Unit 2 Live Session

Discrete Response Model Part 2

Figure 1: South Hall

Class Announcements

- HW 1 is due by end of today
- HW 2 is this week

Roadmap

Rearview Mirror

- Start with the simplest case of discrete response modeling, the Binomial probability model
- Discuss parameter estimation and statistical inference

Today

- Linear Probability Model and Binary Logistic Regression Model
- Estimate and make inferences about a Logistic Regression Model
- The notion of Deviance, Odds ratios, and probability of success

Looking Ahead

- Capture complex relationships by transforming data, including interaction terms and categorical exploratory variables.

Start-up Code

```
# Insert the function to *tidy up* the code when they are printed out
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

# Start with a clean R environment
rm(list = ls())

# Load libraries
## Load a set of packages inclusing: broom, cli, crayon, dbplyr , dplyr, dtplyr, forcats,
## googledrive, googlesheets4, ggplot2, haven, hms, httr, jsonlite, lubridate , magrittr,
## modelr, pillar, purrr, readr, readxl, reprex, rlang, rstudioapi, rvest, stringr, tibble,
## tidyr, xml2
library(tidyverse)

## provide useful functions to facilitate the application and interpretation of regression analysis.
library(car)

## provides  many functions useful for data analysis, high-level graphics, utility operations
library(Hmisc)

## to load SAheart dataset
library(bestglm)

## To assemble multiple plots
library(gridExtra)

## To generate regression results tables
library(finalfit)

## To produces LaTeX code, HTML/CSS code and ASCII text for well-formatted tables
library(stargazer)

# To make nicer y axes in ggplots
library(scales)
```

Case Study: South African Heart Disease

Introduction

High blood pressure, high LDL cholesterol, diabetes, smoking, secondhand smoke exposure, obesity, an unhealthy diet, and physical inactivity are among the leading risk factors for heart disease.

Nearly half of all Americans (47%) have at least one of three key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. (CDC)

Recall the three major modes of model building: prediction, description, and explanation.

- Here, our goal is the description:
  - **How are factors such as blood pressure, smoking, and cholesterol related to heart disease?**
- What are the requirements of explanatory modeling to have a casual interpretation?

Data Description

The data originates from a retrospective sample of men living in a heart disease high-risk region in the Western Cape, South Africa.

Install and load the bestglm library in order to use the SAheart dataset and understand the dataset structure.

Source: Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, South African Medical Journal 64: 430–436.

We summarize some of the variables that we will use:

- sbp: systolic blood pressure
- tobacco: cumulative tobacco use (kg)
- ldl: low density lipoprotein cholesterol ('bad' cholestrol)
- adiposity: Body adiposity index determines body fat percentage(calculated as (HC / (HM)1.5) - 18, where HC = Hip Circumference in Centimetres and HM = Height in meters)
- famhist: family history of heart disease
- typea: A personality type that could raise one's chances of developing coronary heart disease
- obesity: Body Mass Index (BMI) (kg/m^2)
- alcohol: current alcohol consumption
- age: age at onset
- chd: coronary heart disease

Exploratory Analysis

For this case study, we focus on blood pressure, smoking, cholesterol, and age.

- Load the data and answer the following questions:
  - What are the number of variables and number of observations?
  - What is the type of each variable? Do we need to change it?
  - Are there any missing values (in each of the variables)?
  - Are there any abnormal values in each of the variables in the raw data?

```
df <- SAheart %>%
   dplyr::select(tobacco, ldl, sbp, age, chd, obesity)

head(df)%>%
   knitr::kable()
```

tobacco	ldl	sbp	age	chd	obesity
12.00	5.73	160	52	1	25.30
0.01	4.41	144	63	1	28.87
0.08	3.48	118	46	0	29.14
7.50	6.41	170	58	1	31.99
13.60	3.50	134	49	1	25.99
6.20	6.47	132	45	0	30.77

```
#str(df)
#glimpse(df)
#summary(df)
#describe(df)
```

Univariate Analysis

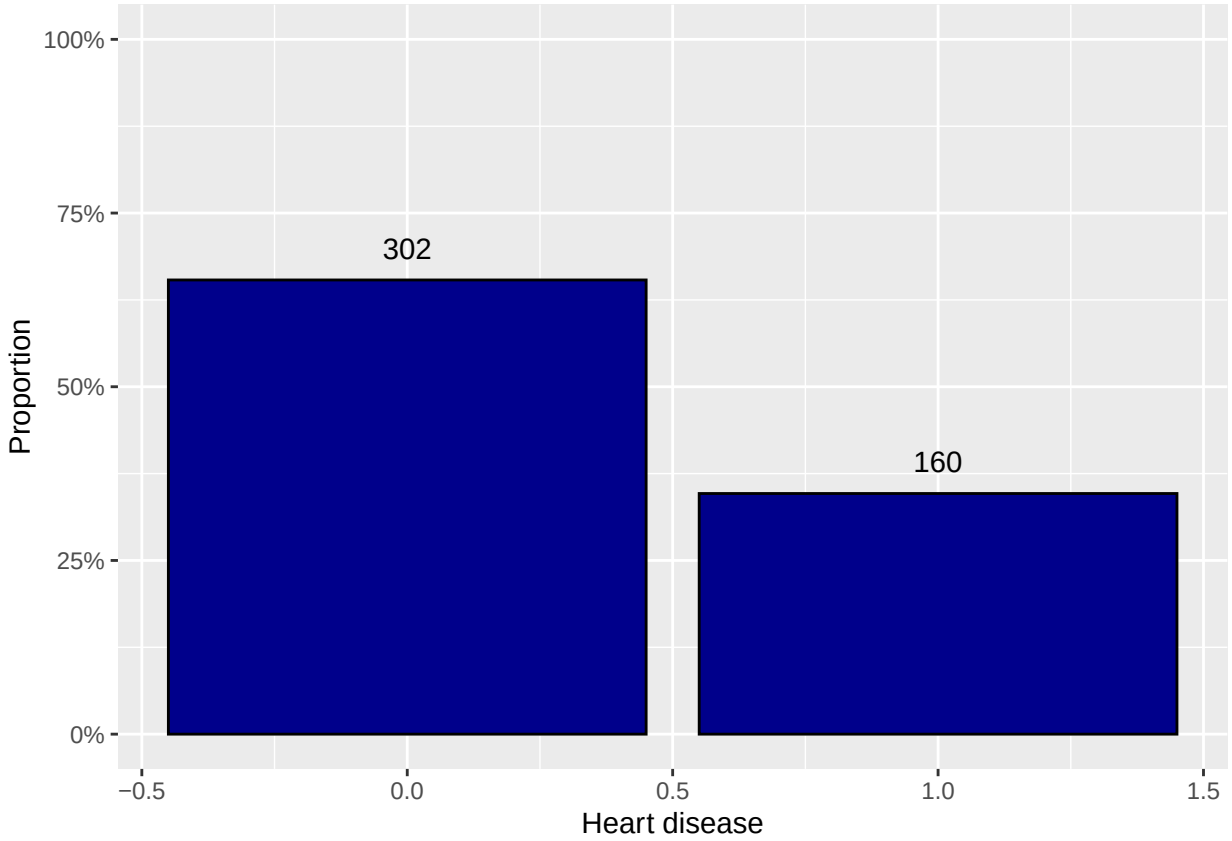
- The response (or dependent) variable of interest, Heart disease, is a binary variable taking the type factor.
- Use a bar chart to explore the distribution of the response variable (chd). What do you learn?

```
df %>%
  count(chd) %>%
  mutate(prop = round(prop.table(n),2)) %>%
  kable(col.names = c('Heart disease', 'N', "Proportion"))
```

Heart disease	N	Proportion
0	302	0.65
1	160	0.35

```
df %>%
  ggplot(aes(x= chd, y = ..prop.., group = 1)) +
  geom_bar(fill = 'DarkBlue', color = 'black') +
  geom_text(stat='count', aes(label=..count..), vjust=-1) +
  xlab("Heart disease") +
  ylab("Proportion") +
  scale_y_continuous(label=percent,limits=c(0,1))
```

```
## Warning: The dot-dot notation (‘..prop..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(prop)‘ instead.
```

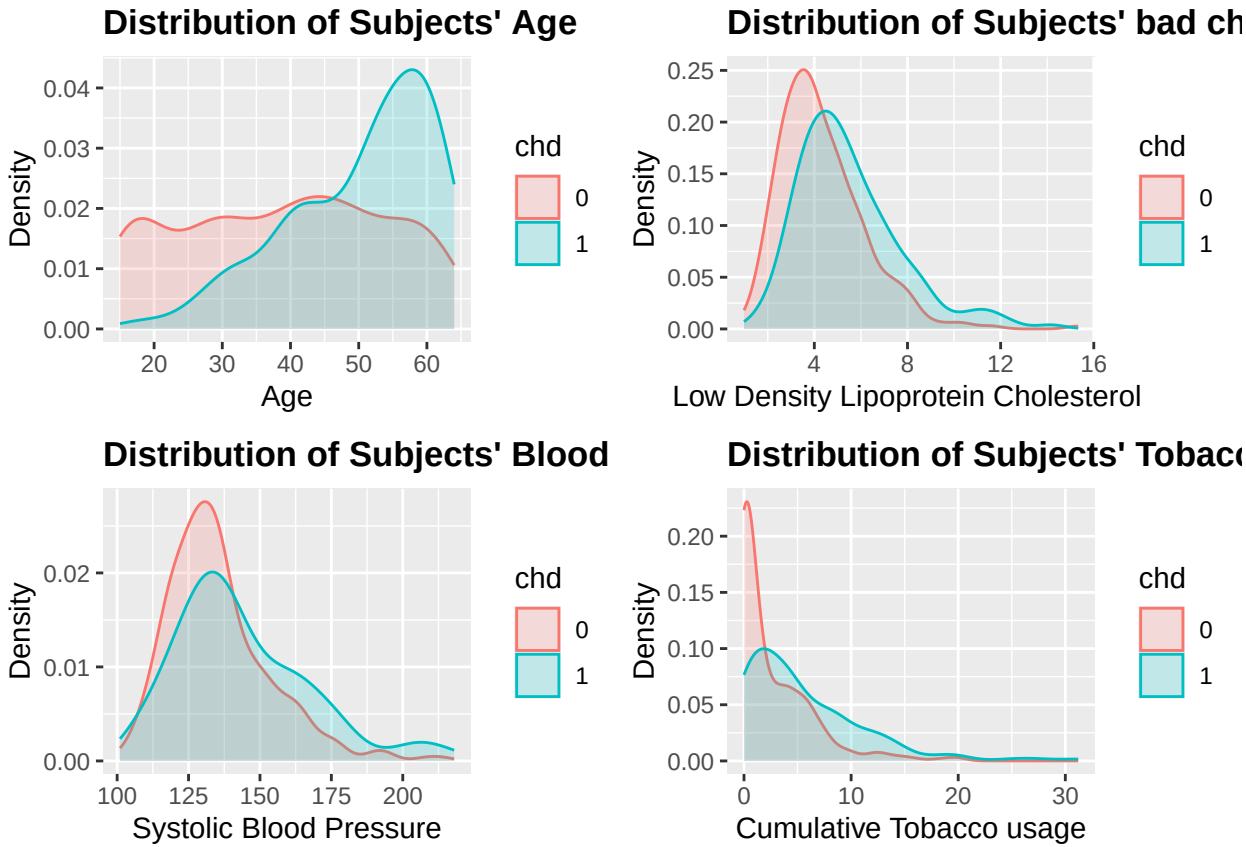


For metric variables, a density plot or histogram allows us to determine the shape of the distribution and look for outliers.

- Use a density plot to explore the distribution of explanatory variables. What do you discover?

```
p1 <- df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(x = age)) +
  geom_density(aes(y = ..density..,  color = chd, fill = chd), alpha = 0.2) +
  ggtitle("Distribution of Subjects' Age") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Age") +
  ylab("Density")
```

```r
p2 <- df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(x = ldl)) +
  geom_density(aes(y = ..density..,  color = chd, fill = chd), alpha = 0.2) +
  ggtitle("Distribution of Subjects' bad cholesterol") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Low Density Lipoprotein Cholesterol ") +
  ylab("Density")

p3 <-df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(x = sbp)) +
  geom_density(aes(y = ..density..,  color = chd, fill = chd), alpha = 0.2) +
  ggtitle("Distribution of Subjects' Blood Pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Systolic Blood Pressure") +
  ylab("Density")

p4 <-df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(x = tobacco)) +
  geom_density(aes(y = ..density..,  color = chd, fill = chd), alpha = 0.2) +
  ggtitle("Distribution of Subjects' Tobacco usage") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Cumulative Tobacco usage") +
  ylab("Density")

grid.arrange(p1, p2,p3,p4, nrow = 2, ncol = 2)
```


Distribution of Subjects' Age
Density
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
Age
chd
0
1
Distribution of Subjects' bad ch
Density
0.25
0.20
0.15
0.10
0.05
0.00
4
8
12
16
Low Density Lipoprotein Cholesterol
chd
0
1
Distribution of Subjects' Blood
Density
0.02
0.01
0.00
100
125
150
175
200
Systolic Blood Pressure
chd
0
1
Distribution of Subjects' Tobac
Density
0.20
0.15
0.10
0.05
0.00
0
10
20
30
Cumulative Tobacco usage
chd
0
1

Bivariate Analysis

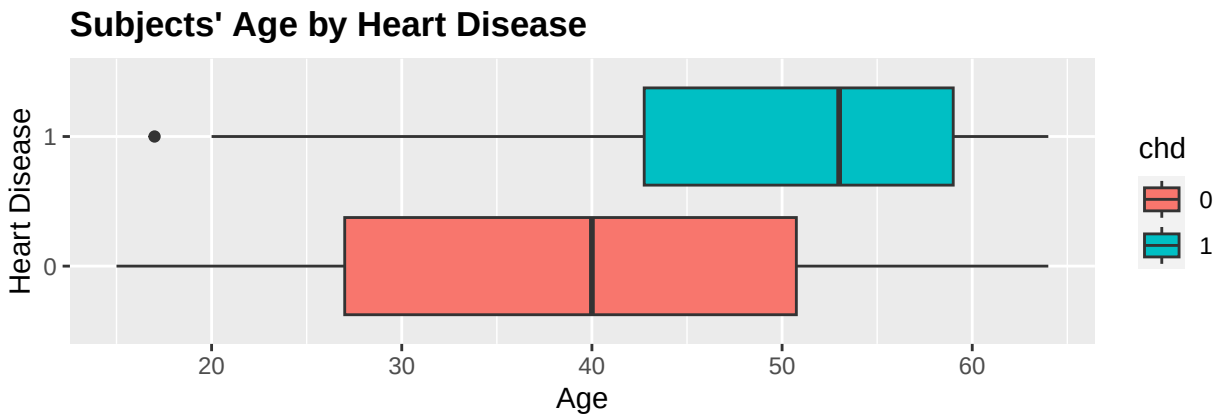
- Prior to moving on to the fully specified model, it is advisable to first examine the simple associations between the response and each explanatory variable.

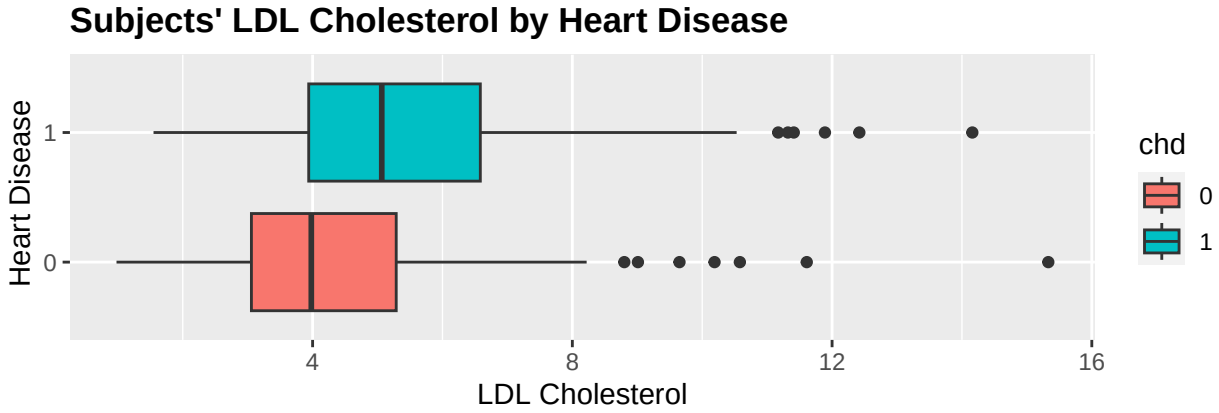
Box plots are useful for exploring the association between a categorical variable and a variable measured on an interval scale.

- Use a boxplot to examine how the explanatory variables are correlated with the response variable (chd)?
  - The coord_flip() function is used to keep the dependent variable on the y-axis.

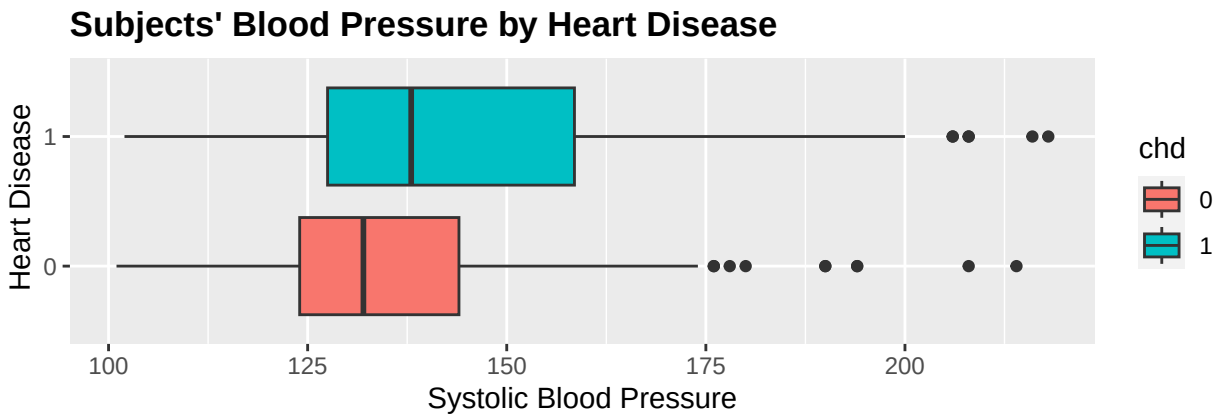
```
p5 <- df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(chd, age)) +
  geom_boxplot(aes(fill = chd)) +
  coord_flip() +
  ggtitle("Subjects' Age by Heart Disease") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Age") +
  xlab("Heart Disease")

p6 <- df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(chd, ldl)) +
  geom_boxplot(aes(fill = chd)) +
  coord_flip() +
  ggtitle("Subjects' LDL Cholesterol by Heart Disease") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("LDL Cholesterol") +
  xlab(" Heart Disease")

p7 <- df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(chd, sbp)) +
  geom_boxplot(aes(fill = chd)) +
  coord_flip() +
  ggtitle("Subjects' Blood Pressure by Heart Disease") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Systolic Blood Pressure") +
  xlab(" Heart Disease")
```

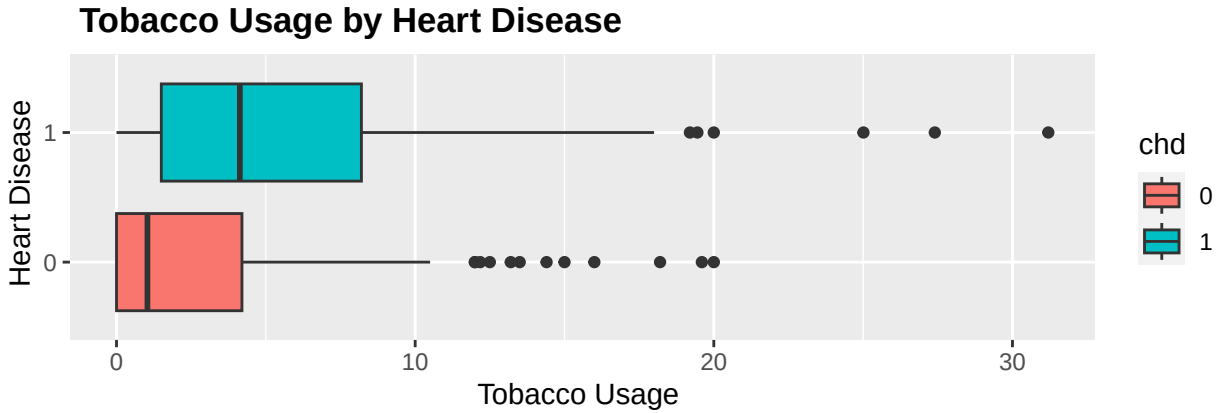
```r
p8 <- df %>%
  mutate(chd=factor(chd)) %>%
  ggplot(aes(chd, tobacco)) +
  geom_boxplot(aes(fill = chd)) +
  coord_flip() +
  ggtitle(" Tobacco Usage by Heart Disease") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Tobacco Usage ") +
  xlab(" Heart Disease")

grid.arrange(p5, p6, nrow = 2, ncol = 1)
```






```
grid.arrange(p7,p8, nrow = 2, ncol = 1)
```


Subjects' Blood Pressure by Heart Disease
Heart Disease
1
0
100
125
150
175
200
Systolic Blood Pressure
chd
0
1


Tobacco Usage by Heart Disease
Heart Disease
1
0
0
10
20
30
Tobacco Usage
chd
0
1

- Use the convenient summary_factorlist() function from the finalfit package to tabulate data.

```
dependent <- "chd"
explanatory <- c("ldl","sbp","tobacco","age")
df %>%
  mutate(chd=as.factor(chd)) %>%
  summary_factorlist(dependent, explanatory, add_dependent_label = TRUE, p = TRUE) %>%
  knitr::kable()
```

Dependent: chd		0	1	p
ldl	Mean (SD)	4.3 (1.9)	5.5 (2.2)	<0.001
sbp	Mean (SD)	135.5 (18.0)	143.7 (23.7)	<0.001
tobacco	Mean (SD)	2.6 (3.6)	5.5 (5.6)	<0.001
age	Mean (SD)	38.9 (14.9)	50.3 (10.6)	<0.001

- According to the plots and the tables, What variable is most important for explaining heart disease? How is that variable correlated with heart disease?

Model Development

Linear probability model

- Is the linear probability model an appropriate choice to study the relationship between heart disease and risk factors?
- Estimate the following linear probability model and interpret the model results.

$$chd = \beta_0 + \beta_1 ldl + \beta_2 sbp + +\beta_3 tobacco + \beta_4 age + u$$

```
#mod.linear <-  # uncomment and replace with your code
```

- What are the advantages and disadvantages of the linear probability model?

Generalized linear model

- Estimate the following logistic regression model and interpret the model results.

$$logit(\pi_i) = \beta_0 + \beta_1 ldl + \beta_2 sbp + + \beta_3 tobacco + \beta_4 age + u$$

```
#mod.logit.h0 <-   # uncomment and replace with your code
```

Interpretation of model results

- Do the "raw" coefficient estimates "directionally make sense"?

```
# Replace with your code
```

- Recall that

$$OR = \frac{Odds_{x_k+c}}{Odds_{x_k}} = exp(c\beta_k)$$

- Compute and interpret the estimated odds ratio for a 10-unit increase in each explanatory variable.

```
# Replace with your code
```

Statistical Inference

Hypothesis Test

- Using the likelihood ratio test (LRT) for hypothesis testing is a common practice in a logistic regression model.

$$-2log(\Lambda) = -2log(\frac{L(\hat{\beta}^{(0)}|y_1,...,y_n)}{L(\hat{\beta}^{(a)}|y_1,...,y_n)}) = -2\sum y_i log(\frac{\hat{\pi}_i^{(0)}}{\hat{\pi}_i^{(a)}}) + (1-y_i)log(\frac{1-\hat{\pi}_i^{(0)}}{1-\hat{\pi}_i^{(a)}})$$

- Explain what LRT measures and when it rejects the Null hypothesis?
- Use LRT to test whether (*obesity*) is associated with heart disease.
  - $H_0 : \beta_{obesity} = 0$
  - $H_a : \beta_{obesity} \neq 0$

Use both *Anova()* or *anova()* functions.

```
#mod.logit.ha <- # uncomment and replace with your code
#anova()
#Anova()
```

Deviance

- From Async, deviance refers to the amount that a particular model deviates from another model as measured by $-2log(\Lambda)$.
- What are the null deviance and residual deviance in the model summary?

For null and residual deviance, the alternative model we use is the saturated model, which has a different coefficent for each data point, leading to perfect prediction, a likelihood of one, and a log likelihood of zero.

- The null deviance therefore measures the performance of the worst model using only an intercept, providing a benchmark.

$$NullDeviance = -2log(L(\hat{\beta}_0|y_1, ..., y_n))$$

- The residual deviance is the deviance of our fitted model. It is always greater than zero unless it is the saturated model / explains the data perfectly.

$$ResidualDeviance = -2log(L(\hat{\beta}|y_1, ..., y_n))$$

Therefore, how much better (smaller) our residual deviance is compared to the null deviance and how close it is to zero is a measure of model fit.

Sometimes people will compute an R squared for logistic regression using $1 - \frac{ResidualDeviance}{NullDeviance}$ since it is bounded between 0 (residual deviance = null deviance) and 1 (residual deviance = saturated model = 0).

Note that we can compute deviance of two separate models by substracting the null model residual deviance and the alternative model residual deviance from separate logistic regression fits. (Why is this?)

- Using deviance, test whether ($obesity$) is associated with heart disease.
  - $H_0 : \beta_{obesity} = 0$
  - $H_a : \beta_{obesity} \neq 0$

```
#degree_freedom <- # uncomment and replace with your code

#test_stat <- # uncomment and replace with your code

#pvalue <- # uncomment and replace with your code
```

Confidence Interval

Confidence Interval for odds ratio Wald Confidence:

$$c * \hat{\beta}_k \pm c * Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_k)}$$

$$exp\left(c * \hat{\beta}_k \pm c * Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_k)}\right)$$

- Calculate Wald CI for odds ratio of 10-unit increase in LDL cholesterol based on the above formula:

```
# Replace with your code
```

- What is the main concern with Wald CI?
- Now calculate the *profile likelihood ratio (LR)* confidence interval using the confint function.

```
# Replace with your code
```

Confidence Interval for the Probability of Success

- Recall that the estimated probability of success is

$$\hat{\pi} = \frac{exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k\right)}{1 + exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k\right)}$$

While backing out the estimated probability of success is straightforward, obtaining its confidence interval is not, as it involves many parameters.

Wald Confidence Interval

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K)}$$

where

$$\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K) = \sum_{i=0}^{K} x_i^2 \widehat{Var}(\hat{\beta}_i) + 2\sum_{i=0}^{K-1}\sum_{j=i+1}^{K} x_i x_j \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_j)$$

So, the Wald Interval for π is

$$\frac{exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k \pm \sqrt{\sum_{i=0}^{K} x_i^2 \widehat{Var}(\hat{\beta}_i) + 2\sum_{i=0}^{K-1}\sum_{j=i+1}^{K} x_i x_j \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_j)}\right)}{1 + exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k\right) \pm \sqrt{\sum_{i=0}^{K} x_i^2 \widehat{Var}(\hat{\beta}_i) + 2\sum_{i=0}^{K-1}\sum_{j=i+1}^{K} x_i x_j \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_j)}}$$

- For an average value of all explanatory variables, compute the Confidence Interval for the Probability of Success given the formula above

```
alpha = 0.5

predict.data <- data.frame(ldl = mean(df$ldl),
                           sbp = mean(df$sbp),
                           tobacco = mean(df$tobacco),
                           age = mean(df$age))
# Obtain the linear predictor
#linear.pred <- # uncomment and replace with your code

# Then, compute pi.hat
#pi.hat <- # uncomment and replace with your code

# Compute Wald Confidence Interval (in 2 steps)
# Step 1
#CI.lin.pred <- # uncomment and replace with your code
#CI.lin.pred

# Step 2
#CI.pi <- # uncomment and replace with your code
#CI.pi

# Store all the components in a data frame
#round(data.frame(pi.hat, lower=CI.pi[1], upper=CI.pi[2]),4)
```

Final Visualization

- Using both the linear probability and logistic regression models, plot the estimated probability of heart disease for different values of cholesterol, holding other variables constant at their average level.
- Discuss which one can better explain this relationship.

```
# uncomment and run the code

#coef <- mod.logit.h0$coefficients

# Effect of income on ldl for a person average age, sbp, and tobacco usage

#xx = c(1, mean(df$ldl), mean(df$sbp), mean(df$tobacco), mean(df$age))

#z = coef[1]*xx[1]+ coef[3]*xx[3] + coef[4]*xx[4] + coef[5]*xx[5]

#x <- df$ldl

# Reproduce the graph overlaying the same result from the linear model as a comparison
#curve(expr = exp(z + coef[2]*x)/(1+exp(z + coef[2]*x)),
#    xlim = c(min(df$ldl), max(df$ldl)),
#    ylim = c(0,2),
#    col = "blue",
#    main = expression(pi == frac(e^{z + coef[inc]*ldl}, 1+e^{z+coef[inc]*ldl})),
#    xlab =  expression(cholesterol), ylab = expression(pi))

#par(new=TRUE)

#lm.coef <- mod.linear$coefficients
#lm.z <- lm.coef[1]*xx[1] + lm.coef[3]*xx[3] + lm.coef[4]*xx[4] + lm.coef[5]*xx[5]
#lines(df$ldl, lm.z + lm.coef[2]*x,col="green")
```

Final Report

- Display both estimated linear and logistic models in a regression table. Is there any significant difference between their results?

```
# uncomment and run the code

#stargazer(mod.linear, mod.logit.h0, type = "text", omit.stat = "f",
#                    star.cutoffs = c(0.05, 0.01, 0.001), title = "Table 1: The estimated relationship between heart disease and  risk fac
```

Reminders

1. Before next live session:
    1. Turn in HW 1 if you have not already
    2. Complete the homework that builds on this unit (HW-2)
    3. Complete all videos and reading for unit 3