Exam 2

MacKenzie Ullman

3/22/2021

Import this dataset into R and inspect the first several rows of your data

```
#setwd("C:/Users/mau0005/Downloads/")
#getwd()
ex2 <- read.csv(file='Exam 2 Data.csv')
head(ex2)
```

```
##    y          x1          x2 x3
## 1  2  1.37034210 -0.66615843  b
## 2  2 -0.70417232 -0.03705622  c
## 3  4 -0.04223752 -1.53148692  c
## 4  9 -0.56711072 -0.06529335  b
## 5  1  0.31189773  0.81650472  a
## 6 13 -0.35994479 -0.80042308  b
```

Fit a Poisson model that assumes your response is a function of x1, x2, and x3. Include an interaction between x1 and x2 only (i.e., do not include an interaction between your categorical variables and any other variables).

```
fit <- glm(y ~ x1 * x2 + x3, family = poisson, data = ex2)
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x1 * x2 + x3, family = poisson, data = ex2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3620  -0.6973  -0.1007   0.5236   2.6779
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.13258    0.08540  13.262  < 2e-16 ***
## x1          -1.03491    0.05019 -20.620  < 2e-16 ***
## x2          -0.90839    0.06977 -13.021  < 2e-16 ***
## x3b          0.37532    0.09246   4.059 4.92e-05 ***
## x3c         -0.88354    0.12072  -7.319 2.50e-13 ***
## x1:x2       -0.28868    0.05142  -5.614 1.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 720.551  on 99  degrees of freedom
## Residual deviance:  89.088  on 94  degrees of freedom
## AIC: 392.86
##
## Number of Fisher Scoring iterations: 5
```

Interpret the effect of variable x1 when x2 = -1

```
b <- coef(fit)
b
```

```
## (Intercept)          x1          x2         x3b         x3c       x1:x2
##   1.1325781  -1.0349058  -0.9083868   0.3753215  -0.8835413  -0.2886813
```

```
b[2] + b[6] * -1
```

```
##         x1
## -0.7462245
```

The log proportional change associated with a one unit increase in x1 decreases by 0.746.

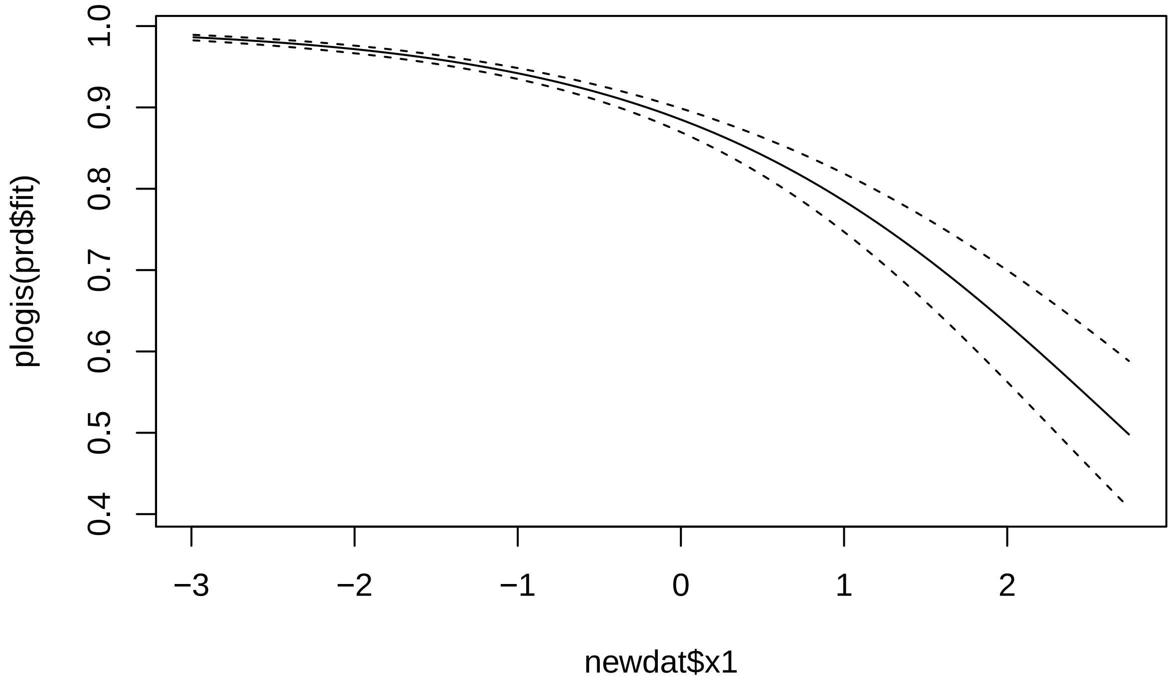
Plot expected counts ±90% confidence intervals over the observed range of variable x1. Assume variable when x2 = -1 and category “a”.

```
newdat <- data.frame(
  x1 = seq(min(ex2$x1), max(ex2$x1), length.out = 100),
  x2 = -1,
  x3 = factor('a', levels = c('a', 'b', 'c'))
)

prd <- predict.glm(fit, newdat, se.fit = T)

low <- plogis(prd$fit - qnorm(0.95) * prd$se.fit)
high <- plogis(prd$fit + qnorm(0.95) * prd$se.fit)

plot(x = newdat$x1, y = plogis(prd$fit), ylim = c(min(low), max(high)), type = 'l')
lines(x = newdat$x1, y = low, lty = 2)
lines(x = newdat$x1, y = high, lty = 2)
```



Interpret the effect of variable x3

```
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x1 * x2 + x3, family = poisson, data = ex2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3620  -0.6973  -0.1007   0.5236   2.6779
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.13258    0.08540  13.262  < 2e-16 ***
## x1          -1.03491    0.05019 -20.620  < 2e-16 ***
## x2          -0.90839    0.06977 -13.021  < 2e-16 ***
## x3b          0.37532    0.09246   4.059 4.92e-05 ***
## x3c         -0.88354    0.12072  -7.319 2.50e-13 ***
## x1:x2       -0.28868    0.05142  -5.614 1.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 720.551  on 99  degrees of freedom
## Residual deviance:  89.088  on 94  degrees of freedom
## AIC: 392.86
##
## Number of Fisher Scoring iterations: 5
```

The difference in log proportions between category b and a is 0.375. The difference in log odds between category c and a is -0.883.

Use contrasts to evaluate the null hypothesis that the difference in log expected count between levels "b" and "c" = 0. Fix x1 and x2 at their means.

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.0.4

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.0.4

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
m <- matrix(c(0, 0, 0, 1, 1, 0), nrow = 1)
cnt <- glht(fit, m)
summary(cnt, test = adjusted('none'))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = y ~ x1 * x2 + x3, family = poisson, data = ex2)
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(>|z|)
## 1 == 0  -0.5082     0.1787  -2.845  0.00445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

we reject the null hypothesis because the p-value (0.004) is less than 0.05. Therefore the difference in log expected count between levels "b" and "c" is different than 0.

Derive the test statistic and p-value associated with the interaction between x1 and x2. What is the null hypothesis? Do we reject or fail to reject this null hypothesis? Defend your answer.

```
#test statistic
s <- summary(fit)[['coefficients']][, 2]
b[6] / s[6]
```

```
##     x1:x2
## -5.614074
```

```
# p-value
pnorm(-1 * abs(b[6] / s[6])) * 2
```

```
##        x1:x2
## 1.976182e-08
```

The null hypothesis is that beta 5 is equal to zero. We reject this null hypothesis. Given the significantly small p-value of 1.976e-08, the effect of variable x1 depends on the level of x2.

assume you have the following realizations of random variable Y : y = (1, 0) Further assume realizations of the random variable Y are Bernoulli distributed: y ~ Bernoulli(p). What is the probability of observing each of these random variables assuming the log odds of success = -2?

```
plogis(-2)
```

```
## [1] 0.1192029
```

```
dbinom(0, size = 1, p = 0.12)
```

```
## [1] 0.88
```

The probability of observing each of these random variable assuming the log odds o success is -2 is 0.88.

What is the "support" of a Bernoulli random variable? What are the acceptable values of it's sole parameter? To which quantity do we apply a link function, and why do we do this? What is the principle link function we use in binomial (i.e., logistic) regression, and what it it's inverse function?

The support is that realizations of the random variable (y) are bounded between 0 and 1. The acceptable values for the probability of success are 0 or 1. We apply a link function to a bounded quantity in order transform it to the real number line. The principal link function is the logit link function.The logit link maps a between 0 and 1 to a real number line. The inverse function of the logit link function is called the inverse link function (plogis()). The inverse logit link maps a number on a real number line to the 0, 1 interval.

What is a fundamental assumption we make to derive inference when comparing two levels of a categorical random variable? Normally distributed with a mean = 0 and the variance = 1.