colorado school of public health

The effects of stress on coronary heart disease as measured by catecholamines

Froelicher J

May 10, 2021

Contents

DATA **2**

METHODS **3**

RESULTS **3**

PREDICTION **4**

APPENDIX **5**

DATA

The data are from a prospective cohort study of 609 white males in Evans County, Georgia who were followed for 7 years. We are interested in the association between coronary heart disease (CHD) and stress, as measured by catecholamine level in the blood (adrenal glands send catecholamines into your blood when you are physically or emotionally stressed). Other potential confounding variables were measured as well. The data dictionary is provided below.

- id: the subject identifier
- chd: the outcome, coronary heart disease status (0 = no chd, 1 = chd)
- cat: catecholamine level, the main exposure of interest (0 = normal, 1 = high)
- age: age in years
- chl: cholesterol level (mg/dL)
- smk: smoking status (0 = nonsmoker, 1 = smoker)
- ecg: ECG status (0 = normal, 1 = abnormal)
- dbp: diastolic blood pressure
- sbp: systolic blood pressure
- hpt: hypertension status (0 = normal, 1 = high blood pressure)

Provide in the table below are counts of categorical variables, and means and standard deviations of continuous variables

label	levels	No CHD	CHD
cat	Normal(%)	443 (82.3)	44 (62.0)
	High(%)	95 (17.7)	27 (38.0)
age	Mean (SD)	53.2 (9.0)	57.3 (10.1)
chl	Mean (SD)	210.4 (39.7)	221.9 (39.8)
smk	Non-smoker(%)	205 (38.1)	17 (23.9)
	smoker(%)	333 (61.9)	54 (76.1)
ecg	Normal(%)	401 (74.5)	42 (59.2)
	Abnormal(%)	137 (25.5)	29 (40.8)
dbp	Mean (SD)	90.6 (14.3)	95.7 (15.4)
sbp	Mean (SD)	144.2 (27.3)	154.9 (27.4)
hpt	Normal(%)	326 (60.6)	28 (39.4)
	High(%)	212 (39.4)	43 (60.6)

Table 1. Descriptive statistics for 609 white males from Evans county, Georgia, who were followed for 7 years.

As seen in Table 1 and and Figure 1 below, coronoray heart disease was not very common amongst the 609 individuals enrolled in the cohort study.

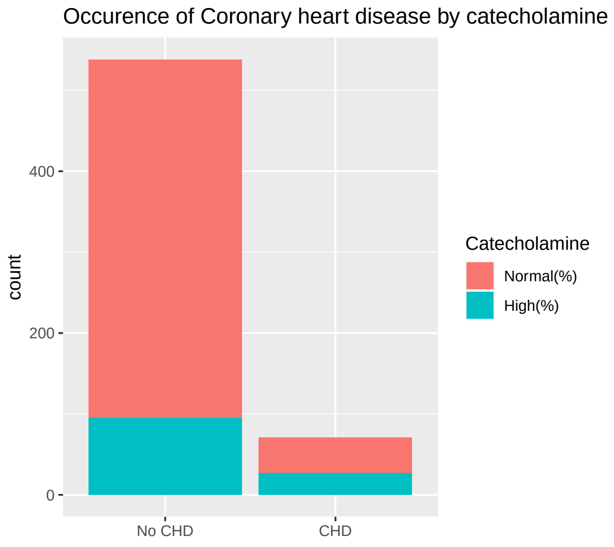


Figure 1. Bar plot of counts of Coronary heart disease for 609 individuals in Evans county, Georgia, colored by Catecholamine levels.

We can see in figure 1 that there may be a relationship between individuals without coronary heart diesase and with normal Catecholamine levels. We should note this as an interesting finding in the pursuit of the question of a relationship between Catecholamine and Coronary heart disease.

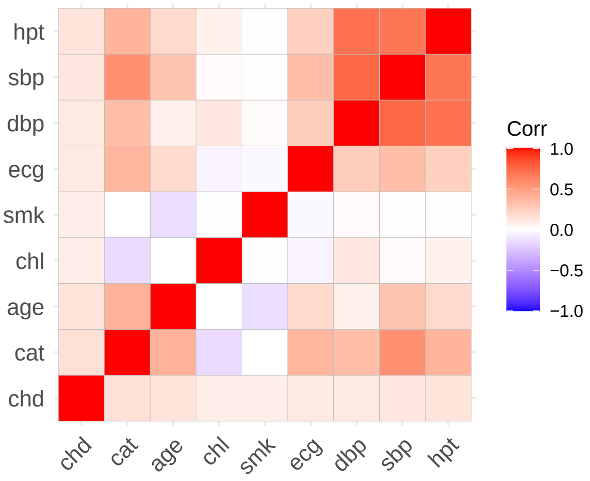


Figure 2. Correlations between covariates of potential interest collected on 609 individuals from Evans county, Georgia. See data dictionary above for descriptions.

From figure 2, note the large magnitude of the correlation between cholesterol and catecholamine levels (negative correlation), this should be identified as a possible interaction for our model.

METHODS

For the Evans county cohort, because the response and outcome variables are both binary, and the investigators have identified an interest in additional covariates, the method of choice selcted is a logistic regression. We are assuming the data are distributed as follows:

$$X_i \sim Bernoulli(p)$$

$$Y = \sum X_i$$

$$Y \sim binomial(n, p)$$

We also need to assume that the observations and errors are independent of eachother, no multicollinearity between out covariates of interest, and a linear relationship between the log odds (link function) and the covariates of interest. Additionally, there should be concern for sufficient sample size, in our case $n = 609$ is sufficient. The link function for logistic regression is as follows:

$$g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$$

$$\mu = E(Y)$$

The researchers have examined four models, a model containing only the outcome and response of interest (traditional model), a model containing all covariates (full model), and a model containing all statistically significant covariates of interest (selection model), and a model containing all statistically significant covariates of interest as well as cholesterol and an interaction term for Chatecholamine levels and cholesterol levels (interaction model). The selection model was chosen using backwards selection ($\alpha = 0.05$). Provided in table 2 are the Akaike information criterion, Bayesian criterion and R^2 values for each model.

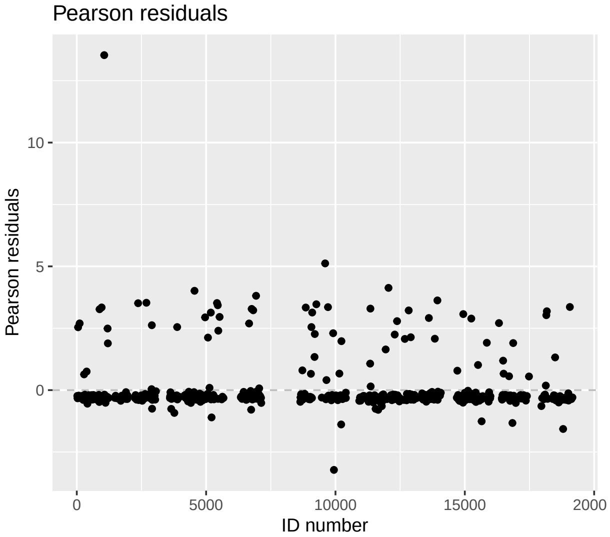
RESULTS

The results logistic regression of four different models is provided in the following section. As we would expect, the full model has the highest BIC, for it has the most parameters in the model.

Model	AIC	BIC	R^2
traditional model	428.4271	437.2507	0.023
full model	416.6001	456.3065	0.064
selection model	414.6566	436.7157	0.054
interaction model	374.9526	401.4235	0.117

Table 2. Model selection criterion for each of the four logistic regression models of the 609 white male patients from Evans county, Georgia.

The model with the lowest AIC, BIC and highest R^2 is the interaction model. This is the chosen model, and provided below in figure 3 are the Pearson residuals of the interaction model. There are a couple of possible outliers of concern in the residual plot, however not enough to warrant further exploration.



Figue 3. Pearson residuals of logistic regression model of 609 white males from Evans county, Georgia. Model includes Catecholamine levels as the response variable, as well as smoking status, cholesterol levels, age and the interaction of Catecholamine levels and cholesterol levels.

	Estimate	SE	Z	P-value
Intercept	-3.914	1.230	-3.181	0.001
Cat	-13.449	3.096	-4.344	0.000
Smk	0.774	0.322	2.406	0.016
Age	0.033	0.016	2.079	0.038
Chl	-0.003	0.004	-0.799	0.424
Inter	0.067	0.014	4.735	0.000

Table 3. Results from the logistic regression of 609 white males from Evans county georgia, including coefficient estimates, standard errors, Z-scores and P-values.

Because the intercept term in this scenario represents an individual with an age of zero, and cholesterol of zero, we will not be interpreting the intercept. Based on the results of the model, we know that high levels of catecholamine was significant in the prediction of coronary heart disease (p = 0.001), however it did not represent a clinically meaningful increase in the odds of coronary heart disease. The odds of coronary heart disease increased by a factor of 2.168 ± 1.88 (p = 0.016) for men who smoke. For a one year increase in age, the odds of coronary heart disease for men increased by a factor of 1.034 ± 1.88 (p = 0.038). Cholesterol was not a significant predictor of coronary heart disease (p = 0.424), however, for those men who had high levels of catecholamine, a unit increase of cholesterol indicates a 1.069 ± 1.88 ($p < 0.001$) times increase in the odds of coronary heart diesease ocurring.

PREDICTION

The investigative team performed a secondary analysis, in which we decided to assess the predictive capacity of our model. Using a threshold of 0.5 on our predicted probabilites of our interaction model, the confusion matrix is as follows:

	Ref. Pos.	Ref. Neg.
Pred. Pos.	532	56
Pred. Neg.	6	15

Table 4. Confusion matrix for the logistic regression interaction model chose for the Evans county coronary heart disease data.

The positive predictive value using a threshold of 0.5 was 0.905, and the negative predictive value was 0.714. In other words, 28.6% of negative predictions were false negatives, and 9.5% of positive predictions were false positives. The research team also examined the reveiver operator characteristic curve for the predictive values vs. the measured values, as a measure of predictive capability of the model. The results of the ROC curve is provided in figure 4.

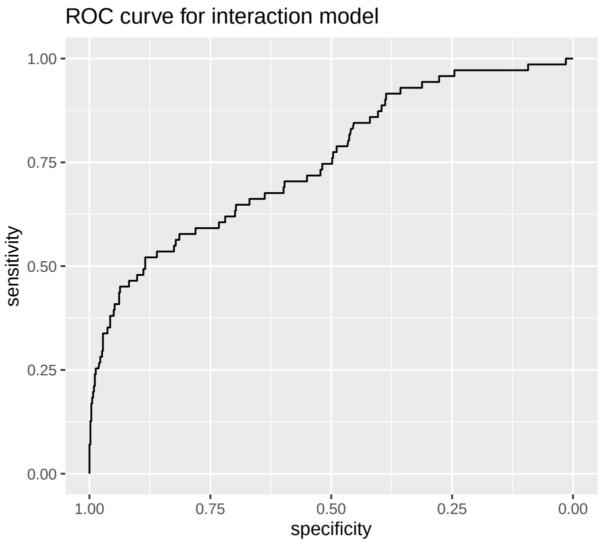


Figure 4. ROC curve for the logistic regression interaction model for the Evans county, Georgia coronary heart disease data.

The area under the ROC curve is 0.754. The area under the curve for the predictive capability of the interaction model is relatively high, we would at lease consider this model as a predictive test for coronary heart disease. This value is simply the probability that a randomly chosen instance of coronary heart disease ranks higher that a randomly chosen instance of not having coronary heart disease. So 75.4% of randomly chosen instances of coronary heart disease rank higher than randomly chosen instances of not having coronary heart disease.

APPENDIX

```r
# libraries
library(caret)
library(finalfit)
library(ggcorrplot)
library(ggplot2)
library(kableExtra)
library(pROC)
library(tableone)
library(tidyverse)

knitr::opts_chunk$set(echo = TRUE)

# data
data = read.csv("evanscounty.csv")

# create table one
data_table1 = data

data_table1$chd <- as.factor(data_table1$chd)
levels(data_table1$chd) = c(
  `0` = "No CHD" ,
  `1` = "CHD"
)

data_table1$cat <- as.factor(data_table1$cat)
levels(data_table1$cat) = c(
  `0` = "Normal(%)",
  `1` = "High(%)"
)

data_table1$smk <- as.factor(data_table1$smk)
levels(data_table1$smk) = c(
  `0` = "Non-smoker(%)" ,
  `1` = "smoker(%)"
)

data_table1$ecg <- as.factor(data_table1$ecg)
levels(data_table1$ecg) = c(
  `0` = "Normal(%)" ,
  `1` = "Abnormal(%)"
)

data_table1$hpt <- as.factor(data_table1$hpt)
levels(data_table1$hpt) = c(
  `0` = "Normal(%)" ,
  `1` = "High(%)"
)

names = dput(names(data_table1))
catvars = names[c(3, 7, 10)]
names = names[c(-1, -2)]
```

```r
table1 = data_table1 %>% summary_factorlist('chd', names, p = FALSE, na_include = TRUE)

kable(as.matrix(table1), booktabs = T)

# figure 1

ggplot(data_table1, aes(x = chd, fill = cat)) +
  geom_bar(position = "stack") +
  ggtitle("Occurence of Coronary heart disease by catecholamine level") +
  theme(axis.title.x = element_blank()) +
  labs(fill = "Catecholamine")

# figure 2
corr = cor(data[-1])
ggcorrplot(corr)

# model with just reponse and outcomes
model0 = glm(chd ~ factor(cat), family = binomial, data = data)
output0 = summary(model0)

# model with all covariates
model_full = glm(chd ~ factor(cat) + factor(smk) + factor(ecg) + factor(hpt) + age + chl + dbp + sbp, fa
output_full = summary(model_full)

# model with insignificant covariates removed'
model1 = glm(chd ~ factor(cat) + factor(smk) + age + chl, family = binomial, data = data)
output1 = summary(model1)

# interaction model
model_int = glm(chd ~ factor(cat) + factor(smk) + age + chl + factor(cat) * chl, family = binomial, data
output_int = summary(model_int)

# residuals
# model 0
data %>%
  mutate(pearson_resid = resid(model0, type = "pearson")) %>%
  ggplot(aes(id, pearson_resid)) +
  geom_hline(yintercept = 0, linetype = 2, color = "gray") +
  geom_point() +
  labs(x = "ID number", y = "Pearson residuals", title = "Pearson residuals")

data %>%
  mutate(deviance_resid = resid(model0, type = "deviance")) %>%
  ggplot(aes(id, deviance_resid)) +
  geom_hline(yintercept = 0, linetype = 2, color = "gray") +
  geom_point() +
  labs(x = "ID number", y = "Deviance residuals", title = "Deviance residuals")

# model 1
data %>%
  mutate(pearson_resid = resid(model1, type = "pearson")) %>%
  ggplot(aes(id, pearson_resid)) +
  geom_hline(yintercept = 0, linetype = 2, color = "gray") +
  geom_point() +
  labs(x = "ID number", y = "Pearson residuals", title = "Pearson residuals")

data %>%
  mutate(deviance_resid = resid(model1, type = "deviance")) %>%
  ggplot(aes(id, deviance_resid)) +
  geom_hline(yintercept = 0, linetype = 2, color = "gray") +
  geom_point() +
  labs(x = "ID number", y = "Deviance residuals", title = "Deviance residuals")

# interaction model
figure3 = data %>%
  mutate(pearson_resid = resid(model_int, type = "pearson")) %>%
  ggplot(aes(id, pearson_resid)) +
  geom_hline(yintercept = 0, linetype = 2, color = "gray") +
  geom_point() +
  labs(x = "ID number", y = "Pearson residuals", title = "Pearson residuals")

data %>%
  mutate(deviance_resid = resid(model_int, type = "deviance")) %>%
  ggplot(aes(id, deviance_resid)) +
  geom_hline(yintercept = 0, linetype = 2, color = "gray") +
  geom_point() +
  labs(x = "ID number", y = "Deviance residuals", title = "Deviance residuals")

# define function to get predictive measures based on what model is input
get_predictive_measures = function(model_of_interest, model_name){
  r_squared = rcompanion::nagelkerke(model_of_interest)

  tibble(
    model = model_name,
    aic = AIC(model_of_interest),
    bic = BIC(model_of_interest),
    gen_Rsq = round(r_squared$Pseudo.R[2], 3),
  )
}

measures1 = get_predictive_measures(model0, "traditional model")
measures2 = get_predictive_measures(model_full, "full model")
measures3 = get_predictive_measures(model1, "selection model")
measures4 = get_predictive_measures(model_int, "interaction model")
table2 = full_join(full_join(full_join(measures1, measures2), measures3), measures4)
colnames(table2) = c("Model", "AIC", "BIC", "R^2")

kable(table2, booktabs = T)

figure3

# table 3
table3 = round(output_int$coefficients, 3)
colnames(table3) = c("Estimate", "SE", "Z", "P-value")
rownames(table3) = c("Intercept", "Cat", "Smk", "Age", "Chl", "Inter")
kable(table3, booktabs = T)
```

```r
# beta coefficients
b2 = round(exp(table3[3, 1]), 3)
b3 = round(exp(table3[4, 1]), 3)
b5 = round(exp(table3[6, 1]), 3)

alpha = 0.05
z = qnorm(1 - (alpha / 2))

b2_ci = round(exp((z * table3[3, 2])), 3)
b4_ci = round(exp((z * table3[4, 2])), 3)
b5_ci = round(exp((z * table3[6, 2])), 3)

# confusion matrix
cmat = caret::confusionMatrix(
  data = factor(as.numeric(predict(model_int, type = "response") > 0.5)),
  reference = factor(as.numeric(model1$y))
)

table4 = cmat$table
colnames(table4) = c("Ref. Pos.", "Ref. Neg.")
rownames(table4) = c("Pred. Pos.", "Pred. Neg.")
kable(table4, booktabs = T)
pvp = round(table4[1, 1] / sum(table4[1, ]), 3)
pvn = round(table4[2, 2] / sum(table4[2, ]), 3)

# roc curve
model_roc = roc(model_int$y, model_int$fitted.values)
pROC::ggroc(model_roc) + ggtitle("ROC curve for interaction model")

auc = model_roc$auc
```