Unit 1 Live Session (Solution)

Figure 1: South Hall

Introduction

Instructor Introduction

- Welcome to the W271!
- Greeting/Introduction

Weekly Workflow

- A "typical" week in the course will proceed in the following way:

1. **Before class**: Complete all asynchronous materials. This means, watch all videos and complete all readings and concept checks.
2. **In live session**: We will work on exercises and coding that will extend your understanding.

3. **After live session**: Complete the homework for the unit. This homework will be due before the next live session.

Resources

- ISVC
- Gradescope
- Github Org
- Slack

Student Introductions

Please take 90 seconds to tell us:

- Where you dial in from
- What kind of work you do / are interested in

What can you learn in this class?

Part-1: Discrete Response Variable

- Cross-sectional data
- Binary, Unordered Multiclass, Ordered Multiclass, and Count data

Binary Data	Unordered and Ordered Multiclass	Count data
Binomial Probability Model: 1 variable	Multinomial Probability Model: 1 variable	Poisson Probability Model: 1 variable
Binomial Probability Model: 2 variables	Multinomial Probability Model: 2 variables	Poisson Probability Model: 2 variables
Binomial Probability Model: N variables	Multinomial Probability Model: N variables	Poisson Probability Model: N variables
Binary Logistic Regression	Multinomial Logistic Regression	Possion Logistic Regression
	Ordinal Logistic Regression	

Part-2: Time Series Data

- Univariate and Multivariate Time Series

Basic Concepts	Modeling Trend and Seasonality	Modeling
Time series	Trend	ARMA
Stochastic process	Seasonality	ARIMA
Stationarity	Both trend and Seasonality	SARIMA
Forecasting		Cointegration, VAR

Part-3: Panel Data

- Data with both temporal and cross-sectional dimensions

Modeling	Modeling	Modeling
OLS: Ignoring the panel structure	Fixed Effect Models	Mixed Effect Models
OLS: For independent cross-sections	Random Effect Models	
Pooled OLS		
First Difference Models		

Roadmap

w203 Review

- Postulate a statistical model that conforms with the underlying (business, policy, scientific, etc.) question being asked
- Estimate the parameter of the statistical model
- Check model assumptions
- Conduct statistical inference

Today

- Computing Odds and examining how odds are related to probabilities
- Estimating the Bernoulli probability model using maximum likelihood estimation (MLE)

Looking Ahead

- Linear Probability Model, its advantages, and its limitations
- Binary Logistic Regression Model: Estimation and Inference

Start-up Code

```
# Insert the function to *tidy up* the code when they are printed out
library(knitr)

# 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)
```

- Why do we need these packages?

Getting Used to Odds and Log-Odds

The odds, o, is an important concept for modeling discrete outcomes. For a Bernoulli random variable with parameter p, the odds are defined as the ratio of the probability of success to the probability of failure, $\frac{p}{1-p}$.

We often work with the log of the odds, or log-odds, written $ln\left(\frac{p}{1-p}\right)$. This is also known as $logit(p)$.

- Suppose you have a fair coin, represented as a Bernoulli random variable with parameter 1/2. Compute the log-odds of success.
  - Now your challenge is to go in reverse. Given log-odds x, what is the probability of heads p?

$$x = log(\frac{p}{1-p})$$

$$exp(x) = \frac{p}{1-p}$$

$$(1-p) * exp(x) = p$$

$$exp(x) = p * (1 + exp(x))$$

$$p = \frac{exp(x)}{1 + exp(x)}$$

- Write an R function that computes the probability of heads, given log-odds.

```
log.odds.to.prob <- function(x){
    p = exp(x)/(1+exp(x))
    return(p)
}

log.odds.to.prob(0)
```

```
## [1] 0.5
```

Your aunt offers a service in which she weights coins to make them unfair. You give her a coin and tell her how much you want the log-odds to change. She returns the modified coin.

- For each of the following orders, use your function to compute the resulting probability of heads:
  - fair coin, increase log-odds by 1.
  - fair coin, increase log-odds by 2.
  - fair coin, increase log-odds by 10.
  - fair coin, decrease log-odds by 1.
  - fair coin, decrease log-odds by 2.
  - fair coin, decrease log-odds by 10.

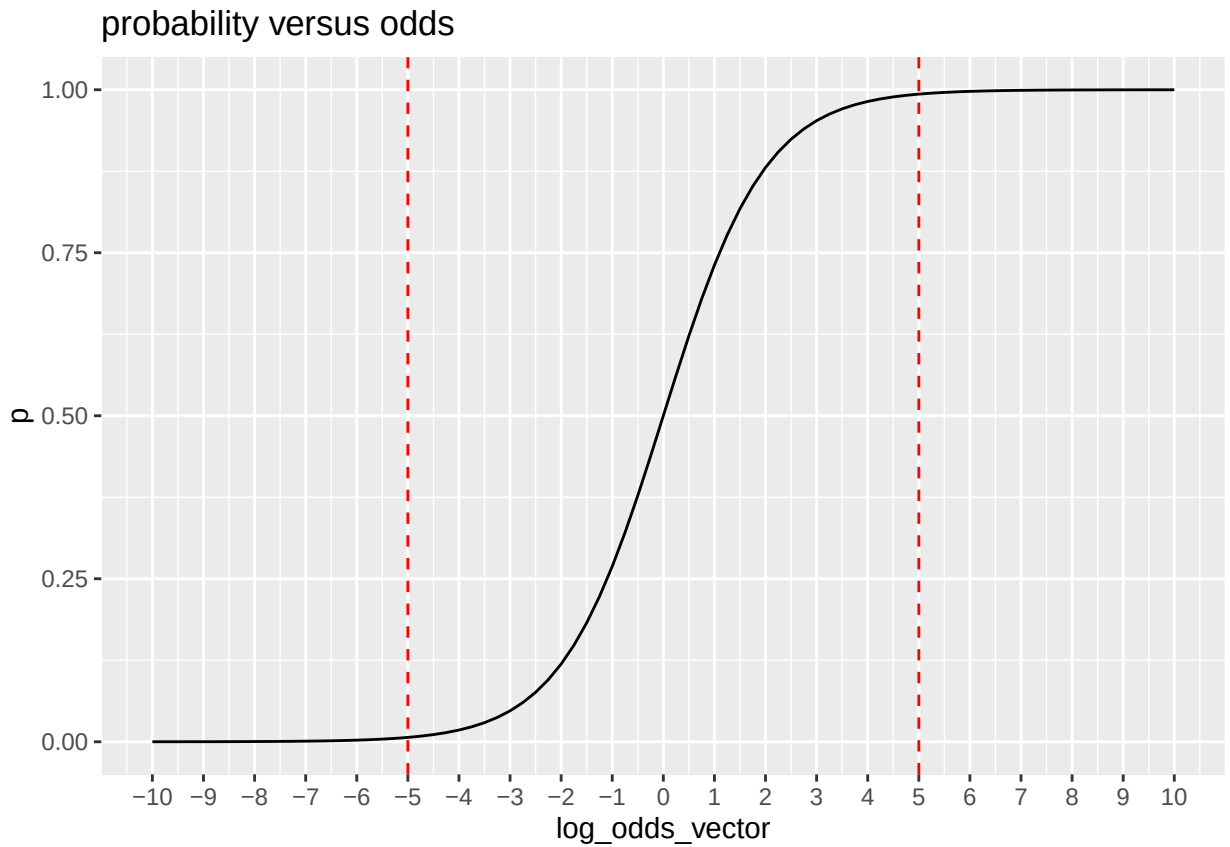
```
log_odds <- c(10,2,1, 0, -1, -2, -10 )

data.frame(log_odds = log_odds,  probability = round(log.odds.to.prob(log_odds),3))
```

```
##   log_odds probability
## 1       10       1.000
## 2        2       0.881
## 3        1       0.731
## 4        0       0.500
## 5       -1       0.269
## 6       -2       0.119
## 7      -10       0.000
```

- In you own words, describe how changes in log-odds translate to changes in probability

```
log_odds_vector = seq(from = -10, to = 10, by = 0.25)
p = log.odds.to.prob(log_odds_vector)
d = data.frame(log_odds_vector, p)
ggplot(d, aes(x = log_odds_vector, y = p)) +
geom_line() +
geom_vline(aes(xintercept = c(-5)), color = "red", linetype = "dashed")+
geom_vline(aes(xintercept = c(5)), color = "red", linetype = "dashed")+
scale_x_continuous(breaks = seq(-10, 10, by = 1)) +
labs(title = "probability versus odds")
```



- You can see in this plot, As log-odds increase, the probability of success increases relative to the probability of failure, and it approaches one. As log-odds decrease probability of success decrease and converges to zero.
- If you get log-odds values that are very very small like -10 the probability of success is almost zero, and if you get log-odds values that are very big like 10 or the probability of success is almost one.
- The relationship between log-odd and probability is not linear, but of s-curve type, and log odds ratios ranging from -5 to +5 create probabilities that range from just above 0 to very close to 1.

Maximum Likelihood Estimation

Suppose your aunt sends you an unfair coin, but you forgot what your order was. To figure out the probability of success, you flip the coin three times and collect the following data (we are defining heads as success here):

HTH

- For a hypothesized Bernoulli parameter π, what is the likelihood of the data? Your answer should be a function of π.
- likelihood function is:

$$\begin{aligned} L(\pi|x_1, x_2, x_3) &= P(X_1 = x_1, X_1 = x_2, X_3 = x_3) \\ &= \prod_{i=1}^{3} P(X = x_i) \\ &= \prod_{i=1}^{3} \pi^{x_i}(1-\pi)^{1-x_i} \\ &= \pi^{\sum_{i=1}^{3} x_i}(1-\pi)^{\sum_{i=1}^{3}(1-x_i)} \end{aligned}$$

- log of the likelihood function

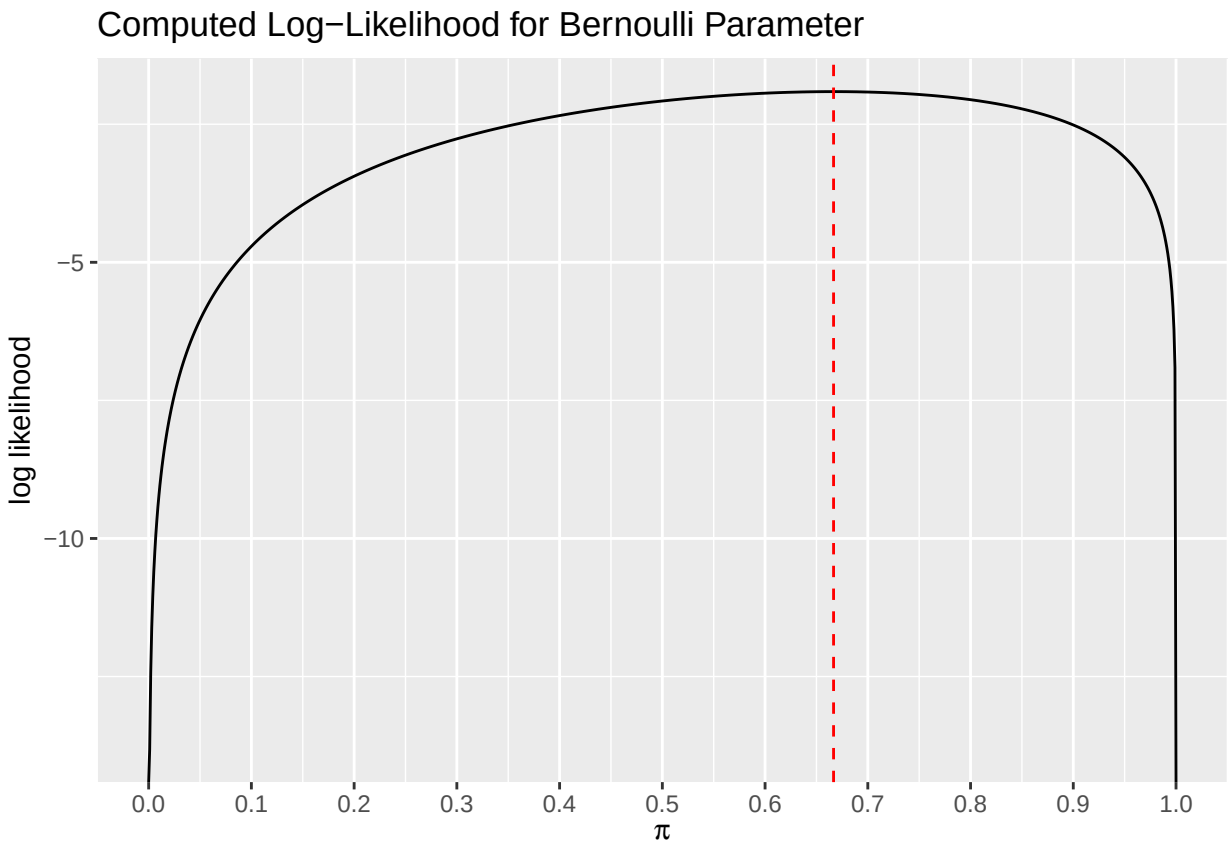
$$Log[L(\pi|x_1, x_2, x_3)] = \left(\sum_{i=1}^{3} x_i\right) log(\pi) + \left(\sum_{i=1}^{3}(1-x_i)\right) log(1-\pi)$$

- What is the natural log of the likelihood of the data? Write an R function that computes the log likelihood.

```
loglikelihood <- function(pi) {
    data <- c(1, 0, 1)
    return(sum(data==1)*log(pi) + (sum(data==0)*log(1-pi)))}
```

- Graph your function and visually estimate what the maximum likelihood estimate for π is.

```
prob = seq(0, 1, by=.001)
d1 <- data.frame(probability = prob, log_likelihood = loglikelihood(prob))
ggplot(d1, aes(x = probability, y = log_likelihood)) +
geom_line() +
geom_vline(aes(xintercept = c(2/3)), color = "red", linetype = "dashed")+
scale_x_continuous(breaks = seq(0, 1, by = 0.1)) +
labs(title = "Computed Log-Likelihood for Bernoulli Parameter",
     x = quote(pi),
     y = 'log likelihood'
     )
```



- We know that MLE of π is:

$$\hat{\pi} = \frac{\sum x_i}{N} = \frac{2}{3}$$

- and in this question, it's:

$$\hat{\pi} = \frac{2}{3}$$

- In the plot, we can see that log-likelihood has a single peak at 2/3.

Reminders

1. Welcome!
2. Before the next live session:
    1. Complete the homework that builds on this unit
    2. Complete all videos and reading for unit 2
    3. Fill out the group assignment questionnaire
    4. Try to look at the live session two plan, install the required packages, and check the case study section, especially the introduction, data description and descriptive Statistics.

Good luck getting started!