M11Guided

Alanna Hazlett

2024-04-14

We will focus on predicting the likelihood of developing coronary heart disease (chd) based on the following predictors: age, systolic blood pressure (sdp), diastolic blood pressure (dbp), number of cigarettes smoked per day (cigs), and behavior type (dibep)- A for aggressive and B for passive.

```
Data<-wcgs
set.seed(6021)
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
##training data frame
train<-Data[sample, ]
##test data frame
test<-Data[-sample, ]
```

Problem 1

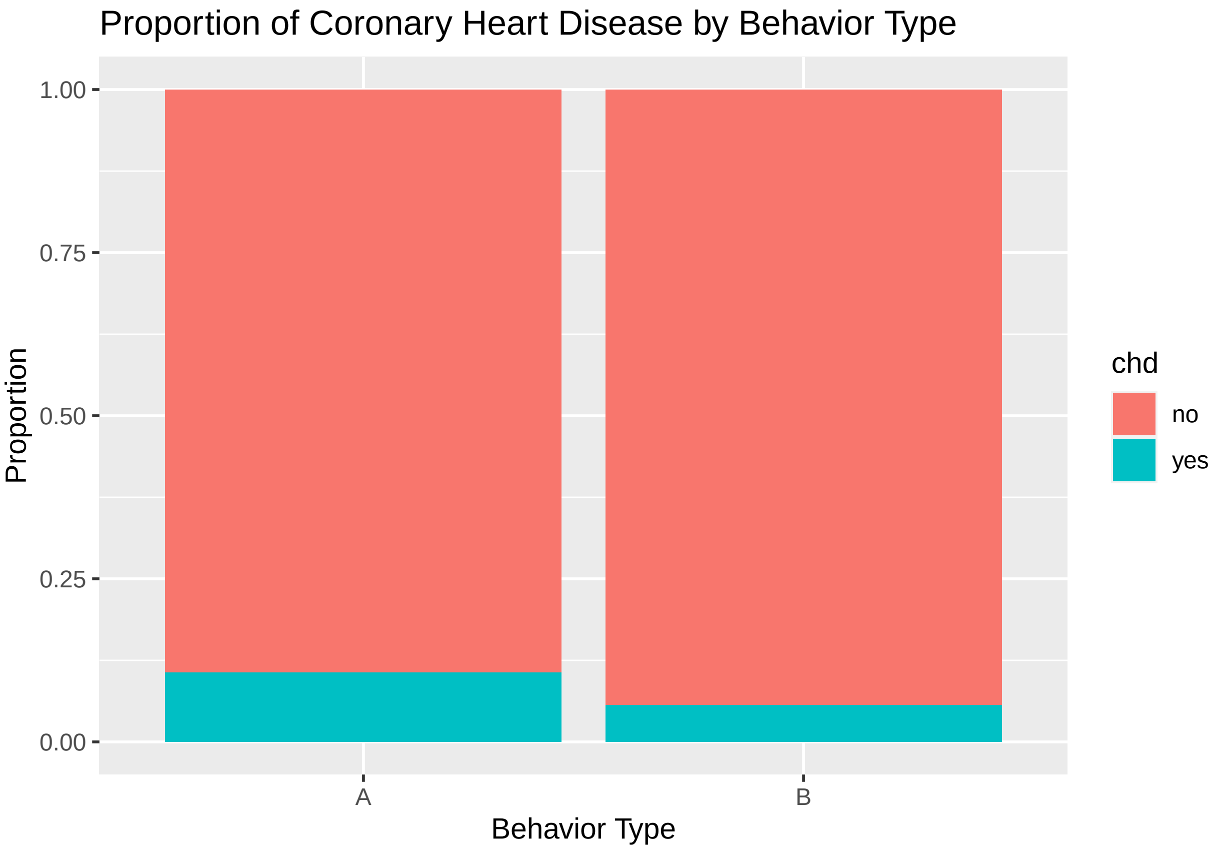
Before fitting a model, create some data visualizations to explore the relationship between these predictors and whether a middle-aged male develops coronary heart disease.

chd is the binary (categorical) response variable.
Categorical predictors: behavior type (dibep)

```
#check dummy coding for dibep
contrasts(train$dibep)
```

```
##   B
## A 0
## B 1
```

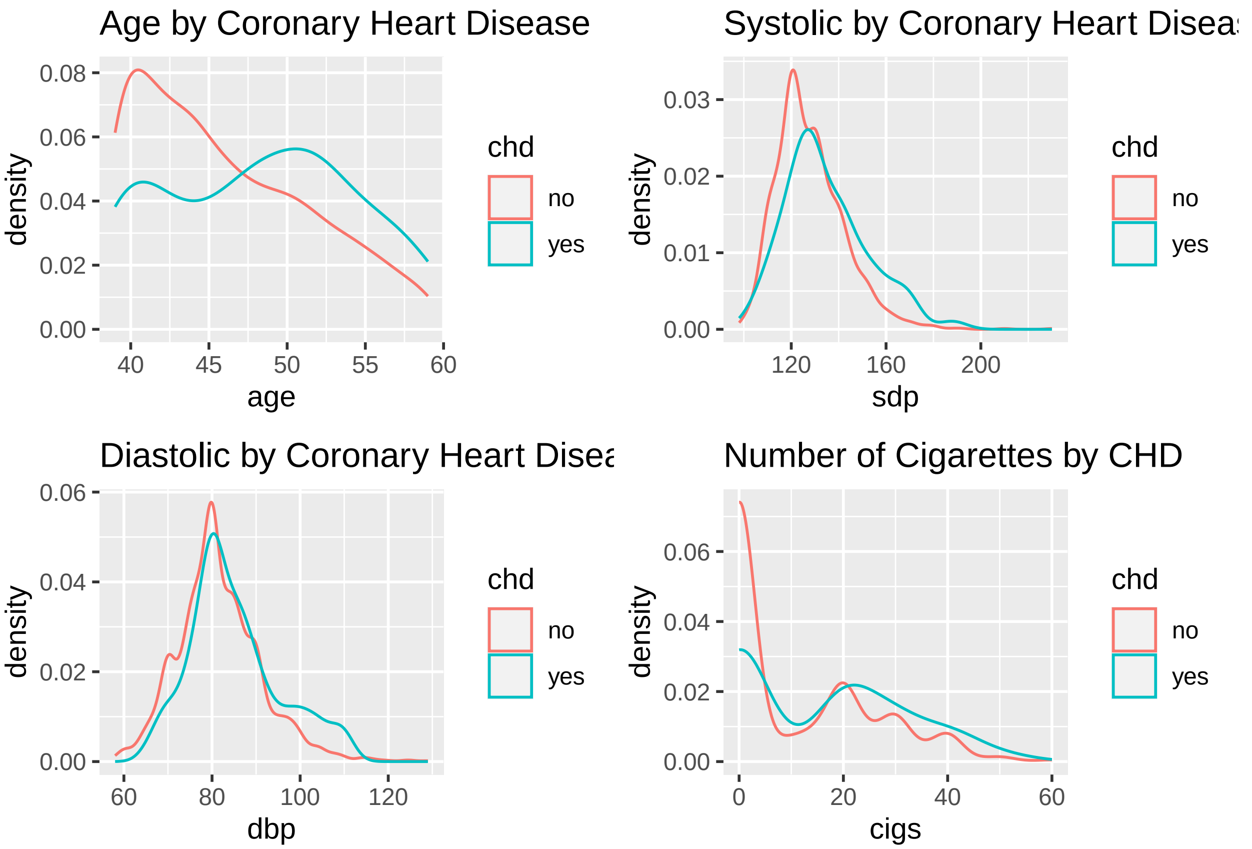
```
ggplot2::ggplot(train, aes(x=dibep, fill=chd))+
  geom_bar(position = "fill")+
  labs(x="Behavior Type", y="Proportion",
       title="Proportion of Coronary Heart Disease by Behavior Type")
```



Proportion of presence of coronary heart diseas is relatively equal between the aggressive and passive behavior types, wiht a slightly higher proportion in behavior type A.

Quatiative Predictors: age, systolic (sdp), diastolic (dbp), number of cigarettes per day (cigs).

```r
dp1<-ggplot2::ggplot(train,aes(x=age, color=chd))+
  geom_density()+
  labs(title="Age by Coronary Heart Disease")
dp2<-ggplot2::ggplot(train,aes(x=sdp, color=chd))+
  geom_density()+
  labs(title="Systolic by Coronary Heart Disease")
dp3<-ggplot2::ggplot(train,aes(x=dbp, color=chd))+
  geom_density()+
  labs(title="Diastolic by Coronary Heart Disease")
dp4<-ggplot2::ggplot(train,aes(x=cigs, color=chd))+
  geom_density()+
  labs(title="Number of Cigarettes by CHD")
gridExtra::grid.arrange(dp1, dp2, dp3, dp4, ncol = 2, nrow = 2)
```


Age by Coronary Heart Disease

Systolic by Coronary Heart Disea

Diastolic by Coronary Heart Disea

Number of Cigarettes by CHD

UL: After age 45 more people have coronary heart disease than not.
UR: Peak density for systolic for people with coronary heart disease is about 10mmHg than those without.
LL: Peak density for for Diastolic people with coronary heart disease is the same as those without, however a higher proportion of those with CHD are present in the range 95-113mmHg. LR: The proportion of of those who smoke cigarettes is higher for those with CHD across almost all number of cigarettes in a day.

Problem 2

Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.

```
result<-glm(chd~age+sdp+dbp+cigs+dibep, family=binomial, data=train)
summary(result)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = binomial,
##     data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.308765   1.080141  -7.692 1.45e-14 ***
## age          0.060212   0.016604   3.626 0.000287 ***
## sdp          0.015119   0.008805   1.717 0.085950 .
## dbp          0.012026   0.014345   0.838 0.401818
## cigs         0.021366   0.006095   3.506 0.000456 ***
## dibepB      -0.526914   0.198429  -2.655 0.007921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 837.55  on 1571  degrees of freedom
## AIC: 849.55
##
## Number of Fisher Scoring iterations: 5
```

Our regression mode is: $log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -8.309 + 0.060X_{age} + 0.015x_{sdp} + 0.012x_{dbp} + 0.021x_{cig} - 0.053I_{dibep}$
Where I is 0 for dibep behavior type A and 1 for behavior type B.

Problem 3

Interpret the estimated coefficient for cigs in context.
Estimated coefficient for cigs is 0.021.
Estimated log(odds) of having coronary heart disease increases by 0.021 for each additional cigarette smoked per day, when controlling for the other predictors.
Estimated odds is multiplied by e^0.021 = 1.021 for each additional cigarette smoked per day, when controlling the other predictors.

Problem 4

Interpret the estimated coefficient for dibep in context.
Estimated coefficient for dibep is -0.053. This indicates the change from behavior type A to behavior type B. The estimated log(odds) is 0.053 lower for behavior type B than behavior type A, when controlling the other predictors.
The estimated odds for behavior type B is e^0.053 = 1.054 times less the odds of type A, when controlling for the other predictors.

Problem 5

What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, does not smoke, and has type B personality? What is this person's corresponding probability of developing heart disease?
Dataset is comprised of all males

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{-8.309+0.060X_{age}+0.015x_{sdp}+0.012x_{dbp}+0.021x_{cig}-0.053I_{dibep}}$$

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{-8.309+0.060(45)+0.015(110)+0.012(70)+0.021(0)-0.053(1)}$$

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{-3.172} = 0.0419$$

Problem 6

Carry out the relevant hypothesis test to **check if this logistic regression model with five predictors is useful** in estimating the odds of heart disease. Clearly state the null and alternative hypotheses, test statistic, and conclusion in context.
Calling for us to use the Likelihood Ratio Test (LRT)

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_a : at\ least\ one\ coefficient \neq 0$$

Looking back at summary(result) we can find the null deviance, the deviance of the intercept only model, and the residual deviance, the deviance of the model. We utilize these to make our test statistic. We then compare this test statistic to a chi-squared distribution χ^2_{df} where df represents the number of parameters you will drop to get the reduced model.

$$\Delta G^2 = D(R) - D(F)$$

$$\Delta G^2 = 893.04 - 837.55 = 55.49$$

```
#p-value for chi squared dist.
1-pchisq(55.49,5)
```

```
## [1] 1.034908e-10
```

```
#critical value
qchisq(0.95,5)
```

```
## [1] 11.0705
```

Our p-value is very very small and our test statistic is larger than our critical value. We reject the null hypothesis, our data supports the alternative hypothesis. This indicates that our model is indeed useful.

Problem 7

Suppose a co-worker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all four predictors.

```
#Create reduced model
reduced<-glm(chd~age+cigs+dibep, family=binomial, data=train)
summary(reduced)
```

```
##
## Call:
## glm(formula = chd ~ age + cigs + dibep, family = binomial, data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.714614   0.804941  -7.099 1.25e-12 ***
## age          0.068802   0.016195   4.248 2.15e-05 ***
## cigs         0.021512   0.006013   3.578 0.000346 ***
## dibepB      -0.571514   0.196930  -2.902 0.003706 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 851.25  on 1573  degrees of freedom
## AIC: 859.25
##
## Number of Fisher Scoring iterations: 5
```

Compared the residual deviance from reduced and full models

$$\Delta G^2 = D(R) - D(F) = 851.25 - 837.55 = 13.7$$

```
#p-value for chi squared dist.
1-pchisq(13.7,2)
```

```
## [1] 0.001059456
```

```
#critical value
qchisq(0.95,2)
```

```
## [1] 5.991465
```

Our p-value is very small and our test statistic is larger than our critical value. We reject the null hypothesis, our data supports the alternative hypothesis. This indicates that we should go with the full model including the blood pressure predictors.

Problem 8

Based on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other predictors are already in the model?
$H_0 : \beta_3 = 0$ $H_a : \beta_3 \neq 0$
Z statistic:

$$Z = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} = \frac{0.012026 - 0}{0.014345} = 0.8383$$

This test statistic is compared to a normal distribution with mean 0 and variance 1. N(0,1)

```
#two tailed test
#p-value
2*(1-pnorm(abs(0.8383)))
```

```
## [1] 0.4018622
```

```
#critical value
```

We could have also just pulled the Z statistic and the p-value from summary(result) above
Our p-value is large enough that we fail to reject the null hypothesis that the coefficient for diastolic blood pressure is zero, our data supports the alternative hypothesis that we should not drop diastolic blood pressure from the model and that it is a significant predictor of heart disease.

Problem 9

Based on all the analysis performed, which of these predictors would you use in your logistic regression model? We should go with our original model in this question set, the full model. chd~age+sdp+dbp+cigs+dibep.