HW3

Nikhil Gopal

2/23/2021

Question 1

A:

```r
setwd("/Users/nikhilgopal/Google Drive/Notability/Categorical Data/psets/HW3")
horseshoe <- read.csv("horseshoe.csv")

crab_model <- lm(y ~ weight, data = horseshoe)
summary(crab_model)
```

```
##
## Call:
## lm(formula = y ~ weight, data = horseshoe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8878 -0.4683  0.1606  0.3704  0.6689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.14487    0.14715  -0.984    0.326
## weight       0.32270    0.05876   5.492 1.42e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4447 on 171 degrees of freedom
## Multiple R-squared:  0.1499, Adjusted R-squared:  0.1449
## F-statistic: 30.16 on 1 and 171 DF,  p-value: 1.421e-07
```

The equation is $\pi(x) = -0.14487 + 0.32270 * weight, \beta = 0.32270$. This means that the slope is equal to 0.32270, suggesting that an increase of 1 in weight unit leads to the crab having 0.32270 more satellites.

B:

For x = 4: $\hat{P}(Y = 1) = 1.1459$ and for x = 5: $\hat{P}(Y = 1) = 1.46863$. This would mean that the probability is greater than 1, which is obviously unrealistic. Since linear predictors can take values along the entire number line, the linear probability model can incorrectly give probabilities greater than 1 or less than zero, an obvious limitation of the model.

C:

```r
logit <- glm(y ~ weight, data = horseshoe, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = y ~ weight, family = "binomial", data = horseshoe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1108  -1.0749   0.5426   0.9122   1.6285
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
## weight        1.8151     0.3767   4.819 1.45e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 195.74  on 171  degrees of freedom
## AIC: 199.74
##
## Number of Fisher Scoring iterations: 4
```

```r
#since its log use exponential func to find true values
exp(coef(logit))
```

```
## (Intercept)      weight
##  0.02485425  6.14196417
```

The prediction equation is $ = -3.6947 + 1.8151 * weight$. In this model, Beta = 1.8151, meaning that the probability of having satellites increases as the weight of the crab increases as the beta coefficient is positive.

D:

$-3.6947 + 1.8151 * 4 = 3.5657, -3.6947 + 1.8151 * 5 = 5.3808$

```r
x4 = exp(3.5657) / (1 + exp(3.5657))
x5 = exp(5.3808) / (1 + exp(5.3808))
x4
```

```
## [1] 0.9725004
```

```r
x5
```

```
## [1] 0.995417
```

The log odds ratio for x = 4 was 3.5657. This means the probability of a crab having at least one satellite is 0.9725004. The ratio for x = 5 was 5.3808, which means a probability of 0.995417. This means that it is extremely likely that crabs that weigh 4/5 units will have satellites.

Question 2

A:

```
credit <- read.csv("credit.csv")
credit_model <- glm(cards ~ income, data = credit, family = "binomial")
```

Prediction equation:

$$log[\frac{\pi(x)}{1-\pi(x)}] = -3.51795 + 0.10541 * income$$

The beta coefficient is positive, which means that the probability of having a travel card increases as one's income increases.

B:

$$log[\frac{0.5}{1-0.5}] = log[\frac{0.5}{0.5}] = -3.51795 + 0.10541 * income$$
$$0 = -3.51795 + 0.10541 * income$$
$$income = 33.37397$$

When income $= 33.37397$, the estimated probablity of having a travel card is 0.5

Question 3

A:

```
poisson.horseshoe <- glm(sat ~ weight, family="poisson", data=horseshoe)
summary(poisson.horseshoe)
```

```
##
## Call:
## glm(formula = sat ~ weight, family = "poisson", data = horseshoe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9307  -1.9981  -0.5627   0.9298   4.9992
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.42841    0.17893  -2.394   0.0167 *
## weight       0.58930    0.06502   9.064   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 560.87  on 171  degrees of freedom
## AIC: 920.16
##
## Number of Fisher Scoring iterations: 5
```

Prediction Equation:

$$log(\hat{\mu}) = -0.42841 + 0.58930 * weight$$

Mean response for avg weight of female crab:

```
log_mu <- -0.42841 + 0.58930*2.44
exp(log_mu)
```

```
## [1] 2.744179
```

The estimated mean response is 2.744179.

B:

$$H_0 = \beta = 0$$
$$H_a = \beta = !0$$

```
library(aod)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(DescTools)

wald.test(b=coef(poisson.horseshoe), Sigma=vcov(poisson.horseshoe), Terms = 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 82.2, df = 1, P(> X2) = 0.0
```

```
lrtest(poisson.horseshoe)
```

```
## Likelihood ratio test
##
## Model 1: sat ~ weight
## Model 2: sat ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -458.08
## 2   1 -494.04 -1 71.925  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The chi-square statistic for the Wald test is 82.2 on 1 df and the LR statistic is 71.925 on 1 df. Both tests have p values below the critical value of 0.05, providing strong evidence that the slope is not equal to zero and thus providing strong evidence that there is a relationship between mean response and weight.

C:

```r
#default to get wald conf int
confint.default(poisson.horseshoe, level = 0.95)
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.7791047 -0.077706
## weight       0.4618742  0.716734
```

```r
#Interval: [0.4618742, 0.716734]

#exponential function to get multiplicative effect

exp(0.4618742)
```

```
## [1] 1.587046
```

```r
exp(0.716734)
```

```
## [1] 2.047734
```

The confidence interval for beta was [0.4618742, 0.716734], meaning we are 95% confident that the true beta lies within that interval. The multiplicative interval was [1.587046, 2.047734], meaning we are 95% confident that the true multiplicative effect increase lies within that interval. Since beta hat is positive and our confidence intervals all contain positive betas, crabs that weigh more likely have more satellites.

Question 4

$$log(\mu_A) = \alpha + \beta * 0 = \alpha$$

$$log(\mu_B) = \alpha + \beta * 1 = \alpha + \beta$$

using properties of log:

$$\beta = log(\mu_B) - log(\mu_A) = log(\mu_B/\mu_A)$$

$$e^{\beta} = \mu_B/\mu_A$$

$$log(\mu_A) = \alpha + \beta * 0 = \alpha$$

$$\mu_A = e^{\alpha}$$

B:

```r
imperf <- c(8, 7, 6, 6, 3, 4, 7, 2, 3, 4, 9, 9, 8, 14,
8, 13, 11, 5, 7, 6)

#create a data frame and assign binary values of 1 for treatment A and 0 for treatment B
treatment <- c(rep(0,10), rep(1,10))

df <- data.frame(imperf, treatment)

model <- glm(imperf~treatment, data = df, family = "poisson")

summary(model)
```

```
##
## Call:
## glm(formula = imperf ~ treatment, family = "poisson", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5280  -0.7622  -0.1699   0.6938   1.5399
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
## treatment     0.5878     0.1764   3.332 0.000861 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 27.857  on 19  degrees of freedom
## Residual deviance: 16.268  on 18  degrees of freedom
## AIC: 94.349
##
## Number of Fisher Scoring iterations: 4
```

Our prediction equation was $log(\hat{\mu}) = 1.6094 + 0.5878x$, giving an alpha of 1.6094 and a beta of 0.5878. This means that our model predicts e^(0.5878) more imperfections using method B than method A. Since method A was assigned a value of 0 in our model, our model predicts that we will have our alpha value of e^(1.6094) imperfections with method A. Since our beta coefficient is positive and method B was coded to 1, method B is predicted to cause more imperfections than method A.

c:

```r
wald.test(b=coef(model), Sigma=vcov(model), Terms = 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 11.1, df = 1, P(> X2) = 0.00086
```

Our wald test returned a chi square statistic of 11.1 on 1 degree of freedom. The p value was 0.00086, which is much less than 0.05. This means that we have strong evidence that the mean number of imperfections is different between groups.

D:

```
confint.default(model, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 1.332258 1.8866181
## treatment   0.242082 0.9334913
```

```
true_vals <- c(exp(0.242082),exp(0.9334913))
true_vals
```

```
## [1] 1.273899 2.543373
```

Our 95% WALD confidence interval is [1.273899, 2.543373]. We are 95% confident that the mean for treatment B is an amount within the interval higher than the mean for treatment A. This is in line with our hypothesis test above, which concluded that there was strong evidence for a difference in means between groups, and our model, which estimated that treatment B would produce more imperfections.