R Lab Guide

Categorical Data Analysis Summer 2017

Tamara van der Does
Based on Kelly Gleason's CDA R lab and Trent Mize's STATA lab

Last Updated: July 7th, 2017

Contents

Preface

1. The R Lab Guide is divided into sections corresponding to class lectures and the Stata Lab Guide. Each section should be reviewed ***before*** starting the assignments.
2. The R Lab Guide makes every effort to follow assignments as they are written, but you are ultimately responnsible for completing all sections of your assignment, specifics of assignment questions should be double-checked.
3. The R lab guide uses the data set *cda_scireview3.dta* available from https://shawnana79.github.io/data/. ***These data cannot be used to complete assingments.***
4. Throughout this lab, the code is writen in shaded boxes and the output is presented starting with "##". We also provide a few examples of interpretations.
5. We recommend using RStudio to write code and Rmarkdown to show code and output. Other methods are accepted as long as you ***turn in the equivalent of a Stata log file.***.

Installation of R and Packages

- If you haven't already done so, install R on your computer (along with your choice of text editor) and verify that it works.
  - R https://cran.r-project.org/
- Suggestions for text editing and presentation:
  - RStudio https://www.rstudio.com/home/
  - Rmarkdown http://rmarkdown.rstudio.com/
- Necessary packages can be downloaded the following way:

```
# Install the `car' package.
install.packages("car")
```

```
# Load `car' package.
library(car)
```

You only need to insall packages once. However, you need to load the library for each new assignment or project.

Section 1: Models for Binary Outcomes: Part 1

For details about models and related Stata commands, see Chapter 4 of L&F (2005).

Resource for R: http://www.ats.ucla.edu/stat/r/dae/logit.htm

1.1.a Set up

To save the outpute and figures, specify the folder you want to work in. This could be anything from a folder, cloud or usb drive. Note that in R, you need to replace the back slashes by forward slashes.

```
# Specify working folder of your choice
setwd("C:/Users/TamaravdD/Box Sync/Teaching/2017-2.5-ICPSR-CA/Rlab")
```

1.1.b Load the Data

```
# Load library that enables you to load Stata data
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.3.3
```

```
# Load the data
data <- as.data.frame(read.dta("https://shawnana79.github.io/data/cda_scireview3.dta",
    convert.factor = TRUE))
```

The "convert.factor=TRUE" keeps the value labels instead of replacing them by numbers

1.1.c Examine the Data and Select Variables

```
# Describe the data
head(data)
```

```
##      id cit1 cit3 cit6 cit9 enrol  fel felclass fellow female mcit3 mcitt
## 1 62352    3    1    4   14     8 1.60   1_Adeq   0_No  1_Yes     2     5
## 2 57249   21   24   20   14     4 4.50   4_Dist  1_Yes   0_No    17    26
## 3 62101    8    5   14   18     7 1.30   1_Adeq   0_No  1_Yes     1    10
## 4 57339   19   23   25   24     5 1.86   1_Adeq   0_No   0_No    34     6
## 5 62083    4    6    3   12     7 4.36   4_Dist   0_No  1_Yes    72     1
## 6 57071    0    1    3    9     5 4.29   4_Dist   0_No  1_Yes    23     1
##   mmale mnas mpub3 nopub1 nopub3 nopub6 nopub9  phd phdclass pub1 pub3
## 1 1_Yes 0_No     2  1_Yes   0_No   0_No   0_No 1.60   1_Adeq    0    1
## 2 1_Yes 0_No    11   0_No   0_No   0_No   0_No 2.58   2_Good    7   12
## 3 1_Yes 0_No    11   0_No   0_No   0_No   0_No 1.30   1_Adeq    5   10
## 4 1_Yes 0_No    25   0_No   0_No   0_No   0_No 1.86   1_Adeq    7    5
## 5 1_Yes 0_No    39   0_No   0_No   0_No   0_No 4.36   4_Dist    1    4
## 6 1_Yes 0_No    11  1_Yes   0_No   0_No  1_Yes 4.29   4_Dist    0    1
##   pub6 pub9      work workadmn worktch workuniv faculty totpub jobimp
## 1    2    4 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      7   1.84
## 2    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     25   1.99
## 3    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     23   1.53
## 4   10    9 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     24   1.74
## 5    2    3 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      9   1.58
```

```
## 6    3    0 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      4   1.47
##   jobprst
## 1  1_Adeq
## 2  1_Adeq
## 3  1_Adeq
## 4  1_Adeq
## 5  1_Adeq
## 6  1_Adeq
```

```
summary(data)
```

```
##        id             cit1             cit3             cit6
##  Min.   :57001   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:57125   1st Qu.:  3.00   1st Qu.:  3.00   1st Qu.:  4.00
##  Median :57265   Median :  5.50   Median :  8.00   Median :  9.00
##  Mean   :58557   Mean   : 11.33   Mean   : 14.69   Mean   : 17.59
##  3rd Qu.:62038   3rd Qu.: 13.00   3rd Qu.: 17.00   3rd Qu.: 21.00
##  Max.   :62420   Max.   :130.00   Max.   :196.00   Max.   :143.00
##       cit9            enrol            fel            felclass
##  Min.   :  0.00   Min.   : 3.00   Min.   :1.000   1_Adeq  :33
##  1st Qu.:  7.00   1st Qu.: 4.00   1st Qu.:2.320   2_Good  :86
##  Median : 14.00   Median : 5.00   Median :3.310   3_Strong:72
##  Mean   : 19.93   Mean   : 5.53   Mean   :3.191   4_Dist  :73
##  3rd Qu.: 23.00   3rd Qu.: 6.00   3rd Qu.:4.290
##  Max.   :214.00   Max.   :14.00   Max.   :4.690
##    fellow      female        mcit3            mcitt          mmale
##  0_No :155   0_No :173   Min.   :  0.00   Min.   :  0.00   0_No :  4
##  1_Yes:109   1_Yes: 91   1st Qu.:  4.00   1st Qu.:  6.00   1_Yes:260
##                          Median : 12.50   Median : 20.00
##                          Mean   : 20.72   Mean   : 43.45
##                          3rd Qu.: 24.00   3rd Qu.: 58.50
##                          Max.   :129.00   Max.   :223.00
##     mnas         mpub3         nopub1      nopub3      nopub6
##  0_No :242   Min.   : 0.00   0_No :197   0_No :213   0_No :212
##  1_Yes: 22   1st Qu.: 4.00   1_Yes: 67   1_Yes: 51   1_Yes: 52
##              Median : 9.00
##              Mean   :11.11
##              3rd Qu.:14.00
##              Max.   :48.00
##    nopub9          phd            phdclass       pub1
##  0_No :213   Min.   :1.000   1_Adeq  :38   Min.   : 0.000
##  1_Yes: 51   1st Qu.:2.260   2_Good  :87   1st Qu.: 0.000
##              Median :3.190   3_Strong:60   Median : 1.000
##              Mean   :3.182   4_Dist  :79   Mean   : 2.322
##              3rd Qu.:4.290                 3rd Qu.: 4.000
##              Max.   :4.660                 Max.   :19.000
##       pub3             pub6             pub9              work
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   1_FacUniv:141
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000   2_ResUniv: 45
##  Median : 2.000   Median : 3.000   Median : 3.000   3_ColTch : 24
##  Mean   : 2.939   Mean   : 3.879   Mean   : 4.254   4_IndRes : 33
##  3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 6.000   5_Admin  : 21
##  Max.   :27.000   Max.   :29.000   Max.   :27.000
##   workadmn    worktch     workuniv    faculty        totpub
##  0_No :243   0_No : 99   0_No : 78   0_No :123   Min.   : 0.00
```

```
##  1_Yes: 21   1_Yes:165   1_Yes:186   1_Yes:141   1st Qu.: 4.00
##                                                  Median : 8.00
##                                                  Mean   :11.07
##                                                  3rd Qu.:15.00
##                                                  Max.   :73.00
##      jobimp          jobprst
##  Min.   :1.010   1_Adeq  : 29
##  1st Qu.:2.413   2_Good  :128
##  Median :2.808   3_Strong: 93
##  Mean   :2.864   4_Dist  : 14
##  3rd Qu.:3.345
##  Max.   :4.690
```

```
# Select variables of interest
myvars <- c("faculty", "fellow", "phd", "mcit3", "mnas")
datasub <- data[myvars]

# Make sure the classs of variables are correct
datasub$faculty <- as.factor(datasub$faculty)
datasub$fellow <- as.factor(datasub$fellow)
datasub$mnas <- as.factor(datasub$mnas)
datasub$mcit3 <- as.numeric(datasub$mcit3)
datasub$phd <- as.numeric(datasub$phd)
```

1.1.d Drop cases with Missing Data and Verify

```
# Only keep non missing variables
dataclean <- na.omit(datasub)

# Check no missing values (both need to return 0)
sum(is.na(dataclean))
```

```
## [1] 0
```

```
dataclean[!complete.cases(dataclean), ]
```

```
## [1] faculty fellow  phd     mcit3   mnas
## <0 rows> (or 0-length row.names)
```

```
# Save newly created dataset
write.dta(dataclean, "cda_scireview3_tvdd.dta")
```

1.2 Describe your Data

```
summary(dataclean)
```

```
##   faculty      fellow         phd            mcit3           mnas
##  0_No :123   0_No :155   Min.   :1.000   Min.   :  0.00   0_No :242
##  1_Yes:141   1_Yes:109   1st Qu.:2.260   1st Qu.:  4.00   1_Yes: 22
##                          Median :3.190   Median : 12.50
##                          Mean   :3.182   Mean   : 20.72
##                          3rd Qu.:4.290   3rd Qu.: 24.00
##                          Max.   :4.660   Max.   :129.00
```

Alternatively:

```
# Library to present data for psychometrics
library(psych)
```

```
describe(dataclean)
```

```
##          vars   n  mean    sd median trimmed   mad min    max  range  skew
## faculty*    1 264  1.53  0.50   2.00    1.54  0.00   1   2.00   1.00 -0.14
## fellow*     2 264  1.41  0.49   1.00    1.39  0.00   1   2.00   1.00  0.35
## phd         3 264  3.18  1.01   3.19    3.21  1.56   1   4.66   3.66 -0.14
## mcit3       4 264 20.72 25.45  12.50   15.33 14.08   0 129.00 129.00  2.28
## mnas*       5 264  1.08  0.28   1.00    1.00  0.00   1   2.00   1.00  3.00
##          kurtosis   se
## faculty*    -1.99 0.03
## fellow*     -1.88 0.03
## phd         -1.24 0.06
## mcit3        5.56 1.57
## mnas*        7.01 0.02
```

1.3 Binary Logit Model

```
# Logit model (format: DV ~ IVs)
mod.log <- glm(faculty ~ fellow + phd + mcit3 + mnas, data = dataclean,
    family = binomial)
# Present the results
summary(mod.log)
```

```
##
## Call:
## glm(formula = faculty ~ fellow + phd + mcit3 + mnas, family = binomial,
##     data = dataclean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4367  -0.9825   0.5350   0.9441   1.5441
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.580603   0.449885  -1.291  0.19686
## fellow1_Yes  1.250155   0.276796   4.517 6.29e-06 ***
## phd         -0.063719   0.147131  -0.433  0.66496
## mcit3        0.020616   0.007126   2.893  0.00381 **
## mnas1_Yes    0.363908   0.557122   0.653  0.51363
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 364.75  on 263  degrees of freedom
## Residual deviance: 327.11  on 259  degrees of freedom
## AIC: 337.11
##
```

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

1.4 Computing Factor Change Coefficients

In R, we need to calculate exp^b manually.

```
# Calculate standard deviation for all continuous variables
sdX <- with(dataclean, c(NA, sd(phd), sd(mcit3), NA))

# Extract beta coefficients and z scores
bHat <- coef(mod.log)[2:5]
zscores <- cbind(summary(mod.log)$coefficients[2:5, 3])

# Put everything in a table
facOdds <- cbind(bHat, exp(bHat), exp(bHat * sdX), sdX, zscores)
colnames(facOdds) <- c("b", "e^b", "e^(b*sd)", "SD of X", "Z-values")
facOdds
```

```
##                       b      e^b  e^(b*sd)  SD of X   Z-values
## fellow1_Yes  1.25015468 3.490883        NA       NA  4.5165126
## phd         -0.06371861 0.938269 0.9379593  1.00518 -0.4330751
## mcit3        0.02061560 1.020830 1.6897343 25.44536  2.8932044
## mnas1_Yes    0.36390817 1.438942        NA       NA  0.6531925
```

1.5 & 1.6 Interpreting Facor Change Coefficients

Interpretations.

- **Unstandardized** Obtaining a post-doctoral fellowship increases the odds of gaining a faculty position by a factor of 3.5, holding other variables constant.
- **X-standardized** A standard deviation incrase in mentor's citations (about 25) increases the odds of gaining a facutly position by a factor of 1.7

1.7.a Computing Predicted Probabilities

To compute predicted probabilities in R and calcualte the discrete change for binary variables, you first need to create a sub-sample with all variables held at chose values (for example, mean for continuous variables and mode for factor variables) and list the factor change for the binary variable of interest. Then, apply the model to this sub-sample using the *predict* command.

```
# Create the sub-sample
s1 <- with(dataclean, data.frame(fellow = factor(1:2, levels = 1:2,
    labels = levels(dataclean$fellow)), phd = mean(phd), mcit3 = mean(mcit3),
    mnas = "0_No"))
s1
```

```
##   fellow      phd    mcit3 mnas
## 1   0_No 3.181894 20.71591 0_No
## 2  1_Yes 3.181894 20.71591 0_No
```

```
# Apply the logit model to the sub-sample
predict(mod.log, s1, type = "response", se.fit = TRUE)$fit
```

```
##         1         2
## 0.4118608 0.7096895
```

Note: 1=Not a fellow and 2=Fellow.

These results are different than in the Stata lab, as in Stata you can hold factor variable "at their mean" while in R they need to be given a specific value

Interpretation. An individual (with a mentor who is not an NAS member) who held a post-doctoral fellowship has a .71 proability of currently holding a faculty position, all other variables held at their mean. A similar individual who did not hold a post-doctoral fellowship has a 0.41 probability of currently holding a faculty position.

1.7.b Use Predicted Probabilities to Compute Factor Change Coefficient

```
a <- predict(mod.log, s1, type = "response", se.fit = TRUE)$fit[1]
b <- predict(mod.log, s1, type = "response", se.fit = TRUE)$fit[2]
(b/(1 - b))/(a/(1 - a))
```

```
##        2
## 3.490883
```

This number is similar to the factor change coefficient for fellow given above (e^b).

1.8-1.9 Compare the Coefficients from Logit and Probit

```
# Probit model
mod.prob <- glm(faculty ~ fellow + phd + mcit3 + mnas, data = dataclean,
    family = binomial(link = probit))
summary(mod.prob)
```

```
##
## Call:
## glm(formula = faculty ~ fellow + phd + mcit3 + mnas, family = binomial(link = probit),
##     data = dataclean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4942  -0.9877   0.5398   0.9534   1.5328
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.345029   0.275006  -1.255  0.20962
## fellow1_Yes  0.763914   0.167520   4.560 5.11e-06 ***
## phd         -0.039268   0.089847  -0.437  0.66207
## mcit3        0.011864   0.004064   2.920  0.00351 **
## mnas1_Yes    0.229953   0.328192   0.701  0.48351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 364.75  on 263  degrees of freedom
## Residual deviance: 327.48  on 259  degrees of freedom
```

```
## AIC: 337.48
##
## Number of Fisher Scoring iterations: 4
```

Table

```
logprob <- cbind(mod.log$coef, summary(mod.log)$coef[, 3], mod.prob$coef,
    summary(mod.prob)$coef[, 3])
colnames(logprob) <- c("logit estimates", "SE", "Probit estimates",
    "SE")
logprob
```

```
##             logit estimates         SE Probit estimates         SE
## (Intercept)     -0.58060305 -1.2905601      -0.34502850 -1.2546210
## fellow1_Yes      1.25015468  4.5165126       0.76391447  4.5601355
## phd             -0.06371861 -0.4330751      -0.03926769 -0.4370520
## mcit3            0.02061560  2.8932044       0.01186414  2.9195762
## mnas1_Yes        0.36390817  0.6531925       0.22995308  0.7006668
```

Section 2: Models for Binary Outcomes: Part 2

2.1.a Set Up

```
# Specify working folder
setwd("C:/Users/TamaravdD/Box Sync/Teaching/2017-2.5-ICPSR-CA/Rlab")
```

```
# Load libraries
library(car)
library(foreign)
library(psych)
```

2.1.b Load the Data

```
# Download the data saved previously
dataclean <- as.data.frame(read.dta("cda_scireview3_tvdd.dta",
    convert.factor = TRUE))
```

2.1.c Re-estimate your Binary Logit Model

Results should match your results in BRM-Part 1

```
mod.logit <- glm(faculty ~ fellow + phd + mcit3 + mnas, data = dataclean,
    family = "binomial")
summary(mod.logit)
```

```
##
## Call:
## glm(formula = faculty ~ fellow + phd + mcit3 + mnas, family = "binomial",
##     data = dataclean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4367  -0.9825   0.5350   0.9441   1.5441
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.580603   0.449885  -1.291  0.19686
## fellow1_Yes  1.250155   0.276796   4.517 6.29e-06 ***
## phd         -0.063719   0.147131  -0.433  0.66496
## mcit3        0.020616   0.007126   2.893  0.00381 **
## mnas1_Yes    0.363908   0.557122   0.653  0.51363
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 364.75  on 263  degrees of freedom
## Residual deviance: 327.11  on 259  degrees of freedom
## AIC: 337.11
##
## Number of Fisher Scoring iterations: 4
```

2.2 Predicted Probabilities for Observed Data

```
# Use 'names' to see the results stored in model
names(mod.logit)
```

```
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "call"              "formula"           "terms"
## [25] "data"              "offset"            "control"
## [28] "method"            "contrasts"         "xlevels"
```

```
# The predicted probabilities are stored as 'fitted.values'
summary(mod.logit$fitted.values)
```

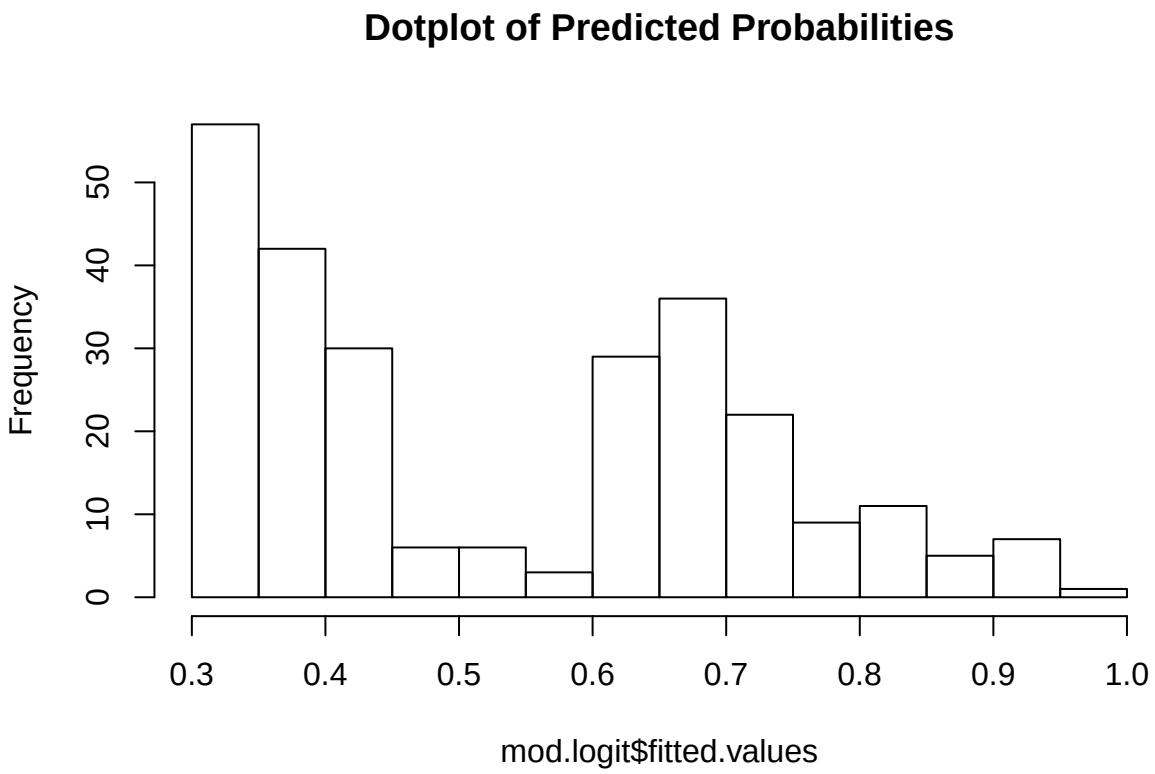
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3036  0.3572  0.4821  0.5341  0.6847  0.9665
```

```
# Table with summary statistics for predicted probabilities
predprob <- cbind(mean(mod.logit$fitted.values), sd(mod.logit$fitted.values),
    length(mod.logit$fitted.values), min(mod.logit$fitted.values),
    max(mod.logit$fitted.values))
colnames(predprob) <- c("mean", "standard dev", "N", "Min", "Max")
rownames(predprob) <- c("prlogit")
predprob
```

```
##              mean standard dev   N       Min       Max
## prlogit 0.5340909    0.1828654 264 0.3035647 0.9665072
```

Histogram of predicted probabilities

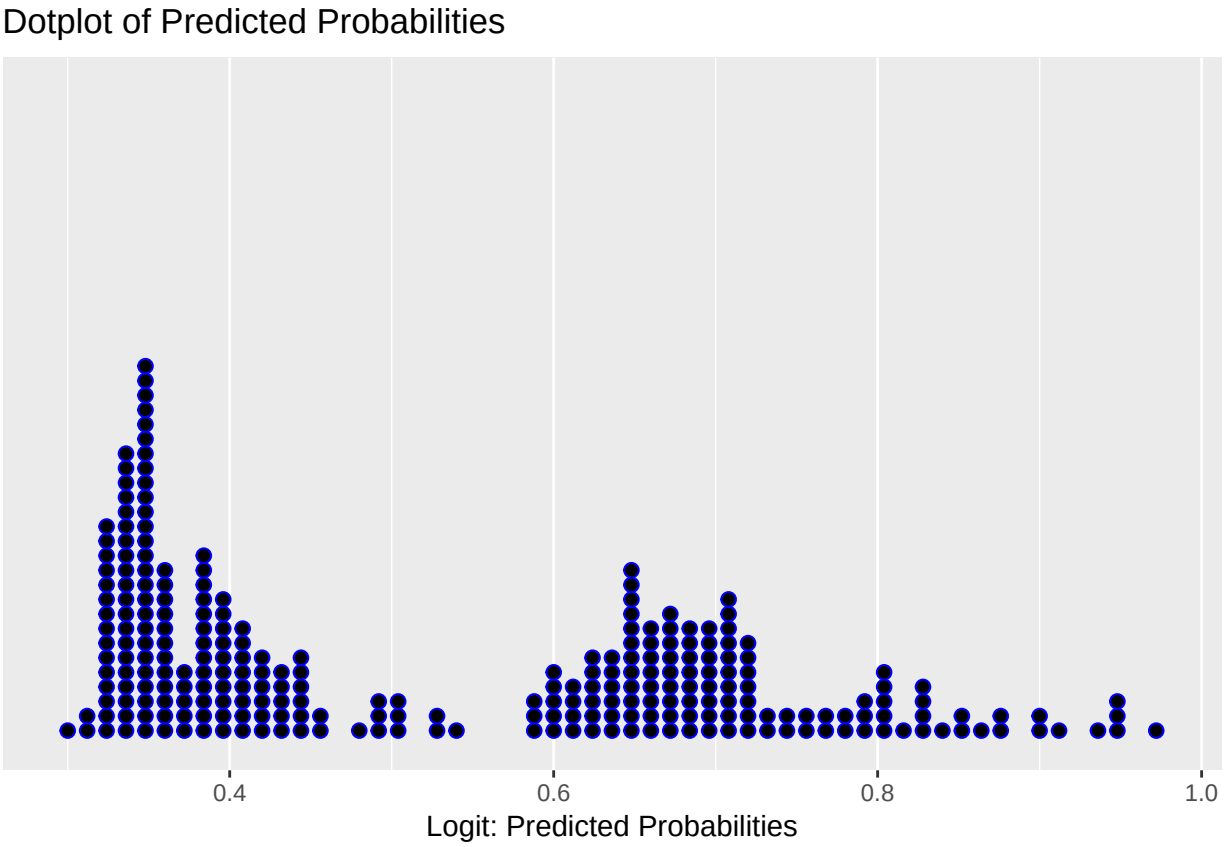
```
# classic version
hist(mod.logit$fitted.values, main = "Dotplot of Predicted Probabilities")
```




```
# GGplot version
library(easyGgplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
ggplot(dataclean, aes(x = mod.logit$fitted.values)) + geom_dotplot(method = "histodot",
    binwidth = 0.012, dotsize = 0.75, color = "blue") + scale_y_continuous(NULL,
    breaks = NULL) + labs(title = "Dotplot of Predicted Probabilities") +
    xlab("Logit: Predicted Probabilities")
```




```
# Save plot
ggsave("icpsrcda02-binary-fig1.png")
```

2.3-2.5 Discrete Change with Confidence Interval

For binary variable

```
# Create sub-sample
s.fel <- with(dataclean, data.frame(fellow = factor(1:2, levels = 1:2,
    labels = levels(dataclean$fellow)), phd = mean(phd), mcit3 = mean(mcit3),
    mnas = "0_No"))

# Verify subsample created correctly
s.fel
```

```
##   fellow      phd    mcit3 mnas
## 1   0_No 3.181894 20.71591 0_No
## 2  1_Yes 3.181894 20.71591 0_No
```

```
# Apply our logit model to the subsample
pred.fel <- predict(mod.logit, s.fel, type = "response", se.fit = TRUE)

# Difference between fellows and non fellows
diff <- pred.fel$fit[2] - pred.fel$fit[1]

# Calculate the confidence intervals
```

```
CI.U <- diff + (1.96 * sqrt(pred.fel$se.fit[1]^2 + pred.fel$se.fit[2]^2))
CI.L <- diff - (1.96 * sqrt(pred.fel$se.fit[1]^2 + pred.fel$se.fit[2]^2))

# Discrete change and confidence interval
cbind(diff, CI.L, CI.U)
```

```
##        diff     CI.L      CI.U
## 2 0.2978287 0.175354 0.4203033
```

These does not exactly match the Stata lab as we are holding factor variables at actual values and not their mean.

Interpretation. A scientist, whose mentor is not an NAS member, who receives a post-doctoral fellowship has a .30 higher probability of being faculty at a university than a scientist who does not receive a fellowship, holding other variables at their mean. This difference is significant (95% CI: 0.18,0.42)

For continuous variable

First, we calculate the values of the contiuous variable we want to predict (in this case, a standard deviation centered around the mean.) Then, we create two sub-samples using these values. Finally, we compute the predicted probabilities using these samples and calculate the discrete change.

```
# Save the values for C
mcit.centL <- mean(dataclean$mcit3) - (0.5 * sd(dataclean$mcit3))
mcit.centU <- mean(dataclean$mcit3) + (0.5 * sd(dataclean$mcit3))

# Create the subsets with other variables at their means
s.U <- with(dataclean, data.frame(fellow = "0_No", mcit3 = mcit.centU,
    phd = mean(phd), mnas = "0_No"))
s.L <- with(dataclean, data.frame(fellow = "0_No", mcit3 = mcit.centL,
    phd = mean(phd), mnas = "0_No"))
s.U
```

```
##   fellow    mcit3      phd mnas
## 1   0_No 33.43859 3.181894 0_No
```

```
s.L
```

```
##   fellow    mcit3      phd mnas
## 1   0_No 7.993227 3.181894 0_No
```

```
# Apply our logit model to the subsample
pred.U <- predict(mod.logit, s.U, type = "response", se.fit = TRUE)
pred.L <- predict(mod.logit, s.L, type = "response", se.fit = TRUE)

diff <- pred.U$fit - pred.L$fit

# Calculate the confidence intervals
CI.U <- diff + (1.96 * sqrt(pred.U$se.fit^2 + pred.L$se.fit^2))
CI.L <- diff - (1.96 * sqrt(pred.U$se.fit^2 + pred.L$se.fit^2))

# Discrete change and confidence interval
cbind(diff, CI.L, CI.U)
```

```
##       diff         CI.L      CI.U
## 1 0.126411 -0.003095407 0.2559174
```

These does not exactly match the Stata lab as we are holding factor variables at actual values and not their mean

Interpretation. For non-fellows with a mentor who is not a member of NAS, a standard deviation increase in the number of mentor's citations (about 25 citations) centered around the mean increases the probability of obtaining a faculty position by .13, holding phd prestige at its mean. However, this difference is not significant (95% CI: -0.003,0.256)

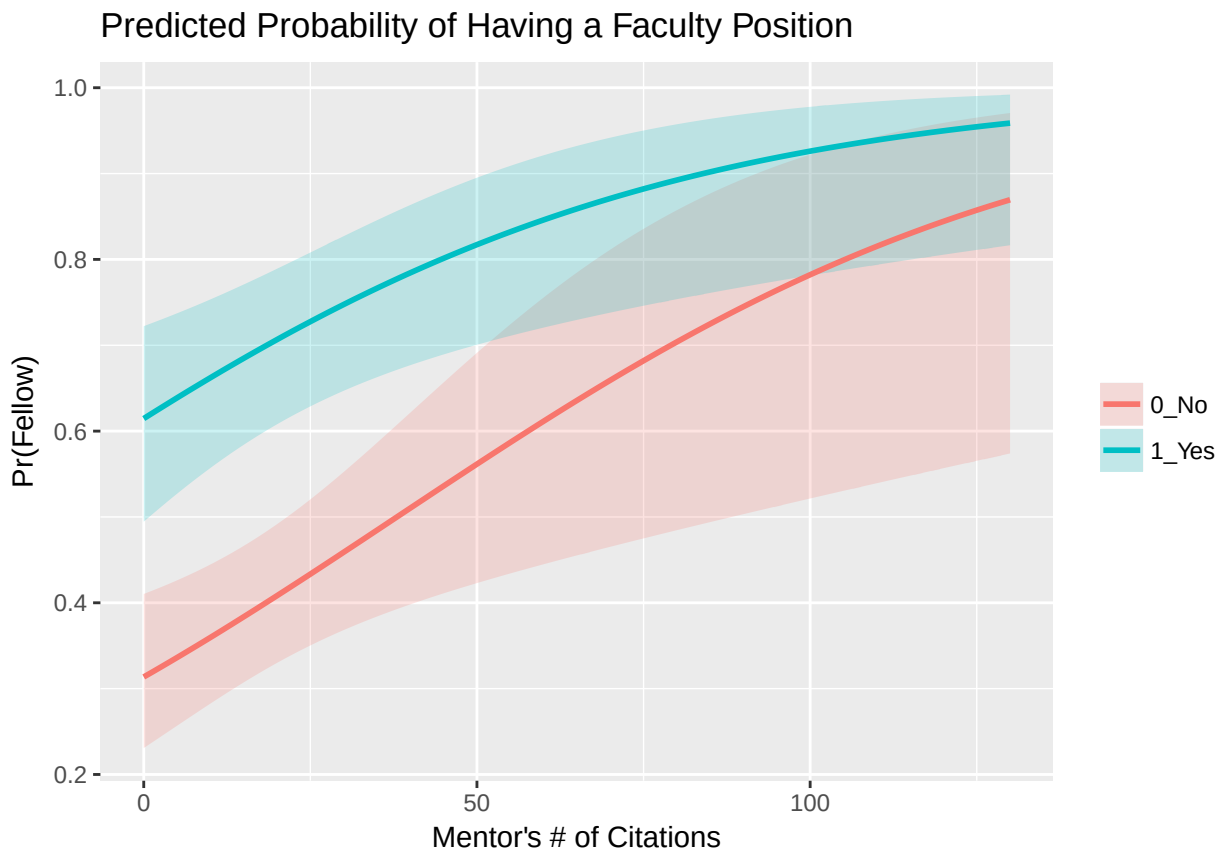
2.6 Plot predicted probabilities

To plot predicted probabilities by B over C, first create a subsample with C as a sequence, B as factor and the rest held at values of interest. Then attach the predicted probabilities for that data. Calculate and add the confidence interval to the data. Finally use the dataset to create a plot.

```
# Create subsample
s.plot <- with(dataclean, data.frame(mcit3 = rep(seq(from = 0,
    to = 130, length.out = 100), 2), phd = mean(phd), mnas = "0_No",
    fellow = factor(rep(1:2, each = 100), labels = levels(dataclean$fellow))))

# Create data with subsample and predicted probabilities
s.plot.data <- cbind(s.plot, predict(mod.logit, newdata = s.plot,
    type = "link", se = TRUE))

# Add confidence interval
s.plot.data <- within(s.plot.data, {
    PredictedProb <- plogis(fit)
    LL <- plogis(fit - (1.96 * se.fit))
    UL <- plogis(fit + (1.96 * se.fit))
})

# Look at data created
head(s.plot.data)
```

```
##      mcit3      phd mnas fellow        fit    se.fit residual.scale
## 1 0.000000 3.181894 0_No   0_No -0.7833489 0.2145753              1
## 2 1.313131 3.181894 0_No   0_No -0.7562779 0.2090872              1
## 3 2.626263 3.181894 0_No   0_No -0.7292070 0.2038809              1
## 4 3.939394 3.181894 0_No   0_No -0.7021360 0.1989786              1
## 5 5.252525 3.181894 0_No   0_No -0.6750650 0.1944033              1
## 6 6.565657 3.181894 0_No   0_No -0.6479940 0.1901786              1
##          UL        LL PredictedProb
## 1 0.4102865 0.2307792     0.3135986
## 2 0.4142394 0.2375627     0.3194549
## 3 0.4183378 0.2443801     0.3253688
## 4 0.4225930 0.2512167     0.3313388
## 5 0.4270164 0.2580563     0.3373636
## 6 0.4316202 0.2648825     0.3434417
```

```
## Plot of Predicted Probabilities using ggplot
ggplot(s.plot.data, aes(x = mcit3, y = PredictedProb)) + geom_ribbon(aes(ymin = LL,
    ymax = UL, fill = fellow), alpha = 0.2) + geom_line(aes(colour = fellow),
    size = 1) + labs(title = "Predicted Probability of Having a Faculty Position",
    x = "Mentor's # of Citations", y = "Pr(Fellow)") + theme(legend.title = element_blank())
```




```
# Save plot
ggsave("icpsrcda02-binary-fig2.png")
```

Interpretation. For a scientist whose mentor is not an NAS member and who is average on other charecteristics, receiving a fellowship increases the probability of being employed as a faculty member when mentor's citations are below 50 or so, by about .30. Above 50 mentor citations, there is no significant difference between scientists who received a fellowship and those who did not. As such, fellowships seem to be particularly useful when mentor's citations are low. For both fellow and non-fellows, the probability of having a faculty position increases as the number of mentor's citation increases. This effect is largest for non-fellows, whose predicted probability of being a faculty increases from .37 at 10 mentor citations to .79 at 100 mentor citations, and increase of .42. This change is significant (95% cI: 0.20, 0.64).

Section 3: Testing and Assessing Fit

For a fuller discussion of testing and assessing fit in Stata, see Chapter 3 of L&F (2005).

3.1.a Set up

```
# Specify working folder
setwd("C:/Users/TamaravdD/Box Sync/Teaching/2017-2.5-ICPSR-CA/Rlab")
```

```
# Load libraries
library(car)
library(foreign)
library(psych)
library(easyGgplot2)
```

3.1.b Load the Data

```
data <- as.data.frame(read.dta("https://shawnana79.github.io/data/cda_scireview3.dta",
    convert.factor = TRUE))
```

3.1.c Examine Data, keep variables, drop missing and verify.

```
# Describe the data
head(data)
```

```
##      id cit1 cit3 cit6 cit9 enrol  fel felclass fellow female mcit3 mcitt
## 1 62352    3    1    4   14     8 1.60   1_Adeq   0_No  1_Yes     2     5
## 2 57249   21   24   20   14     4 4.50   4_Dist  1_Yes   0_No    17    26
## 3 62101    8    5   14   18     7 1.30   1_Adeq   0_No  1_Yes     1    10
## 4 57339   19   23   25   24     5 1.86   1_Adeq   0_No   0_No    34     6
## 5 62083    4    6    3   12     7 4.36   4_Dist   0_No  1_Yes    72     1
## 6 57071    0    1    3    9     5 4.29   4_Dist   0_No  1_Yes    23     1
##   mmale mnas mpub3 nopub1 nopub3 nopub6 nopub9  phd phdclass pub1 pub3
## 1 1_Yes 0_No     2  1_Yes   0_No   0_No   0_No 1.60   1_Adeq    0    1
## 2 1_Yes 0_No    11   0_No   0_No   0_No   0_No 2.58   2_Good    7   12
## 3 1_Yes 0_No    11   0_No   0_No   0_No   0_No 1.30   1_Adeq    5   10
## 4 1_Yes 0_No    25   0_No   0_No   0_No   0_No 1.86   1_Adeq    7    5
## 5 1_Yes 0_No    39   0_No   0_No   0_No   0_No 4.36   4_Dist    1    4
## 6 1_Yes 0_No    11  1_Yes   0_No   0_No  1_Yes 4.29   4_Dist    0    1
##   pub6 pub9      work workadmn worktch workuniv faculty totpub jobimp
## 1    2    4 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      7   1.84
## 2    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     25   1.99
## 3    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     23   1.53
## 4   10    9 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     24   1.74
## 5    2    3 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      9   1.58
## 6    3    0 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      4   1.47
##   jobprst
## 1  1_Adeq
## 2  1_Adeq
## 3  1_Adeq
```

```
## 4  1_Adeq
## 5  1_Adeq
## 6  1_Adeq
```

```
# Select variables of interest
myvars <- c("faculty", "female", "fellow", "phd", "mcit3", "mnas")
datasub <- data[myvars]

# Make sure the classs of variables are correct
datasub$faculty <- as.factor(datasub$faculty)
datasub$fellow <- as.factor(datasub$fellow)
datasub$female <- as.factor(datasub$female)
datasub$mnas <- as.factor(datasub$mnas)
datasub$mcit3 <- as.numeric(datasub$mcit3)
datasub$phd <- as.numeric(datasub$phd)

# Only keep non missing variables
dataclean <- na.omit(datasub)

# Verify data is clean
sum(is.na(dataclean))
```

```
## [1] 0
```

```
dataclean[!complete.cases(dataclean), ]
```

```
## [1] faculty female  fellow  phd     mcit3   mnas
## <0 rows> (or 0-length row.names)
```

```
summary(dataclean)
```

```
##   faculty      female      fellow         phd            mcit3
##  0_No :123   0_No :173   0_No :155   Min.   :1.000   Min.   :  0.00
##  1_Yes:141   1_Yes: 91   1_Yes:109   1st Qu.:2.260   1st Qu.:  4.00
##                                      Median :3.190   Median : 12.50
##                                      Mean   :3.182   Mean   : 20.72
##                                      3rd Qu.:4.290   3rd Qu.: 24.00
##                                      Max.   :4.660   Max.   :129.00
##     mnas
##  0_No :242
##  1_Yes: 22
##
##
##
##
```

```
describe(dataclean)
```

```
##          vars   n  mean    sd median trimmed   mad min    max  range  skew
## faculty*    1 264  1.53  0.50   2.00    1.54  0.00   1   2.00   1.00 -0.14
## female*     2 264  1.34  0.48   1.00    1.31  0.00   1   2.00   1.00  0.65
## fellow*     3 264  1.41  0.49   1.00    1.39  0.00   1   2.00   1.00  0.35
## phd         4 264  3.18  1.01   3.19    3.21  1.56   1   4.66   3.66 -0.14
## mcit3       5 264 20.72 25.45  12.50   15.33 14.08   0 129.00 129.00  2.28
## mnas*       6 264  1.08  0.28   1.00    1.00  0.00   1   2.00   1.00  3.00
##          kurtosis   se
## faculty*    -1.99 0.03
```

```
## female*     -1.58 0.03
## fellow*     -1.88 0.03
## phd         -1.24 0.06
## mcit3        5.56 1.57
## mnas*        7.01 0.02
```

3.2 Computing a Z-test

Z-statistics are produced with the standard estimation commands. In the output below, the z-statistics are in the 4th column (z value).

```
mod.log <- glm(faculty ~ female + fellow + phd + mcit3 + mnas,
    data = dataclean, family = binomial)
summary(mod.log)
```

```
##
## Call:
## glm(formula = faculty ~ female + fellow + phd + mcit3 + mnas,
##     family = binomial, data = dataclean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4592  -1.0051   0.4996   0.9258   1.6503
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.500484   0.453908  -1.103  0.27020
## female1_Yes -0.586900   0.291194  -2.015  0.04385 *
## fellow1_Yes  1.118336   0.284461   3.931 8.44e-05 ***
## phd          0.002004   0.152130   0.013  0.98949
## mcit3        0.019081   0.007258   2.629  0.00857 **
## mnas1_Yes    0.353710   0.565277   0.626  0.53149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 364.75  on 263  degrees of freedom
## Residual deviance: 323.03  on 258  degrees of freedom
## AIC: 335.03
##
## Number of Fisher Scoring iterations: 4
```

3.3.a Single coefficient Wald test

```
# Package for Wald tests
library(aod)
```

```
# Run the test on 2nd coefficient of model
wald.test(b = coef(mod.log), Sigma = vcov(mod.log), Terms = 2)
```

```
## Wald test:
## ----------
```

```
##
## Chi-squared test:
## X2 = 4.1, df = 1, P(> X2) = 0.044
```

Interpretation. The effect of female is significant at the .05 level *(χ^2=4.06, df=1, p=0.04)*

3.4 Single Coefficient LR Test

```
# First, run a new model without your variable
mod2.log <- glm(faculty ~ fellow + phd + mcit3 + mnas, data = dataclean,
    family = binomial)

# Then, test the two models using the anova command
anova(mod2.log, mod.log, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: faculty ~ fellow + phd + mcit3 + mnas
## Model 2: faculty ~ female + fellow + phd + mcit3 + mnas
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       259     327.11
## 2       258     323.03  1   4.0804  0.04338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation. The effect of female is significant at the .05 level *(LRχ^2=4.08, df=1, p=0.04)*

3.5 Multiple Coefficient Wald Test

```
# Run a Wald test on coefficients 5 and 6
wald.test(b = coef(mod.log), Sigma = vcov(mod.log), Terms = 5:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.8, df = 2, P(> X2) = 0.02
```

Interpretation. The hypothesis that the effects of mentor's citations and mentor's status as an NAS member are simultaneously equal to zero can be rejected at the .05 level *(χ^2=7.78, df=2, p=0.02)*.

3.6 Multiple Coefficient LR test

```
# Create estimates of model without the two varialbes
mod3.log <- glm(faculty ~ female + fellow + phd, data = dataclean,
    family = binomial)

# Test against the full model
anova(mod3.log, mod.log, test = "Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: faculty ~ female + fellow + phd
## Model 2: faculty ~ female + fellow + phd + mcit3 + mnas
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       260     332.22
## 2       258     323.03  2   9.1929  0.01009 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation. The hypothesis that the effects of mentor's citation and mentor's status as an NAS member are simultaneously equal to zero can be rejected at the .01 level *($LR\chi^2$=9.19,df=2, p=.01)*

3.7 Wald Test All Coefficients are Zero

```
wald.test(b = coef(mod.log), Sigma = vcov(mod.log), Terms = 2:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 33.8, df = 5, P(> X2) = 2.6e-06
```

Interpretation. We can reject he hypothesis that all coefficients execept the intercept are zero at the .01 level *(χ^2=33.8, df=5, p<.001)*

3.8 LR test All Coefficients are Zero

```
# Use LR test code
LRtest <- mod.log$null.deviance - mod.log$deviance
LRtest
```

```
## [1] 41.72321
```

```
# calculate p-value of test
pvalue <- dchisq(LRtest, mod.log$df.null - mod.log$df.residual)
pvalue
```

```
## [1] 3.12087e-08
```

Interpretation. We can reject he hypothesis that all coefficients execept the intercept are zero at the .01 level *($LR\chi^2$=41.72, df=5, p<.001)*

3.9 More Complicated Wald Tests

Two examples are presented bellow: we test that the coefficients for mcit3 and mnas are equal, and that the effect of female is twice the effect of mcit3.

Coefficients equal to each other

```
# Create a vector matching the estimates
vect1 <- cbind(0, 0, 0, 0, 1, -1)
```

```r
# Conduct Wald test with vector
wald.test(b = coef(mod.log), Sigma = vcov(mod.log), L = vect1)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.35, df = 1, P(> X2) = 0.55
```

Effect of female is twice the effect of mcit3.

```r
vect2 <- cbind(0, -1,0,0,2,0)
wald.test(b=coef(mod.log), Sigma= vcov(mod.log), L=vect2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 4.6, df = 1, P(> X2) = 0.031
```

Interpretations.

- The hypothesis that the effect of mentor's citations is equal to the effect of mentor's status as an NAS member cannot be rejected *(χ^2=.35, df=1, p=.55)*.
- The hypothesis that the effect of gender is equal to twice the effect of mentor's citation is rejected at the 0.05 level *(χ^2=4.64, df=1, p=0.03)*.

3.10.a Compare BIC and AIC Statistics Across Non-nested Models

BIC & AIC allow comparison of non-nested models. Here we estimate a series of four non-nested models and then present the BIC and AIC statistics of all models.

```r
# Run and save each models
mod.loga <- glm(faculty ~ female + fellow, data = dataclean,
    family = binomial)
mod.logb <- glm(faculty ~ mcit3, data = dataclean, family = binomial)
mod.logc <- glm(faculty ~ phd + mnas, data = dataclean, family = binomial)
mod.logd <- glm(faculty ~ phd + mnas + female + fellow, data = dataclean,
    family = binomial)

# Save AIC and BIC for each model
mods.aic <- c(AIC(mod.loga), AIC(mod.logb), AIC(mod.logc), AIC(mod.logd))
mods.bic <- c(BIC(mod.loga), BIC(mod.logb), BIC(mod.logc), BIC(mod.logd))

# Present test statistics
mods.tab <- cbind(mods.aic, mods.bic)
colnames(mods.tab) <- c("AIC", "BIC")
rownames(mods.tab) <- c("m1", "m2", "m3", "m4")
mods.tab
```

```
##         AIC      BIC
## m1 340.5262 351.2541
## m2 353.9361 361.0880
```

```
## m3 365.3197 376.0475
## m4 341.2474 359.1271
```

3.11 Compute and List Residuals

```
# Model and predicted probabilities
predprob <- mod.log$fitted.values
data2 <- cbind(dataclean, predprob)

# Residuals
res <- residuals(mod.log, "pearson")/(sqrt(1 - hatvalues(mod.log)))

# add to data
data3 <- cbind(data2, res)

# Sort and present 10 highest and lowest
sort1.data3 <- data3[order(res), ]
sort1.data3[1:10, ]
```

```
##     faculty female fellow  phd mcit3  mnas  predprob       res
## 259    0_No   0_No  1_Yes 4.54   123  0_No 0.9513833 -4.485054
## 74     0_No   0_No  1_Yes 4.62    32 1_Yes 0.8308162 -2.264443
## 64     0_No   0_No  1_Yes 3.54    43  0_No 0.8092961 -2.074258
## 177    0_No   0_No   0_No 4.66    77 1_Yes 0.7911424 -2.014804
## 179    0_No   0_No  1_Yes 3.69    35  0_No 0.7846702 -1.920928
## 222    0_No  1_Yes  1_Yes 4.29    41 1_Yes 0.7641545 -1.856190
## 120    0_No   0_No  1_Yes 3.60    27  0_No 0.7577278 -1.778848
## 228    0_No   0_No  1_Yes 3.36     2 1_Yes 0.7342824 -1.720090
## 249    0_No   0_No  1_Yes 4.54    19  0_No 0.7289883 -1.656825
## 62     0_No   0_No  1_Yes 1.73    17  0_No 0.7202499 -1.620594
```

```
sort2.data3 <- data3[order(-res), ]
sort2.data3[1:10, ]
```

```
##     faculty female fellow  phd mcit3  mnas  predprob      res
## 3     1_Yes  1_Yes   0_No 1.30     1  0_No 0.2562222 1.725802
## 18    1_Yes  1_Yes   0_No 1.42     1  0_No 0.2562680 1.724085
## 131   1_Yes  1_Yes   0_No 2.37     1  0_No 0.2566310 1.714515
## 1     1_Yes  1_Yes   0_No 1.60     2  0_No 0.2599911 1.705497
## 247   1_Yes  1_Yes   0_No 4.36     2  0_No 0.2610567 1.700555
## 241   1_Yes  1_Yes   0_No 2.86     9  0_No 0.2870161 1.586056
## 135   1_Yes  1_Yes   0_No 2.00    12  0_No 0.2985106 1.547463
## 178   1_Yes  1_Yes   0_No 4.29    14  0_No 0.3075393 1.514375
## 16    1_Yes  1_Yes   0_No 4.29     3 1_Yes 0.3389824 1.457150
## 107   1_Yes  1_Yes   0_No 4.29     3 1_Yes 0.3389824 1.457150
```

These results are different than in Stata as they are no know ways to calculate Pearson residuals in R. Any suggestions are more than welcomed.

Section 4: Models for Ordinal Outcomes

For a fuller discussion of testing and assessing fit in Stata, see Chapter 5 of L&F (2005).

Resource for R: http://www.ats.ucla.edu/stat/r/dae/ologit.htm

4.1.a Set Up

```
# Specify working folder
setwd("C:/Users/TamaravdD/Box Sync/Teaching/2017-2.5-ICPSR-CA/Rlab")
```

```
# Load libraries
library(car)
library(foreign)
library(psych)
library(easyGgplot2)
library(aod)
```

4.1.b Load the Data

```
data <- as.data.frame(read.dta("https://shawnana79.github.io/data/cda_scireview3.dta",
    convert.factor = TRUE))
```

4.1.c Examine Data, Keep Variables, Drop Missing and Verify.

```
head(data)
```

```
##      id cit1 cit3 cit6 cit9 enrol  fel felclass fellow female mcit3 mcitt
## 1 62352    3    1    4   14     8 1.60   1_Adeq   0_No  1_Yes     2     5
## 2 57249   21   24   20   14     4 4.50   4_Dist  1_Yes   0_No    17    26
## 3 62101    8    5   14   18     7 1.30   1_Adeq   0_No  1_Yes     1    10
## 4 57339   19   23   25   24     5 1.86   1_Adeq   0_No   0_No    34     6
## 5 62083    4    6    3   12     7 4.36   4_Dist   0_No  1_Yes    72     1
## 6 57071    0    1    3    9     5 4.29   4_Dist   0_No  1_Yes    23     1
##   mmale mnas mpub3 nopub1 nopub3 nopub6 nopub9  phd phdclass pub1 pub3
## 1 1_Yes 0_No     2  1_Yes   0_No   0_No   0_No 1.60   1_Adeq    0    1
## 2 1_Yes 0_No    11   0_No   0_No   0_No   0_No 2.58   2_Good    7   12
## 3 1_Yes 0_No    11   0_No   0_No   0_No   0_No 1.30   1_Adeq    5   10
## 4 1_Yes 0_No    25   0_No   0_No   0_No   0_No 1.86   1_Adeq    7    5
## 5 1_Yes 0_No    39   0_No   0_No   0_No   0_No 4.36   4_Dist    1    4
## 6 1_Yes 0_No    11  1_Yes   0_No   0_No  1_Yes 4.29   4_Dist    0    1
##   pub6 pub9      work workadmn worktch workuniv faculty totpub jobimp
## 1    2    4 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      7   1.84
## 2    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     25   1.99
## 3    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     23   1.53
## 4   10    9 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     24   1.74
## 5    2    3 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      9   1.58
## 6    3    0 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      4   1.47
##   jobprst
## 1  1_Adeq
```

```
## 2  1_Adeq
## 3  1_Adeq
## 4  1_Adeq
## 5  1_Adeq
## 6  1_Adeq
```

```
myvars <- c("jobprst", "pub1", "phd", "female")
datasub <- data[myvars]
datasub$jobprst <- as.factor(datasub$jobprst)
datasub$female <- as.factor(datasub$female)
datasub$phd <- as.numeric(datasub$phd)
datasub$pub1 <- as.numeric(datasub$pub1)
dataclean <- na.omit(datasub)
sum(is.na(dataclean))
```

```
## [1] 0
```

```
dataclean[!complete.cases(dataclean), ]
```

```
## [1] jobprst pub1    phd     female
## <0 rows> (or 0-length row.names)
```

4.2 Produce Table including Distribution of Output Variable

```
describe(dataclean)
```

```
##          vars   n mean   sd median trimmed  mad min   max range  skew
## jobprst*    1 264 2.35 0.74   2.00    2.37 1.48   1  4.00  3.00  0.11
## pub1        2 264 2.32 2.58   1.00    1.91 1.48   0 19.00 19.00  2.08
## phd         3 264 3.18 1.01   3.19    3.21 1.56   1  4.66  3.66 -0.14
## female*     4 264 1.34 0.48   1.00    1.31 0.00   1  2.00  1.00  0.65
##          kurtosis   se
## jobprst*    -0.31 0.05
## pub1         7.32 0.16
## phd         -1.24 0.06
## female*     -1.58 0.03
```

```
# Example on how to do it manually
coltitles <- cbind("Name", "Mean", "Std Dev", "Min", "Max", "Description")
firstcol <- rbind("jobprst", "pub1", "female", "phd")
seccol <- rbind("", "2.32", "0.35", "3.18")
thirdcol <- rbind("", "2.58", "", "1.01")
fourcol <- rbind(1, min(dataclean$pub1), 0, min(dataclean$phd))
fifthcol <- rbind(4, max(dataclean$pub1), 1, "4.66")
sixcol <- rbind("1=Adeq, 2=Good, 3=Strong, 4=Dist", "Publications",
    "1=Female, 0=Male", "Phd program prestige")
allcol <- cbind(firstcol, seccol, thirdcol, fourcol, fifthcol,
    sixcol)
allcol
```

```
##      [,1]      [,2]   [,3]   [,4] [,5]
## [1,] "jobprst" ""     ""     "1"  "4"
## [2,] "pub1"    "2.32" "2.58" "0"  "19"
## [3,] "female"  "0.35" ""     "0"  "1"
## [4,] "phd"     "3.18" "1.01" "1"  "4.66"
```

```
##      [,6]
## [1,] "1=Adeq, 2=Good, 3=Strong, 4=Dist"
## [2,] "Publications"
## [3,] "1=Female, 0=Male"
## [4,] "Phd program prestige"
```

```
# Example of publication ready table
kable(allcol, col.names=coltitles, digits=c(0,3,3,0,0,0))
```

Name	Mean	Std Dev	Min	Max	Description
jobprst			1	4	1=Adeq, 2=Good, 3=Strong, 4=Dist
pub1	2.32	2.58	0	19	Publications
female	0.35		0	1	1=Female, 0=Male
phd	3.18	1.01	1	4.66	Phd program prestige

4.3 Estimate the Ordered Logit

```
# Load library to run polr
library(MASS)
```

```
# Present model
mod.olog <- polr(jobprst ~ pub1 + female + phd, data = dataclean,
    Hess = T)
summary(mod.olog)
```

```
## Call:
## polr(formula = jobprst ~ pub1 + female + phd, data = dataclean,
##     Hess = T)
##
## Coefficients:
##                Value Std. Error t value
## pub1          0.1079    0.04811   2.242
## female1_Yes -0.6974    0.26171  -2.665
## phd           1.1300    0.14440   7.825
##
## Intercepts:
##                   Value   Std. Error t value
## 1_Adeq|2_Good      0.9275  0.4268     2.1729
## 2_Good|3_Strong    4.0032  0.4997     8.0117
## 3_Strong|4_Dist    7.0346  0.6297    11.1719
##
## Residual Deviance: 509.0304
## AIC: 521.0304
```

4.4-6 Odds Ratios

```
# Save odds ratios, CI, stx
bHat <- coef(mod.olog)
ci <- exp(cbind(confint.default(mod.olog)))
sdX <- c(sd(dataclean$pub1), NA, sd(dataclean$phd))
```

```r
# Calculate precentage changes
perc.bHat <- (exp(bHat) - 1) * 100
perc.sdx <- (exp(bHat * sdX) - 1) * 100

# Put everything together and round to 4 digits
facOdds <- round(cbind(bHat, ci, exp(bHat), exp(bHat * sdX),
    sdX, perc.bHat, perc.sdx), digits = 3)
colnames(facOdds) <- c("b", "LL", "UL", "e^b", "e^(b*sd)", "SD of X",
    "%", "%StdX")
facOdds
```

```
##                   b    LL    UL   e^b e^(b*sd) SD of X       %   %StdX
## pub1          0.108 1.014 1.224 1.114    1.321   2.581  11.391  32.102
## female1_Yes -0.697 0.298 0.832 0.498       NA      NA -50.210      NA
## phd           1.130 2.333 4.108 3.096    3.114   1.005 209.574 211.392
```

Interpretation.

- The odds of receiving a higher ranked job are 0.50 times smaller for women than men, holding other variables constant *(p<0.01)*.
- For a standard deviation increase in publications, about 2.6, the odds of receiving a higher ranked job increases by 32 percent, holder other variables constant *(p<0.0001)*.

4.7 Compute Discrete Change for C and B

C: Pub1

```r
# Set up the values of the changes for C (+1 centered)
pub.centL <- mean(dataclean$pub1) - 0.5
pub.centU <- mean(dataclean$pub1) + 0.5

# Create subsets and check
s.U <- with(dataclean, data.frame(female = "0_No", phd = mean(phd),
    pub1 = pub.centU))
s.U
```

```
##   female      phd    pub1
## 1   0_No 3.181894 2.82197
```

```r
s.L <- with(dataclean, data.frame(female = "0_No", phd = mean(phd),
    pub1 = pub.centL))
s.L
```

```
##   female      phd    pub1
## 1   0_No 3.181894 1.82197
```

```r
# Apply our model to subsample
pred.U <- predict(mod.olog, s.U, type = "probs", se.fit = TRUE)
pred.L <- predict(mod.olog, s.L, type = "probs", se.fit = TRUE)

# Discrete change
diff <- pred.U - pred.L
diff
```

```
##       1_Adeq       2_Good     3_Strong       4_Dist
```

```
## -0.005246125 -0.021551520  0.022693271  0.004104374
```

This is different than in Stata because female is held at typical value and not at mean

Interpretation. For a male scientist, an additional publication centered at the mean increases the probability of receiving a strong job by 0.02, other variables held at their mean.

B: Female

```
# Subsample
s.fem <- with(dataclean, data.frame(pub1 = mean(pub1), female = factor(1:2,
    levels = 1:2, labels = levels(female)), phd = mean(phd)))

# Prediction
pred.fem <- predict(mod.olog, s.fem, type = "probs", se.fit = TRUE)

# Discrete change
diff <- pred.fem[2, ] - pred.fem[1, ]
diff
```

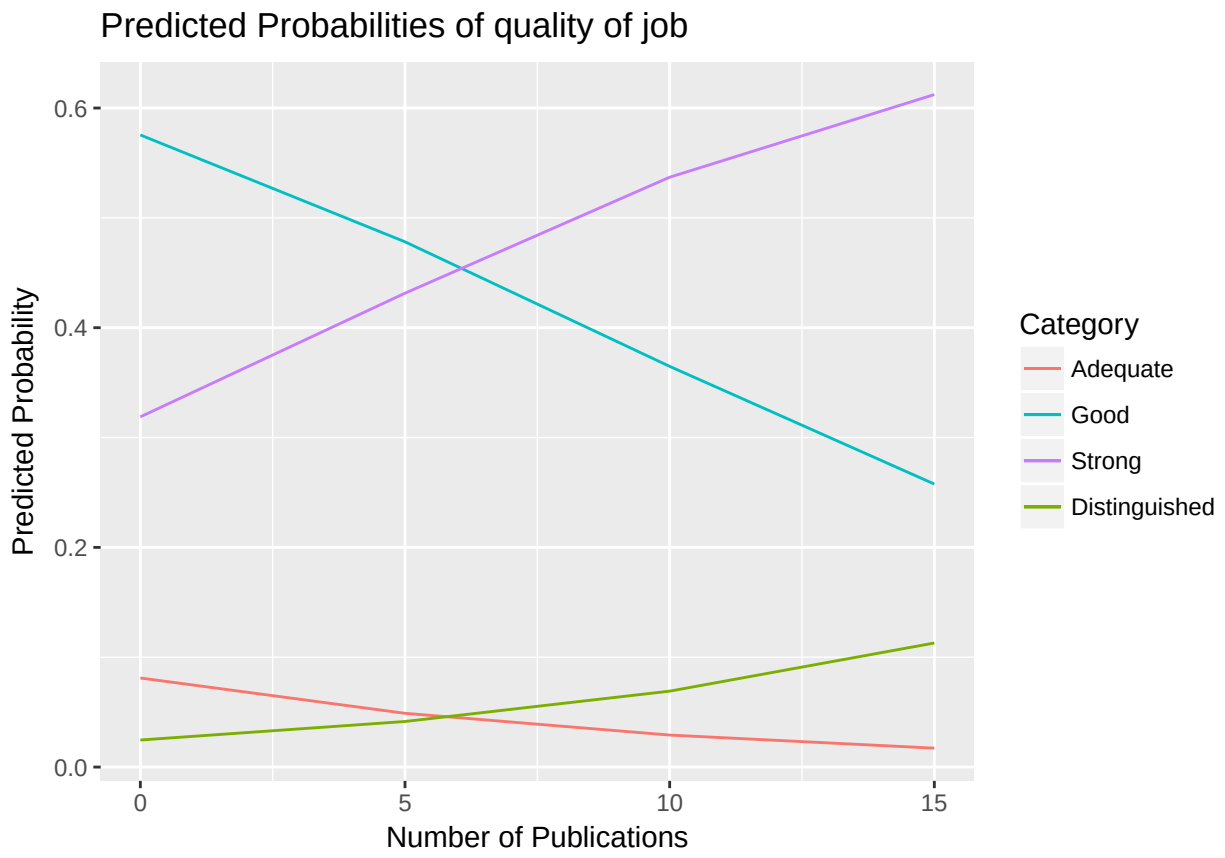
```
##      1_Adeq      2_Good    3_Strong      4_Dist
##  0.04661480  0.11569375 -0.14282526 -0.01948329
```

4.8 Graph Predcited Probabilities

```
# Load necessary libraries
library(effects)
```

```
## Warning: package 'effects' was built under R version 3.3.3
```

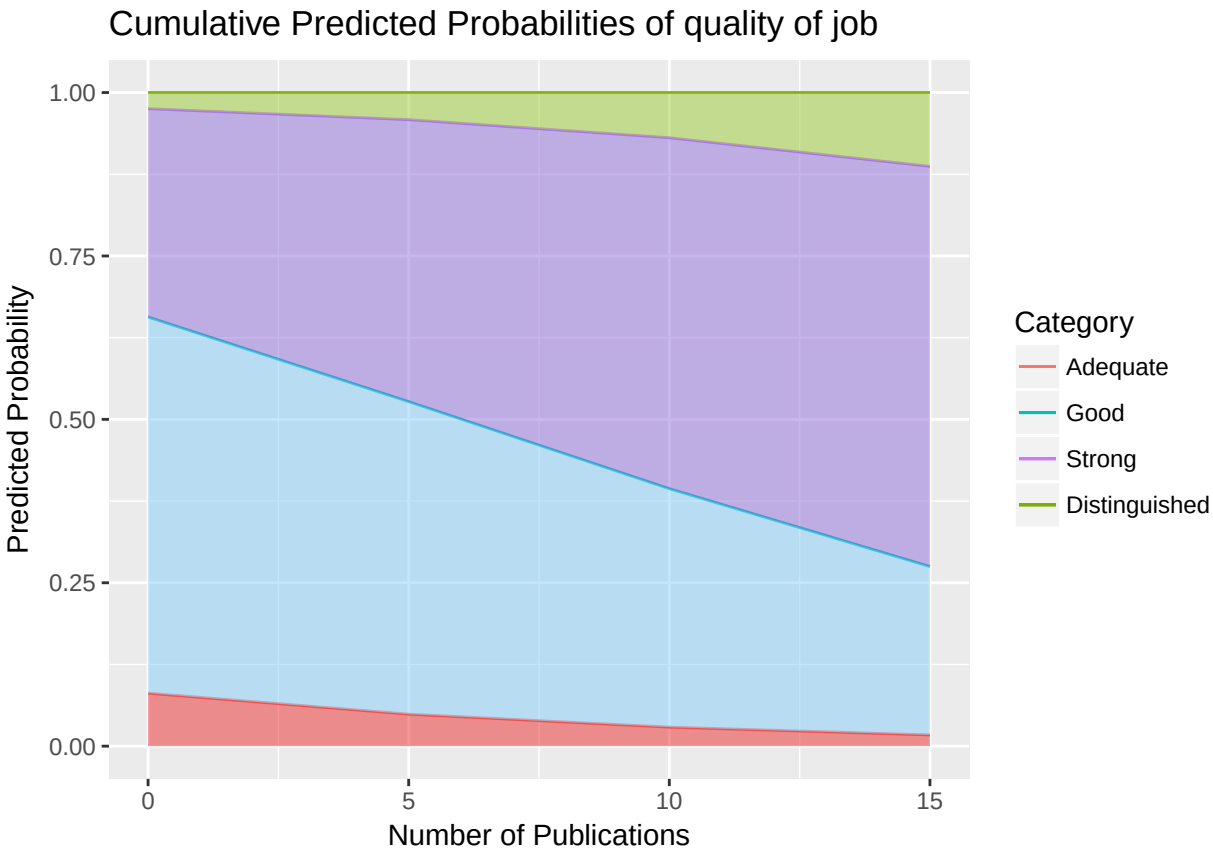
```
# Create a data frame of the data
effect1 <- effect("pub1", mod.olog, typical = mean)
effect1 <- data.frame(effect1)

# Plot Graph
ggplot(effect1, aes(pub1)) + geom_line(aes(y = prob.X1_Adeq,
    colour = "Adequate")) + geom_line(aes(y = prob.X2_Good, colour = "Good")) +
    geom_line(aes(y = prob.X3_Strong, colour = "Strong")) + geom_line(aes(y = prob.X4_Dist,
    colour = "Distinguished")) + scale_colour_hue(breaks = c("Adequate",
    "Good", "Strong", "Distinguished")) + labs(title = "Predicted Probabilities of quality of job",
    x = "Number of Publications", y = "Predicted Probability",
    colour = "Category")
```




```
# Save Graph
ggsave("icpsrcda04-ordinal-fig2.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
# Create cummulative probabilities
effect1$prob.X1X2 <- effect1$prob.X1_Adeq + effect1$prob.X2_Good
effect1$prob.X1X2X3 <- effect1$prob.X1X2 + effect1$prob.X3_Strong
effect1$prob.X1X2X3X4 <- effect1$prob.X1X2X3 + effect1$prob.X4_Dist

# Plot
ggplot(effect1, aes(pub1)) + geom_line(aes(y = prob.X1_Adeq,
    colour = "Adequate")) + geom_line(aes(y = prob.X1X2, colour = "Good")) +
    geom_line(aes(y = prob.X1X2X3, colour = "Strong")) + geom_line(aes(y = prob.X1X2X3X4,
    colour = "Distinguished")) + scale_colour_hue(breaks = c("Adequate",
    "Good", "Strong", "Distinguished")) + geom_ribbon(data = effect1,
    aes(x = pub1, ymin = 0, ymax = prob.X1_Adeq), fill = "firebrick2",
    alpha = "0.5") + geom_ribbon(data = effect1, aes(x = pub1,
    ymin = prob.X1_Adeq, ymax = prob.X1X2), fill = "lightskyblue",
    alpha = "0.5") + geom_ribbon(data = effect1, aes(x = pub1,
    ymin = prob.X1X2, ymax = prob.X1X2X3), fill = "mediumpurple",
    alpha = "0.5") + geom_ribbon(data = effect1, aes(x = pub1,
    ymin = prob.X1X2X3, ymax = prob.X1X2X3X4), fill = "olivedrab3",
    alpha = "0.5") + labs(title = "Cumulative Predicted Probabilities of quality of job",
    x = "Number of Publications", y = "Predicted Probability",
    colour = "Category")
```



Interpret. First, the probabiliy of obtaining a job in the lowest ranking category (adequate) is quite low for scientists regardless of the number of publications, peaking at aonly around .10. Similarly, the probability of attaining a distinguished position is quite low, even at the highest level of publication (around.20). However, individuals with more publications have a much higher probability of attaining a job that is either strong or dinstinguished compared to those who have published at lower levels. As such,the most dramatic change accross the range of publications appears to be in the probability of obtaining a good job, which decreases by 0.389 as publications increase (CI=-0.658,-0.120) and is offset by an increase by 0.320 (CI:0.207,0.433) in the probability of obtaining a prestigious job

4.14 Testing the Parallel Regression Assumption

```
# load libraries
library(ordinal)
```

```
# Create models with forced prallel assumption for each
# variable
ologit.parallel <- clm(jobprst ~ pub1 + female + phd, data = dataclean,
    link = "logit")
ologit.test1 <- clm(jobprst ~ female + phd, nominal = ~pub1,
    data = dataclean, link = "logit")
ologit.test2 <- clm(jobprst ~ pub1 + female, nominal = ~phd,
    data = dataclean, link = "logit")
ologit.test3 <- clm(jobprst ~ pub1 + phd, nominal = ~female,
    data = dataclean, link = "logit")
```

```r
# Test these models with regular model using LR test
anova(ologit.parallel, ologit.test1)
```

```
## Likelihood ratio tests of cumulative link models:
##
##                 formula:                       nominal: link: threshold:
## ologit.parallel jobprst ~ pub1 + female + phd ~1       logit flexible
## ologit.test1    jobprst ~ female + phd        ~pub1    logit flexible
##
##                 no.par    AIC  logLik LR.stat df Pr(>Chisq)
## ologit.parallel      6 521.03 -254.52
## ologit.test1         8 522.04 -253.02  2.9926  2      0.224
```

```r
anova(ologit.parallel, ologit.test2)
```

```
## Likelihood ratio tests of cumulative link models:
##
##                 formula:                       nominal: link: threshold:
## ologit.parallel jobprst ~ pub1 + female + phd ~1       logit flexible
## ologit.test2    jobprst ~ pub1 + female       ~phd     logit flexible
##
##                 no.par    AIC  logLik LR.stat df Pr(>Chisq)
## ologit.parallel      6 521.03 -254.52
## ologit.test2         8 507.70 -245.85  17.327  2  0.0001728 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(ologit.parallel, ologit.test3)
```

```
## Likelihood ratio tests of cumulative link models:
##
##                 formula:                       nominal: link: threshold:
## ologit.parallel jobprst ~ pub1 + female + phd ~1       logit flexible
## ologit.test3    jobprst ~ pub1 + phd          ~female  logit flexible
##
##                 no.par    AIC  logLik LR.stat df Pr(>Chisq)
## ologit.parallel      6 521.03 -254.52
## ologit.test3         8 513.61 -248.80  11.422  2   0.003309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation. The variables phd *($LR\chi^2$=17.327, df=2, p<.001)* and female *($LR\chi^2$=11.422, df=2, p<.001)* both are problematic and lead to violation of the parallel assumption.

Section 5: Models for Nominal Outcomes

For a fuller discussion of testing and assessing fit in Stata, see Chapter 6 of L&F (2005).

Resource for R: http://www.ats.ucla.edu/stat/r/dae/mlogit.htm

5.1.a Set Up

```
# Specify working folder
setwd("C:/Users/TamaravdD/Box Sync/Teaching/2017-2.5-ICPSR-CA/Rlab")
```

```
# Load libraries
library(car)
library(foreign)
library(psych)
library(easyGgplot2)
library(aod)
library(MASS)
library(effects)
library(ordinal)
```

5.1.b Load the Data

```
data <- as.data.frame(read.dta("https://shawnana79.github.io/data/cda_scireview3.dta",
    convert.factor = TRUE))
```

5.1.c Examine Data, Keep Variables, Drop Missing

```
head(data)
```

```
##      id cit1 cit3 cit6 cit9 enrol  fel felclass fellow female mcit3 mcitt
## 1 62352    3    1    4   14     8 1.60   1_Adeq   0_No  1_Yes     2     5
## 2 57249   21   24   20   14     4 4.50   4_Dist  1_Yes   0_No    17    26
## 3 62101    8    5   14   18     7 1.30   1_Adeq   0_No  1_Yes     1    10
## 4 57339   19   23   25   24     5 1.86   1_Adeq   0_No   0_No    34     6
## 5 62083    4    6    3   12     7 4.36   4_Dist   0_No  1_Yes    72     1
## 6 57071    0    1    3    9     5 4.29   4_Dist   0_No  1_Yes    23     1
##   mmale mnas mpub3 nopub1 nopub3 nopub6 nopub9  phd phdclass pub1 pub3
## 1 1_Yes 0_No     2  1_Yes   0_No   0_No   0_No 1.60   1_Adeq    0    1
## 2 1_Yes 0_No    11   0_No   0_No   0_No   0_No 2.58   2_Good    7   12
## 3 1_Yes 0_No    11   0_No   0_No   0_No   0_No 1.30   1_Adeq    5   10
## 4 1_Yes 0_No    25   0_No   0_No   0_No   0_No 1.86   1_Adeq    7    5
## 5 1_Yes 0_No    39   0_No   0_No   0_No   0_No 4.36   4_Dist    1    4
## 6 1_Yes 0_No    11  1_Yes   0_No   0_No  1_Yes 4.29   4_Dist    0    1
##   pub6 pub9      work workadmn worktch workuniv faculty totpub jobimp
## 1    2    4 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      7   1.84
## 2    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     25   1.99
## 3    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     23   1.53
## 4   10    9 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     24   1.74
## 5    2    3 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      9   1.58
```

```
## 6    3    0 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      4   1.47
##   jobprst
## 1  1_Adeq
## 2  1_Adeq
## 3  1_Adeq
## 4  1_Adeq
## 5  1_Adeq
## 6  1_Adeq
```

```
myvars <- c("jobprst", "female", "mcit3", "pub1", "phd")
datasub <- data[myvars]
datasub$jobprst <- as.factor(datasub$jobprst)
datasub$pub1 <- as.numeric(datasub$pub1)
datasub$phd <- as.numeric(datasub$phd)
datasub$female <- as.factor(datasub$female)
datasub$mcit3 <- as.numeric(datasub$mcit3)
dataclean <- na.omit(datasub)
sum(is.na(dataclean))
```

```
## [1] 0
```

```
dataclean[!complete.cases(dataclean), ]
```

```
## [1] jobprst female  mcit3   pub1    phd
## <0 rows> (or 0-length row.names)
```

5.2 Verify Data is Clean using Descriptives

```
summary(dataclean)
```

```
##      jobprst      female        mcit3             pub1
##  1_Adeq  : 29   0_No :173   Min.   :  0.00   Min.   : 0.000
##  2_Good  :128   1_Yes: 91   1st Qu.:  4.00   1st Qu.: 0.000
##  3_Strong: 93               Median : 12.50   Median : 1.000
##  4_Dist  : 14               Mean   : 20.72   Mean   : 2.322
##                             3rd Qu.: 24.00   3rd Qu.: 4.000
##                             Max.   :129.00   Max.   :19.000
##       phd
##  Min.   :1.000
##  1st Qu.:2.260
##  Median :3.190
##  Mean   :3.182
##  3rd Qu.:4.290
##  Max.   :4.660
```

```
describe(dataclean)
```

```
##          vars   n  mean    sd median trimmed   mad min    max  range  skew
## jobprst*    1 264  2.35  0.74   2.00    2.37  1.48   1   4.00   3.00  0.11
## female*     2 264  1.34  0.48   1.00    1.31  0.00   1   2.00   1.00  0.65
## mcit3       3 264 20.72 25.45  12.50   15.33 14.08   0 129.00 129.00  2.28
## pub1        4 264  2.32  2.58   1.00    1.91  1.48   0  19.00  19.00  2.08
## phd         5 264  3.18  1.01   3.19    3.21  1.56   1   4.66   3.66 -0.14
##          kurtosis   se
## jobprst*    -0.31 0.05
```

```
## female*     -1.58 0.03
## mcit3        5.56 1.57
## pub1         7.32 0.16
## phd         -1.24 0.06
```

5.4 Multinomial Logit

```
# load library for multinom
library(nnet)
```

```
# Chose reference category for data
dataclean$jobprst <- relevel(dataclean$jobprst, ref = "4_Dist")

# Run multinomial logit using 'multinom'
mlogit.mod <- multinom(jobprst ~ pub1 + female + phd + mcit3,
    data = dataclean, Hess = TRUE)
```

```
## # weights:  24 (15 variable)
## initial  value 365.981711
## iter  10 value 277.317447
## iter  20 value 236.449067
## final  value 236.389133
## converged
```

```
summary(mlogit.mod)
```

```
## Call:
## multinom(formula = jobprst ~ pub1 + female + phd + mcit3, data = dataclean,
##     Hess = TRUE)
##
## Coefficients:
##          (Intercept)        pub1 female1_Yes        phd       mcit3
## 1_Adeq      8.320968 -0.14747811    1.695229 -1.9187033 -0.01206232
## 2_Good     10.226742 -0.20968703    2.601098 -2.0747491 -0.02144588
## 3_Strong    5.498688 -0.09374101    1.390095 -0.6595403 -0.02294845
##
## Std. Errors:
##          (Intercept)       pub1 female1_Yes       phd       mcit3
## 1_Adeq      2.349763 0.12320133    1.189386 0.6037841 0.011628090
## 2_Good      2.295593 0.10920367    1.127409 0.5772616 0.010335927
## 3_Strong    2.262484 0.09154044    1.108390 0.5595469 0.008537916
##
## Residual Deviance: 472.7783
## AIC: 502.7783
```

We present here the factor changes in odds for each category compared to the reference category (Distinguished job):

```
# Save coefficients, z-scores, and standard deviation
bHat <- coef(mlogit.mod)[, 2:5]
z <- summary(mlogit.mod)$coefficients[, 2:5]/summary(mlogit.mod)$standard.errors[,
    2:5]
sdX <- cbind(cbind(replicate(3, sd(dataclean$pub1))), cbind(replicate(3,
    NA)), cbind(replicate(3, sd(dataclean$phd))), cbind(replicate(3,
    sd(dataclean$mcit3))))
```

```
# Create tables for each covariate
facOdds.pub <- cbind(bHat[, 1], exp(bHat[, 1]), exp(bHat[, 1] *
    sdX[, 1]), z[, 1], sdX[, 1])
facOdds.fem <- cbind(bHat[, 2], exp(bHat[, 2]), exp(bHat[, 2] *
    sdX[, 2]), z[, 2], sdX[, 2])
facOdds.phd <- cbind(bHat[, 3], exp(bHat[, 3]), exp(bHat[, 3] *
    sdX[, 3]), z[, 3], sdX[, 3])
facOdds.mcit3 <- cbind(bHat[, 4], exp(bHat[, 4]), exp(bHat[,
    4] * sdX[, 4]), z[, 4], sdX[, 4])

# Name the columns
colnames(facOdds.pub) <- c("b", "e^b", "e^(b*sd)", "z-score",
    "SD of X")
colnames(facOdds.fem) <- c("b", "e^b", "e^(b*sd)", "z-score",
    "SD of X")
colnames(facOdds.phd) <- c("b", "e^b", "e^(b*sd)", "z-score",
    "SD of X")
colnames(facOdds.mcit3) <- c("b", "e^b", "e^(b*sd)", "z-score",
    "SD of X")

# Present the factor changes in odds
facOdds.pub
```

```
##                    b       e^b  e^(b*sd)   z-score  SD of X
## 1_Adeq   -0.14747811 0.8628813 0.6834498 -1.197050 2.580736
## 2_Good   -0.20968703 0.8108380 0.5820803 -1.920146 2.580736
## 3_Strong -0.09374101 0.9105186 0.7851183 -1.024039 2.580736
```

```
facOdds.fem
```

```
##                  b       e^b e^(b*sd)  z-score SD of X
## 1_Adeq   1.695229  5.447895       NA 1.425298      NA
## 2_Good   2.601098 13.478533       NA 2.307147      NA
## 3_Strong 1.390095  4.015233       NA 1.254157      NA
```

```
facOdds.phd
```

```
##                   b       e^b  e^(b*sd)   z-score SD of X
## 1_Adeq   -1.9187033 0.1467972 0.1453454 -3.177797 1.00518
## 2_Good   -2.0747491 0.1255879 0.1242454 -3.594123 1.00518
## 3_Strong -0.6595403 0.5170890 0.5153254 -1.178704 1.00518
```

```
facOdds.mcit3
```

```
##                    b       e^b  e^(b*sd)   z-score  SD of X
## 1_Adeq   -0.01206232 0.9880101 0.7357020 -1.037343 25.44536
## 2_Good   -0.02144588 0.9787824 0.5794371 -2.074887 25.44536
## 3_Strong -0.02294845 0.9773129 0.5577013 -2.687828 25.44536
```

5.5 Single Coefficient Wald and LR Test

2-tailed Z test

The code to test that all coefficients associated with a given variable are equal to zero was not found. Below are the results of a 2-tailed z test for each coefficient.

```
z <- summary(mlogit.mod)$coefficients/summary(mlogit.mod)$standard.errors
chi2 <- z^2
chi2
```

```
##          (Intercept)     pub1 female1_Yes       phd    mcit3
## 1_Adeq     12.540059 1.432928    2.031474 10.098395 1.076081
## 2_Good     19.846532 3.686962    5.322929 12.917717 4.305156
## 3_Strong    5.906729 1.048656    1.572909  1.389343 7.224420
```

```
p <- (1 - pnorm(abs(z), 0, 1))*2
p
```

```
##           (Intercept)       pub1 female1_Yes          phd       mcit3
## 1_Adeq   3.983196e-04 0.23128718  0.15407113 0.0014839855 0.299575887
## 2_Good   8.391482e-06 0.05483941  0.02104661 0.0003254866 0.037997009
## 3_Strong 1.508315e-02 0.30581675  0.20978505 0.2385160103 0.007191841
```

Both gender and number of citations have at least on coefficient that is significantly different than zero ($p<0.05$).

LR Test

```
Anova(mlogit.mod)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: jobprst
##        LR Chisq Df Pr(>Chisq)
## pub1      4.204  3  0.2402369
## female   16.623  3  0.0008449 ***
## phd      69.401  3  5.736e-15 ***
## mcit3     8.140  3  0.0432030 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation. The effect of gender on job prestige is significant at the .01 level *($LR\chi^2$=16.62, df=3, p=.001)*

5.6 Plot Odds Ratios

There is no easy way to plot odds ratios in R. Instead, you need to first run the multinomial logit with each different reference category and save the confidence intevals of the coefficients.

```
# Get info on for all comparisons
dataclean$jobprst <- relevel(dataclean$jobprst, ref = "1_Adeq")
mlogit.mod.1 <- multinom(jobprst ~ pub1 + female + phd + mcit3,
    data = dataclean, Hess = TRUE)
ci.1 <- confint(mlogit.mod.1, level = 0.95)

dataclean$jobprst <- relevel(dataclean$jobprst, ref = "2_Good")
mlogit.mod.2 <- multinom(jobprst ~ pub1 + female + phd + mcit3,
    data = dataclean, Hess = TRUE)
ci.2 <- confint(mlogit.mod.2, level = 0.95)
```

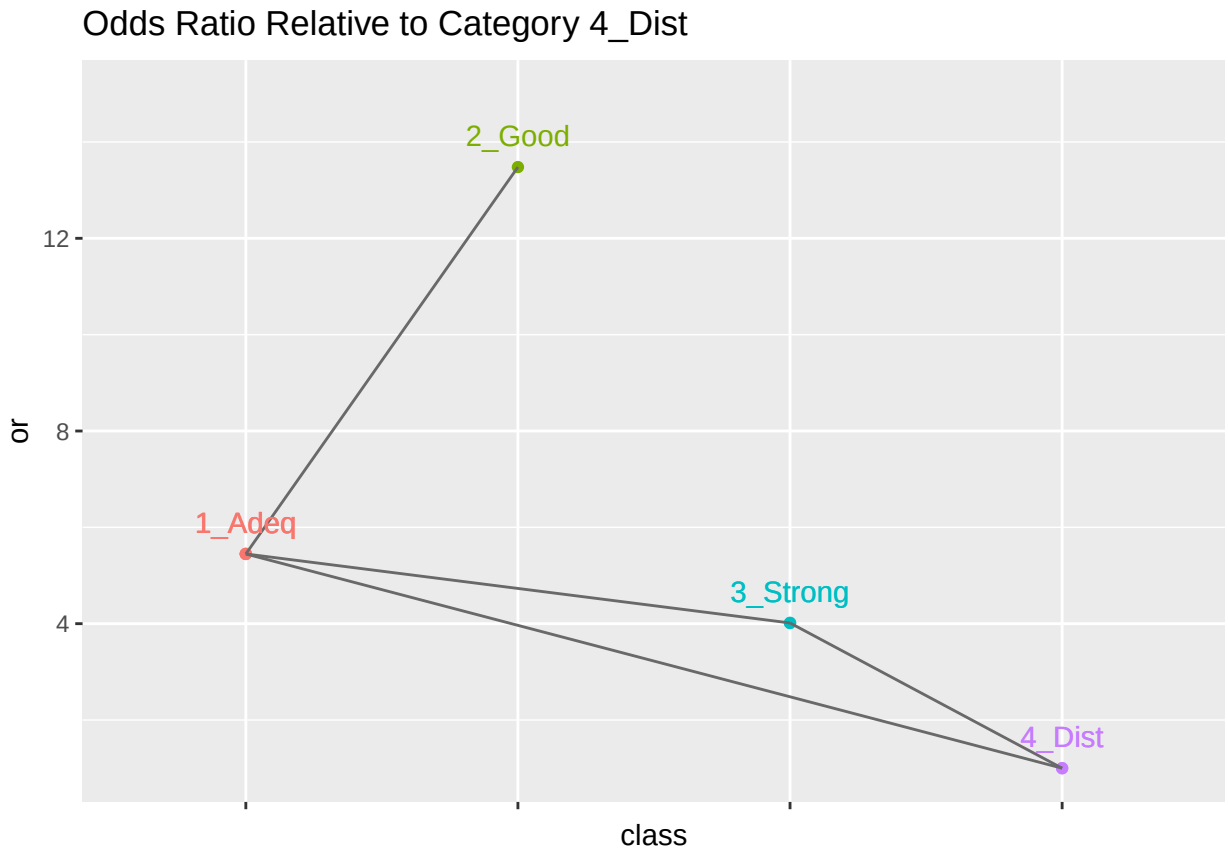
```
dataclean$jobprst <- relevel(dataclean$jobprst, ref = "3_Strong")
mlogit.mod.3 <- multinom(jobprst ~ pub1 + female + phd + mcit3,
    data = dataclean, Hess = TRUE)
ci.3 <- confint(mlogit.mod.3, level = 0.95)

dataclean$jobprst <- relevel(dataclean$jobprst, ref = "4_Dist")
mlogit.mod.4 <- multinom(jobprst ~ pub1 + female + phd + mcit3,
    data = dataclean, Hess = TRUE)
ci.4 <- confint(mlogit.mod.4, level = 0.95)

# Look at conference intervals for all comparisons

# CI for OR compared to having Distinguished job
Dist.ci <- cbind(cbind(ci.4[3, 1:2, 1]), cbind(ci.4[3, 1:2, 2]),
    cbind(ci.4[3, 1:2, 3]))
colnames(Dist.ci) <- c("3_Strong", "2_Good", "1_Adeq")
Dist.ci

# CI for OR compared to having Strong job
Strong.ci <- cbind(cbind(ci.3[3, 1:2, 1]), cbind(ci.3[3, 1:2,
    2]))
colnames(Strong.ci) <- c("2_Good", "1_Adeq")
Strong.ci

# CI for OR compared to having Good job
Good.ci <- cbind(cbind(ci.2[3, 1:2, 1]))
colnames(Good.ci) <- c("1_Adeq")
Good.ci
```

Using the information above, we see that the OR are not significant between categories 4 and 3, 4 and 1, 3 and 1, and 2 and 1.

```
# Save odd ratios compared to distinguished
or.fem <- exp(bHat[, 2])
```

To create lines between odd ratios, we need to create groups and have all values in the same group connect to each other.

```
# Create groups to have line for significance
sig <- as.factor(rbind(1, 2, 3, 1, 2, 3, 4, 4))
# Create dataframe matching groups with labels and OR
dataorplot <- data.frame(class = c("1_Adeq", "1_Adeq", "1_Adeq",
    "2_Good", "3_Strong", "4_Dist", "3_Strong", "4_Dist"), or = c(or.fem[1],
    or.fem[1], or.fem[1], or.fem[2], or.fem[3], 1, or.fem[3],
    1))
dataorplot$sig <- sig

# Plot
plot <- ggplot(dataorplot, aes(x = class, y = or, ymax = 15,
    group = sig, col = class)) + geom_point() + geom_line(col = "dimgrey") +
    geom_text(aes(label = class), hjust = 0.5, vjust = -1) +
    labs(title = "Odds Ratio Relative to Category 4_Dist")
plot + theme(axis.text.x = element_blank(), legend.position = "none")
```



Interpretation. The effect of mentor's citations is small, with a standard deviation increase in publications increasing the odds of obtaining a distinguished (4) job compared to either a good (2) or strong (3) job, but does not distinguish between a distinguished (4) job and an adequate (1) job. Females have larger odds of obtaining a good (2) job relative to either strong (3) or distinguished (4) jobs, but gender does not distinguish between obtaining a good (2) job and an adequate (1) job, or between a strong (3) and adistinguished (4) job.

5.8 Calculating Discrete Change

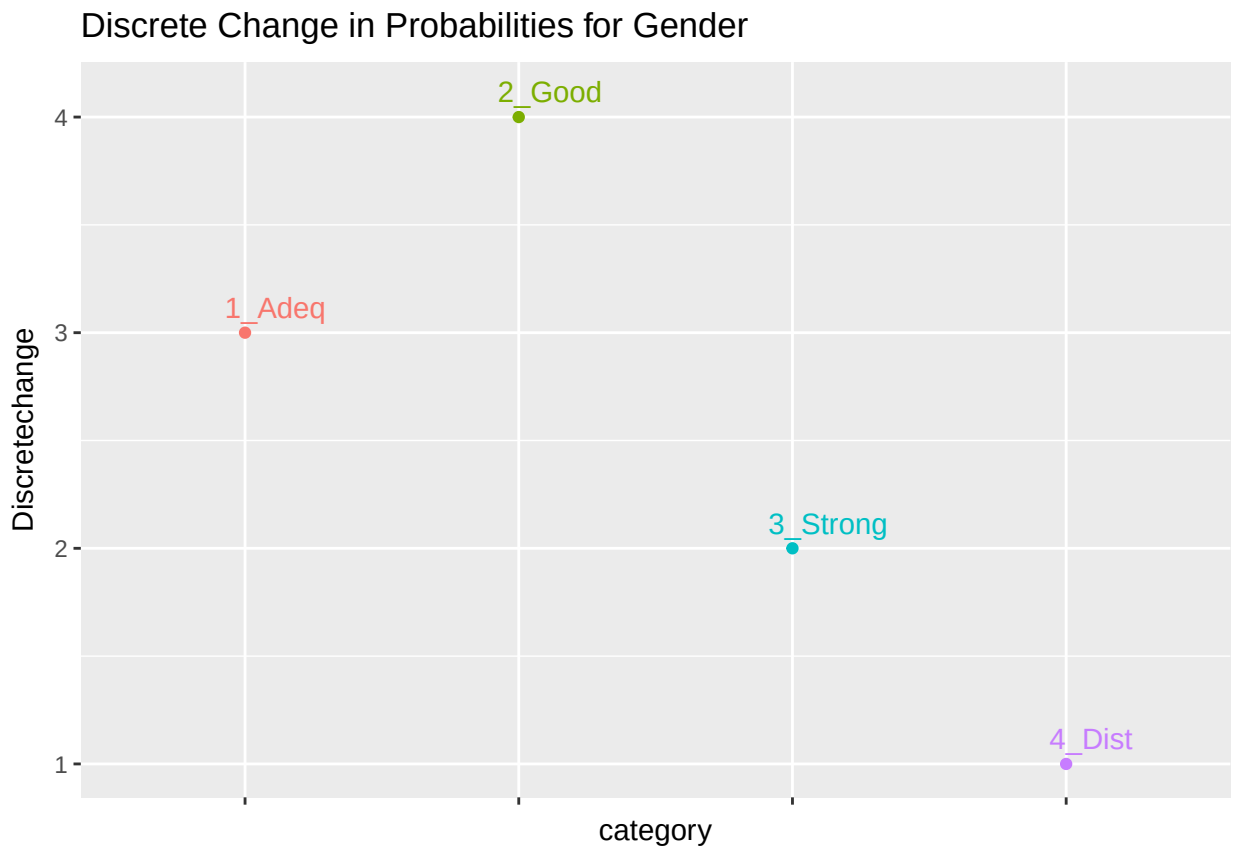
```
# For B
s.fem <- with(dataclean, data.frame(pub1 = mean(pub1), female = factor(1:2,
    levels = 1:2, labels = levels(dataclean$female)), phd = mean(phd),
    mcit3 = mean(mcit3)))
head(s.fem)

fem.pred <- predict(mlogit.mod, s.fem, type = "probs", se.fit = T)
diff.fem <- fem.pred[2, ] - fem.pred[1, ]
diff.fem

# For C
mcit3.U <- mean(dataclean$mcit3) + (0.5 * sd(dataclean$mcit3))
mcit3.L <- mean(dataclean$mcit3) - (0.5 * sd(dataclean$mcit3))

s.mcitU <- with(dataclean, data.frame(pub1 = mean(pub1), female = "0_No",
    phd = mean(phd), mcit3 = mcit3.U))
head(s.mcitU)
```

```
s.mcitL <- with(dataclean, data.frame(pub1 = mean(pub1), female = "0_No",
    phd = mean(phd), mcit3 = mcit3.L))
head(s.mcitL)

mcitU.pred <- predict(mlogit.mod, s.mcitU, type = "probs", se.fit = T)
mcitL.pred <- predict(mlogit.mod, s.mcitL, type = "probs", se.fit = T)
diff.mcit <- mcitU.pred - mcitL.pred
diff.mcit
```

5.8 Plotting Discrete Change

Example of the plot for the effect of being female.

```
# Save data
mnlpred1 <- as.numeric(diff)

# Add labels for the four outcomes
mnlpred1 <- as.data.frame(cbind(rbind("1_Adeq", "2_Good", "3_Strong",
    "4_Dist"), mnlpred1))

# Add column names and change as numeric
colnames(mnlpred1) <- c("category", "Discretechange")
mnlpred1$Discretechange <- as.numeric(mnlpred1$Discretechange)

# Plot
plot2 <- ggplot(mnlpred1, aes(x = category, y = Discretechange,
    ymax = 4.1, col = category)) + geom_point() + geom_text(aes(label = category),
    hjust = 0.2, vjust = -0.7) + labs(title = "Discrete Change in Probabilities for Gender")
plot2 + theme(axis.text.x = element_blank(), legend.position = "none")
```



Interpretation. A standard deviation increase in mentor's citations generally has a small effect on predcited probabilities for male students, increase the probability of obtaining an adequate (1) job by 0.03 and decreasing the probability of obtaining a strong (3) job by about 0.03 other variables held at their mean. Females have a higher predicted probability than men of bieng in a good (2) job of about .28 but have a .21 lower predicted probability than men of having a strong(3) job, for average scientists. Neither of the variables have much of an impact on the probability of obtaining an adequate (1) or distinguished (4) job.

5.8 Calculating Discrete Change II

We will repeat steps in 5.8, this time selection a new location for the other variables.

```
# For B
s.fem2 <- with(dataclean, data.frame(pub1 = 4, female = factor(1:2,
    levels = 1:2, labels = levels(dataclean$female)), phd = 4,
    mcit3 = mean(mcit3)))
head(s.fem2)
```

```
##   pub1 female phd    mcit3
## 1    4   0_No   4 20.71591
## 2    4  1_Yes   4 20.71591
```

```
fem.pred2 <- predict(mlogit.mod, s.fem2, type = "probs", se.fit = T)
diff.fem2 <- fem.pred2[2, ] - fem.pred2[1, ]
diff.fem2
```

```
##        4_Dist        1_Adeq        2_Good      3_Strong
## -0.0729789005 -0.0001926537  0.2498195291 -0.1766479749
```

```
# For C

s.mcitU2 <- with(dataclean, data.frame(pub1 = 4, female = "0_No",
    phd = 4, mcit3 = mcit3.U))
head(s.mcitU2)
```

```
##   pub1 female phd    mcit3
## 1    4   0_No   4 33.43859
```

```
s.mcitL2 <- with(dataclean, data.frame(pub1 = 4, female = "0_No",
    phd = 4, mcit3 = mcit3.L))
head(s.mcitL2)
```

```
##   pub1 female phd    mcit3
## 1    4   0_No   4 7.993227
```

```
mcitU.pred2 <- predict(mlogit.mod, s.mcitU2, type = "probs",
    se.fit = T)
mcitL.pred2 <- predict(mlogit.mod, s.mcitL2, type = "probs",
    se.fit = T)
diff.mcit2 <- mcitU.pred2 - mcitL.pred2
diff.mcit2
```

```
##       4_Dist       1_Adeq       2_Good     3_Strong
##  0.045366012  0.014502600 -0.007007129 -0.052861483
```

Other way to Plot Predicted Probabilities

```
# Load library to reshape long-wide
library(reshape2)
```

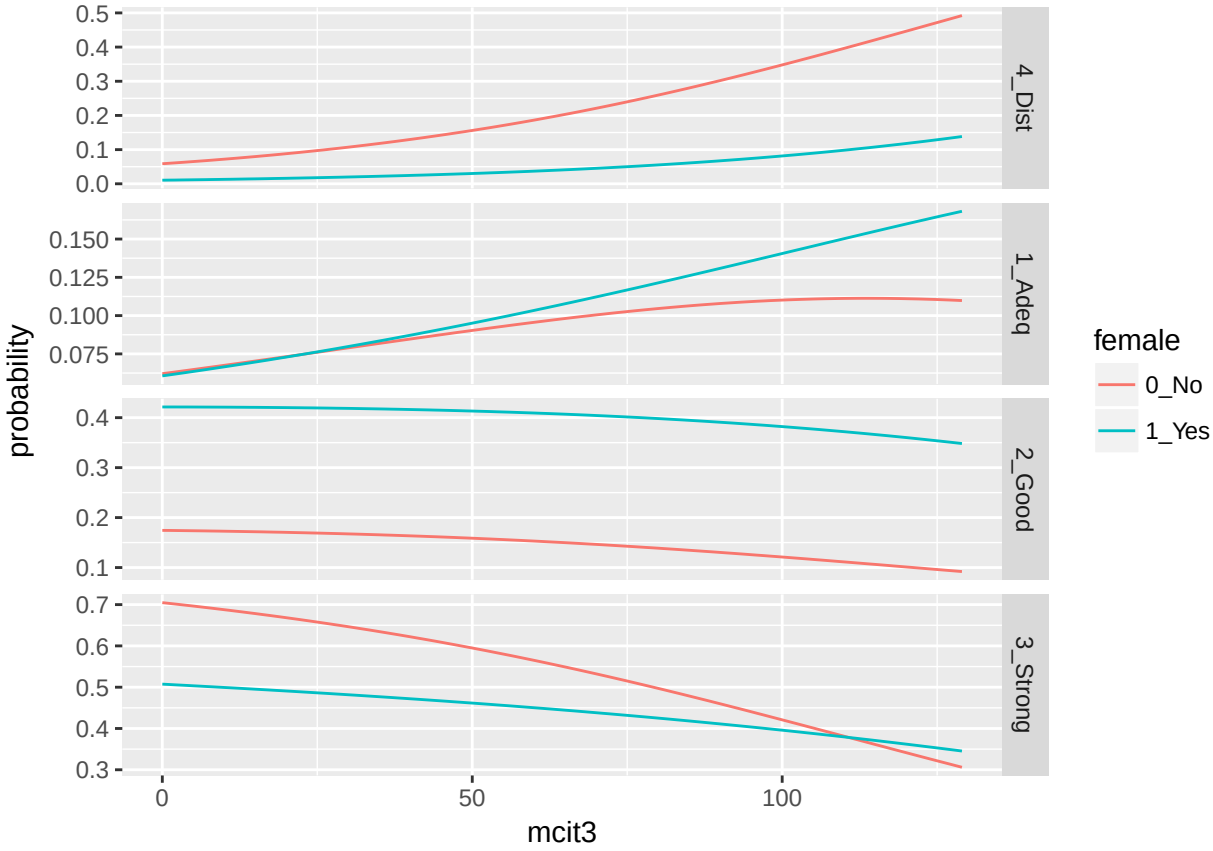
```
# create dataframe
mnlplot <- with(dataclean, data.frame(female = rep(c("1_Yes",
    "0_No"), each = 130), pub1 = 4, phd = 4, mcit3 = rep(c(0:129),
    2)))

# Store the predicted probabilities
pp <- cbind(mnlplot, predict(mlogit.mod, newdata = mnlplot, type = "probs",
    se = TRUE))

# calculate the mean probabilities for each level of gender
# and mcit3
by(pp[, 5:8], pp$female, colMeans)
```

```
## pp$female: 0_No
##     4_Dist     1_Adeq     2_Good   3_Strong
## 0.22811466 0.09383634 0.14399934 0.53404965
## ------------------------------------------------------------
## pp$female: 1_Yes
##    4_Dist    1_Adeq    2_Good  3_Strong
## 0.0518053 0.1099411 0.3996390 0.4386146
```

```
lpp <- melt(pp, id.vars = c("female", "mcit3", "pub1", "phd"),
    value.name = "probability")
head(lpp)  # view first few rows
```

```
##   female mcit3 pub1 phd variable probability
## 1  1_Yes     0    4   4   4_Dist  0.01052342
## 2  1_Yes     1    4   4   4_Dist  0.01075113
## 3  1_Yes     2    4   4   4_Dist  0.01098364
## 4  1_Yes     3    4   4   4_Dist  0.01122105
## 5  1_Yes     4    4   4   4_Dist  0.01146346
## 6  1_Yes     5    4   4   4_Dist  0.01171097
```

```
# Plot probability for each outcome by gender and citations
ggplot(lpp, aes(x = mcit3, y = probability, colour = female)) +
    geom_line() + facet_grid(variable ~ ., scales = "free")
```



Interpretation. For highly productive male scientists from high-prestige PhD programs, a standard deviation increase in mentor's citations generally has only a small effect on predicted probabilities, increasing the probability of obtaining a distinguished (4) job by .05 and decreasing the probability of obtaining a strong (3) job by about .05. Highly-productive females from high-prestige universities have a predicted probability of attaining a good (2) job that is about .25 higher than their male counterparts, and similarly a .18 lower predicted probability of attaining a strong (3) job, as well as a .07 lower predicted probability of obtaining a distinguished (4) job. Neither ofthe variables have much of an impact on the probability of obtaining an adequate (1) job.

Section 6: Models for Count Outcomes

For a fuller discussion of testing and assessing fit in Stata, see Chapter 7 of L&F (2005).

Resource for R: * Poisson: http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm * Negative Binomial: http://www.ats.ucla.edu/stat/r/dae/nbreg.htm * Zero-inflated Poisson: http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm * Zero-inflated Negative Binomial: http://www.ats.ucla.edu/stat/r/dae/zinbreg.htm

6.1.a Set Up

```
# Specify working folder
setwd("C:/Users/TamaravdD/Box Sync/Teaching/2017-2.5-ICPSR-CA/Rlab")
```

```
# Load libraries
library(car)
library(foreign)
library(psych)
library(easyGgplot2)
library(aod)
library(MASS)
library(effects)
library(ordinal)
library(reshape2)
library(nnet)
```

6.1.b Load the Data

```
data <- as.data.frame(read.dta("https://shawnana79.github.io/data/cda_scireview3.dta",
    convert.factor = TRUE))
```

6.1.c Examine Data, Keep Variables, Drop Missing, and Verify

```
head(data)
```

```
##      id cit1 cit3 cit6 cit9 enrol  fel felclass fellow female mcit3 mcitt
## 1 62352    3    1    4   14     8 1.60   1_Adeq   0_No  1_Yes     2     5
## 2 57249   21   24   20   14     4 4.50   4_Dist  1_Yes   0_No    17    26
## 3 62101    8    5   14   18     7 1.30   1_Adeq   0_No  1_Yes     1    10
## 4 57339   19   23   25   24     5 1.86   1_Adeq   0_No   0_No    34     6
## 5 62083    4    6    3   12     7 4.36   4_Dist   0_No  1_Yes    72     1
## 6 57071    0    1    3    9     5 4.29   4_Dist   0_No  1_Yes    23     1
##   mmale mnas mpub3 nopub1 nopub3 nopub6 nopub9  phd phdclass pub1 pub3
## 1 1_Yes 0_No     2  1_Yes   0_No   0_No   0_No 1.60   1_Adeq    0    1
## 2 1_Yes 0_No    11   0_No   0_No   0_No   0_No 2.58   2_Good    7   12
## 3 1_Yes 0_No    11   0_No   0_No   0_No   0_No 1.30   1_Adeq    5   10
## 4 1_Yes 0_No    25   0_No   0_No   0_No   0_No 1.86   1_Adeq    7    5
## 5 1_Yes 0_No    39   0_No   0_No   0_No   0_No 4.36   4_Dist    1    4
## 6 1_Yes 0_No    11  1_Yes   0_No   0_No  1_Yes 4.29   4_Dist    0    1
##   pub6 pub9      work workadmn worktch workuniv faculty totpub jobimp
## 1    2    4 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      7   1.84
```

```
## 2    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     25   1.99
## 3    6    7 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     23   1.53
## 4   10    9 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes     24   1.74
## 5    2    3 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      9   1.58
## 6    3    0 1_FacUniv     0_No   1_Yes    1_Yes   1_Yes      4   1.47
##   jobprst
## 1  1_Adeq
## 2  1_Adeq
## 3  1_Adeq
## 4  1_Adeq
## 5  1_Adeq
## 6  1_Adeq
```

```
myvars <- c("female", "enrol", "pub6", "phd")
datasub <- data[myvars]
datasub$female <- as.factor(datasub$female)
datasub$phd <- as.numeric(datasub$phd)
datasub$enrol <- as.numeric(datasub$enrol)
datasub$pub6 <- as.numeric(datasub$pub6)
dataclean <- na.omit(datasub)
sum(is.na(dataclean))
```

```
## [1] 0
```

```
dataclean[!complete.cases(dataclean), ]
```

```
## [1] female enrol  pub6   phd
## <0 rows> (or 0-length row.names)
```

```
summary(dataclean)
```

```
##    female        enrol            pub6             phd
##  0_No :173   Min.   : 3.00   Min.   : 0.000   Min.   :1.000
##  1_Yes: 91   1st Qu.: 4.00   1st Qu.: 1.000   1st Qu.:2.260
##              Median : 5.00   Median : 3.000   Median :3.190
##              Mean   : 5.53   Mean   : 3.879   Mean   :3.182
##              3rd Qu.: 6.00   3rd Qu.: 6.000   3rd Qu.:4.290
##              Max.   :14.00   Max.   :29.000   Max.   :4.660
```

```
describe(dataclean)
```

```
##         vars   n mean   sd median trimmed  mad min   max range  skew
## female*    1 264 1.34 0.48   1.00    1.31 0.00   1  2.00  1.00  0.65
## enrol      2 264 5.53 1.44   5.00    5.39 1.48   3 14.00 11.00  1.24
## pub6       3 264 3.88 4.31   3.00    3.13 2.97   0 29.00 29.00  2.18
## phd        4 264 3.18 1.01   3.19    3.21 1.56   1  4.66  3.66 -0.14
##         kurtosis   se
## female*    -1.58 0.03
## enrol       3.72 0.09
## pub6        6.69 0.27
## phd        -1.24 0.06
```

6.3 Estimate the Poisson and NBreg Regression Models

```
mod.poi <- glm(pub6~female+phd+enrol, family= "poisson", data=dataclean)
summary(mod.poi)
```

```
##
## Call:
## glm(formula = pub6 ~ female + phd + enrol, family = "poisson",
##     data = dataclean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5791  -1.6869  -0.5368   0.8345   8.1336
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.53259    0.16993   9.019  < 2e-16 ***
## female1_Yes -0.24081    0.06900  -3.490 0.000483 ***
## phd          0.18825    0.03218   5.849 4.94e-09 ***
## enrol       -0.13255    0.02408  -5.505 3.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1090.2  on 263  degrees of freedom
## Residual deviance: 1011.5  on 260  degrees of freedom
## AIC: 1687.6
##
## Number of Fisher Scoring iterations: 5
```

```
mod.neg <- glm.nb(pub6~female+phd+enrol, data=dataclean)
summary(mod.neg)
```

```
##
## Call:
## glm.nb(formula = pub6 ~ female + phd + enrol, data = dataclean,
##     init.theta = 1.225897159, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1430  -0.9642  -0.2824   0.3952   3.1238
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.60742    0.33496   4.799  1.6e-06 ***
## female1_Yes -0.28223    0.13820  -2.042  0.04114 *
## phd          0.19959    0.06505   3.068  0.00215 **
## enrol       -0.15089    0.04679  -3.225  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2259) family taken to be 1)
##
##     Null deviance: 320.20  on 263  degrees of freedom
## Residual deviance: 298.71  on 260  degrees of freedom
## AIC: 1295.4
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.226
##           Std. Err.:  0.154
##
##  2 x log-likelihood:  -1285.446
```

```
# Store values of estimates and ratios
varname <- rbind("female", "phd", "enrol")
prm <- cbind(summary(mod.poi)$coef[2:4, 1])
nbrm <- cbind(summary(mod.neg)$coef[2:4, 1])
ratio1 <- prm/nbrm
prmz <- cbind(summary(mod.poi)$coef[2:4, 3])
nbrmz <- cbind(summary(mod.neg)$coef[2:4, 3])
ratio2 <- prmz/nbrmz

# Bind, round, and add column names
poinbreg <- cbind(varname, round(prm, digits = 3), round(nbrm,
    digits = 3), round(ratio1, digits = 3), round(prmz, digits = 3),
    round(nbrmz, digits = 3), round(ratio2, digits = 3))
rownames(poinbreg) <- NULL
coltitles <- cbind("Name", "PRM", "NBRM", "Ratio coef", "PRM Z-score",
    "NBRM Z-score", "Ratio Z-scores")
```

Name	PRM	NBRM	Ratio coef	PRM Z-score	NBRM Z-score	Ratio Z-scores
female	-0.241	-0.282	0.853	-3.49	-2.042	1.709
phd	0.188	0.2	0.943	5.849	3.068	1.906
enrol	-0.133	-0.151	0.878	-5.505	-3.225	1.707

6.7 Testing NBRM vs. the PRM

Use LR test to compare models

```
X2 <- 2 * (logLik(mod.neg) - logLik(mod.poi))
X2
```

```
## 'log Lik.' 394.115 (df=5)
```

```
pchisq(X2, df = 1, lower.tail=FALSE)
```

```
## 'log Lik.' 1.052059e-87 (df=5)
```

Interpretation. Negative binomial is strongly preferred ($G^2(5)$=394, p<0.001)

Alternatively: Overdispersion tests

```
library(AER)
```

```
## Warning: package 'survival' was built under R version 3.3.3
```

```
dispersiontest(mod.poi, trafo=0)
```

```
##
##  Overdispersion test
##
## data:  mod.poi
## z = 3.6884, p-value = 0.0001128
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 10.43601
```

Interpretation. There is evidence of overdispersion of the poisson model (z-3.7, p<0.001)

6.8 Factor changes

Since the NBRM model is preferred, we only show the output for the NBRM.

```
bHat <- cbind(coef(mod.neg)[2:4])
Z <- cbind(summary(mod.neg)$coefficients[2:4, 3])
Pz <- cbind(summary(mod.neg)$coefficients[2:4, 4])
sdx <- rbind(NA, sd(dataclean$phd), sd(dataclean$enrol))

facodds <- cbind(round(bHat, digits = 4), round(Z, digits = 4),
    round(Pz, digits = 4), round(exp(bHat), digits = 4), round(exp(bHat *
        sdx), digits = 4), round(sdx, digits = 4))

colnames(facodds) <- cbind("b", "z", "P>|z|", "e^b", "e^bStdX",
    "StdX")
facodds
```

```
##                   b       z  P>|z|    e^b e^bStdX   StdX
## female1_Yes -0.2822 -2.0421 0.0411 0.7541      NA     NA
## phd          0.1996  3.0681 0.0022 1.2209  1.2222 1.0052
## enrol       -0.1509 -3.2252 0.0013 0.8599  0.8043 1.4432
```

Interpretation. Expected publications by female scientists are decreased by a factor of .75 compared to expected publications by male scientists, holding all other variables constant. A standard deviation increase in the number of years from enrollment to completion of PhD, about 1.4 years, decreases the expected number of publications by 14%, holding other variables constant.

6.10 Calculating Discrete Channge

```
# Discrete change for B
s.fem <- with(dataclean, data.frame(female = factor(1:2, levels = 1:2,
    labels = levels(dataclean$female)), phd = mean(phd), enrol = mean(enrol)))
predsfem <- predict(mod.neg, s.fem, type = "response", se.fit = TRUE)
diff.fem <- predsfem$fit[2] - predsfem$fit[1]
diff.fem
```

```
##         2
## -1.005179
```

```
# Discrete change for C
phd.U <- mean(dataclean$phd) + (0.5 * sd(dataclean$phd))
phd.L <- mean(dataclean$phd) - (0.5 * sd(dataclean$phd))
```

```r
s.phdU <- with(dataclean, data.frame(female = "0_No", phd = phd.U,
    enrol = mean(enrol)))
head(s.phdU)
```

```
##   female      phd    enrol
## 1   0_No 3.684484 5.530303
```

```r
s.phdL <- with(dataclean, data.frame(female = "0_No", phd = phd.L,
    enrol = mean(enrol)))
head(s.phdL)
```

```
##   female      phd    enrol
## 1   0_No 2.679304 5.530303
```

```r
phdU.pred <- predict(mod.neg, s.phdU, type = "response", se.fit = T)
phdL.pred <- predict(mod.neg, s.phdL, type = "response", se.fit = T)
diff.phd <- phdU.pred$fit - phdL.pred$fit
diff.phd
```

```
##         1
## 0.8214844
```

Interpretation. On average, female scientists have expected productivity that is approximately 1 publication lower than their male counterparts. A standard deviation increase (centered around the mean) in the number of years from enrollment to completion of PhD, about 1.4 yearas, for men, decreases the expected rate of productivty by .82 publications, other variables held at their mean.

6.11 Add Confidence Intervals

```r
# For B
CIfem.U <- diff.fem + (1.96 * sqrt(predsfem$se.fit[2]^2 + predsfem$se.fit[1]^2))
CIfem.L <- diff.fem - (1.96 * sqrt(predsfem$se.fit[2]^2 + predsfem$se.fit[1]^2))
cbind(diff.fem, CIfem.L, CIfem.U)
```

```
##    diff.fem   CIfem.L     CIfem.U
## 2 -1.005179 -1.935613 -0.07474541
```

```r
# For C
CIphd.U <- diff.phd + (1.96 * sqrt((phdU.pred$se.fit)^2 + (phdL.pred$se.fit)^2))
CIphd.L <- diff.phd - (1.96 * sqrt((phdU.pred$se.fit)^2 + (phdL.pred$se.fit)^2))
cbind(diff.phd, CIphd.L, CIphd.U)
```

```
##    diff.phd    CIphd.L  CIphd.U
## 1 0.8214844 -0.1586296 1.801598
```

Missing: calculating discrete change for each outcome count value

Interpretation. On average, female scientists are expected to have approximately one fewer publications than male scientists, with estimated bounds for the 95% confidence interval at (-1.94,-0.07).

6.12 ZIP and ZINP Models

```r
# Load libraries
library(pscl)
```

```r
# ZIP
mod.zip <- zeroinfl(pub6 ~ female + phd + enrol | phd, data = dataclean)
summary(mod.zip)
```

```
##
## Call:
## zeroinfl(formula = pub6 ~ female + phd + enrol | phd, data = dataclean)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.7565 -1.0719 -0.3974  0.5845  9.3104
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.83897    0.17492  10.513  < 2e-16 ***
## female1_Yes -0.12106    0.07108  -1.703   0.0886 .
## phd          0.14003    0.03348   4.182 2.89e-05 ***
## enrol       -0.13068    0.02502  -5.224 1.75e-07 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7539     0.5333  -1.414    0.157
## phd          -0.2383     0.1658  -1.437    0.151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood:  -758 on 6 Df
```

```r
# ZINP
mod.zinp <- zeroinfl(pub6 ~ female + phd + enrol | phd, data = dataclean,
    dist = "negbin", EM = TRUE)
summary(mod.zinp)
```

```
##
## Call:
## zeroinfl(formula = pub6 ~ female + phd + enrol | phd, data = dataclean,
##     dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.0621 -0.7184 -0.2603  0.4531  6.5742
##
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.73989    0.34986   4.973 6.59e-07 ***
## female1_Yes -0.27087    0.13719  -1.974  0.04833 *
## phd          0.17456    0.06954   2.510  0.01207 *
## enrol       -0.15272    0.04703  -3.247  0.00117 **
## Log(theta)   0.35164    0.21068   1.669  0.09510 .
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4576     2.0809  -0.700    0.484
## phd          -0.5433     0.8650  -0.628    0.530
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 1.4214
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -642.2 on 7 Df
```

6.14 Factor Change for ZINB

```
# Factor change for those not always zero
bHat <- cbind(coef(mod.zinp)[2:4])
Z <- cbind(summary(mod.zinp)$coefficients$count[2:4, 3])
Pz <- cbind(summary(mod.zinp)$coefficients$count[2:4, 4])
sdx <- rbind(NA, sd(dataclean$phd), sd(dataclean$enrol))
facodds <- cbind(round(bHat, digits = 4), round(Z, digits = 4),
    round(Pz, digits = 4), round(exp(bHat), digits = 4), round(exp(bHat *
        sdx), digits = 4), round(sdx, digits = 4))
colnames(facodds) <- cbind("b", "z", "P>|z|", "e^b", "e^bStdX",
    "StdX")
facodds
```

```
##                         b       z  P>|z|    e^b e^bStdX   StdX
## count_female1_Yes -0.2709 -1.9745 0.0483 0.7627      NA     NA
## count_phd          0.1746  2.5102 0.0121 1.1907  1.1918 1.0052
## count_enrol       -0.1527 -3.2473 0.0012 0.8584  0.8022 1.4432
```

```
# Factor change in odds of always being zero
bHat.0 <- cbind(coef(mod.zinp)[6])
Z.0 <- cbind(summary(mod.zinp)$coefficients$zero[2, 3])
Pz.0 <- cbind(summary(mod.zinp)$coefficients$zero[2, 4])
sdx.0 <- rbind(sd(dataclean$phd))
facodds.0 <- cbind(round(bHat.0, digits = 4), round(Z.0, digits = 4),
    round(Pz.0, digits = 4), round(exp(bHat.0), digits = 4),
    round(exp(bHat.0 * sdx.0), digits = 4), round(sdx.0, digits = 4))
colnames(facodds.0) <- cbind("b", "z", "P>|z|", "e^b", "e^bStdX",
    "StdX")
facodds.0
```

```
##                b       z  P>|z|    e^b e^bStdX   StdX
## zero_phd -0.5433 -0.6281 0.5299 0.5808  0.5792 1.0052
```

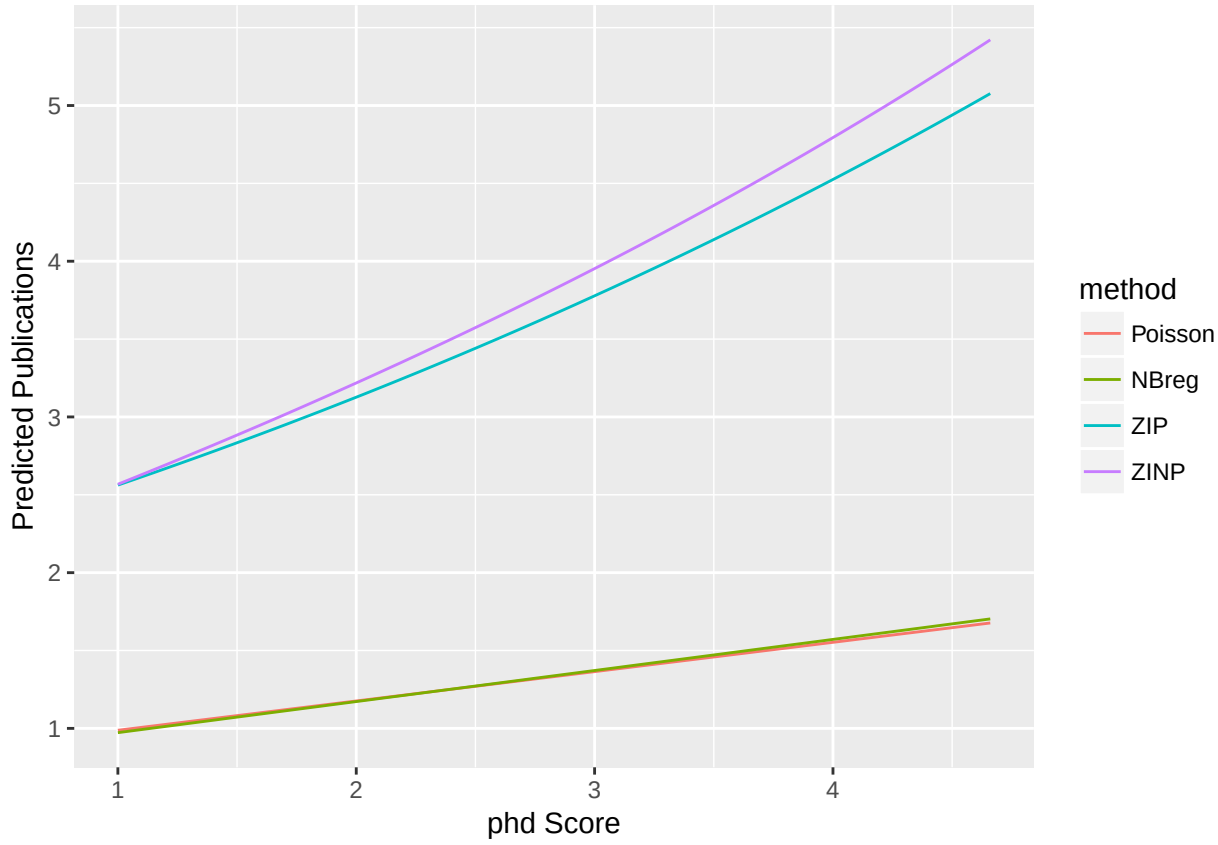
Interpretation. Among those who have the opportunity to publish, a standard deviation increase in PhD prestige increases the expected rate of publication by a factor of 1.2, for men, holding other variables constant. A standard deviation increase in PhD prestige decreases the odds of not having the opportunity to publish by a factor of .54, altho this is not significant (z=-0.63, p=0.52)

6.15 Plot Pr(y=0) across the range of C

We did not find a way to predict speicif count outcome. Instead, I present the predicted probabilities of expected outcome by model accross phd score.

```
s.plot <- data.frame(enrol = mean(dataclean$enrol), phd = rep(seq(from = min(dataclean$phd),
    to = max(dataclean$phd), length.out = 100), 2), female = "0_No")
```

```r
# Attach all predictions to subsample
s.plotall1 <- cbind(s.plot, predict(mod.poi, newdata = s.plot,
    type = "link"), rep("Poisson", 100))
colnames(s.plotall1) <- c("enrol", "phd", "female", "predict",
    "method")
s.plotall2 <- cbind(s.plot, predict(mod.neg, newdata = s.plot,
    type = "link"), rep("NBreg", 100))
colnames(s.plotall2) <- c("enrol", "phd", "female", "predict",
    "method")
s.plotall3 <- cbind(s.plot, predict(mod.zip, newdata = s.plot),
    rep("ZIP", 100))
colnames(s.plotall3) <- c("enrol", "phd", "female", "predict",
    "method")
s.plotall4 <- cbind(s.plot, predict(mod.zinp, newdata = s.plot),
    rep("ZINP", 100))
colnames(s.plotall4) <- c("enrol", "phd", "female", "predict",
    "method")
s.plot.all <- rbind(s.plotall1, s.plotall2, s.plotall3, s.plotall4)

ggplot(s.plot.all, aes(phd, predict)) + geom_line(aes(colour = method)) +
    labs(x = "phd Score", y = "Predicted Publications")
```



6.16 Compare Models

Use Bic and AIC from section 3

```r
# Vuong comparing ZINP and ZIP to non inflated
vuong(mod.poi, mod.zip)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -4.255825 model2 > model1 1.0414e-05
## AIC-corrected         -4.151742 model2 > model1 1.6498e-05
## BIC-corrected         -3.965643 model2 > model1 3.6599e-05
```

```r
vuong(mod.neg, mod.zinp)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                  -0.5063918 model2 > model1    0.30629
## AIC-corrected         1.4396717 model1 > model2    0.07498
## BIC-corrected         4.9191837 model1 > model2 4.3453e-07
```
