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
- Teams for Lab-1 will be created soon. Please fill out the survey in the slack channel.

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)
```

- Why do we need these packages?

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 are related to heart disease?

- What are the requirements of explanatory modeling to have a casual interpretation?
- Is logistic regression useful for prediction? if yes, How?

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

Descriptive Statistics

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)
```

```
## 'data.frame':    462 obs. of  6 variables:
##  $ tobacco: num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
##  $ ldl    : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
##  $ sbp    : int  160 144 118 170 134 132 142 114 114 132 ...
##  $ age    : int  52 63 46 58 49 45 38 58 29 53 ...
##  $ chd    : int  1 1 0 1 1 0 0 1 0 1 ...
##  $ obesity: num  25.3 28.9 29.1 32 26 ...
```

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

Univariate Analysis

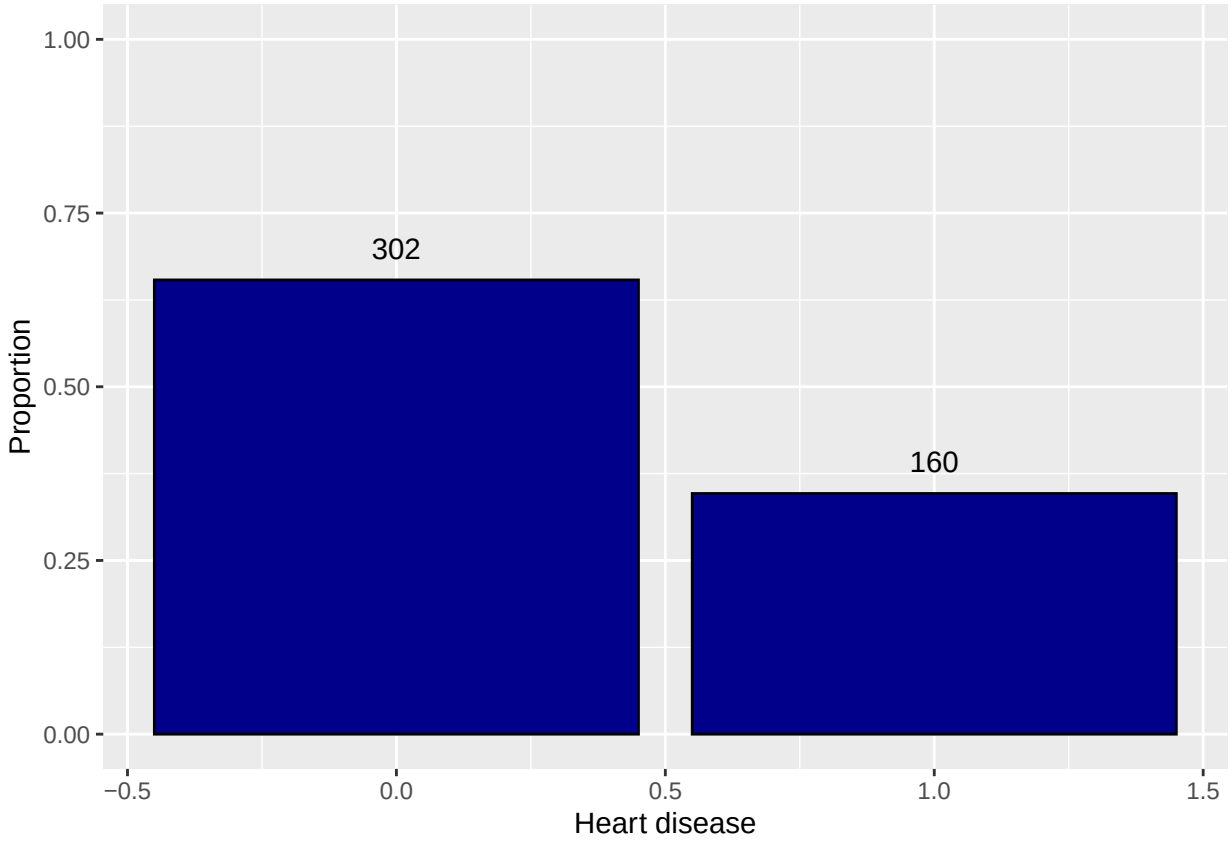
- The response (or dependent) variable of interest, Heart disease, is a binary variable taking the type factor.
- Use a bar chart to explore the distribution of the response variable (chd). What do you learn?

```
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") +
  ylim(0,1)
```

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



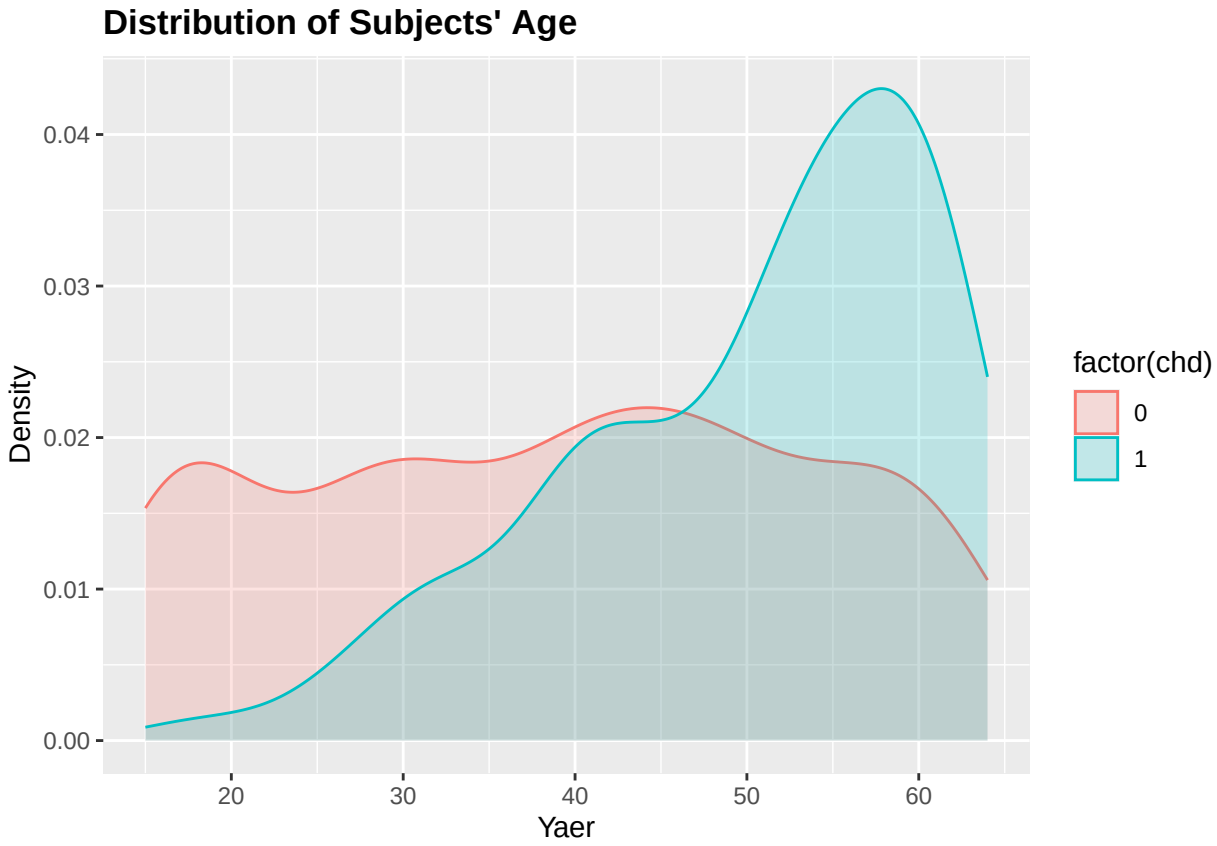
According to the bar plot and the table, we learn that 35% of subjects in our sample suffered heart disease, which is pretty high.

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?

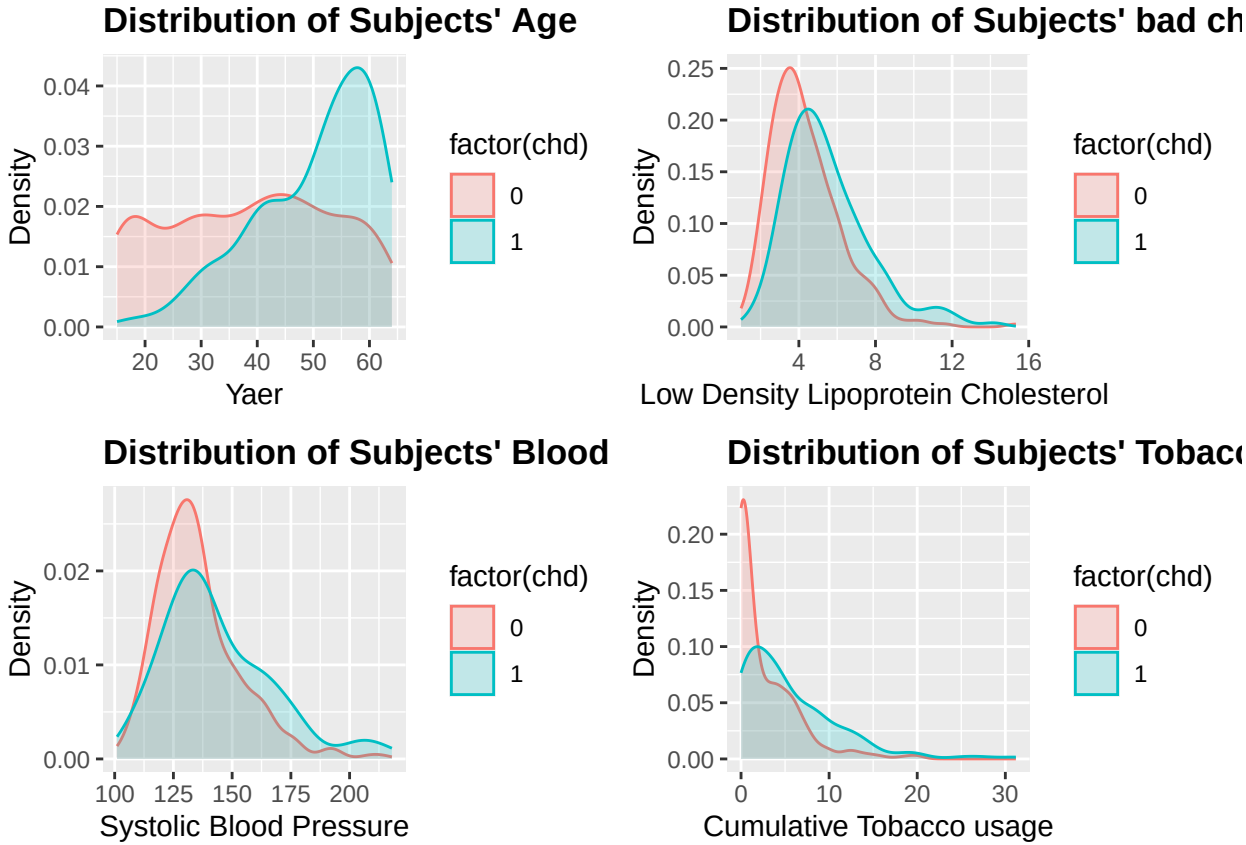
```
p1 <- df %>%
  ggplot(aes(x = age)) +
  geom_density(aes(y = ..density.., color = factor(chd), fill = factor(chd)), alpha = 0.2) +
  ggtitle("Distribution of Subjects' Age") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
```

```
  xlab("Yaer") +
  ylab("Density")
p1
```




```
p2 <- df %>%
  ggplot(aes(x = ldl)) +
  geom_density(aes(y = ..density.., color = factor(chd), fill = factor(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")
```

```r
p3 <-df %>%
  ggplot(aes(x = sbp)) +
  geom_density(aes(y = ..density.., color = factor(chd), fill = factor(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 %>%
  ggplot(aes(x = tobacco)) +
  geom_density(aes(y = ..density..,  color = factor(chd), fill = factor(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)
```



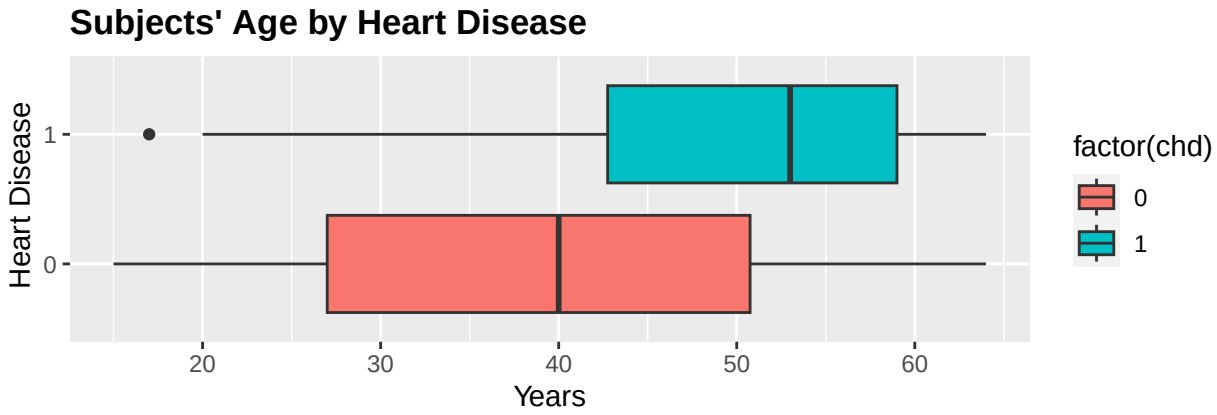
Based on the density plots, men in our sample are, on average, 43 years old, with average high blood pressure stage-1 (130-139), and they have on average 4.8 units of LDL and smoke 3.7 kg of tobacco.

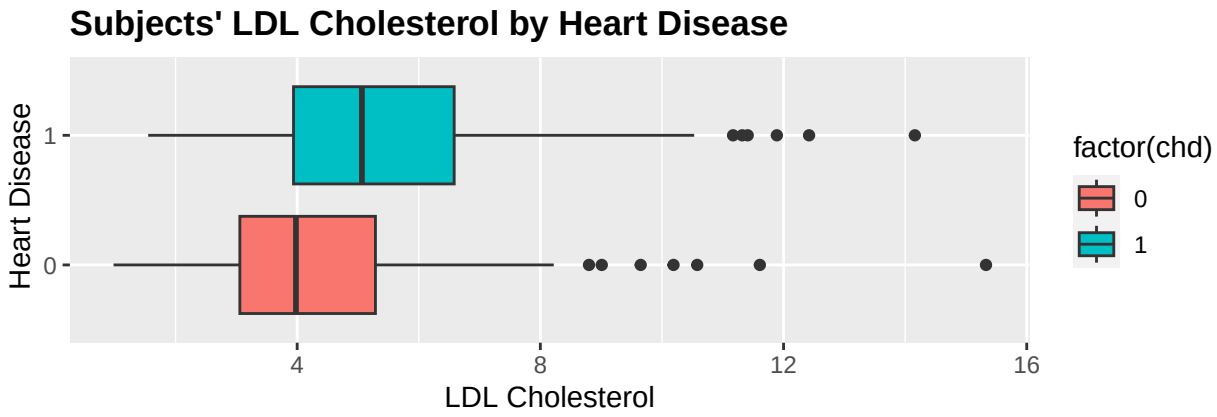
Bivariate Analysis

- Before moving on to the fully specified model, it is advisable to examine the simple associations between the response and each explanatory variable.
- Box plots help explore 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.

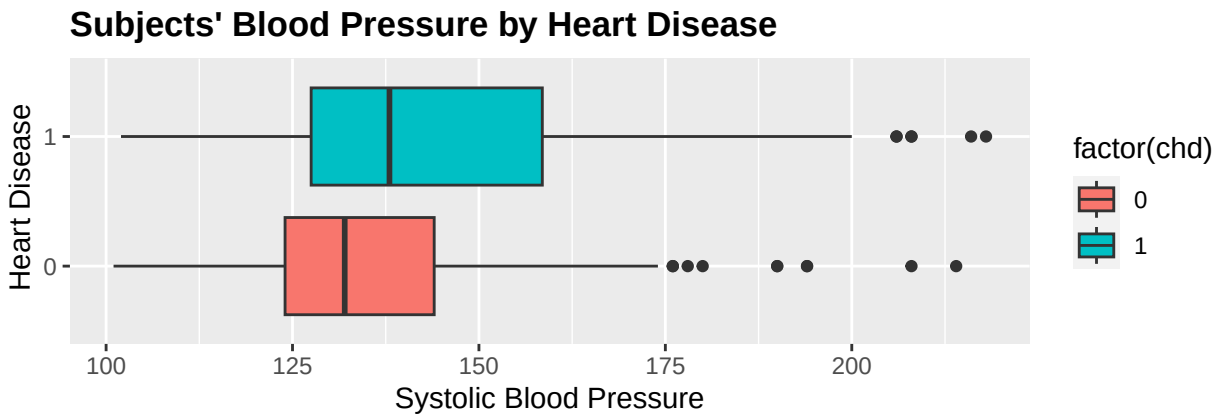
```
p5 <- df %>%
  ggplot(aes(factor(chd), age)) +
  geom_boxplot(aes(fill = factor(chd))) +
  coord_flip() +
  ggtitle("Subjects' Age by Heart Disease") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Years") +
  xlab(" Heart Disease")

p6 <- df %>%
 ggplot(aes(factor(chd), ldl)) +
  geom_boxplot(aes(fill = factor(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 %>%
 ggplot(aes(factor(chd), sbp)) +
  geom_boxplot(aes(fill = factor(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")

p8 <- df %>%
 ggplot(aes(factor(chd), tobacco)) +
  geom_boxplot(aes(fill = factor(chd))) +
```

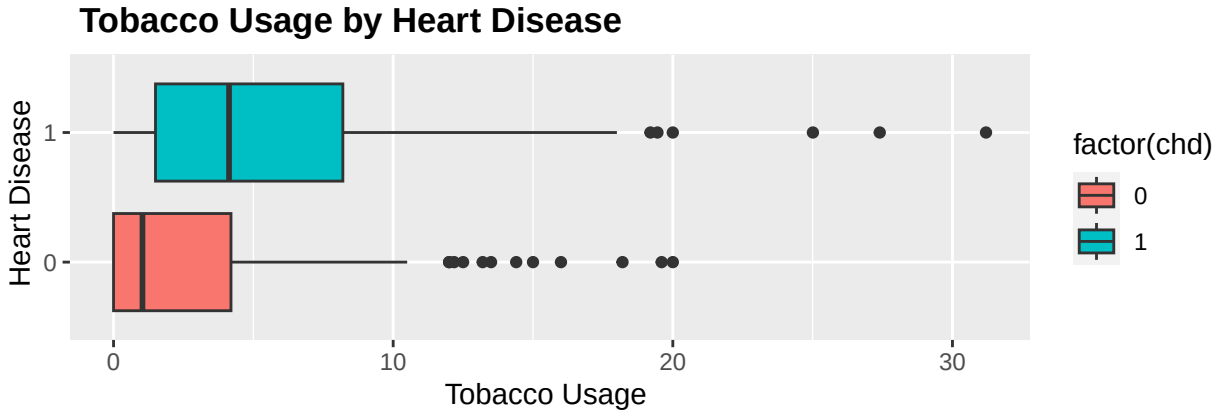
```
  coord_flip() +
  ggtitle(" Tobacco Usage by Heart Disease") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Tobacco Usage ") +
  xlab(" Heart Disease")

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






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





- 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) %>%
  knitr::kable()
```

Dependent: chd		0	1
ldl	Mean (SD)	4.3 (1.9)	5.5 (2.2)

Dependent: chd		0	1
sbp	Mean (SD)	135.5 (18.0)	143.7 (23.7)
tobacco	Mean (SD)	2.6 (3.6)	5.5 (5.6)
age	Mean (SD)	38.9 (14.9)	50.3 (10.6)

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

  From the boxplots and the table, we can conclude that the men who suffered from heart disease are older with higher LDL cholesterol, blood pressure, and tobacco usage. There is a positive correlation between these risk factors and 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?

  The linear probability model could lead to probabilities less than 0 or greater than 1, which is not desirable.

- 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 <- lm(chd ~ ldl + sbp + tobacco + age, data = df)

summary(mod.linear)
```

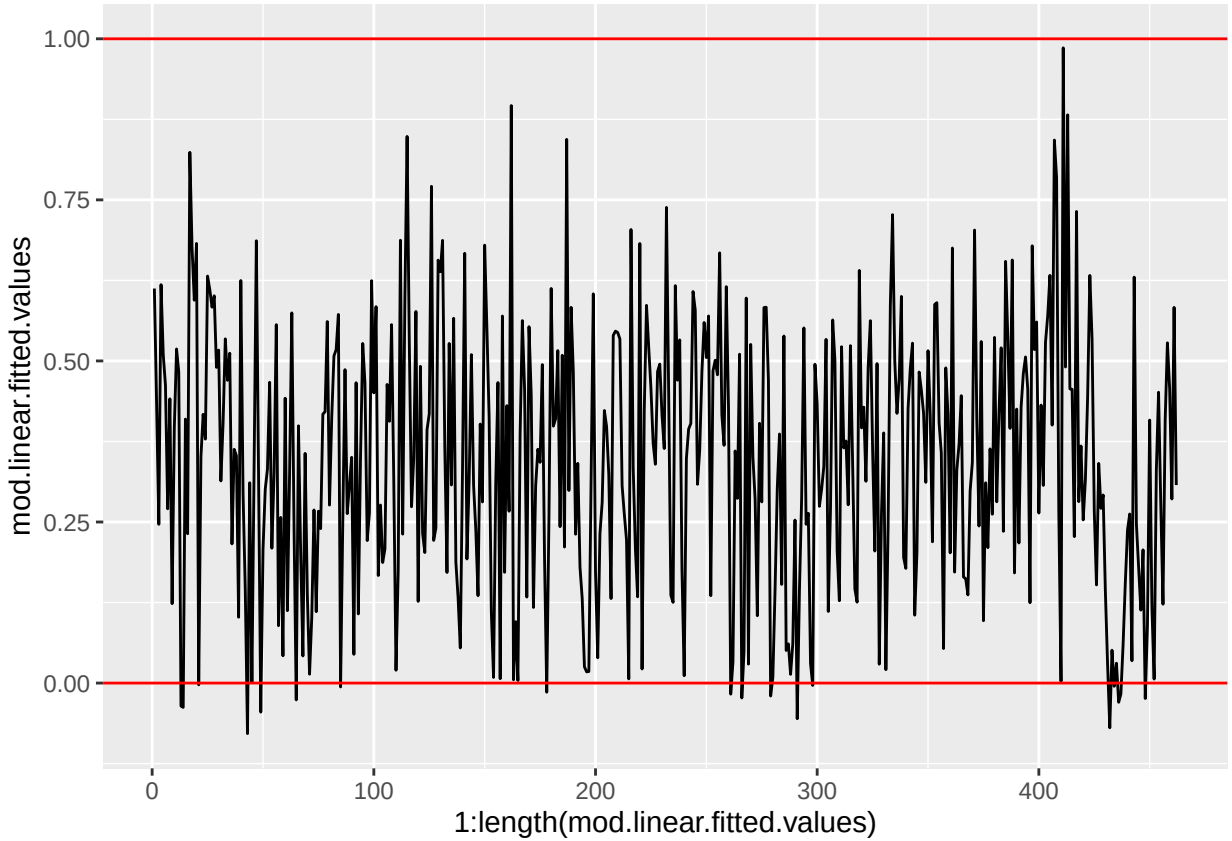
```
##
## Call:
## lm(formula = chd ~ ldl + sbp + tobacco + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8439 -0.3405 -0.1250  0.4365  1.0172
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.3493578  0.1405912  -2.485 0.013315 *
## ldl          0.0362419  0.0102322   3.542 0.000438 ***
## sbp          0.0009739  0.0010670   0.913 0.361839
## tobacco      0.0165577  0.0049101   3.372 0.000809 ***
## age          0.0076831  0.0016886   4.550 6.89e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4318 on 457 degrees of freedom
## Multiple R-squared:  0.1853, Adjusted R-squared:  0.1781
## F-statistic: 25.98 on 4 and 457 DF,  p-value: < 2.2e-16
```

All the explanatory variables except blood pressure are statistically significant and positively related to the probability of heart disease.

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

The main disadvantage of the linear probability model is that estimated probabilities could be less than zero or greater than one. If we plot the model fitted values or estimated probability in this model, we can see that some are less than zero, which is not desirable.

```
fitted_values <- data.frame(mod.linear$fitted.values)

fitted_values %>%
  ggplot(aes(x= 1:length(mod.linear.fitted.values), y = mod.linear.fitted.values)) +
  geom_line() +
  geom_hline(aes(yintercept = 0), color = "red") +
  geom_hline(aes(yintercept = 1), color = "red")
```



The advantage of the linear probability model is that it's easy to understand and communicate with a less technical audience.

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 <- glm(chd ~ ldl + sbp + tobacco + age, family = binomial(link = logit), data = df)
```

Interpretation of model results

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

```
summary(mod.logit.h0)
```

```
##
## Call:
## glm(formula = chd ~ ldl + sbp + tobacco + age, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9457  -0.8595  -0.4999   1.0238   2.3906
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.535524   0.781360  -5.805 6.45e-09 ***
## ldl          0.185131   0.054121   3.421 0.000625 ***
## sbp          0.004307   0.005394   0.798 0.424623
## tobacco      0.075982   0.025616   2.966 0.003016 **
## age          0.046264   0.009852   4.696 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 502.19  on 457  degrees of freedom
## AIC: 512.19
##
## Number of Fisher Scoring iterations: 4
```

Again, all of the explanatory variables except blood pressure are statistically significant and positively correlated with the probability of heart disease, same as the linear probability model.

- 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.

```
round(exp(10*coef(mod.logit.h0)),2)
```

```
## (Intercept)         ldl         sbp     tobacco         age
##        0.00        6.37        1.04        2.14        1.59
```

The estimated odds of success or having a heart disease change by 6.37 times for every 10-unit increase in LDL or 'bad' cholesterol.

Interestingly, the odds of having a heart disease is almost 1 for every 10-unit increase in blood pressure, which means an increase in blood pressure doesn't change the odds of having heart disease, and it's consistent with its insignificant coefficient.

Statistical Inference

Hypothesis Test

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?

  When the null hypothesis is that one or a couple of coefficients is zero, the numerator is the likelihood function of the model that excludes those variables (whose coefficients are set to 0 in the null hypothesis), but the denominator in this equation is the likelihood function of the model containing all coefficients.

  LRT statistic has an approximate chi-squared distribution when the null hypothesis is true, and we reject the null hypothesis if the LRT statistic has an unusually large observed value for this chi-squared distribution. The LRT statistic is bounded between 0 and infinity since if the null model is as good as the alternative model (coefficients really are zero) the LRT is 0 since log(1)=0, and if the null model is really bad i.e. likelihood is 0 then log(0) is -inifinity and times negative 2 is then really large, leading a large chi squared test statistic and rejecting the null hypothesis.

- Use LRT to test whether (*obesity*) is associated with heart disease.
  - $H_0 : \beta_{obesity} = 0$
  - $H_a : \beta_{obesity} \neq 0$

Hint use both *Anova()* or *anova()* functions.

```
mod.logit.ha <- glm(chd ~ ldl + sbp +tobacco + age + obesity, family = binomial(link = logit), data = df)

anova(mod.logit.h0, mod.logit.ha, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: chd ~ ldl + sbp + tobacco + age
## Model 2: chd ~ ldl + sbp + tobacco + age + obesity
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       457     502.19
## 2       456     501.07  1   1.1191   0.2901
```

```r
Anova(mod.logit.ha, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: chd
##         LR Chisq Df Pr(>Chisq)
## ldl      13.3932  1  0.0002525 ***
## sbp       0.8640  1  0.3526279
## tobacco   9.4670  1  0.0020920 **
## age      24.3447  1  8.055e-07 ***
## obesity   1.1191  1  0.2901078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both these functions report the same p-value of 2.9 for obesity. We fail to reject the null hypothesis that obesity is not associated with heart disease in this sample.

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 hypothesis model residual deviance and the alternative model residual deviance from separate logistic regression fits. (Why is this?)

This is because the difference of the residual deviance of two models is the same as the LRT test statistic for those two models, assuming they are nested. This means the difference in residual deviance is the LRT test. The degrees of freedom is just the difference in number of parameters i.e. non zero coefficients in the two models.

- Again, test whether (*obesity*) is associated with heart disease. But this time, use deviance.
  - $H_0 : \beta_{obesity} = 0$
  - $H_a : \beta_{obesity} \neq 0$

```
test_stat <- mod.logit.h0$deviance - mod.logit.ha$deviance

degree_freedom <- mod.logit.h0$df.residual - mod.logit.ha$df.residual

pvalue <- 1-pchisq(test_stat, df = degree_freedom)
```

Compute the p-value, using pchisq()

We get a p-value of 0.29, the same as what we got from both anova() and Anova() functions, and again we fail to reject the null hypothesis that obesity is not correlated with heart disease given this data set.

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

```
vcov(mod.logit.h0)
```

```
##              (Intercept)           ldl           sbp       tobacco
## (Intercept)  0.610523787 -9.955527e-03 -3.315082e-03  1.122271e-03
## ldl         -0.009955527  2.929029e-03 -1.336470e-05 -4.923675e-06
## sbp         -0.003315082 -1.336470e-05  2.909849e-05 -1.506782e-06
## tobacco      0.001122271 -4.923675e-06 -1.506782e-06  6.562050e-04
## age         -0.001828353 -6.178475e-05 -1.503794e-05 -7.699037e-05
##                       age
## (Intercept) -1.828353e-03
## ldl         -6.178475e-05
## sbp         -1.503794e-05
## tobacco     -7.699037e-05
## age          9.706566e-05
```

```
round(exp(10*mod.logit.h0$coefficients[2] +10*qnorm(p=c(0.025, 0.975))*
            sqrt(vcov(mod.logit.h0)[2,2])),2)
```

```
## [1]  2.20 18.39
```

With 95% confidence, the odds of having a heart disease change between 2.20 to 18.4 times for every 10-unit increase in LDL or 'bad' cholesterol.

- What is the main concern with Wald CI?

  Wald confidence interval has a true confidence level close to the 95% only when we have large samples. When the sample size is not large, profile LR confidence intervals generally perform better.

- Now calculate the *profile likelihood ratio (LR)* confidence interval.

Note that conceptually the profile likelihood ratio models the LRT test statistic and finds the coefficient values i.e. those for the numerator model and those for the alternative model in the denominator that lead to a chi squared test statistic at the 95th quantile. You can think of this as finding the coefficient values that lead to a 95% range in the LRT test statistic and therefore cover 95% probability.

```
beta_ci <- confint(mod.logit.h0)
```

```
## Waiting for profiling to be done...
```

```
odds_ci <- exp(10*beta_ci)

round(cbind(odds_ci ),2)
```

```
##              2.5 % 97.5 %
## (Intercept)  0.00   0.00
## ldl          2.24  18.84
## sbp          0.94   1.16
## tobacco      1.31   3.59
## age          1.31   1.93
```

Since we have a large sample, 462 observations, the profile likelihood ratio (LR) confidence interval is pretty close to the Wald CI.

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

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

- 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 = predict(object = mod.logit.h0, newdata = predict.data, type = "link", se = TRUE)

# Then, compute pi.hat
pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit))

# Compute Wald Confidence Interval (in 2 steps)
# Step 1
CI.lin.pred = linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se.fit
```

```r
#CI.lin.pred

# Step 2
CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred))
#CI.pi

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

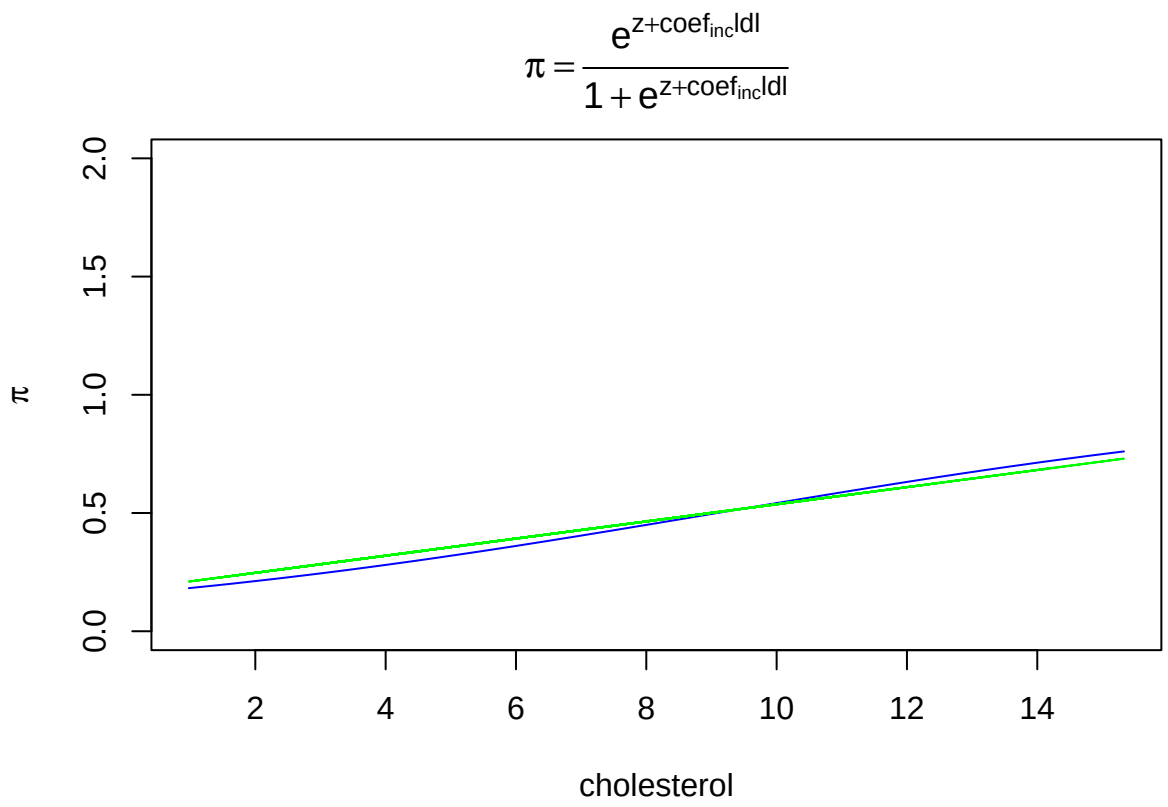
```
##   pi.hat  lower  upper
## 1 0.3089 0.2925 0.3259
```

The 95% Wald confidence interval for the probability of having a heart disease is between 0.293 and 0.326, So the probability of having a heart disease is not that high for a man with average age, LDL, tobacco usage, and blood pressure.

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.

```
coef <- mod.logit.h0$coefficients

# Effect of income on LDL for a person's 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")
```



Surprisingly, both linear probability and logistic model have a pretty close estimated probability but note the differences when the predicted probability approaches zero and one.

Final Report

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

```
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 factors")
```

```
##
## Table 1: Th
## ================================================
##                           Dependent variable:
##                       -----------------------------
##                                   chd
##                             OLS          logistic
##                             (1)             (2)
## ------------------------------------------------
## ldl                        0.036***       0.185***
##                            (0.010)         (0.054)
##
## sbp                         0.001           0.004
##                            (0.001)         (0.005)
##
## tobacco                    0.017***        0.076**
##                            (0.005)         (0.026)
##
## age                        0.008***       0.046***
##                            (0.002)         (0.010)
##
## Constant                   -0.349*        -4.536***
##                            (0.141)         (0.781)
##
## ------------------------------------------------
## Observations                462             462
## R2                         0.185
## Adjusted R2                0.178
## Log Likelihood                           -251.093
## Akaike Inf. Crit.                         512.187
## Residual Std. Error  0.432 (df = 457)
## ================================================
```

```
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

In both models, all the coefficients except blood pressure are statistically significant and positively associated with the probability of having heart disease. Also, LDL is the most correlated variable with the probability of heart disease in both models.

Reminders

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