Unit 3 Live Session

Discrete Response Model Part 3

Figure 1: South Hall

Class Announcements

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

Roadmap

Rearview Mirror

- Discuss why the classical linear regression model is not 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 terms
- 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 logistic regression model

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)

## to load glow500 from "Applied Logistic Regression" by D.W. Hosmer, S. Lemeshow, and R.X. Sturdivant (3rd ed., 2013)
library(aplore3)

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

## to work with "grid" graphics
library(gridExtra)

## To generate regression results, tables, and plots
library(finalfit)

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

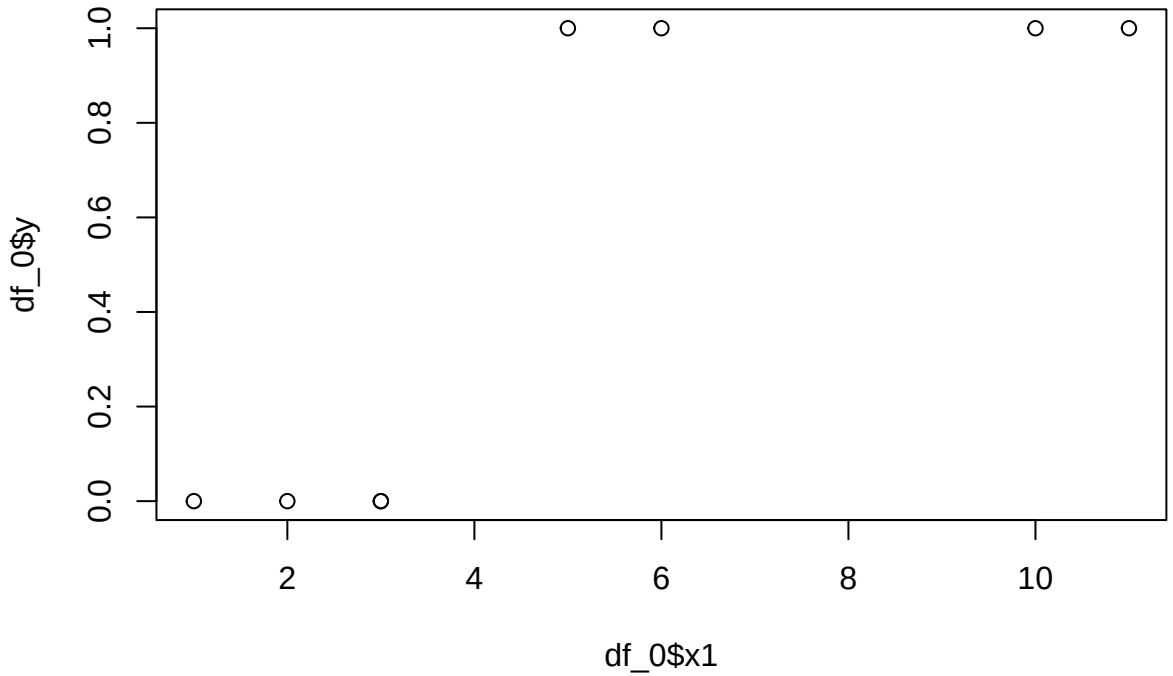
Discusion: Complete Separation

- What is complete separation in logistic regression?

A complete separation or sometimes also referred to as perfect prediction in logistic regression, happens when the response variable is completely separated by an explanatory variable or a linear combination of them.

- 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)
```



Here, observations with $Y = 0$ all have values of $X1 \leq 3$ and observations with $Y = 1$ all have values of $X1 > 3$. In other words X1 predicts Y perfectly, and we have complete separation

- 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)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial(link = logit), data = df_0)
##
## Deviance Residuals:
##          1           2           3           4           5           6
## -2.110e-08  -1.404e-05  -2.522e-06  -2.522e-06   1.564e-05   2.110e-08
##          7           8
##  2.110e-08   2.110e-08
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -66.098 183471.722   0.000        1
## x1              15.288  27362.843   0.001        1
## x2               6.241  81543.720   0.000        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.1090e+01  on 7  degrees of freedom
## Residual deviance: 4.5454e-10  on 5  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 24
```

glm() report a Warning message: "glm.fit: fitted probabilities numerically 0 or 1 occurred", which signals perfect separation as we can predict cases perfectly. Also, estimated coefficients have extremely large estimated standard deviations, and we can not interpret them reliably. This is due to the likelihood function not having a well defined optimum (we can increase the likelihood and therefore fit by increasing the coefficient on X1 to infinity).

If we're only interested in classifying response levels of y, complete separation is not necessarily a bad thing. But The problem is that the estimated model by maximum likelihood does not have a good interpretation or stable estimate.

- What are the techniques to deal with complete separation?

There are a few possible options: (1) collect more data to potentially find examples that remove complete separation, (2) modify the likelihood function by adding a penalized term akin to lasso regression for the binary case, or (3) drop the variables causing complete separation. Since finding which exact variables are causing the complete separation, the better solution is (2) and adding a penalized term to shrink the coefficients as in lasso or ridge regression.

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 that are post menopause, are at the greatest risk. A healthy diet, weight-bearing exercises, and medications can strengthen weak bones or prevent their loss. (Mayo Clinic)

Here, Our goal is description of the data:

- **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 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)
```

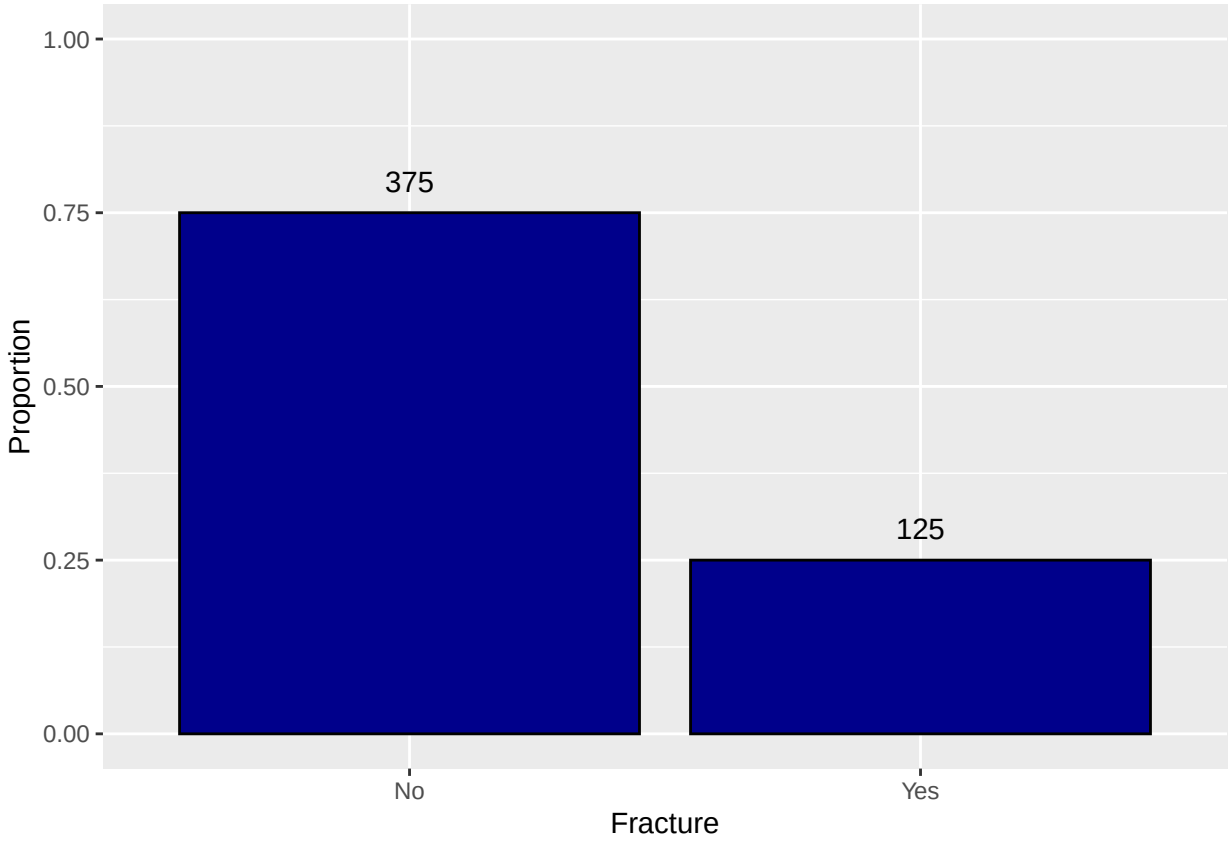
Univariate Analysis

- 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)
```



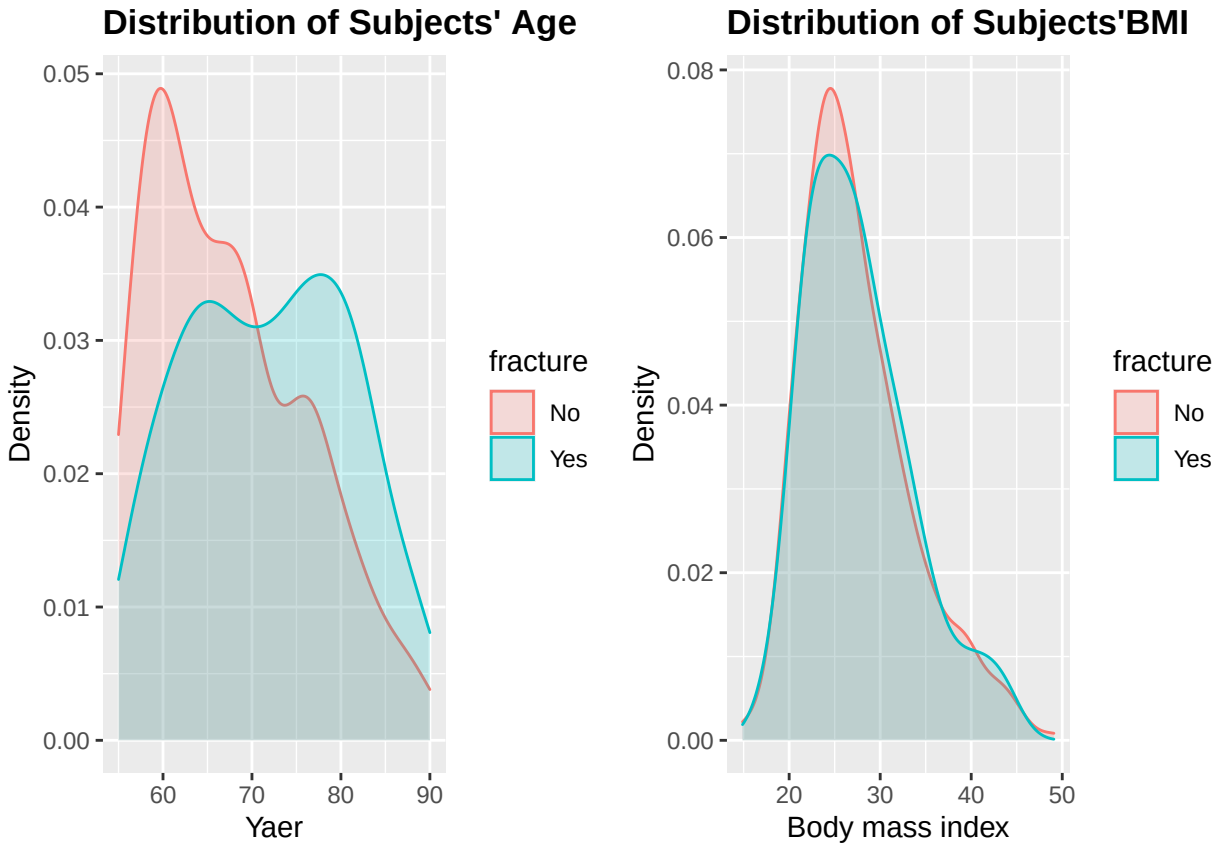
From the bar plot and the table, 25% of subjects in our sample suffered a fracture. If we didn't have any other information, 25% would be MLE of the probability of having a fracture in older women.

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)
```

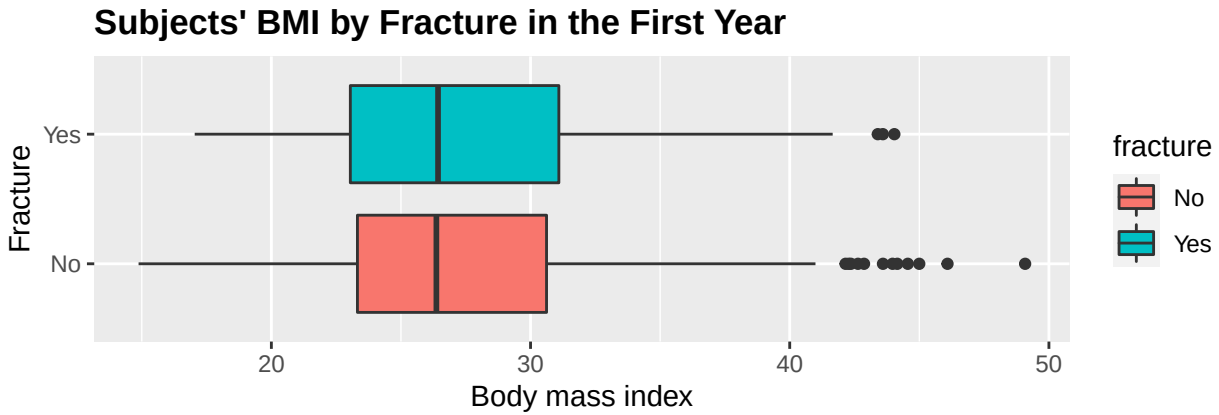


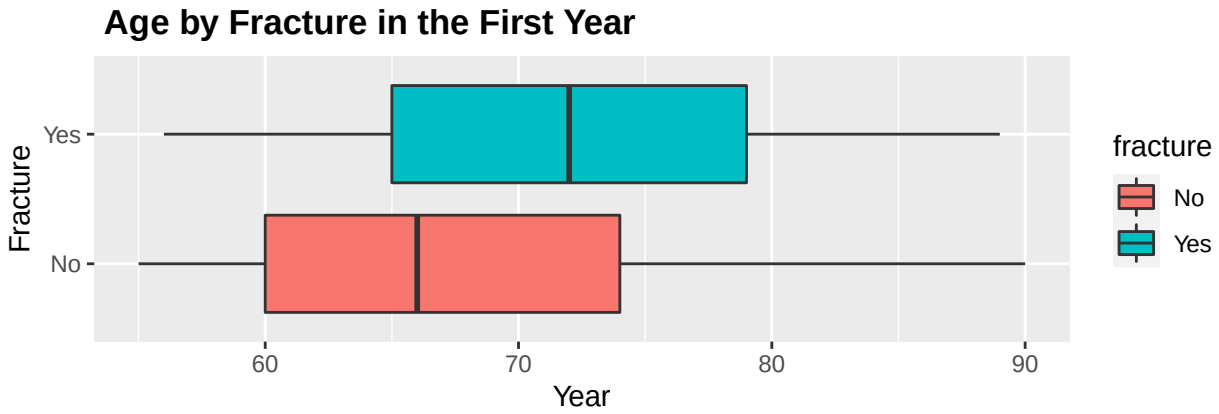
Age has a higher age in women with fractures than women without fractures. BMI distributions have almost the same mean and same variance in both groups with and without fracture, so probably BMI is not a useful variable to classify these two groups

Bivariate Analysis

- Use boxplots to examine how the fracture is correlated with age and BMI.
  - The coord_flip() function keeps 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)
```

From these box plots, we can see the women who suffered from a fracture are older, but both groups have the same distribution of BMI.

From the plots above, we see that the women with a history of prior fracture, and a high self-reported risk of fracture, have a higher probability of having a fracture in the first year of study. But, smokers and no smokers and women with or without menopause before 45 have the same probability of having a fracture. so smokers and menopause do not help classify these two groups, and we're not going to use them for modeling

- 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 <- glm(fracture ~ bmi + age, family = binomial(link = logit),
    data = df)

summary(mod.logit.1)
```

```
##
## Call:
## glm(formula = fracture ~ bmi + age, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.21426  -0.77408  -0.62995  -0.07905   2.02854
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.83441    1.10792  -5.266 1.39e-07 ***
## bmi          0.02692    0.01817   1.482    0.138
## age          0.05736    0.01211   4.735 2.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 562.34  on 499  degrees of freedom
## Residual deviance: 538.89  on 497  degrees of freedom
## AIC: 544.89
##
## Number of Fisher Scoring iterations: 4
```

As we expected from EDA, only the age coefficient is statistically significant. Age is positively correlated with the probability of having a fracture, and holding BMI constant, For one year increase in age, the log odds of having a fracture increases by 0.05 or 5%.

- 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.

```
round(cbind(exp(10 * coef(mod.logit.1)[3])), 2)
```

```
##     [,1]
## age 1.77
```

The estimated odds of having a fracture change by 1.77 times for every 10-year increase in age, or it's 77% higher

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)
```

```
## [1] "No"  "Yes"
```

```
levels(df$raterisk)
```

```
## [1] "Less"    "Same"    "Greater"
```

```
# 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 <- glm(fracture ~ bmi + age + priorfrac + raterisk,
    family = binomial(link = logit), data = df)

summary(mod.logit.2)
```

```
##
## Call:
## glm(formula = fracture ~ bmi + age + priorfrac + raterisk, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.42657  -0.75467  -0.59626  -0.04718   2.28787
##
## Coefficients:
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -6.17617    1.20026  -5.146 2.67e-07 ***
## bmi              0.02906    0.01877   1.549  0.12145
## age              0.05138    0.01302   3.945 7.98e-05 ***
## priorfracYes     0.66171    0.24306   2.722  0.00648 **
## rateriskSame     0.53943    0.27529   1.959  0.05005 .
## rateriskGreater  0.91817    0.28955   3.171  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 562.34  on 499  degrees of freedom
## Residual deviance: 516.53  on 494  degrees of freedom
## AIC: 528.53
##
## Number of Fisher Scoring iterations: 4
```

Here, six parallel lines are being estimated, one for each combination of priorfrac and raterisk, and each line has a different intercept but the same slope for age and BMI.

1- priorfrac = 0 and raterisk = Less

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

$\beta_0 = -6.17617$ **is log-odds in women with no prior fracture and less risk when age and BMI are zero**

2- priorfrac = 0 and raterisk = Same

$$logit(\pi_i) = (\beta_0 + \beta_4) + \beta_1 bmi + \beta_2 age$$

$\beta_0 + \beta_4$ **is log-odds in women with no prior fracture and the same risk when age and BMI are zero**

3- priorfrac = 0 and raterisk = Greater

$$logit(\pi_i) = (\beta_0 + \beta_5) + \beta_1 bmi + \beta_2 age$$

$\beta_0 + \beta_5$ **is log-odds in women with no prior fracture and greater risk, when age and BMI are zero**

4- priorfrac = 1 and raterisk =less

$$logit(\pi_i) = (\beta_0 + \beta_3) + \beta_1 bmi + \beta_2 age$$

$\beta_0 + \beta_3$ **is log-odds in women with prior fracture and less risk, when age and BMI are zero**

5- priorfrac = 1 and raterisk = Same

$$logit(\pi_i) = (\beta_0 + \beta_3 + \beta_4) + \beta_1 bmi + \beta_2 age$$

$\beta_0 + \beta_3 + \beta_4$ **is log-odds in women with prior fracture and same risk, when age and BMI are zero**

6- priorfrac = 1 and raterisk = Greater

$$logit(\pi_i) = (\beta_0 + \beta_3 + \beta_5) + \beta_1 bmi + \beta_2 age$$

$\beta_0 + \beta_3 + \beta_5$ **is log-odds in women with prior fracture and greater risk, when age and BMI are zero**

In all six models, for a one year increase in age, the log odds of having a fracture increase by 0.05 or 5%

- 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.

```
# since all except for last odds ratio compare to
# reference, we can estimate them using exponentiated
# coefficients from the glm output directly
round(exp(coef(mod.logit.2)), 2)
```

```
##     (Intercept)             bmi             age    priorfracYes    rateriskSame
##            0.00            1.03            1.05            1.94            1.72
## rateriskGreater
##            2.50
```

```
oods_rateriskGreater_Same <- round(exp(coef(mod.logit.2)[6] -
    coef(mod.logit.2)[5]), 2)
oods_rateriskGreater_Same
```

```
## rateriskGreater
##            1.46
```

The estimated odds of fracture are 94% higher in women with prior fracture v.s. women without prior fracture, where other variables are held constant

The estimated odds of having a fracture is 72% higher in women with the same risk v.s. women with less risk hold other variables constant.

The estimated odds of fracture change by 2.5 times for women with greater risk level v.s. less risk level, hold other variables constant

The estimated odds of having a fracture is 46% higher in women with greater risk vs. women with the same risk, holding other variables constant.

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.

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

```
mod.logit.3 <- glm(fracture ~ age + bmi + priorfrac + raterisk +
    age:priorfrac + age:raterisk, family = binomial(link = logit),
    data = df)


summary(mod.logit.3)
```

```
##
## Call:
## glm(formula = fracture ~ age + bmi + priorfrac + raterisk + age:priorfrac +
##     age:raterisk, family = binomial(link = logit), data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.32424  -0.77308  -0.57525   0.05465   2.54712
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -9.22115    2.08991  -4.412 1.02e-05 ***
## age                  0.09451    0.02658   3.555 0.000378 ***
## bmi                  0.02781    0.01913   1.454 0.145937
## priorfracYes         4.61668    1.85077   2.494 0.012615 *
## rateriskSame         2.45845    2.30002   1.069 0.285122
## rateriskGreater      3.51314    2.33771   1.503 0.132887
## age:priorfracYes    -0.05485    0.02558  -2.145 0.031991 *
## age:rateriskSame    -0.02626    0.03159  -0.831 0.405854
```

```
## age:rateriskGreater -0.03606    0.03239  -1.113 0.265630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 562.34  on 499  degrees of freedom
## Residual deviance: 510.01  on 491  degrees of freedom
## AIC: 528.01
##
## Number of Fisher Scoring iterations: 5
```

Here, we estimated six lines with six different intercepts and six different slopes of age, one for each combination of priorfrac and raterisk.

1- priorfrac = 0 and raterisk = Less

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

β_2 **is an increase in log-odds for a year increase in age in women with no prior fracture and less risk**

2- priorfrac = 0 and raterisk = Same

$$logit(\pi_i) = (\beta_0 + \beta_4) + \beta_1 bmi + (\beta_2 + \beta 7) age$$

$\beta_2 + \beta_7$ **is an increase in log-odds for a year increase in age in women with no prior fracture and the same risk**

3- priorfrac = 0 and raterisk = Greater

$$logit(\pi_i) = (\beta_0 + \beta_5) + \beta_1 bmi + (\beta_2 + \beta_8) age$$

$\beta_2 + \beta_8$ **is an increase in log-odds for a year increase in age in women with no prior fracture and greater risk**

4- priorfrac = 1 and raterisk =less

$$logit(\pi_i) = (\beta_0 + \beta_3) + \beta_1 bmi + (\beta_2 + \beta_6) age$$

$\beta_2 + \beta_6$ **is an increase in log-odds for a year increase in age in women with prior fracture and less risk**

5- priorfrac = 1 and raterisk = Same

$$logit(\pi_i) = (\beta_0 + \beta_3 + \beta_4) + \beta_1 bmi + (\beta_2 + \beta_6 + \beta_7) age$$

$\beta_2 + \beta_6 + \beta_7$ **is an increase in log-odds for a year increase in age in women with prior fracture and same risk**

6- priorfrac = 1 and raterisk = Greater

$$logit(\pi_i) = (\beta_0 + \beta_3 + \beta_5) + \beta_1 bmi + (\beta_2 + \beta_6 + \beta_8) age$$

$\beta_2 + \beta_6 + \beta_8$ **is an increase in log-odds for a year increase in age in women with prior fracture and Greater risk**

In all six models, BMI has the same effects on log-odds

- 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 a 10-year increase in age for people with and without prior fracture.

```
beta.hat <- mod.logit.3$coefficients
# beta.hat

c <- 10
prior_fracture <- c(0, 1)
log.OR.age <- c * (beta.hat[2] + beta.hat[7] * prior_fracture)
OR.age <- exp(log.OR.age)
round(data.frame(prior_fracture = prior_fracture, OR.hat = OR.age),
    2)
```

```
##   prior_fracture OR.hat
## 1              0   2.57
## 2              1   1.49
```

The odds of having a fracture change by 2.58 times for a 10-year increase in age in women without prior fracture and by 1.52 times in women with the previous fracture. The odds ratio of 10 years increase in age is smaller in women with the previous fracture.

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 <- glm(fracture ~ age + priorfrac + raterisk + age:priorfrac,
    family = binomial(link = logit), data = df)
summary(mod.logit.4)
```

```
##
## Call:
## glm(formula = fracture ~ age + priorfrac + raterisk + age:priorfrac,
##     family = binomial(link = logit), data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.21673  -0.77435  -0.56079  -0.00669   2.36207
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.55156    1.13990  -5.747 9.06e-09 ***
## age               0.06831    0.01578   4.329 1.50e-05 ***
## priorfracYes      5.01720    1.83182   2.739  0.00616 **
## rateriskSame      0.56411    0.27621   2.042  0.04112 *
## rateriskGreater   0.88764    0.28809   3.081  0.00206 **
## age:priorfracYes -0.06015    0.02534  -2.374  0.01762 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 562.34  on 499  degrees of freedom
## Residual deviance: 513.27  on 494  degrees of freedom
## AIC: 525.27
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
anova(mod.logit.4, mod.logit.3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: fracture ~ age + priorfrac + raterisk + age:priorfrac
## Model 2: fracture ~ age + bmi + priorfrac + raterisk + age:priorfrac +
##     age:raterisk
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       494     513.27
## 2       491     510.01  3   3.2585   0.3535
```

As the p-value is large, exceeding 0.05, we fail to reject the null hypothesis that BMI and interaction terms between age and risk risk are different from zero. Because BMI and this interaction term are both individually and together insignificant, we remove them from the model

Confidence Interval

- Recall when:

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

- Then $(1-\alpha)$ 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 <- mod.logit.4$coefficients
# c is 1 since is either prior fraction = Yes or prior
# fracture = No
c <- 1
# fixed age levels we are interested in i.e. x1
age <- seq(from = 55, to = 85, by = 10)

log.OR.prior_fracture <- c * (beta.hat[3] + beta.hat[6] * age)
OR.prior_fracture <- exp(log.OR.prior_fracture)

cov.mat <- vcov(mod.logit.4)
var.log.OR <- c^2 * (cov.mat[3, 3] + age^2 * cov.mat[6, 6] +
    2 * age * cov.mat[3, 6])

ci.log.OR.low <- exp(log.OR.prior_fracture - qnorm(p = 0.975,
    ) * sqrt(var.log.OR))
```

```r
ci.log.OR.up <- exp(log.OR.prior_fracture + qnorm(p = 0.975,
    ) * sqrt(var.log.OR))

round(data.frame(age = age, OR.hat = OR.prior_fracture, OR.low = ci.log.OR.low,
    OR.up = ci.log.OR.up), 2)
```

```
##   age OR.hat OR.low OR.up
## 1  55   5.52   2.14 14.27
## 2  65   3.03   1.71  5.35
## 3  75   1.66   1.01  2.71
## 4  85   0.91   0.41  2.04
```

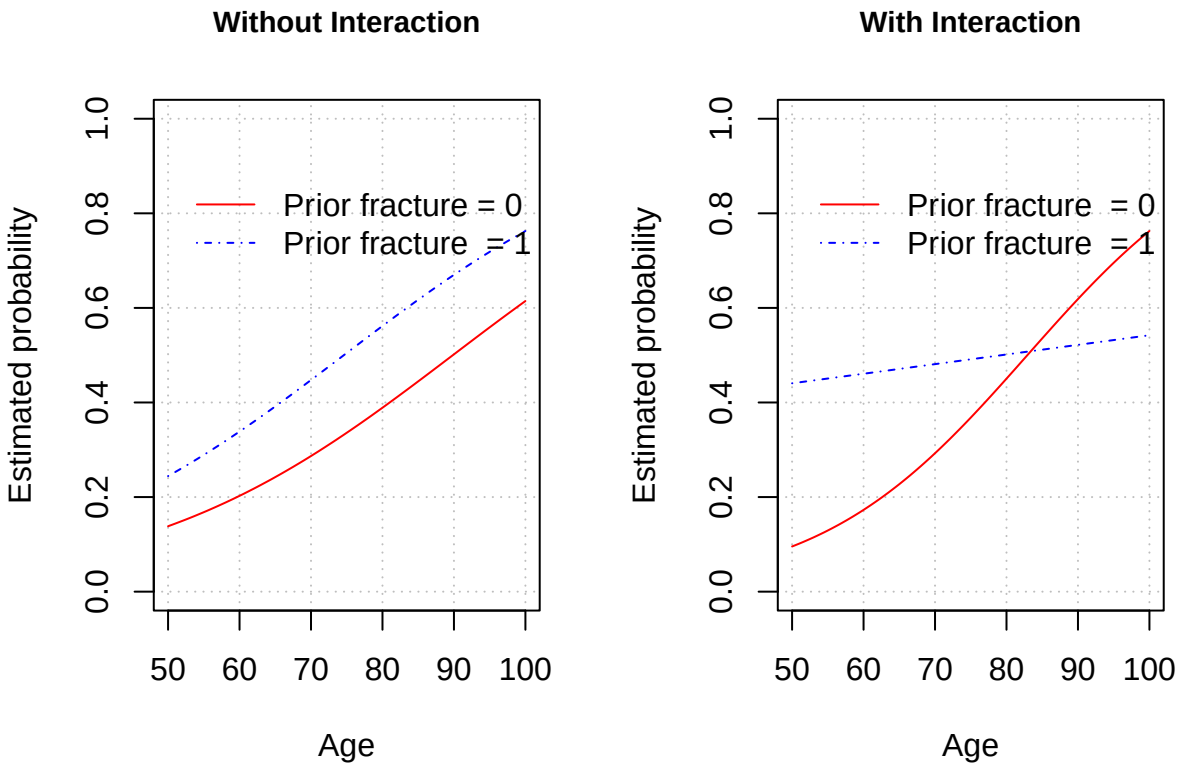
With 95% confidence, the odds of having a fracture change by an amount between 2.14 to 14.27 times in women with prior fractures verse women without previous fractures for 55 years-old women. The odd ratio of previous fracture and their CI decrease as women get older, which indicate that prior fracture has a smaller effect on a fracture as women get older.

Final Visualization

Plot the estimated logistic regression model with and without age and prior fracture interaction for women with greater self-reported risk. Are there any interesting differences between the logistic regression model with and without the interaction term?

```
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")
```



In the left plot without interaction term, the estimated probability of having a fracture is always greater for women with prior fractures. But, in the right plot, we can see that fracture probability is higher for women with a previous fracture below the age of 84 or 85. But the red curve has a higher slope, the probability of fracture increases faster in women without the prior fracture, and as a result, they have a higher probability of fracture after age 84 or 85.

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")
```

```
##
## Table 1
## ============================================================
##                                Dependent variable:
##                     ---------------------------------------
##                                      fracture
##                         (1)       (2)       (3)       (4)
## ------------------------------------------------------------
## bmi                    0.027     0.029     0.028
##                       (0.018)   (0.019)   (0.019)
##
## age                  0.057***  0.051***  0.095***  0.068***
##                       (0.012)   (0.013)   (0.027)   (0.016)
##
## priorfracYes                    0.662**   4.617*    5.017**
##                                 (0.243)   (1.851)   (1.832)
##
## rateriskSame                     0.539     2.458    0.564*
##                                 (0.275)   (2.300)   (0.276)
##
## rateriskGreater                 0.918**    3.513    0.888**
##                                 (0.290)   (2.338)   (0.288)
##
## age:priorfracYes                          -0.055*   -0.060*
##                                           (0.026)   (0.025)
##
## age:rateriskSame                          -0.026
##                                           (0.032)
```

```
##
## age:rateriskGreater                      -0.036
##                                          (0.032)
##
## Constant            -5.834*** -6.176*** -9.221*** -6.552***
##                      (1.108)   (1.200)   (2.090)   (1.140)
##
## ------------------------------------------------------------
## Observations           500       500       500       500
## Log Likelihood      -269.447  -258.264  -255.006  -256.635
## Akaike Inf. Crit.    544.894   528.529   528.012   525.270
## ============================================================
## Note:                          *p<0.05; **p<0.01; ***p<0.001
```

In all models, coefficients of age and priorfracYes are statistically significant with positive coefficients, which is a sign of the robustness of these effects.

Reminders

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