Final

Nikhil Gopal

4/15/2021

```
rm(list = ls())
#setwd("C:/Users/d/Google Drive/Notability/Categorical Data/psets/final")
```

Question 1

```
msparrownest <- read.csv("msparrownest.csv", header=T)


sparrow <- glm(y~x, data = msparrownest, family = "binomial")

summary(sparrow)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = msparrownest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8431  -1.0382   0.4252   0.9201   1.8275
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.6657     5.4952  -2.669  0.00761 **
## x             1.1644     0.4278   2.722  0.00649 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 58.466  on 42  degrees of freedom
## Residual deviance: 47.999  on 41  degrees of freedom
## AIC: 51.999
##
## Number of Fisher Scoring iterations: 4
```

```
plot(jitter(y, .2) ~ x, data=msparrownest, main = "Nesting Success Probablity vs Wingspan"
     ,ylab="Nesting Success Probability", xlab= "Wingspan",
     pch=19, xlim=c(8, 17))

x <- seq(10, 15, .125)
```

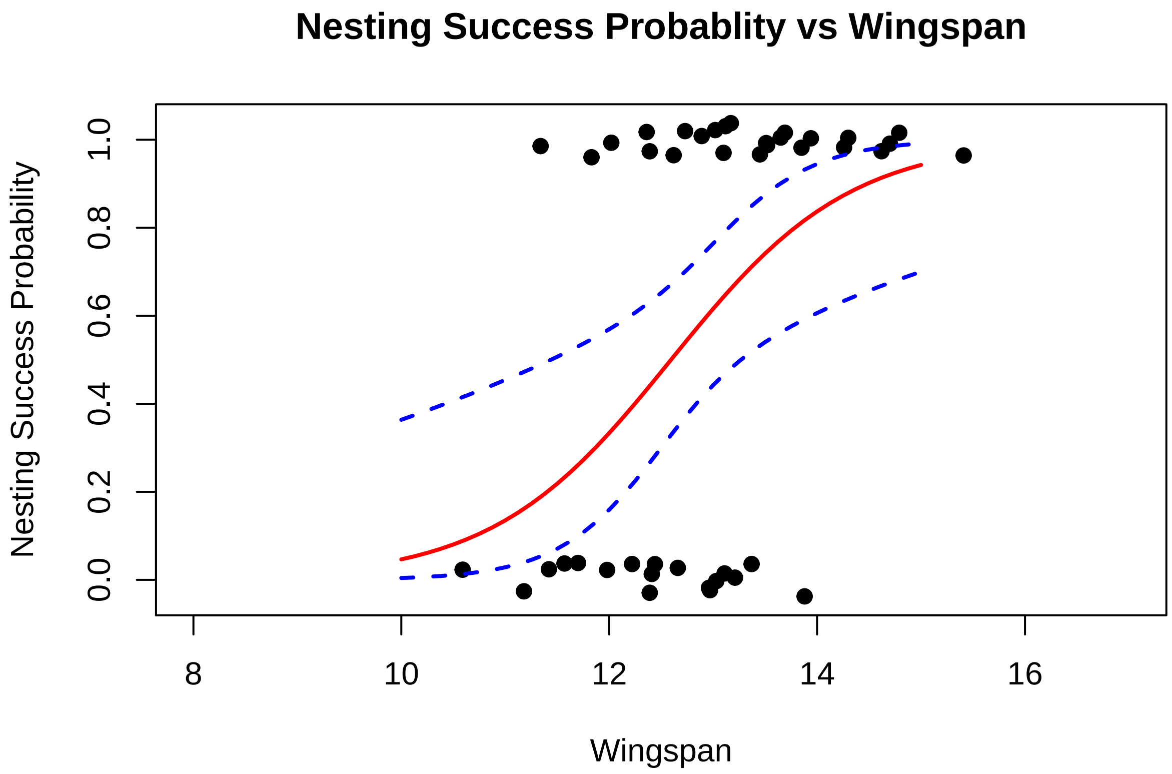
```
preds <- predict(sparrow, data.frame(width=x), se.fit=T)

logistic <- function(u){ exp(u) / (1 + exp(u)) }

lines(x, logistic(preds$fit), col="red", lwd=2)

lines(x, logistic(preds$fit - 1.96 * preds$se.fit), lty=2,col="blue", lwd=2)

lines(x, logistic(preds$fit + 1.96 * preds$se.fit), lty=2,col="blue", lwd=2)
```



The curve implies that the probability of nesting success increases as wingspan increases. However, the confidence interval is quite wide, meaning the probability could lie within a large range of values. Additionally, the data points at the top and the bottom of the graph seem to suggest that there is no real correlation between wingspan and nesting success probability, as there are only a few points with larger wingspan that have a higher nesting success probability. The logistic function also results in the confidence bands not being symmetrical, as it cannot output negative values. Wider confidence bands would mean that the possible values could be further away from the curve. I would say the curve is a little bit deceiving.

Question 2

```
azdiabetes <- read.csv("azdiabetes.csv", header=T)
```

2a:

```
azdiabetes$diabetes <- ifelse(azdiabetes$diabetes == "Yes", 1, 0)


mod2 <- glm(diabetes~age+bmi+bp+npreg, data = azdiabetes, family = "binomial")

summary(mod2)
```

```
##
## Call:
## glm(formula = diabetes ~ age + bmi + bp + npreg, family = "binomial",
##     data = azdiabetes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9197  -0.8257  -0.5201   1.0164   2.2163
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.112557   0.783272  -7.804 6.00e-15 ***
## age          0.049747   0.012324   4.037 5.42e-05 ***
## bmi          0.106890   0.016952   6.305 2.88e-10 ***
## bp          -0.001186   0.009217  -0.129   0.8976
## npreg        0.075969   0.037952   2.002   0.0453 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.79  on 531  degrees of freedom
## Residual deviance: 573.13  on 527  degrees of freedom
## AIC: 583.13
##
## Number of Fisher Scoring iterations: 4
```

$$logit(Diabetes) = -6.112557 + 0.049747(age) + 0.106890(bmi) - 0.001186(bp) + 0.075969(npreg)$$

2b:

```
CI <- exp(confint(mod2))
```

```
## Waiting for profiling to be done...
```

```
CI
```

```
##                     2.5 %      97.5 %
## (Intercept) 0.0004539432 0.009832503
## age         1.0262831800 1.077273604
## bmi         1.0773693898 1.151562832
## bp          0.9808319383 1.017018626
## npreg       1.0020163093 1.163214478
```

```
#Age 40-25 = 15
1.0262831800^15;1.077273604^15
```

```
## [1] 1.475734
```

```
## [1] 3.054149
```

```
# [1.475734, 3.054149]

#BMI 35-30 = 5
1.0773693898^5;1.151562832^5
```

```
## [1] 1.45152
```

```
## [1] 2.025061
```

```
# [1.45152,2.025061]
```

The confidence interval for diabetes was [1.475734, 3.054149] and for BMI was [1.45152,2.025061]. We are 95% confident that the true conditional odds ratios lies within those intervals.

We are 95% confident that the true increase in conditional odds for a 5 unit increase in BMI is between the values [1.45152,2.025061]. We are also 95% confident that the true increase in conditional odds for a 15 unit increase in age is between [1.475734, 3.054149]. This would make sense intuitively, as increases in Age and BMI should make one more likely to have diabetes.

Question 2c

```
anova(mod2, glm(diabetes~age+bmi, data = azdiabetes, family = binomial), test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: diabetes ~ age + bmi + bp + npreg
## Model 2: diabetes ~ age + bmi
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       527     573.13
## 2       529     577.20 -2  -4.0728   0.1305
```

Ho: coefficients on BP/npreg = 0 Ho: coefficients on BP/npreg != 0

The anova above tested the hypothesis that the coefficients of bp and npreg were simultaneously zero. Had the coefficients been zero, these variables would not have an effect on the outcome, and thus there would be no difference between the models. The hypothesis test returned a p value of 0.1305. Since the p value is above 0.05, we fail to reject the null hypothesis, and cannot be sure that the cp and npreg coefficients are simultaneously not equal to zero.

Question 3

```
prayer <- read.csv("prayer.csv", header=T)

prayer$prayer <- as.factor(prayer$prayer)
prayer$female <- as.factor(prayer$female)
```

```r
library(MASS)

simple <- polr(prayer ~ female + vocab, data = prayer, method = "logistic")

comparison <- polr(prayer ~ 1, data = prayer, method = "logistic")

anova(simple, comparison)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: prayer
##            Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1              1       941   3002.671
## 2 female + vocab       939   2947.861 1 vs 2     2 54.81028 1.253442e-12
```

Add interaction term and test GOF:

```r
interaction <- polr(prayer ~ female + vocab + vocab:female, data = prayer, method = "logistic")

anova(simple, interaction)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: prayer
##                           Model Resid. df Resid. Dev   Test    Df LR stat.
## 1                female + vocab       939   2947.861
## 2 female + vocab + vocab:female       938   2942.327 1 vs 2     1  5.53465
##      Pr(Chi)
## 1
## 2 0.01864349
```

```r
summary(interaction)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = prayer ~ female + vocab + vocab:female, data = prayer,
##     method = "logistic")
##
## Coefficients:
##                   Value Std. Error t value
## female1         1.59665    0.36552   4.368
## vocab          -0.02726    0.04293  -0.635
## female1:vocab -0.13533    0.05758  -2.350
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -3.4595   0.3335   -10.3727
## 2|3  -1.1269   0.2726    -4.1339
## 3|4  -0.6769   0.2705    -2.5024
## 4|5  -0.0504   0.2704    -0.1862
## 5|6   1.3968   0.2753     5.0738
##
## Residual Deviance: 2942.327
## AIC: 2958.327
```

Our first anova for the simple model returned a p value 1.253442eˆ-12, meaning that we can say with confidence that our model with gender/vocab as parameters fits the data better than the simple model. Our second anova between the simple model and the interaction term returned a p value of 0.01864349, meaning we can also say with confidence that our interaction model fits the data better than the simple model

Our model has the following prediction equations (for all of these subsitute the proper intercept for the requisite value of Prayer:

female:

$$logit(P(\hat{Prayer} \leq J_i)) = a_i + 1.59665 - (-0.16259(vocab))$$

male:

$$logit(P(\hat{Prayer} \leq J_i)) = a_i - (-0.02726(vocab))$$

An ordinal logistic regression model was chosen for this data since the outcome variable (Prayer) was ordinal. Simple logistic regression would not work since the outcome variable was not binary (6 categories), and a loglinear model would not work as the outcome variable does not have a poisson distribution (it is ordinal). This model will allow us to predict an individual's frequency of prayer using their gender and number of items correct on a vocab test.

Our model had a coefficients of 1.59665 for female, -0.02726 for vocab score and -0.13533 for the interaction term (female:vocab). The odds of being in a higher category for females decreases by a factor 0.8499396 for each question answered correctly on a vocab test. The odds of being in a higher category for males decreases by a factor 0.9731082 for each question answered correctly on a vocab test. Given this difference in odds ratios, we are certain that the vocab scores effect on prayer is differs across genders.