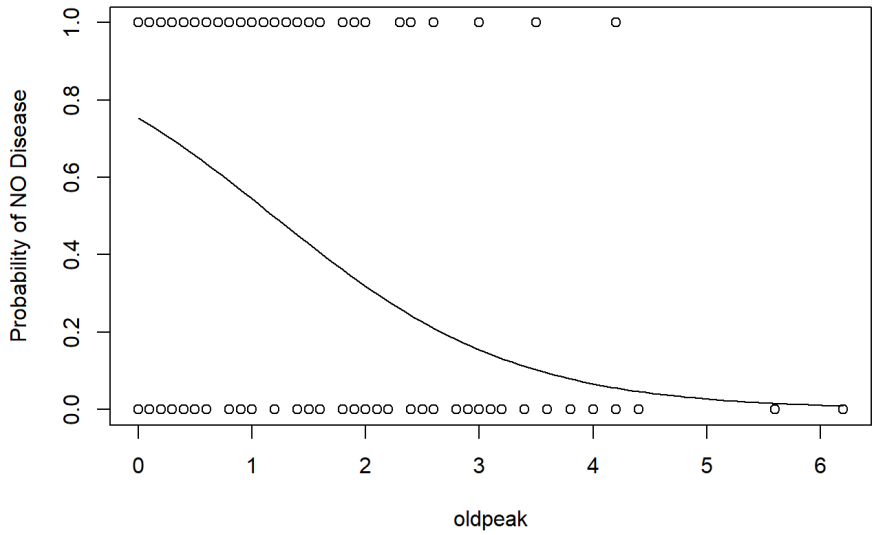
```
plot(Heart$ï..age,Heart$target,xlab="age",ylab="Probability of NO Disease")
g=(glm(target ~ ï..age, data = Heart, family = binomial))

curve(predict(g,data.frame(ï..age=x), type='resp'),add=TRUE) # draws a curve based on prediction from logistic regression mo
del
```



Probability of no disease drops as

oldpeak increase

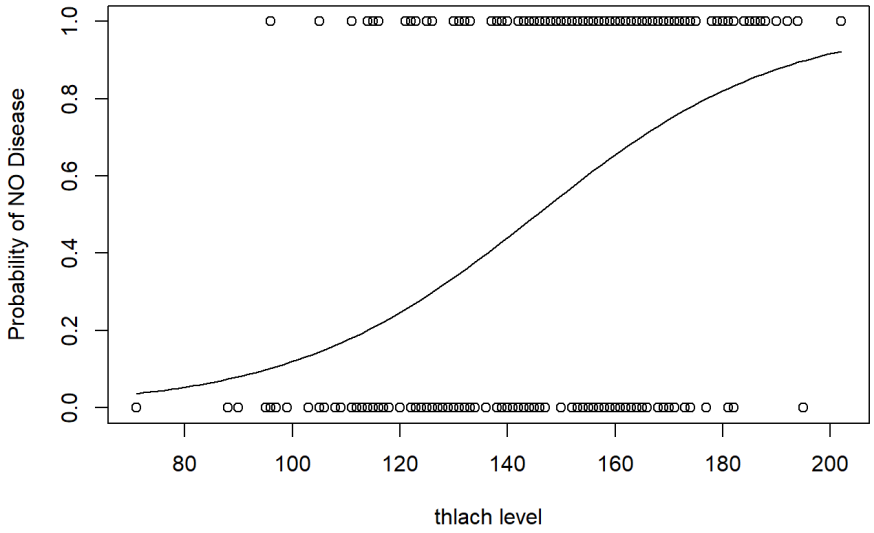
```
plot(Heart$oldpeak,Heart$target,xlab="oldpeak",ylab="Probability of NO Disease")
g=(glm(target ~ oldpeak, data = Heart, family = binomial)) # run a logistic regression model (in this case, generalized line
ar model with logit link). see ?glm
curve(predict(g,data.frame(oldpeak=x), type='resp'),add=TRUE) # draws a curve based on prediction from logistic regression m
odel
```



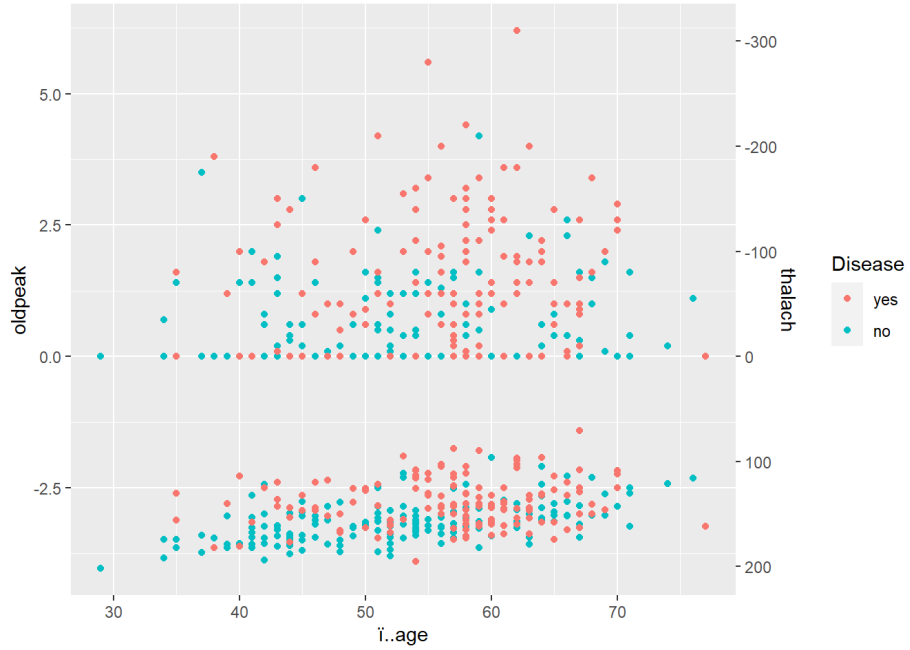
Probability of NO disease increases

as thalach level increase

```
plot(Heart$thalach,Heart$target,xlab="thlach level",ylab="Probability of NO Disease")
g=(glm(target ~ thalach, data = Heart, family = binomial))
curve(predict(g,data.frame(thalach=x), type='resp'),add=TRUE) # draws a curve based on prediction from logistic regression m
odel
```




```
ggplot(Heart, aes(x= ï..age,col= Disease)) +
  geom_point(aes(y=oldpeak)) +
  geom_point(aes(y=thalach/-50))+
  scale_y_continuous(sec.axis = sec_axis(~.*-50, name = "thalach"))+
  geom_abline()
```




```
  labs( x="age", y="oldpeak")
```

```
## $x
## [1] "age"
##
## $y
## [1] "oldpeak"
##
## attr(,"class")
## [1] "labels"
```

The above graph shows both oldpeak(y-axis on the left) and oldpeak (y axis on the right) in terms of age. The data points for oldpeak are on the top half of the graph, while the datapoints for thalach are on the bottom portion of the graph,

Finally we take a look at a combined model featuring all the categorical and continuous variables available. We then try a model which only uses the statstically signinficant variables we found earlier, we use this settling on a model that takes into account interaction between thal and ca

The resulting model is: 3.395429-
1.354833(thal)-2.371512(ca)+0.665827(slope)-0.910823(exang)+0.093924(cp)-1.621221(sex)-0.588705(oldpeak)+0.023139(thalach)-0.021921(trstbps)+0.69807

```
summary(glm(target ~ thal+ca+slope+exang+cp+restecg+fbs+sex+oldpeak+thalach+ï..age+trestbps+chol, data = Heart, family = bin
omial))
```

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + restecg +
##     fbs + sex + oldpeak + thalach + ï..age + trestbps + chol,
##     family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5849  -0.3872   0.1551   0.5863   2.6249
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.450472   2.571479   1.342 0.179653
## thal        -0.900432   0.290098  -3.104 0.001910 **
## ca          -0.773349   0.190885  -4.051 5.09e-05 ***
## slope        0.579288   0.349807   1.656 0.097717 .
## exang       -0.979981   0.409784  -2.391 0.016782 *
## cp           0.859851   0.185397   4.638 3.52e-06 ***
## restecg      0.466282   0.348269   1.339 0.180618
## fbs          0.034888   0.529465   0.066 0.947464
## sex         -1.758181   0.468774  -3.751 0.000176 ***
## oldpeak     -0.540274   0.213849  -2.526 0.011523 *
## thalach      0.023211   0.010460   2.219 0.026485 *
## ï..age      -0.004908   0.023175  -0.212 0.832266
## trestbps    -0.019477   0.010339  -1.884 0.059582 .
## chol        -0.004630   0.003782  -1.224 0.220873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 211.44  on 289  degrees of freedom
## AIC: 239.44
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(target ~ thal+ca+slope+exang+cp+sex+oldpeak+thalach+trestbps, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + sex +
##     oldpeak + thalach + trestbps, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5150  -0.3981   0.1670   0.5841   2.6249
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.423391   1.916185   1.265 0.205980
## thal        -0.916021   0.279482  -3.278 0.001047 **
## ca          -0.755279   0.183549  -4.115 3.87e-05 ***
## slope        0.604485   0.341428   1.770 0.076651 .
## exang       -0.947169   0.400644  -2.364 0.018073 *
## cp           0.854141   0.180253   4.739 2.15e-06 ***
## sex         -1.588807   0.433237  -3.667 0.000245 ***
## oldpeak     -0.531327   0.207396  -2.562 0.010410 *
## thalach      0.022843   0.009320   2.451 0.014251 *
## trestbps    -0.021043   0.009876  -2.131 0.033110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 215.77  on 293  degrees of freedom
## AIC: 235.77
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(target ~ thal+ca+slope+exang+cp+sex+oldpeak+thalach+trestbps+thal*ca, data = Heart, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + sex +
##     oldpeak + thalach + trestbps + thal * ca, family = binomial,
##     data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6293  -0.4109   0.1444   0.5500   2.5856
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.395429   1.998530   1.699 0.089326 .
## thal        -1.354833   0.359013  -3.774 0.000161 ***
## ca          -2.371512   0.802078  -2.957 0.003109 **
## slope        0.665827   0.345518   1.927 0.053975 .
## exang       -0.910823   0.410931  -2.216 0.026658 *
## cp           0.939324   0.190802   4.923 8.52e-07 ***
## sex         -1.621221   0.446693  -3.629 0.000284 ***
## oldpeak     -0.588705   0.211950  -2.778 0.005477 **
## thalach      0.023139   0.009659   2.396 0.016588 *
## trestbps    -0.021921   0.010132  -2.164 0.030499 *
## thal:ca      0.698071   0.326046   2.141 0.032272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 211.11  on 292  degrees of freedom
## AIC: 233.11
##
## Number of Fisher Scoring iterations: 6
```

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.