Unit 3 Live Session

Figure 1: South Hall

Discrete Response Model Part 3

Class Announcements

- HW 3 is this week
- Lab-1 due in 3 weeks

Roadmap

Rearview Mirror

- Discuss why the classical linear regression model is the best choice for the binary response model
- Discuss logistic regression models, the most important special case of generalized linear models (GLMs).

Today

- Variable transformation: interactions among explanatory variables and quadratic term
- Categorical explanatory variables
- Convergence criteria and complete separation

Looking Ahead

- Multinomial probability distribution,
- IJ contingency tables and inference using contingency tables
- Nominal response models
- Ordinal logistical regression model

Start-up Code

```
# Insert the function to *tidy up* the code when they are printed out
if(!"knitr"%in%rownames(installed.packages())) {install.packages("knitr")}
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
if(!"tidyverse"%in%rownames(installed.packages())) {install.packages("tidyverse")}
library(tidyverse)

## to load glow500 from "Applied Logistic Regression" by D.W. Hosmer, S. Lemeshow and R.X. Sturdivant (3rd ed., 2013)
if(!"aplore3"%in%rownames(installed.packages())) {install.packages("aplore3")}
library(aplore3)

## provides  many functions useful for data analysis, high-level graphics, utility operations like describe()
if(!"Hmisc"%in%rownames(installed.packages())) {install.packages("Hmisc")}
library(Hmisc)

## to work with "grid" graphics
if(!"gridExtra"%in%rownames(installed.packages())) {install.packages("gridExtra")}
library(gridExtra)

## To generate regression results tables and plots
if(!"finalfit"%in%rownames(installed.packages())) {install.packages("finalfit")}
library(finalfit)

## To produces LaTeX code, HTML/CSS code and ASCII text for well-formatted tables
if(!"stargazer"%in%rownames(installed.packages())) {install.packages("stargazer")}
library(stargazer)
```

Discusion: Complete Separation

- What is complete separation?
- Is there any problem with the following data set?

```
df_0 <- data.frame(y =  c(0,0,0,0,1,1,1,1), x1 = c(1,2,3,3,5,6,10,11), x2 = c(3,2,-1,-1,2,4,1,0))
#plot(df_0$x1, df_0$y)
```

- What happens when we try to fit a logistic regression model of Y on X1 and X2?

```
mod.logit.complete<- glm(y~ x1+x2, family=binomial(link = logit), data = df_0)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
#summary(mod.logit.complete)
```

- What are the techniques to deal with complete separation?

Case Study: Osteoporosis in Women

Introduction

In osteoporosis, bones become weak and brittle, so weak that even bending over or coughing can fracture them. Hip, wrist, and spine fractures are the most common osteoporosis-related fractures.

All races of people are at risk for osteoporosis. However, white and Asian women, particularly those who that past menopause, are at the greatest risk. A healthy diet, weight-bearing exercises, and medications can strengthen already weak bones or prevent their loss. (Mayo Clinic)

Here, Our goal is the description:

- **How factors such as age and weight are related to the fracture rates among older women?**

Data Description

This sample comes from the Global Longitudinal Study of Osteoporosis in Women (GLOW).

The data set includes information on 500 subjects enrolled in this study.

Install and load the aplore3 library in order to use the glow500 dataset and understand the structure dataset.

We summarize some of the variables that we will use:

- PRIORFRAC: History of prior fracture
- AGE: Age at enrollment
- WEIGHT: Weight at enrollment (Kilograms)
- HEIGHT: Height at enrollment (Centimeters)
- BMI: Body mass index (kg/m^2)
- PREMENO: Menopause before age 45
- FRACTURE: Any fracture in first year of follow up
- RATERISK: Self-reported risk of fracture
- SMOKE: Former or current smoker

Descriptive Statistics

- First, load and check the data set.

```
df = glow500 %>%
  dplyr::select(fracture, age, priorfrac, premeno, raterisk, smoke, bmi)


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

fracture	age	priorfrac	premeno	raterisk	smoke	bmi
No	62	No	No	Same	No	28.16055
No	65	No	No	Same	No	34.02344
No	88	Yes	No	Less	No	20.60936
No	82	No	No	Less	No	24.25781
No	61	No	No	Same	No	29.43213
No	67	Yes	No	Same	Yes	26.23356

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

Univariate Analysis

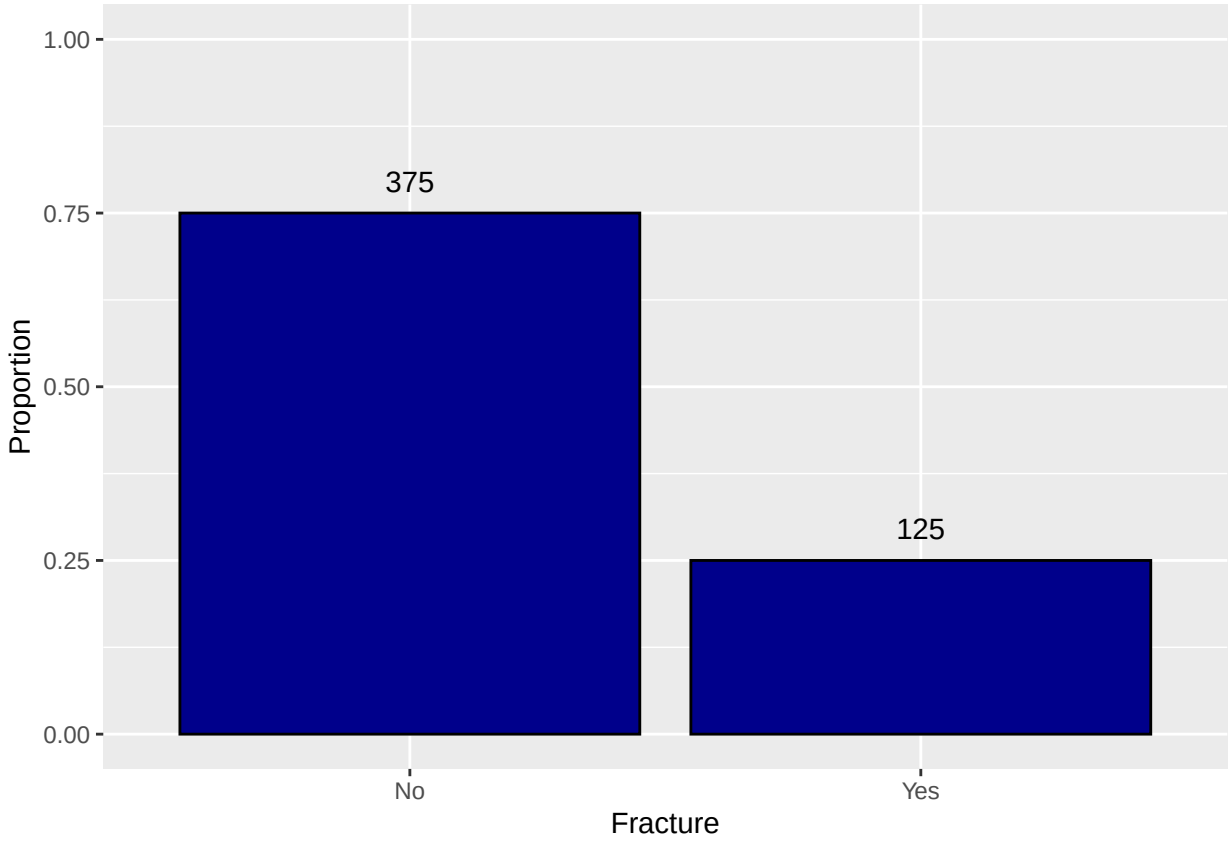
- The response (or dependent) variable of interest, fracture in the first year of follow-up as FRACTURE, is a binary variable taking the type "factor".
- Use the following code to review the distribution of the response variable (FRACTURE). What do you discover?

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

Fracture	N	Proportion
No	375	0.75
Yes	125	0.25

```
df %>%
  ggplot(aes(x= fracture, y = ..prop.., group = 1)) +
  geom_bar(fill = 'DarkBlue', color = 'black') +
  geom_text(stat='count', aes(label=..count..), vjust=-1) +
  xlab("Fracture") +
  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.
```

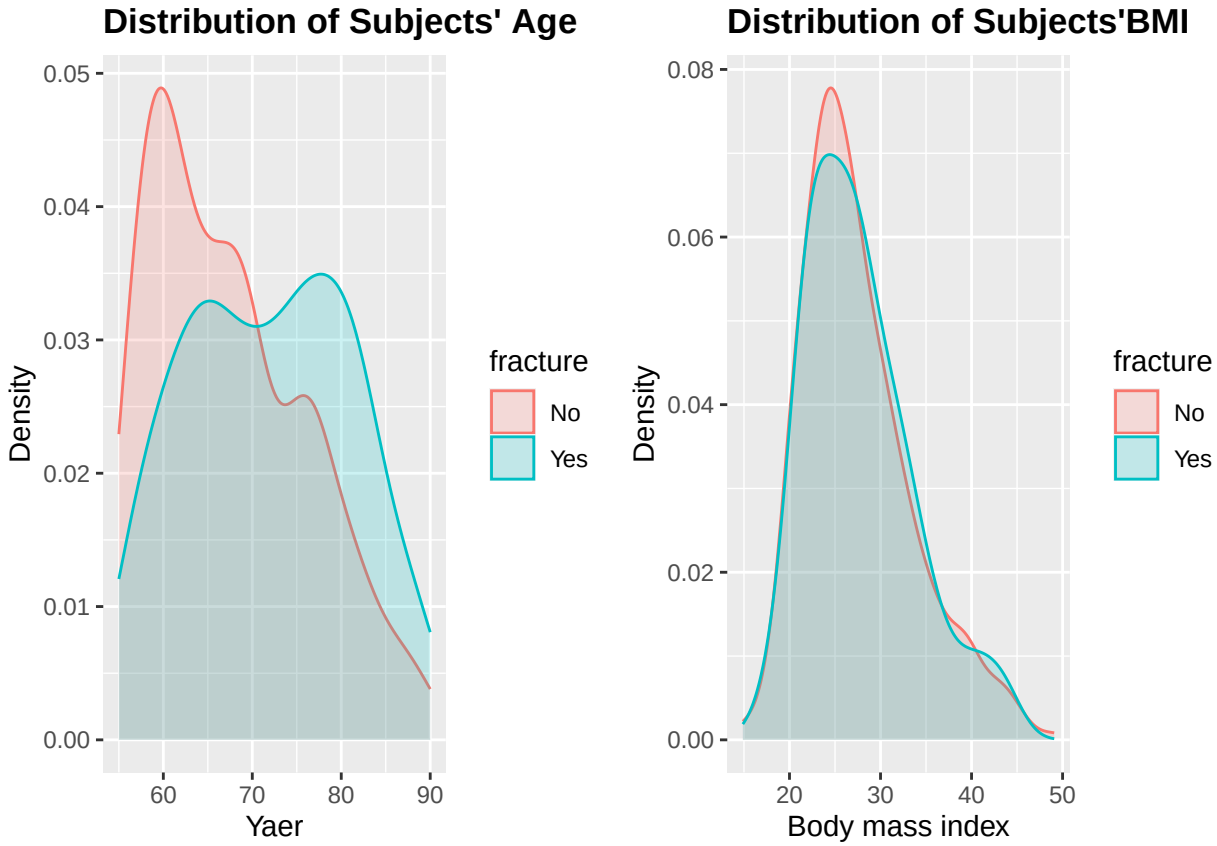


For metric variables, histograms allow us to determine the shape of the distribution and look for outliers.

- Use a density plot to examine the distribution of age and BMI. What do you learn?

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

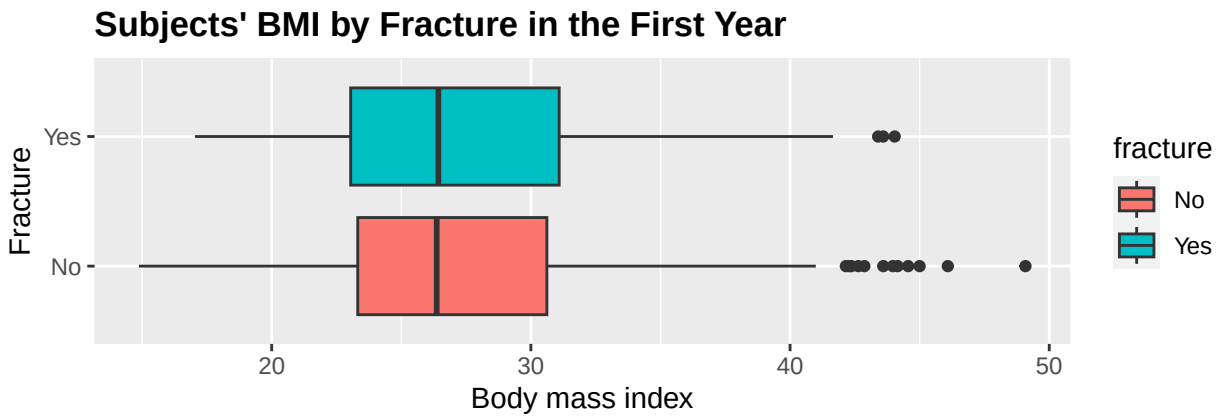
```
p2 <-df %>%
  ggplot(aes(x = bmi)) +
  geom_density(aes(y = ..density.., color = fracture, fill = fracture),alpha=0.2) +
  ggtitle("Distribution of Subjects'BMI") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Body mass index") +
  ylab("Density")

grid.arrange(p1, p2, nrow = 1, ncol = 2)
```

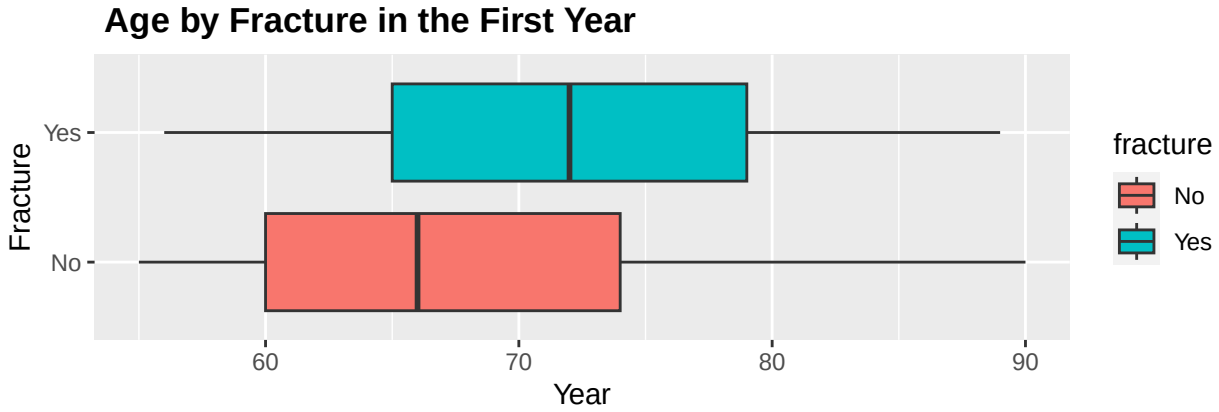


Bivariate Analysis

- Use boxplots to examine how the fracture is correlated with age and BMI.
  - The coord_flip() function is used to keep the dependent variable on the y-axis.

```
p3 <- df %>%
 ggplot(aes(fracture, bmi)) +
  geom_boxplot(aes(fill = fracture)) +
  coord_flip() +
  ggtitle("Subjects' BMI by Fracture in the First Year") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Body mass index") +
  xlab("Fracture")

p4 <- df %>%
 ggplot(aes(fracture, age)) +
  geom_boxplot(aes(fill = fracture)) +
  coord_flip() +
  ggtitle(" Age by Fracture in the First Year") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Year") +
  xlab("Fracture")

grid.arrange(p3, p4, nrow = 2, ncol = 1)
```

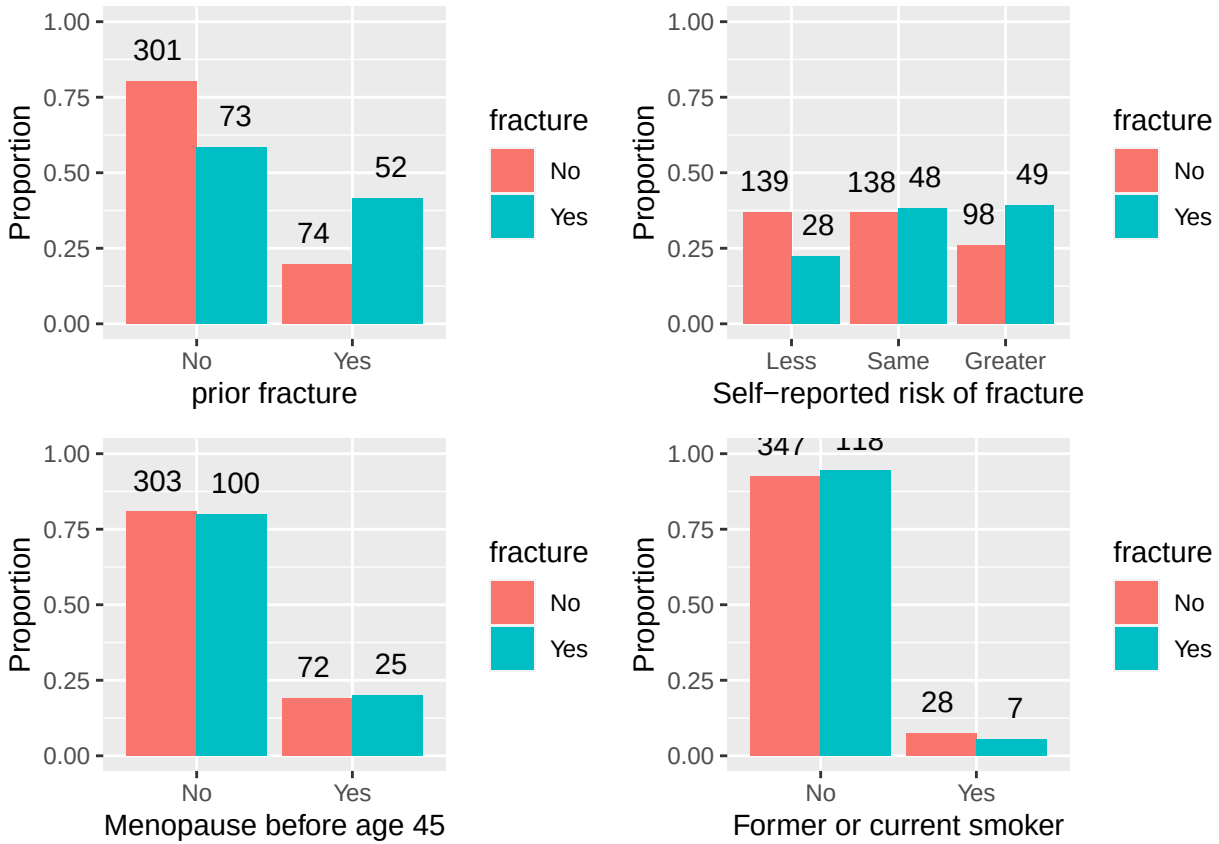





```
p5 <- df %>%
  ggplot(aes(x=priorfrac,
             y = ..prop..,
             group = fracture,
             fill = fracture)) +
  geom_bar( position = 'dodge') +
  geom_text(stat='count',
            aes(label=..count..),
            vjust=-1,
            position = position_dodge(width = 1)) +
  xlab("prior fracture") +
```

```
  ylab("Proportion") +
  ylim(0,1) +
  labs(fill = "fracture")

p6 <- df %>%
  ggplot(aes(x=raterisk,
             y = ..prop..,
             group = fracture,
             fill = fracture)) +
  geom_bar( position = 'dodge') +
  geom_text(stat='count',
            aes(label=..count..),
            vjust=-1,
            position = position_dodge(width = 1)) +
  xlab("Self-reported risk of fracture") +
  ylab("Proportion") +
  ylim(0,1) +
  labs(fill = "fracture")

p7 <- df %>%
  ggplot(aes(x= premeno,
             y = ..prop..,
             group = fracture,
             fill = fracture)) +
  geom_bar( position = 'dodge') +
  geom_text(stat='count',
            aes(label=..count..),
            vjust=-1,
            position = position_dodge(width = 1)) +
  xlab("Menopause before age 45") +
  ylab("Proportion") +
  ylim(0,1) +
  labs(fill = "fracture")

p8 <- df %>%
  ggplot(aes(x= smoke,
             y = ..prop..,
             group = fracture,
```

```
             fill = fracture)) +
  geom_bar( position = 'dodge') +
  geom_text(stat='count',
            aes(label=..count..),
            vjust=-1,
            position = position_dodge(width = 1)) +
  xlab("Former or current smoker") +
  ylab("Proportion") +
  ylim(0,1) +
  labs(fill = "fracture")

grid.arrange(p5, p6, p7, p8, nrow = 2, ncol = 2)
```


1.00
0.75
0.50
0.25
0.00
Proportion
301
73
74
52
No
Yes
prior fracture
fracture
No
Yes
139
28
138
48
98
49
Less
Same
Greater
Self−reported risk of fracture
303
100
72
25
Menopause before age 45
347
118
28
7
Former or current smoker

- Use the convenient summary_factorlist() function from the finalfit package to tabulate data. What do you learn from the EDA?

```
dependent <- "fracture"
explanatory <- c("bmi","age", "priorfrac", "premeno", "raterisk", "smoke")
df %>%
  summary_factorlist(dependent, explanatory, add_dependent_label = TRUE) %>%
  knitr::kable()
```

Dependent: fracture		No	Yes
bmi	Mean (SD)	27.5 (6.0)	27.7 (5.9)
age	Mean (SD)	67.5 (8.7)	71.8 (9.1)
priorfrac	No	301 (80.3)	73 (58.4)
	Yes	74 (19.7)	52 (41.6)
premeno	No	303 (80.8)	100 (80.0)
	Yes	72 (19.2)	25 (20.0)
raterisk	Less	139 (37.1)	28 (22.4)
	Same	138 (36.8)	48 (38.4)
	Greater	98 (26.1)	49 (39.2)
smoke	No	347 (92.5)	118 (94.4)
	Yes	28 (7.5)	7 (5.6)

Model Development

Simple Binary Logistic Regression

- Estimate the following base model and interpret the results.

$$logit(\pi_i) = \beta_0 + \beta_1 bmi + \beta_2 age + u$$

```
#mod.logit.1 <- # uncomment and replace with your code
#summary(mod.logit.1) # uncomment
```

- Recall:

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

- Find and interpret the estimated odds ratios for a 10-unit increase in age.

```
# Replace with your code
```

Categorical explantory variables

- First, check the levels attribute of priorfrac and raterisk
- Estimate the following model with three categorical variables and interpret the results.

$$logit(\pi_i) = \beta_0 + \beta_1 bmi + \beta_2 age + \beta_3 priorfrac + \beta_4 rateriskSame + \beta_5 rateriskGreater + u$$

```
#levels(df$priorfrac)
#levels(df$raterisk)

#set reference levels in factors to make interpretation easier
#df$priorfrac<-relevel(df$priorfrac, ref="No")
#df$raterisk<-relevel(df$raterisk, ref="Less")

#mod.logit.2 <- # uncomment and replace with your code
#summary(mod.logit.2) # uncomment
```

- Recall that for categorical explanatory variable:
  - Odds ratio comparing k level to reference level is:

$$OR = \frac{Odds_{x_k}}{Odds_{x_0}} = exp(\beta_k)$$

- and odds ratio comparing k level to another level like k-1 is:

$$OR = \frac{Odds_{x_k}}{Odds_{x_{k-1}}} = exp(\beta_k - \beta_{k-1})$$

- Find and interpret the estimated all odds ratios for prior risk and raterisk variable.

```
# uncomment and replace with your code

#odds_priorfracYes_No <- # uncomment and replace with your code
#odds_priorfracYes_No # uncomment

#oods_rateriskSame_Less <- # uncomment and replace with your code
#oods_rateriskSame_Less # uncomment

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

```
#oods_rateriskGreater_Less # uncomment

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

Interaction Terms

- What is the purpose of an interaction term?
- Estimate the following model with interaction terms between age and categorical variables and interpret the results.

$$logit(\pi_i) = \beta_0 + \beta_1 bmi + \beta_2 age + \beta_3 priorfrac + \beta_4 rateriskSame + \beta_5 rateriskGreater + \beta_6 age * priorfrac + \beta_7 age \cdot rateriskSame + \beta_8 age * rateriskGreater + u$$

```
#mod.logit.3 <-  # uncomment and replace with your code
#summary(mod.logit.3)
```

- Recall that for the following model with interaction term :

$$y = \beta_0 + \beta_1 * x_1 + ... + \beta_k * x_k + \beta_{k+1} * x_1 * x_k + u$$

$$OR = \frac{Odds_{x_k+c}}{Odds_{x_k}} = exp(c * (\beta_k + \beta_{k+1} * x_1))$$

- Find and interpret the odds ratio of 10 year increase in age for people with and without prior fracture.

```
#beta.hat <-  # uncomment and replace with your code

c <- 10
prior_fracture <- c(0,1)
#OR.age <-  # uncomment and replace with your code
```

Statistical Inference

Hypothesis Test

- Perform the likelihood ratio test comparing two models with and without BMI and age:raterisk.
  - $H_0 : \beta_{bmi} = \beta_{age:raterisk} = 0$
  - $H_a : \beta_{bmi} or \beta_{age:raterisk} \neq 0$

```
#mod.logit.4 <-  # uncomment and replace with your code
#summary(mod.logit.4)  # uncomment
#anova()  # uncomment and replace with your code
```

Confidence Interval

- Recall when:

$$OR = \frac{Odds_{x_k+c}}{Odds_{x_k}} = exp(c * (\beta_k + \beta_{k+1} * x_1))$$

- Then the $(1-\alpha)$ wald confidence interval is:

$$exp\left(c * (\widehat{\beta_k} + \widehat{\beta_{k+1}} * x_1) \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(c * (\widehat{\beta_k} + \widehat{\beta_{k+1}} * x_1))}\right)$$

- with

$$\widehat{Var}(c * (\hat{\beta}_k + \hat{\beta}_{k+1}x_1)) = c^2\widehat{Var}(\hat{\beta_k}) + c^2 * x_1^2 * \widehat{Var}(\widehat{\beta_{k+1}}) + c^2 * 2 * x_1 * \widehat{Cov}(\hat{\beta}_k, \hat{\beta}_{k+1})$$

- Use model.logit.4 and compute the odds ratio and wald confidence interval of prior fracture for 55, 65, 75, 85 years old women.

```
#beta.hat <-  # uncomment and replace with your code
c <- 1
age <- seq(from = 55, to = 85, by =10)

#OR.prior_fracture <-  # uncomment and replace with your code

#cov.mat <-  # uncomment and replace with your code

#var.log.OR <-  # uncomment and replace with your code
#ci.log.OR.low <-  # uncomment and replace with your code
#ci.log.OR.up <-  # uncomment and replace with your code
```

Final Visualization

- For women with great self-reported risk, plot the estimated logistic regression model with and without age and prior fracture interaction. Are there any interesting differences between the logistic regression model with and without the interaction term?

```
 # uncomment and run the code

#par(mfrow = c(1,2))

## models
# mod.logit.without <- glm(fracture ~ age + priorfrac + raterisk ,
#                          family = binomial(link = logit), data = df)
# mod.logit.with <- glm(fracture ~ age + priorfrac + raterisk + age:priorfrac,
#                       family = binomial(link = logit), data = df)
#

#### Without interaction term
# curve(expr = predict(object = mod.logit.without,
#                      newdata = data.frame(age = x, priorfrac = "No", raterisk= "Greater" ),
#                      type = "response"), col = "red", lty = "solid", xlim = c(50,100),
#      ylim = c(0,1), ylab = "Estimated probability", main = "Without Interaction",
#      xlab = "Age", panel.first = grid(col = "gray", lty = "dotted"),
#      cex.main = 0.9, lwd = 1)

# curve(expr = predict(object = mod.logit.without,
#                      newdata = data.frame(age = x, priorfrac = "Yes", raterisk= "Greater" ),
#                      type = "response"),col = "blue", lty = "dotdash", lwd = 1, add = TRUE)

# legend(x = 50, y = 0.9, legend = c("Prior fracture = 0", "Prior fracture  = 1"),
#        lty = c("solid", "dotdash"), col = c("red", "blue"), lwd = c(1,1), bty = "n")
#

##### with interaction term
# curve(expr = predict(object = mod.logit.with,
#                      newdata = data.frame(age = x, priorfrac = "No", raterisk= "Greater" ),
#                      type = "response"), col = "red", lty = "solid", xlim = c(50,100),
#      ylim = c(0,1), ylab = "Estimated probability", main = "With Interaction",
#      xlab = "Age", panel.first = grid(col = "gray", lty = "dotted"), cex.main = 0.9, lwd = 1)

# curve(expr = predict(object = mod.logit.with,
```

```
#                       newdata = data.frame(age = x, priorfrac = "Yes", raterisk= "Greater" ),
#                       type = "response"), col = "blue", lty = "dotdash", lwd = 1, add = TRUE)

# legend(x = 50, y = 0.9, legend = c("Prior fracture  = 0", "Prior fracture  = 1"),
#        lty = c("solid", "dotdash"), col = c("red", "blue"), lwd = c(1,1), bty = "n")
```

Final Report

- Display all estimated logistic models in a regression table. How robust are your results?

```
# uncomment and run the code
#
# stargazer(mod.logit.1, mod.logit.2, mod.logit.3, mod.logit.4,type = "text", omit.stat = "f",
#                     star.cutoffs = c(0.05, 0.01, 0.001), title = "Table 1:
#           The estimated relationship between risk of fracture and risk factors")
```

Reminders

1. Before next live session:
    1. Complete the homework that builds on this unit (HW-3)
    2. Complete all videos and reading for unit 4