ECO663
Data Analysis
of
Experiment/Survey

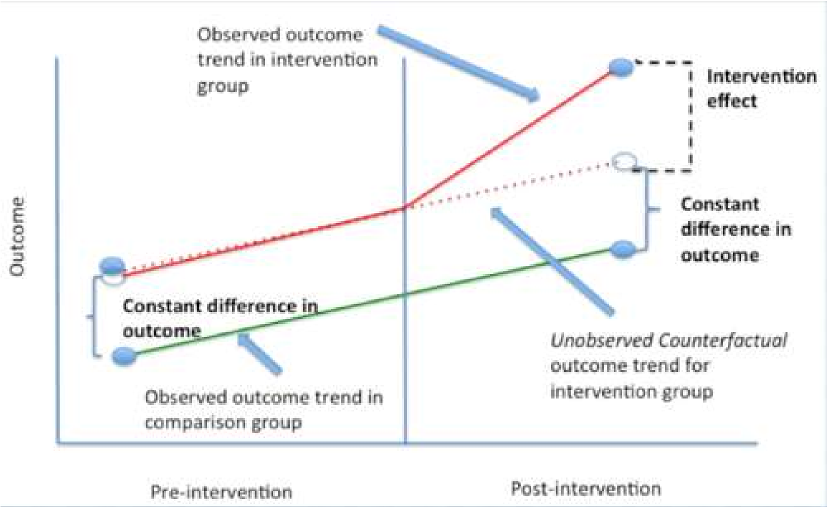
Week 11

ANOVA (Analysis of Variance)

- One-Way ANOVA
- Two-Way ANOVA
- Three-Way ANOVA
- Tukey’s Test

- Refer to the handout.

Difference in Differences


Observed outcome trend in intervention group
Intervention effect
Constant difference in outcome
Outcome
Constant difference in outcome
Unobserved Counterfactual outcome trend for intervention group
Observed outcome trend in comparison group
Pre-intervention
Post-intervention

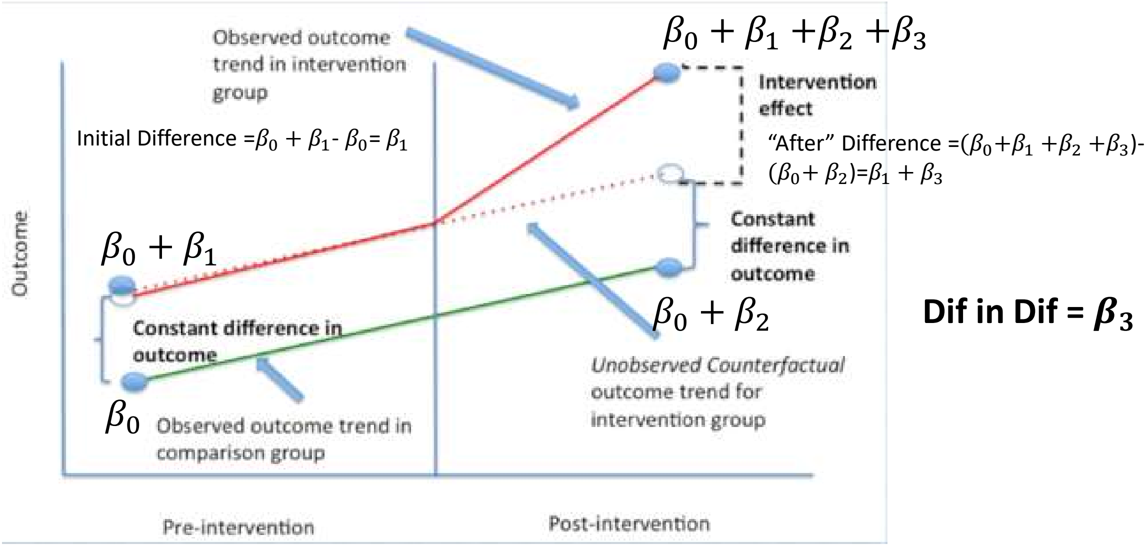
$$y = \beta_0 + \beta_1 dtreat + \beta_2 dtime + \beta_3 dtreat * dtime + \boldsymbol{\delta X} + \epsilon$$

dtreat: dummy 1: treatment group, 0: control group

dtime: dummy 1: after, 0: before

X: other determinants

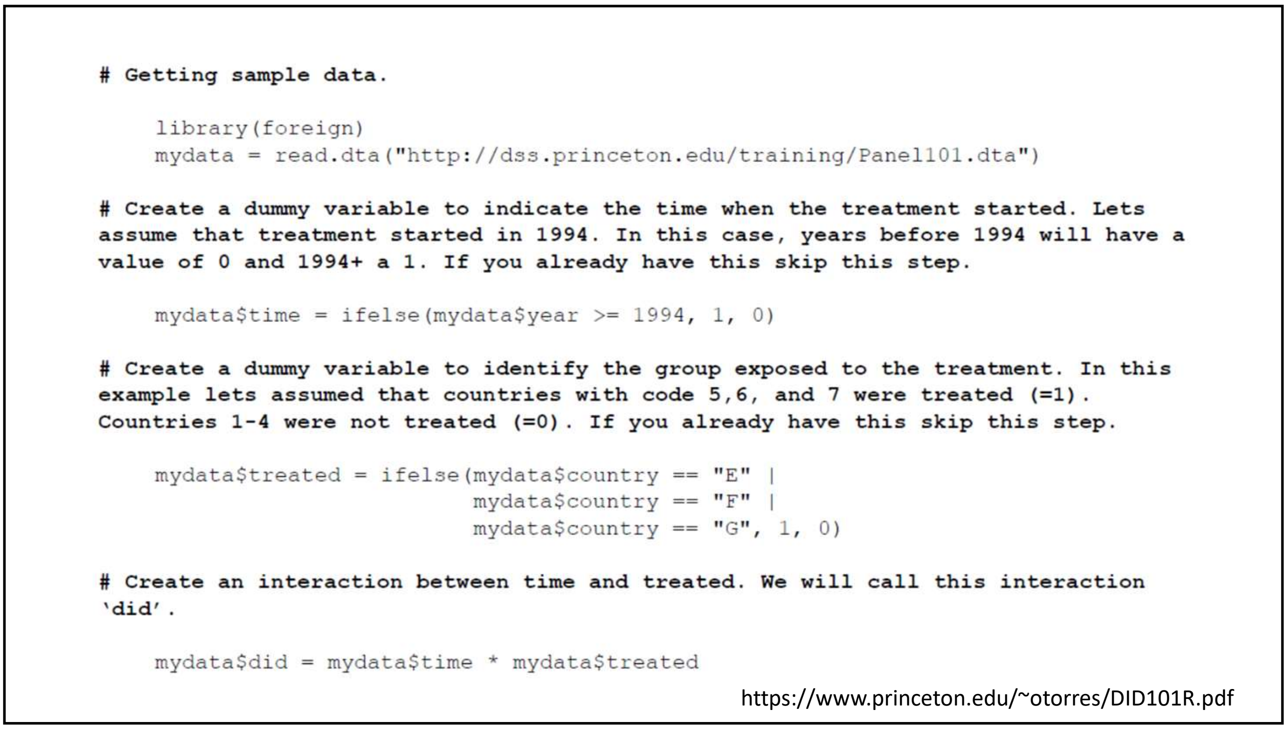
$$y = \beta_0 + \beta_1 dtreat + \beta_2 dtime + \beta_3 dtreat * dtime(+\boldsymbol{\delta X}) + \epsilon$$


Observed outcome trend in intervention group
$\beta_0 + \beta_1 + \beta_2 + \beta_3$
Intervention effect
Initial Difference =$\beta_0 + \beta_1$- β_0= β_1
"After" Difference =(β_0+β_1 +β_2 +β_3)-(β_0+ β_2)=$\beta_1 + \beta_3$
Outcome
$\beta_0 + \beta_1$
Constant difference in outcome
$\beta_0 + \beta_2$
Constant difference in outcome
β_0
Observed outcome trend in comparison group
Unobserved Counterfactual outcome trend for intervention group
Pre-intervention
Post-intervention

Dif in Dif = $\boldsymbol{\beta_3}$

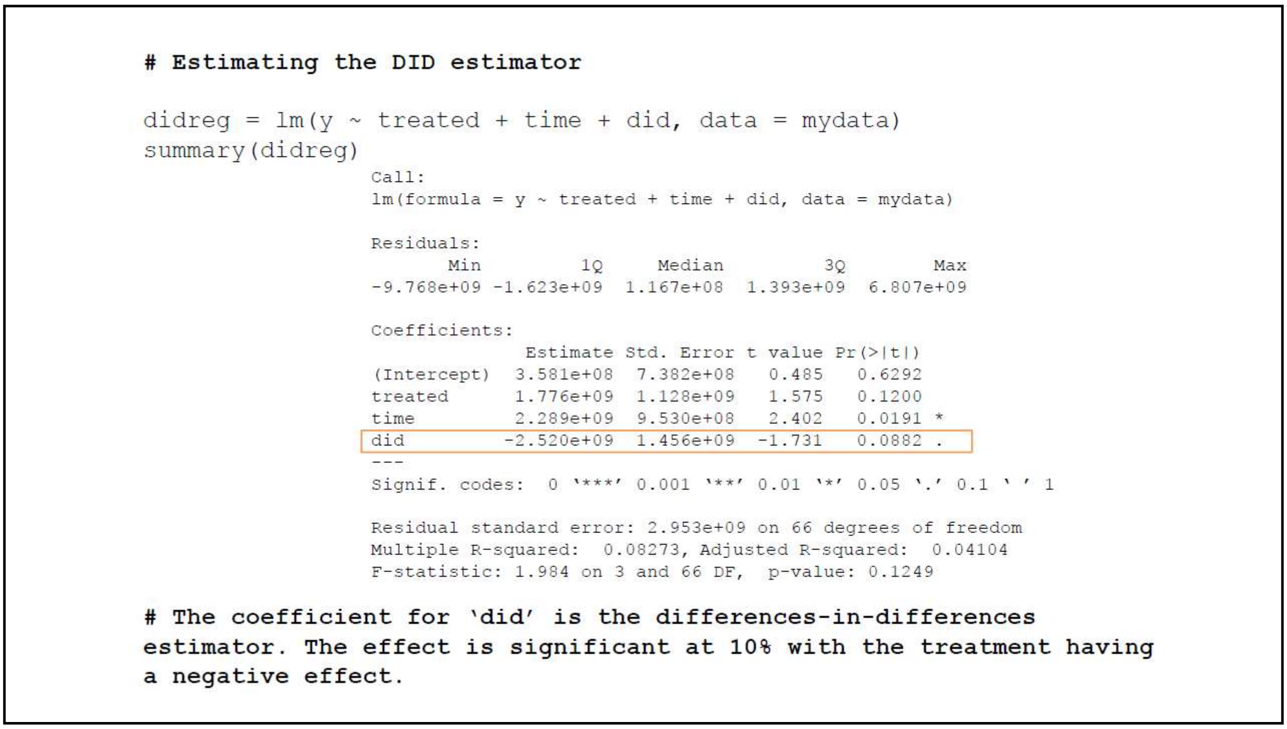
Dif in Dif Estimation using R

- https://dss.princeton.edu/training/
- For DifinDif and Logit models, lecture notes from the above site was referenced.



```
# Getting sample data.

    library(foreign)
    mydata = read.dta("http://dss.princeton.edu/training/Panel101.dta")

# Create a dummy variable to indicate the time when the treatment started. Lets
assume that treatment started in 1994. In this case, years before 1994 will have a
value of 0 and 1994+ a 1. If you already have this skip this step.

    mydata$time = ifelse(mydata$year >= 1994, 1, 0)

# Create a dummy variable to identify the group exposed to the treatment. In this
example lets assumed that countries with code 5,6, and 7 were treated (=1).
Countries 1-4 were not treated (=0). If you already have this skip this step.

    mydata$treated = ifelse(mydata$country == "E" |
                            mydata$country == "F" |
                            mydata$country == "G", 1, 0)

# Create an interaction between time and treated. We will call this interaction
'did'.

    mydata$did = mydata$time * mydata$treated
```

https://www.princeton.edu/~otorres/DID101R.pdf

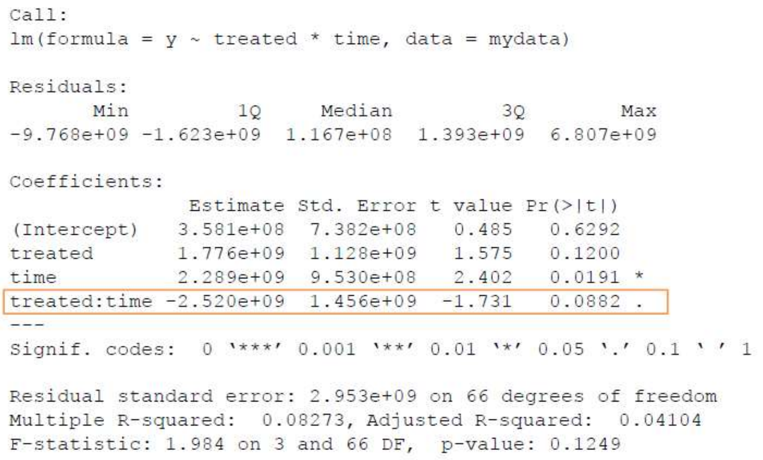


```
# Estimating the DID estimator

didreg = lm(y ~ treated + time + did, data = mydata)
summary(didreg)
```

```
Call:
lm(formula = y ~ treated + time + did, data = mydata)

Residuals:
       Min         1Q     Median         3Q        Max 
-9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  3.581e+08  7.382e+08   0.485   0.6292  
treated      1.776e+09  1.128e+09   1.575   0.1200  
time         2.289e+09  9.530e+08   2.402   0.0191 *
did         -2.520e+09  1.456e+09  -1.731   0.0882 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.953e+09 on 66 degrees of freedom
Multiple R-squared:  0.08273, Adjusted R-squared:  0.04104 
F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
```

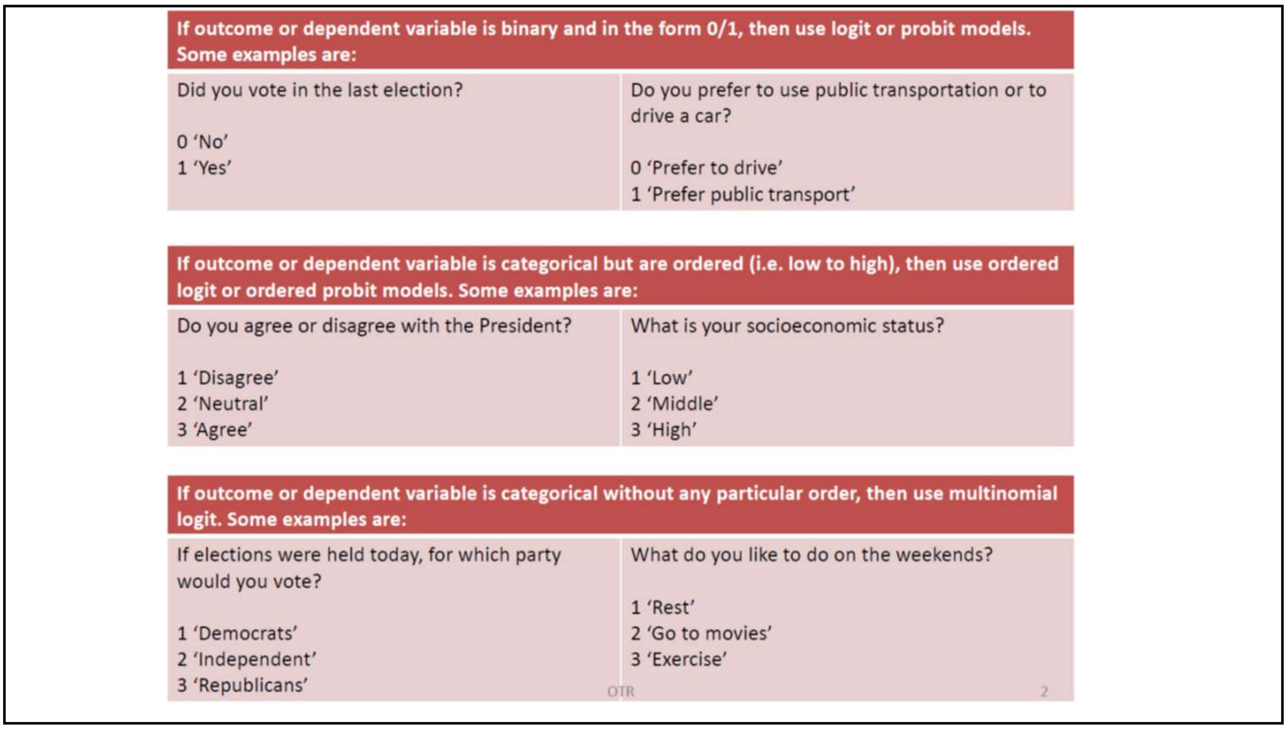
```
# The coefficient for 'did' is the differences-in-differences
estimator. The effect is significant at 10% with the treatment having
a negative effect.
```

```
# Estimating the DID estimator (using the multiplication method, no
need to generate the interaction)

didreg1 = lm(y ~ treated*time, data = mydata)
summary(didreg1)
```



```
Call:
lm(formula = y ~ treated * time, data = mydata)

Residuals:
       Min         1Q     Median         3Q        Max
-9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.581e+08  7.382e+08   0.485   0.6292
treated       1.776e+09  1.128e+09   1.575   0.1200
time          2.289e+09  9.530e+08   2.402   0.0191 *
treated:time -2.520e+09  1.456e+09  -1.731   0.0882 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.953e+09 on 66 degrees of freedom
Multiple R-squared:  0.08273, Adjusted R-squared:  0.04104
F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
```

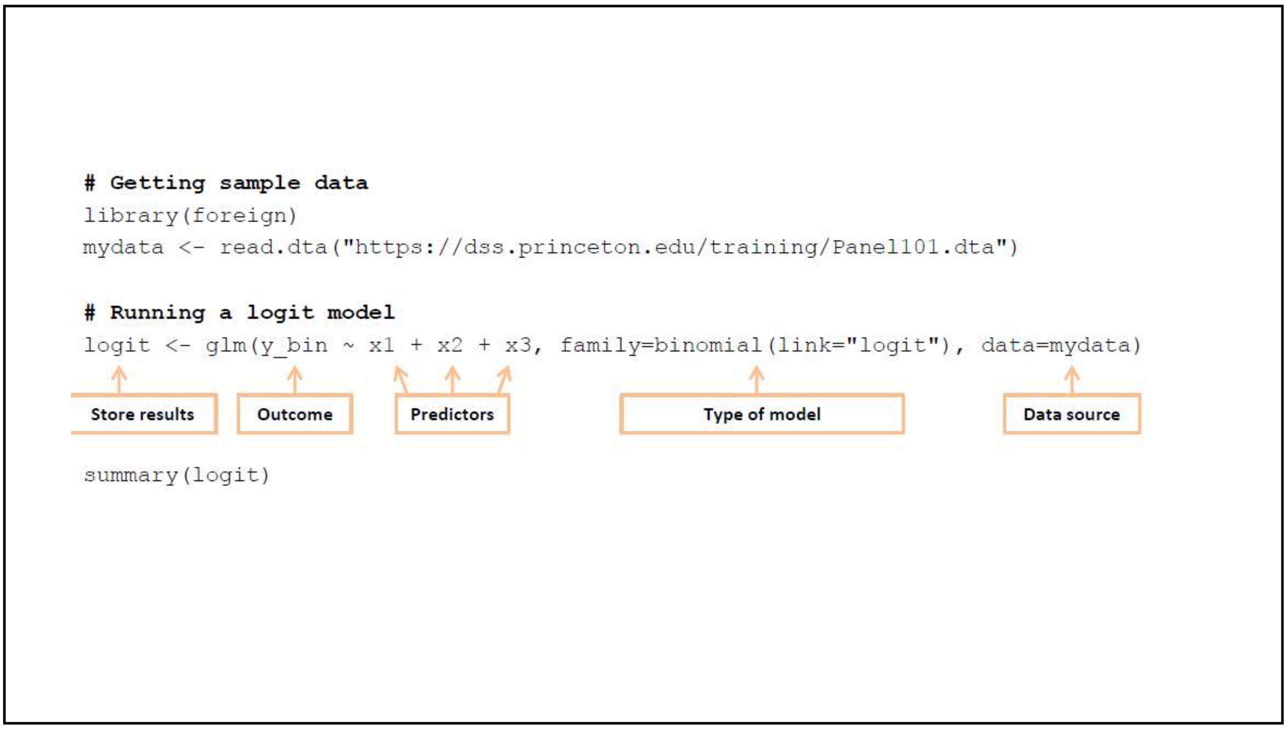
Logit/Probit
(Discrete Dependent Variable Models)



If outcome or dependent variable is binary and in the form 0/1, then use logit or probit models. Some examples are:	
Did you vote in the last election? 0 'No' 1 'Yes'	Do you prefer to use public transportation or to drive a car? 0 'Prefer to drive' 1 'Prefer public transport'

If outcome or dependent variable is categorical but are ordered (i.e. low to high), then use ordered logit or ordered probit models. Some examples are:	
Do you agree or disagree with the President? 1 'Disagree' 2 'Neutral' 3 'Agree'	What is your socioeconomic status? 1 'Low' 2 'Middle' 3 'High'

If outcome or dependent variable is categorical without any particular order, then use multinomial logit. Some examples are:	
If elections were held today, for which party would you vote? 1 'Democrats' 2 'Independent' 3 'Republicans'	What do you like to do on the weekends? 1 'Rest' 2 'Go to movies' 3 'Exercise'

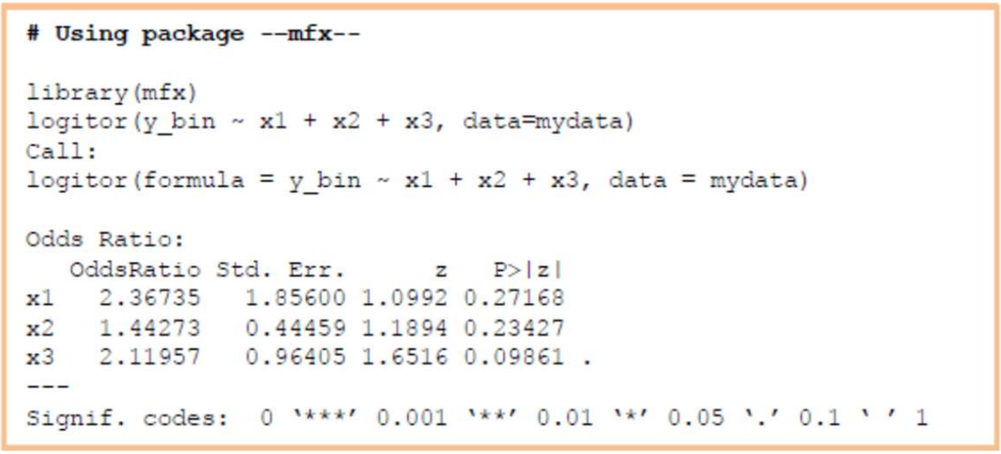


```
# Getting sample data
library(foreign)
mydata <- read.dta("https://dss.princeton.edu/training/Panel101.dta")

# Running a logit model
logit <- glm(y_bin ~ x1 + x2 + x3, family=binomial(link="logit"), data=mydata)
```

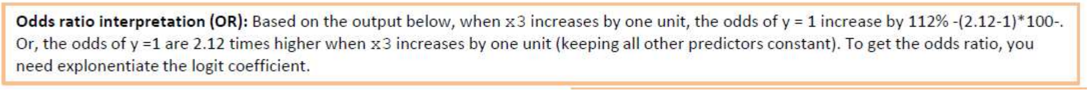
Store results
Outcome
Predictors
Type of model
Data source

```
summary(logit)
```

Odds ratio



```
# Using package --mfx--

library(mfx)
logitor(y_bin ~ x1 + x2 + x3, data=mydata)
Call:
logitor(formula = y_bin ~ x1 + x2 + x3, data = mydata)

Odds Ratio:
   OddsRatio Std. Err.      z   P>|z|
x1   2.36735   1.85600 1.0992 0.27168
x2   1.44273   0.44459 1.1894 0.23427
x3   2.11957   0.96405 1.6516 0.09861 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

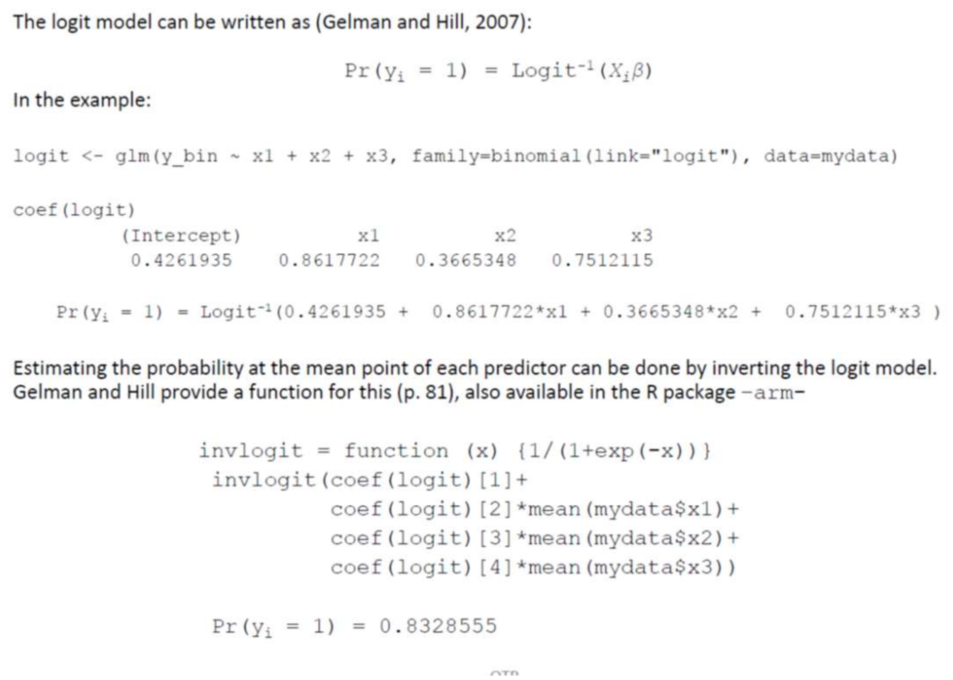


Odds ratio interpretation (OR): Based on the output below, when x3 increases by one unit, the odds of y = 1 increase by 112% -(2.12-1)*100-. Or, the odds of y =1 are 2.12 times higher when x3 increases by one unit (keeping all other predictors constant). To get the odds ratio, you need explonentiate the logit coefficient.

- Odds Ratio

e.g. When x3 increases by one unit, the odds of y=1 increases by 112%.

e.g. If x3 is age, and odds ratio was 220, then we can interpret the result as “When the age of the respondent increases by one (year), the probability of saying “yes (y=1)” increases by 120%”.



The logit model can be written as (Gelman and Hill, 2007):

$$\Pr(y_i = 1) = \text{Logit}^{-1}(X_i\beta)$$

In the example:

```
logit <- glm(y_bin ~ x1 + x2 + x3, family=binomial(link="logit"), data=mydata)

coef(logit)
        (Intercept)          x1          x2          x3
         0.4261935   0.8617722   0.3665348   0.7512115
```

$$\Pr(y_i = 1) = \text{Logit}^{-1}(0.4261935 + 0.8617722*x1 + 0.3665348*x2 + 0.7512115*x3)$$

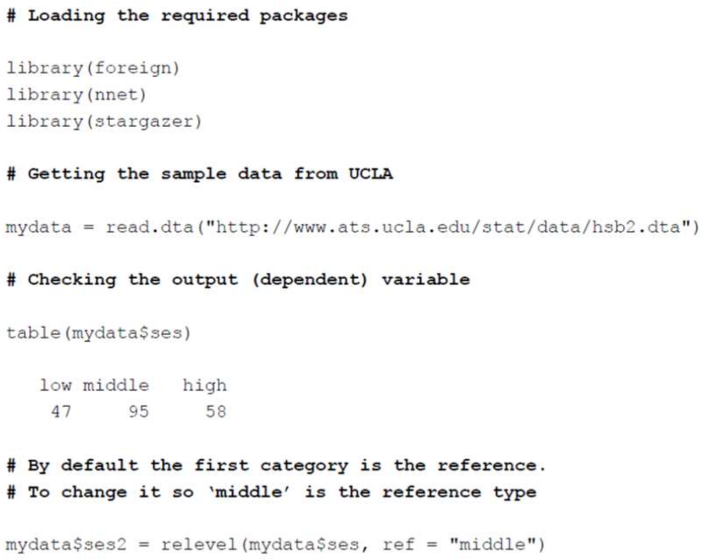
Estimating the probability at the mean point of each predictor can be done by inverting the logit model. Gelman and Hill provide a function for this (p. 81), also available in the R package -arm-

```
invlogit = function (x) {1/(1+exp(-x))}
 invlogit(coef(logit)[1]+
          coef(logit)[2]*mean(mydata$x1)+
          coef(logit)[3]*mean(mydata$x2)+
          coef(logit)[4]*mean(mydata$x3))
```

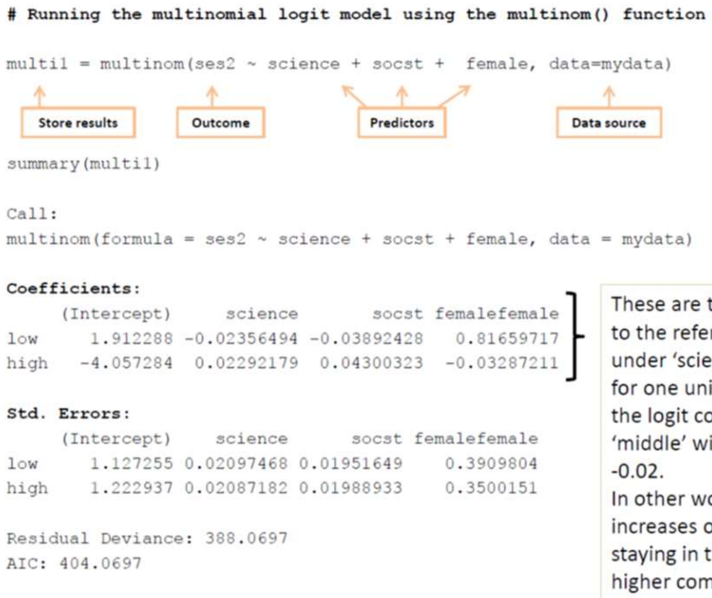
$$\Pr(y_i = 1) = 0.8328555$$

Marginal effects show the change in probability when the predictor or independent variable increases by one unit. For continuous variables this represents the instantaneous change given that the 'unit' may be very small. For binary variables, the change is from 0 to 1, so one 'unit' as it is usually thought.

Multinomial Logit



```
# Loading the required packages

library(foreign)
library(nnet)
library(stargazer)

# Getting the sample data from UCLA

mydata = read.dta("http://www.ats.ucla.edu/stat/data/hsb2.dta")

# Checking the output (dependent) variable

table(mydata$ses)

   low middle   high
    47     95     58

# By default the first category is the reference.
# To change it so 'middle' is the reference type

mydata$ses2 = relevel(mydata$ses, ref = "middle")
```



```
# Running the multinomial logit model using the multinom() function

multi1 = multinom(ses2 ~ science + socst +  female, data=mydata)
```

Store results

Outcome

Predictors

Data source

```
summary(multi1)

Call:
multinom(formula = ses2 ~ science + socst + female, data = mydata)

Coefficients:
     (Intercept)     science       socst femalefemale
low     1.912288 -0.02356494 -0.03892428   0.81659717
high   -4.057284  0.02292179  0.04300323  -0.03287211

Std. Errors:
     (Intercept)    science      socst femalefemale
low     1.127255 0.02097468 0.01951649    0.3909804
high    1.222937 0.02087182 0.01988933    0.3500151

Residual Deviance: 388.0697
AIC: 404.0697
```

These are the logit coefficients relative to the reference category. For example, under 'science', the -0.02 suggests that for one unit increase in 'science' score, the logit coefficient for 'low' relative to 'middle' will go down by that amount, -0.02.
In other words, if your science score increases one unit, your chances of staying in the middle ses category are higher compared to staying in low ses.