# Stat 184 Final Project

Yuming Sun

2022-12-15

## Introduction

The Western Collaborative Group Study (WCGS), a prospective cohort study, recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary Heart disease (CHD). A number of other risk factors were also measured to provide the best possible assessment of the CHD risk associated with behavior type. Additional variables collected include age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, smoking, and corneal arcus. The data set is available at https://rdrr.io/cran/epitools/man/wcgs.html.

The dataset includes the following columns (columns irrelevant are removed):

- chd69: Coronary heart disease event (0 = none; 1 = yes), dependent variable
- age: Age (age in years)
- arcus: Corneal arcus (0 = none; 1 = yes)
- chol: Cholesterol (mg/100 ml)
- dbp: Diastolic blood pressure (mm Hg)
- dibpat: Dichotomous behavior pattern
- height: Height (height in inches)
- ncigs: Smoking (Cigarettes/day)
- sbp: Systolic blood pressure (mm Hg)
- smoke: Smoking state
- weight: Weight (weight in pounds)

The main objective of this report is to find the important factors which affects the risk of Coronary heart disease event. Then, using these factors to construct a predictive model to predict the risk of suffering from CHD.

# Methods and Result

## Data cleaning

Table 1: Counts of NAs in each column

| age | arcus | chd69 | chol | dbp | dibpat | height | ncigs | sbp | smoke | weight |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

The data contains some missing values as the figure above shown. The variable `chol` contains 12 missing values and `arcus` contains 2 missing values. The total number of observations are 3154, which indicates that only a few observations contains missing values. Thus observations with missing values are removed.

## Descriptive statistic

Table 2: Descriptive Statistic

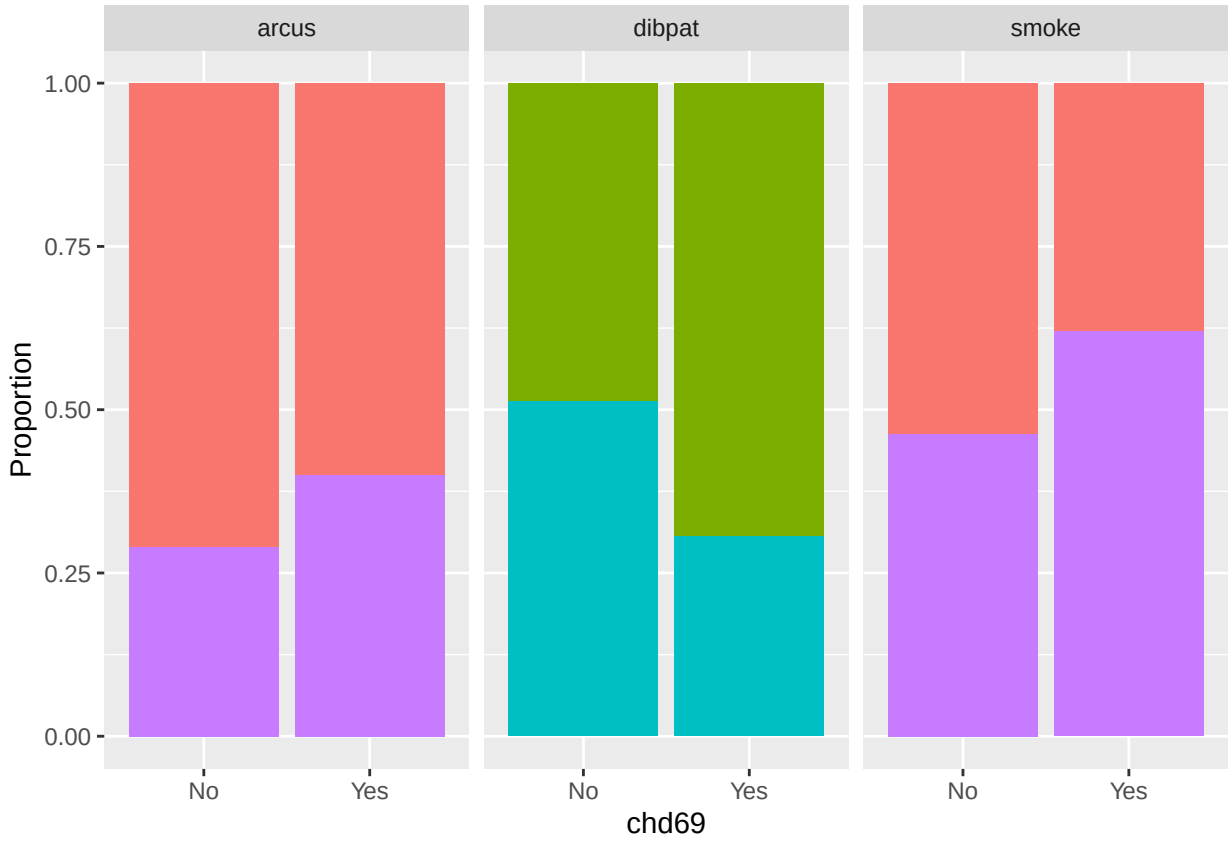
| | No | Yes | p | test |
|---|---|---|---|---|
| n | 2885 | 255 | | |
| age (mean (SD)) | 46.08 (5.45) | 48.51 (5.78) | <0.001 | |
| arcus = 1 (%) | 836 (29.0) | 102 ( 40.0) | <0.001 | |
| chd69 = Yes (%) | 0 ( 0.0) | 255 (100.0) | <0.001 | |
| dbp (mean (SD)) | 81.68 (9.52) | 85.25 (10.32) | <0.001 | |
| dibpat = Type B (%) | 1479 (51.3) | 78 ( 30.6) | <0.001 | |
| height (mean (SD)) | 69.77 (2.53) | 69.92 (2.41) | 0.362 | |
| ncigs (median [IQR]) | 0.00 [0.00, 20.00] | 20.00 [0.00, 30.00] | <0.001 | nonnorm |
| sbp (mean (SD)) | 128.01 (14.67) | 135.36 (17.54) | <0.001 | |
| smoke = Yes (%) | 1336 (46.3) | 158 ( 62.0) | <0.001 | |
| weight (mean (SD)) | 169.54 (20.99) | 174.49 (21.65) | <0.001 | |

In this section, the summarily descriptive statistics are constructed, in which the categorical variables such as `behpat`, `dibpat` and `smoke` are summarized as counts and proportion of each level, the approximately normally distributed numeric variables are summarized as mean and standard deviance. The ncigs is serious right skewed distributed, thus it is summarized as median and IQR. In the single variable statistic, to test whether these predictors affect the response variable, the $\chi^2$ test is performed for categorical predictors, t-test is performed for numerical predictor distributed normally and non-parameter test is performed for numerical predictors not normally distributed.

## Exploratory Analysis

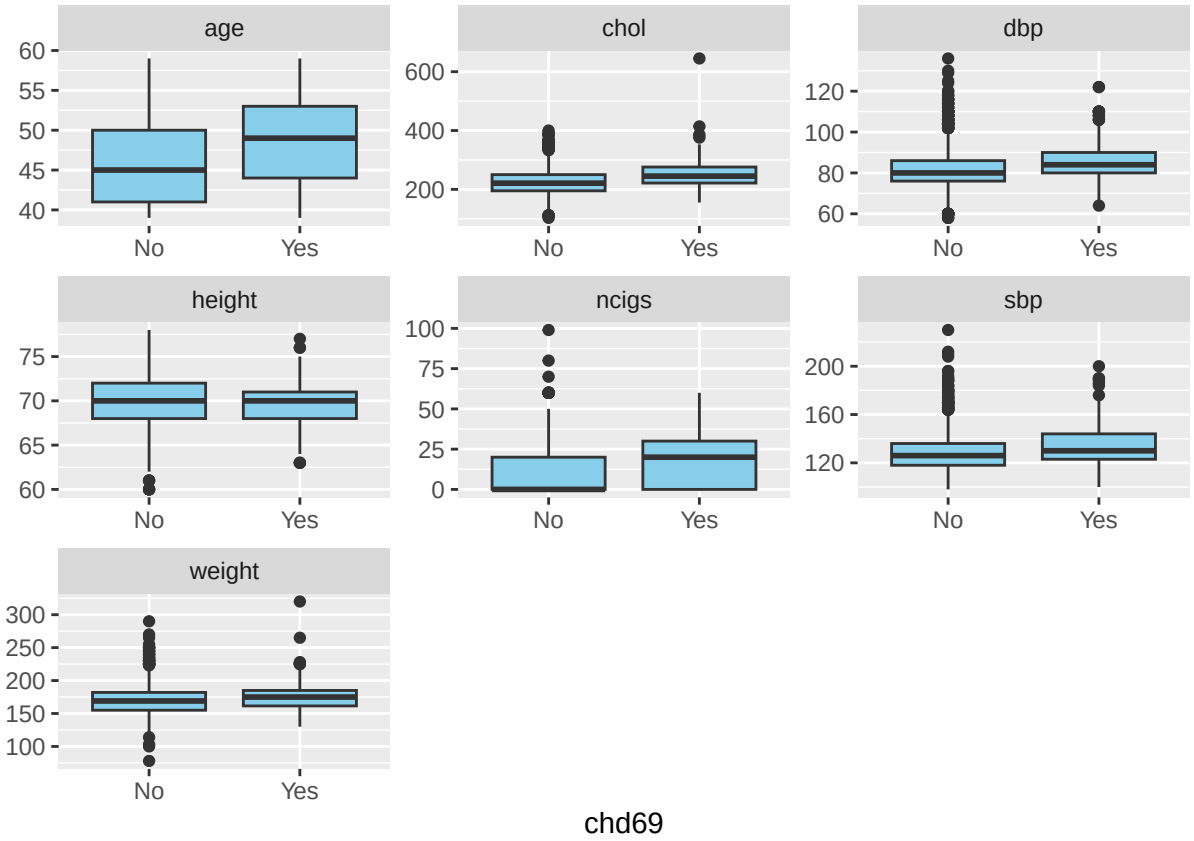
### Single Variable Analysis

According to the statistic table, proportion of Corneal arcus is different obviously between with and without CHD. Besides, behpat, dibpat and smoke can also affect the proportion of CHD according to the table above.Thus to show the relationship between these categorical variable and CHD, proportion bar plots are constructed as:




As a result, the proportion of each levels for these categorical are quite different between observations with CHD and without CHD. Thus these categorical may affect the risk of CHD.

Besides, numerical variables are also the potential risk factors affecting CHD, which are also explored and visualized as box plots.




As a result, the age of people with CDH is older than those without CDH, which indicates that the old people may suffer from higher risk for CHD. Similarly, according to the figure, the people with high cholesterol and blood pressure (both dpb and spb) are more likely suffer from CHD. Besides, Smoking is also a risk factor of CHD, which indicates that more cigarettes used, high probability of CHD is.

## Model Construction

### Full model

Due to the response variable is binary, logistic regression model is used. In this section, the full model is constructed by all of these predictors with significant effects on risk of CHD shown in Table2. In this way, all predictors are added into the full model except `height`.

Table 3: Result of full model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -11.856 | 0.988 | -12.004 | 0.000 |
| age | 0.061 | 0.012 | 4.967 | 0.000 |
| arcus | 0.209 | 0.144 | 1.456 | 0.145 |
| chol | 0.011 | 0.002 | 6.980 | 0.000 |
| dbp | -0.001 | 0.011 | -0.110 | 0.912 |
| dibpatType B | -0.663 | 0.146 | -4.537 | 0.000 |
| ncigs | 0.017 | 0.007 | 2.412 | 0.016 |
| sbp | 0.018 | 0.006 | 2.837 | 0.005 |
| smokeYes | 0.153 | 0.237 | 0.644 | 0.520 |

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| weight | 0.009 | 0.003 | 2.838 | 0.005 |

According to the result, the p-values of `arcus`, `dbp` and `smoke` are larger than 0.05, which indicates that these predictors are irrelevance about risk of CHD. Thus to select important predictors among all of these predictors significant in Table2, step-wise method is used with AIC criteria. **Reduced model**

The reduced model is constructed with step-wise method by `both` direction, in which predictors can both be removed and added to reach the optimalized model. The result is shown as:

Table 4: Comparison between full and reduced models

| | *Dependent variable:* | | |
|---|---|---|---|
| | chd69 | | |
| | (1) | (2) | (3) |
| age | 0.061*** | 0.061*** | 0.065*** |
| | (0.012) | (0.012) | (0.012) |
| arcus | 0.209 | 0.213 | |
| | (0.144) | (0.143) | |
| chol | 0.011*** | 0.011*** | 0.011*** |
| | (0.002) | (0.002) | (0.002) |
| dbp | −0.001 | | |
| | (0.011) | | |
| dibpatType B | −0.663*** | −0.658*** | −0.659*** |
| | (0.146) | (0.146) | (0.146) |
| ncigs | 0.017** | 0.021*** | 0.021*** |
| | (0.007) | (0.004) | (0.004) |
| sbp | 0.018*** | 0.018*** | 0.017*** |
| | (0.006) | (0.004) | (0.004) |
| smokeYes | 0.153 | | |
| | (0.237) | | |
| weight | 0.009*** | 0.009*** | 0.009*** |
| | (0.003) | (0.003) | (0.003) |
| Constant | −11.856*** | −11.813*** | −11.904*** |
| | (0.988) | (0.974) | (0.973) |
| Observations | 3,140 | 3,140 | 3,140 |
| Log Likelihood | −784.552 | −784.767 | −785.856 |
| Akaike Inf. Crit. | 1,589.105 | 1,585.533 | 1,585.712 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

As a result, the predictors `dbp` and `smoke` are removed, and the AIC is decreased from 1589.105 to 1585.533.

The deviance test is performed to compare the fitness of the two model. In the deviance test, the null hypothesis and alternative hypothesis can be considered as:

H0: The deviance of the two models are same

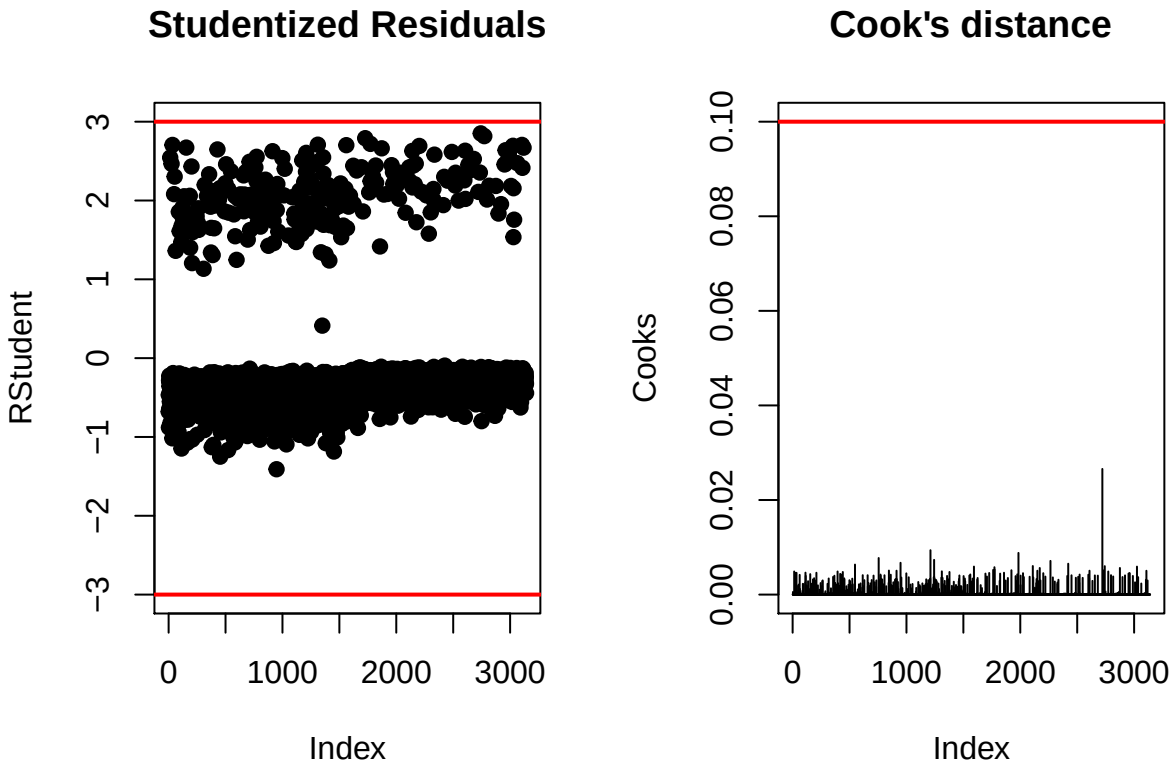
H1: The deviance of the two models are not same

According to the Table4, the $\chi^2$ statistic can be calculated as $\chi^2 = 2 * (-784.552 + 784.767) = 0.43$. Thus according to the $\chi^2$ distribution with df=2, the p-value can be calculated as 0.806. Thus the null hypothesis cannot be rejected, which indicates that the deviance of the two models are not different significantly. It reveals after removing the `dbp` and `smoke` would not decrease the fitness of the model.

Besides, according to the result, the `arcus` is also insignificant in the reduced model. Thus, the optimal reduced model is also constructed by remove predictor `arcus` furtherly. The result of optimal reduced model is also shown in Table4, in which p-values of all predictors are less than 0.05. The deviance test is also performed similarly in which $\chi^2 = 2 * (-784.552 + 785.856) = 2.608$ and $p - value = 0.456$. Thus the optimal reduced model is also as same as full model in fitness but the complexity of optimal reduced model is obviously lower than full model (3 irrelevance predictors are removed), by which this model is selected as best model finally.

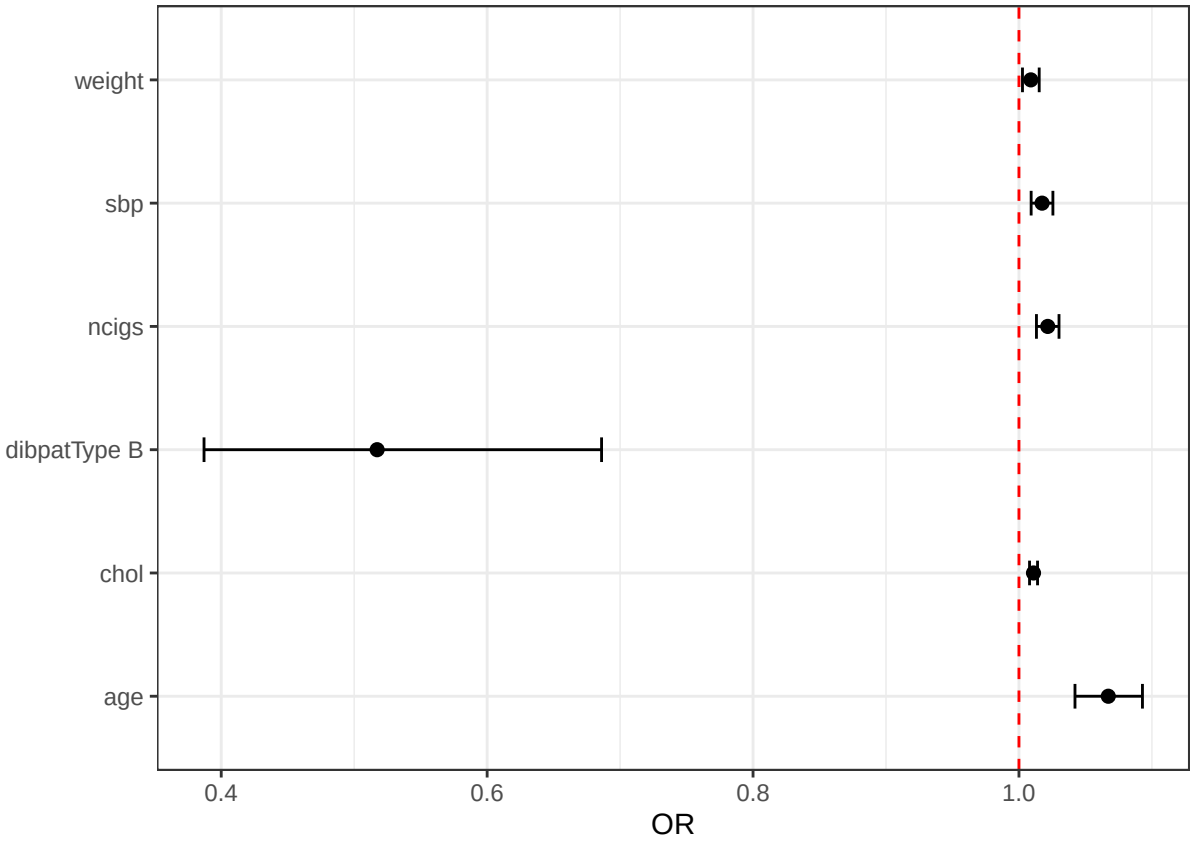
## Model Diagnose

In generalized models, the normality assumption is not required now. Thus the studentized residuals are used to check the fitness of the model and Cooks distances are used to detect the influential points. As a result, almost content of all studentized residuals are less than 3 and cook's distances of these observations are less than 0.1, which indicates that the model fitted well and no influential points existed.



### Model explaination

The OR values of these predictors and their 95% CIs are visualized as following:




As shown above, the Dichotomous behavior pattern as Type B is a protect factor for CHD of which the OR is 0.517 [0.387, 0.686]. It reveals that while the Dichotomous behavior pattern is Type B, the odds ratio would be decrease to 51.7% comparing with Type A. The other predictors are all risk factor due to the whole 95% CI are larger than 1. Among of these risk factor, the OR of age is 1.067 [1.042, 1.093], which indicates that while age increase by one, the odds ratio would increase to 1.067 times; the OR of `chol` is 1.010 [1.001, 1.014], which indicates that while cholesterol increases by one, the odds ratio would increase to 1.067 times; the OR of `ncigs` is 1.022 [1.013, 1.030], which indicates that while number of cigarettes per day increases by one, the odds ratio would increase to 1.022 times; the OR of `sbp` is 1.017 [1.009, 1.026], which indicates that while systolic blood pressure increases by one, the odds ratio would increase to 1.017 times; the OR of `weight` is 1.009 [1.003, 1.015], which indicates that while weight increases by one, the odds ratio would increase to 1.017 times.

## Conclusion

In this report, the objective is to find the factors affecting the risk of CHD and then construct a predictive model for CHD. As a result, the height is removed by single-variable analysis due to the height is not different significant between observations with and without CHD. Then all predictors are all added into the model as full model. To reduce the complexity of full model, reduced model is constructed by step-wise method and predictor `arcus` is also removed due to insignificance. Finally, the risk factors are revealed as weight, systolic blood pressure, number of cigarettes per day, cholesterol and age. The protect factor is Type B Dichotomous behavior. This suggest us that to avoid CHD, the Type B Dichotomous behavior should be

more prefered. The systolic blood pressure, weight and systolic blood pressure should be paid more attention and the smoking should be keep away.

## Codes appendix

```
library(tidyverse)
library(tableone)
library(stargazer)

dat<-read.csv('wcgs.csv')
dat<-dat%>%select(-id,-t1,-time169,-typchd69,-uni,-behpat)
knitr::kable(t(apply(is.na(dat),2,sum)),caption='Counts of NAs in each column')
dat<-dat%>%na.omit()

res<-CreateTableOne(colnames(dat)[-4],strata="chd69",data=dat,
                    factorVars = 'arcus')
kableone(res,caption='Descriptive Statistic',digits = 3,nonnormal='ncigs',
         align='c')

dat%>%select(arcus,dibpat,smoke,chd69)%>%
  mutate(arcus=ifelse(arcus==1,'Yes','No'))%>%
  pivot_longer(cols=!chd69)%>%
  ggplot(aes(chd69,fill=value))+geom_bar(position = 'fill')+
  labs(y='Proportion')+
  facet_wrap(vars(name))+
  guides(fill='none')

dat%>%select(-arcus,-dibpat,-smoke,chd69)%>%
  pivot_longer(cols=!chd69)%>%
  ggplot(aes(x=chd69,y=value))+geom_boxplot(fill='skyblue')+
  labs(y='')+
  facet_wrap(vars(name),scales = 'free')

dat$chd69<-ifelse(dat$chd69=='Yes',1,0)
full_model<-glm(chd69~.-height,data=dat,family=binomial(),
                control=list(maxit=100))
knitr::kable(summary(full_model)$coefficients,digits = 3,
             caption='Result of full model')

redu_model<-step(full_model,trace=0)

mod3<-glm(chd69 ~ age + chol + dibpat + ncigs + sbp + weight,
          family = binomial(), data = dat, control = list(maxit = 100))

stargazer(full_model,redu_model,mod3,
          title='Comparison between full and reduced models',
          header=F)

layout(t(1:2))
RStudent <- rstudent(model = mod3)
plot(RStudent,pch=19,ylim=range(RStudent,3,-3),
```

```
     main='Studentized Residuals')
abline(h = c(3,-3),col="red",lwd=2)

Cooks <- cooks.distance(model = mod3)
plot(Cooks,type="h",ylim=range(Cooks,0.1),main="Cook's distance")
abline(h = c(0.1),col="red",lwd=2)

ci<-exp(confint(mod3)[-1,])

ci%>%data.frame()%>%rename(lower=X2.5..,upper=X97.5..)%>%
  mutate(OR=exp(coef(mod3)[-1]),
         var=rownames(ci))%>%
  ggplot(aes(y=var,x=OR))+geom_point(size=2)+
  geom_errorbar(aes(xmin=lower,xmax=upper),width=0.2)+
  geom_vline(xintercept = 1,colour='red',linetype=2)+
  theme_bw()+labs(y='')
```