HW4

Nikhil Gopal

3/14/2021

Question 1

```
#setwd("/Users/nikhilgopal/Google Drive/Notability/Categorical Data/psets/HW4")

LI <- c( 8, 8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,22,22,24,26,28,32,34,38,38,38)
y  <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)



logit_model <- glm(y ~ LI, family = "binomial")


summary(logit_model)
```

```
##
## Call:
## glm(formula = y ~ LI, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9448  -0.6465  -0.4947   0.6571   1.6971
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.77714    1.37862  -2.740  0.00615 **
## LI           0.14486    0.05934   2.441  0.01464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.372  on 26  degrees of freedom
## Residual deviance: 26.073  on 25  degrees of freedom
## AIC: 30.073
##
## Number of Fisher Scoring iterations: 4
```

1a:

```r
#1a

new.LI <- data.frame(
  LI = c(12)
)

predict.glm(logit_model, newdata = new.LI, type = "response")
```

```
##         1
## 0.1151908
```

probablity = 0.1151908

1b:

```r
library(chemCal)
```

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

```r
inverse.predict(logit_model, as.numeric(data.frame(LI = 0)),
                ws, alpha=0.05, var.s = "auto")
```

```
## $Prediction
## [1] 26.07384
##
## $`Standard Error`
## [1] 3.289526
##
## $Confidence
## [1] 6.774906
##
## $`Confidence Limits`
## [1] 19.29894 32.84875
```

The percentage of labeled cells at which the probability of remission will be 0.5 is 26.07384.

1c:

```r
#calculate odds ratio
exp(coef(logit_model))
```

```
## (Intercept)          LI
##  0.02288805  1.15588143
```

A 1 unit increase in LI increases the odds of remission by a factor of 1.15588143

1d:

```r
new.LI <- data.frame(
  LI = c(12, 25)
)

predict.glm(logit_model, newdata = new.LI, type = "response")
```

```
##         1         2
## 0.1151908 0.4611882
```

The probability of remission changes from 0.1151908 to 0.4611882 (changes by 0.3459974) with a change from the lower to upper quartiles values of the labeling index.

Question 2

2a

```
0.14486 * 0.1151869*(1- 0.1151869)
```

```
## [1] 0.01476397
```

At x=12, the estimated ROC is 0.01476397

2b:

```
#wald test
summary(logit_model)
```

```
##
## Call:
## glm(formula = y ~ LI, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9448  -0.6465  -0.4947   0.6571   1.6971
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.77714    1.37862  -2.740  0.00615 **
## LI           0.14486    0.05934   2.441  0.01464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.372  on 26  degrees of freedom
## Residual deviance: 26.073  on 25  degrees of freedom
## AIC: 30.073
##
## Number of Fisher Scoring iterations: 4
```

```
library(epitools)
```

```
## Warning: package 'epitools' was built under R version 4.0.3
```

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.0.3
```

```r
#confidence interval
OddsRatio(logit_model)
```

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

##
## Call:
## glm(formula = y ~ LI, family = "binomial")
##
## Odds Ratios:
##                or or.lci or.uci Pr(>|z|)
## (Intercept) 0.023  0.001  0.244   0.0061 **
## LI          1.156  1.043  1.329   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Brier Score: 0.162     Nagelkerke R2: 0.368
```

The z statistic for the wald test on the LI effect was 2.44 and p value was 0.01464. This means that the coefficient value is likely non-zero, since our p value is low enough to reject the null hypothesis that the beta coefficient is zero. Thus, there is likely an effect of LI on remissions.

The 95% confidence interval was [1.043,1.329], meaning that we are 95% confident that the true odds ratio is within that interval.

2c:

```r
library(lmtest)
```

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

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##
## Attaching package: 'zoo'

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

```r
lrtest(logit_model)
```

```
## Likelihood ratio test
##
## Model 1: y ~ LI
## Model 2: y ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -13.037
## 2   1 -17.186 -1 8.2988   0.003967 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
exp(confint(logit_model, level = 0.95))
```

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

```
##                    2.5 %    97.5 %
## (Intercept) 0.0009162778 0.2441813
## LI          1.0434402672 1.3293190
```

The likelihood ratio test returned a chisquare statistic of 8.2988 and a p-value of 0.003967. This means that we can reject our null hypothesis, and that our model using 1 as a predictor of y fits the data better than our logistic regression model using the LI effect as a predictor.

The confidence interval for the odds ratio was [1.0434402672,1.3293190]. This means that we are 95% confident that the true odds ratio is within the interval.

2d:

```r
dataa <- read.csv("dataa.csv")

grouped_model <-glm(formula = as.factor(remissions) ~ dataa$ï..LI + cases, family = "binomial", data = d

summary(grouped_model)
```

```
##
## Call:
## glm(formula = as.factor(remissions) ~ dataa$ï..LI + cases, family = "binomial",
##     data = dataa)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95682  -0.70732  -0.05869   0.85226   1.44172
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.68387    3.48168  -1.058    0.290
## dataa$ï..LI  0.16842    0.10799   1.560    0.119
## cases        0.04954    0.85775   0.058    0.954
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19.408  on 13  degrees of freedom
## Residual deviance: 14.411  on 11  degrees of freedom
## AIC: 20.411
##
## Number of Fisher Scoring iterations: 4
```

```r
lrtest(grouped_model)
```

```
## Likelihood ratio test
##
## Model 1: as.factor(remissions) ~ dataa$ï..LI + cases
## Model 2: as.factor(remissions) ~ 1
```

```
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -7.2056
## 2   1 -9.7041 -2 4.9969    0.08221 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The parameter estimates do change slightly (LI went from 0.14 to 0.16) as did the standard errors (0.0593 to 0.10699) and the null deviance (34.372 on 26 df to 19.408 on 13 df).

The likelihood ratio test also does change. Our p value became 0.08, meaning that it is no longer statistically significant. Thus, we cannot be sure that the given model fits the data better than our new model, a positive result for the grouped data.

Question 3

```r
crabs <- read.csv("crabs.csv")

crabs_model <- glm(y~weight, data = crabs, family = "binomial")

summary(crabs_model)
```

```
##
## Call:
## glm(formula = y ~ weight, family = "binomial", data = crabs)
##
## 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
```

3a:

ML Prediction equation:

Logit(Y=1) = 1.8151(weight)-3.6947

3b

```r
#wald test
summary(crabs_model)
```

```
##
## Call:
## glm(formula = y ~ weight, family = "binomial", data = crabs)
##
## 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
```

```
#lr test
lrtest(crabs_model)
```

```
## Likelihood ratio test
##
## Model 1: y ~ weight
## Model 2: y ~ 1
##   #Df   LogLik Df  Chisq Pr(>Chisq)
## 1   2  -97.869
## 2   1 -112.879 -1 30.021  4.273e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The wald test returned a Z statistic of 4.819, and a p value of 1.45eˆ-6. The lr test returned a chi square stat of 30.021 and a p value of 4.273e-08 Thus, we can reject both null hypotheses and be sure with statistical significance that the coefficient on weight is not zero, and, that weight is a significant predictor of having a satellite (compared to the standard model).

3c:

```
new_weight <- data.frame(
  weight = c(1.2, 2.44, 5.20)
)

predict(crabs_model, new_weight, type = "response")
```

```
##         1         2         3
## 0.1799697 0.6757320 0.9968084
```

$$\hat{\pi(1.20)} = 0.1799697$$

$$\hat{\pi(2.44)} = 0.6757320$$
$$\hat{\pi(5.2)} = 0.9968084$$

3d

```
library(chemCal)
inverse.predict(crabs_model, 0.5)
```

```
## $Prediction
## [1] 2.31096
##
## $`Standard Error`
## [1] 0.2577086
##
## $Confidence
## [1] 0.5086998
##
## $`Confidence Limits`
## [1] 1.80226 2.81966
```

Weight = 2.31096

3e

```
1.8151 * 0.5 * 0.5
```

```
## [1] 0.453775
```

```
1.8151 * 0.5 * 0.5 * 0.1
```

```
## [1] 0.0453775
```

```
1.8151 * 0.5 * 0.5 * 0.58
```

```
## [1] 0.2631895
```

The estimated effect of a 1kg increase is 0.453775, a 0.1kg increase is 0.0453775 and a 0.58 kg increase is 0.2631895.

3f

```
exp(confint(crabs_model, level = 0.95))
```

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

```
##                   2.5 %     97.5 %
## (Intercept) 0.004062602  0.1300676
## weight      3.045879867 13.4274964
```

Our confidence interval for the weight coefficient was [3.045879867 , 13.4274964]. We are 95% confident that the true value of weight is within the interval. The coefficient is very likely positive, implying a positive association between weight and satellites.

Question 4

```
new_crab_model <- glm(y~color, data = crabs, family = "binomial")
```

4a:

```
new_crab_model_factor <- glm(y~as.factor(color), data = crabs, family = "binomial")

summary(new_crab_model_factor)
```

```
##
## Call:
## glm(formula = y ~ as.factor(color), family = "binomial", data = crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6651  -1.3370   0.7997   0.7997   1.5134
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.0986     0.6667   1.648   0.0994 .
## as.factor(color)2  -0.1226     0.7053  -0.174   0.8620
## as.factor(color)3  -0.7309     0.7338  -0.996   0.3192
## as.factor(color)4  -1.8608     0.8087  -2.301   0.0214 *
## ---
## 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: 212.06  on 169  degrees of freedom
## AIC: 220.06
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation was logit(Y) = -0.1226(color2)-0.7309(color3)-1.8608(color4)+1.0986

To compare the effects of the first vs fourth color on weight, simply plug into the prediction equation the desired color and get the output value to see how it will affect the number of satellite. If the color is the first color the intercept term will be returned. If it is one color then the value of 0 will be inputted into the other variables in the model:

```
new_color <- data.frame(
  color = c(1,4)
)

predict(new_crab_model_factor, new_color)
```

```
##          1          2
##  1.0986123 -0.7621401
```

Our model predicts that crabs with the 1st color

4b

```
library(lmtest)
lrtest(new_crab_model_factor)
```

```
## Likelihood ratio test
##
## Model 1: y ~ as.factor(color)
## Model 2: y ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -106.03
## 2   1 -112.88 -3 13.698   0.003347 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test returned a Chi square statistic of 13.698 and a p value of 0.003347. This means that we can conclude with statistical significance that our model fits the data better than the simpler default one, and that one of the colors has a significant effect on having satellites.

4c:

```
new_crab_model_numeric <- glm(y~as.numeric(color), data = crabs, family = "binomial")

summary(new_crab_model_numeric)
```

```
##
## Call:
## glm(formula = y ~ as.numeric(color), family = "binomial", data = crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9103  -1.2719   0.8142   0.8142   1.3937
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.3635     0.5551   4.257 2.07e-05 ***
## as.numeric(color)  -0.7147     0.2095  -3.412 0.000645 ***
## ---
## 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: 213.30  on 171  degrees of freedom
## AIC: 217.3
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation was y = -0.7147(color) + 2.3635. Since the coefficient is negative, this model predicts that the likelihood of the crab having satellites decreases by -0.7147 units as the color increases.

4d:

```
summary(new_crab_model_numeric)
```

```
##
## Call:
## glm(formula = y ~ as.numeric(color), family = "binomial", data = crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9103  -1.2719   0.8142   0.8142   1.3937
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.3635     0.5551   4.257 2.07e-05 ***
## as.numeric(color)  -0.7147     0.2095  -3.412 0.000645 ***
## ---
## 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: 213.30  on 171  degrees of freedom
## AIC: 217.3
##
## Number of Fisher Scoring iterations: 4
```

```
lrtest(new_crab_model_numeric)
```

```
## Likelihood ratio test
##
## Model 1: y ~ as.numeric(color)
## Model 2: y ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -106.65
## 2   1 -112.88 -1 12.461  0.0004156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The wald hypothesis test for the color coefficient returned a pvalue of 0.000645. This means that we can conclude with statistical significance that the coefficient is not zero and that color does have an effect on the crabs having satellites.

4e:

When color is quantitative there is only 1 instead of 4 factors in the model, which decreases bias and reduces standard error, increasing power.

However, color does not exist on a scale and is really qualitative, so a linear model is probably not appropriate and thus has worse fit.

Question 5

```
alcohol_data <- read.csv("alcohol.csv")
```

5a

```
alcohol_data$personality <- paste(alcohol_data$ï..EI, alcohol_data$SN, alcohol_data$TF, alcohol_data$JP

alcohol_data$prop_that_drinks <- alcohol_data$drink/alcohol_data$n

a <- sort(alcohol_data$prop_that_drinks)


#prop = 0.19047619, personality = ESTP
```

ESTP has the highest percentage that report drinking alcohol.

5b:

```
alc_model <- glm(prop_that_drinks ~ as.factor(alcohol_data$ï..EI)+
                   as.factor(SN)+ as.factor(TF) + as.factor(JP)
                 , data = alcohol_data)

summary(alc_model)
```

```
##
## Call:
## glm(formula = prop_that_drinks ~ as.factor(alcohol_data$ï..EI) +
##     as.factor(SN) + as.factor(TF) + as.factor(JP), data = alcohol_data)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.057539  -0.019371  -0.007566   0.024059   0.058732
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.10610    0.02188   4.849 0.000512 ***
## as.factor(alcohol_data$ï..EI)i -0.05323    0.01957  -2.720 0.019929 *
## as.factor(SN)s                 -0.02120    0.01957  -1.083 0.301862
## as.factor(TF)t                  0.03988    0.01957   2.038 0.066351 .
## as.factor(JP)p                  0.02267    0.01957   1.158 0.271321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001532124)
##
##     Null deviance: 0.038405  on 15  degrees of freedom
## Residual deviance: 0.016853  on 11  degrees of freedom
## AIC: -52.287
##
## Number of Fisher Scoring iterations: 2
```

For this model, I used the personality types as predictor variables, and treated them as factors since they are obviously categorical. The prediction equation was:

logit(% who drink) = -0.05323(I) + -0.02120(S) + 0.03988(T) + 0.02267(P) + 0.10610

If trying to predict for a personality that doesn't use one of the above ^ letters, just substitute zero in for that value into the prediction equation.

This model appears to not fit the data very well. The p values for the coefficients were all above 0.05 except for the SN data, which did appear to be statistically significant. Thus we can only be certain with statistical significance that the coefficient of this variable is not zero, and thus that is has an effect on the percentage of people who drink.

5c:

```
lrtest(alc_model)
```

```
## Likelihood ratio test
##
## Model 1: prop_that_drinks ~ as.factor(alcohol_data$ï..EI) + as.factor(SN) +
##     as.factor(TF) + as.factor(JP)
## Model 2: prop_that_drinks ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 32.143
## 2   2 25.554 -4 13.178    0.01044 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test returned a statistically significant p value of 0.01044. This means that we can definitely say that our model fits the data better than the saturated model.

5d:

Our model predicts that ENTP has the highest probability of drinking frequently. This is not the same as part A, which was ESTP. This is because our model considers the 4 different personality variables when fitting the model.

Question 6

6a:

```
soret_data <- read.csv("soret.csv")

main_effects <- glm(as.factor(soret_data$Y) ~ soret_data$D+as.factor(soret_data$T), family = "binomial")

summary(main_effects)
```

```
##
## Call:
## glm(formula = as.factor(soret_data$Y) ~ soret_data$D + as.factor(soret_data$T),
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3802  -0.5358   0.3047   0.7308   1.7821
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.41734    1.09457  -1.295  0.19536
## soret_data$D                0.06868    0.02641   2.600  0.00931 **
## as.factor(soret_data$T)1 -1.65895    0.92285  -1.798  0.07224 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46.180  on 34  degrees of freedom
## Residual deviance: 30.138  on 32  degrees of freedom
## AIC: 36.138
##
## Number of Fisher Scoring iterations: 5
```

The parameter estimate for the log odds duration was 0.06868, suggesting that longer surgeries are more likely to result in a sore throat. The parameter estimate for the type of device was -1.65895. This means that type of device is predicted to have a much larger influence on the probability of having a sore throat, decreasing it if a treacheal tube is used or increasing it if a mask airway is used. However, this parameter was not statistically significant, so we cannot be sure that its true value is not zero.

B:

The p value for the D effect was 0.00931, meaning that we can be sure that it is significant and that the true value of the parameter is not zero.

6c:

```
interaction_model <- glm(as.factor(soret_data$Y) ~ soret_data$D+as.factor(soret_data$T)+soret_data$D*as

summary(interaction_model)
```

```
##
## Call:
## glm(formula = as.factor(soret_data$Y) ~ soret_data$D + as.factor(soret_data$T) +
##     soret_data$D * as.factor(soret_data$T), family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9707  -0.3779   0.3448   0.7292   1.9961
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           0.04979    1.46940   0.034   0.9730
## soret_data$D                          0.02848    0.03429   0.831   0.4062
## as.factor(soret_data$T)1             -4.47224    2.46707  -1.813   0.0699 .
## soret_data$D:as.factor(soret_data$T)1  0.07460    0.05777   1.291   0.1966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46.180  on 34  degrees of freedom
## Residual deviance: 28.321  on 31  degrees of freedom
## AIC: 36.321
##
## Number of Fisher Scoring iterations: 6
```

Prediction equation:

logit(Y) = 0.02848(D)-4.47224(T) + 0.07460(D*T)

When T = 0, just input 0 in for T and those terms and the interaction term will have no weight on the model. When T = 1 input 1 into the equation. Interestingly, it appears that T = 1 has a large negative effect on Y with a coefficient of -4.47224, meaning that is has a large effect on the probability of having a sore throat, and decreases it. The interaction term has a positive coefficient, but since its value is small the interaction term will not have a large effect.

6d:

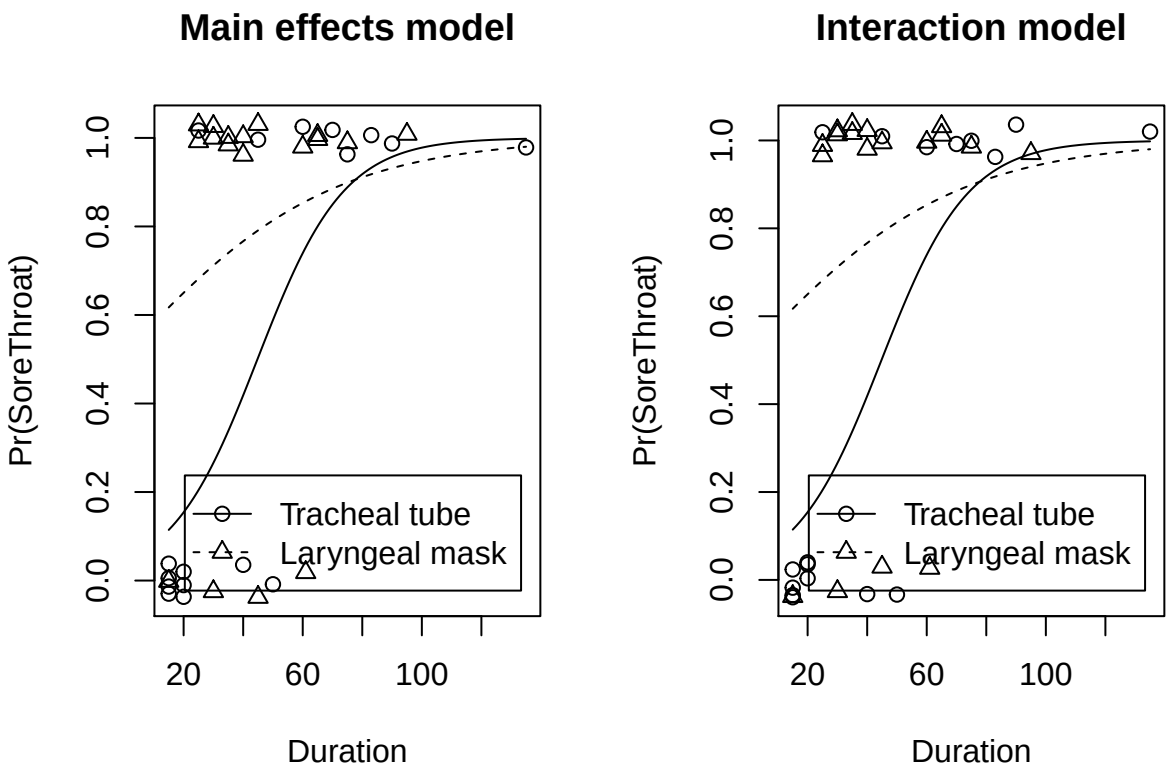
```
library(lmtest)
lrtest(main_effects, interaction_model)
```

```
## Likelihood ratio test
##
## Model 1: as.factor(soret_data$Y) ~ soret_data$D + as.factor(soret_data$T)
## Model 2: as.factor(soret_data$Y) ~ soret_data$D + as.factor(soret_data$T) +
##     soret_data$D * as.factor(soret_data$T)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -15.069
## 2   4 -14.161  1 1.8169     0.1777
```

The likelihood ratio test returned a p value of 0.1777, meaning that we cannot be certain that our interaction model fits the data better than the previous model. It is unlikely that the interaction term is needed.

6e:

```
# Assume data saved as "SoreThroat" dataframe
# Main effects model fit saved as m1, interaction model saved as m2

sorethroat <- read.table(file="http://users.stat.ufl.edu/~aa/cat/data/SoreThroat.dat", header=T)

m1 <- glm(Y~D+T, data=sorethroat, family="binomial")
m2 <- glm(Y~D+T+D*T, data=sorethroat, family="binomial")

x <- range(sorethroat$D)
x <- seq(x[1], x[2])

par(mfrow=c(1,2)); set.seed(21406);

plot(jitter(Y,.2) ~ D, pch=2-T, data=sorethroat, ylab="Pr(SoreThroat)",xlab="Duration", main="Main effe

curve(predict(m1, data.frame(D=x,T=1), type="response"), lty=1, add=T)
curve(predict(m2, data.frame(D=x,T=0), type="response"), lty=2, add=T)
legend("bottomright", inset=.05, pch=1:2, lty=1:2,legend=c("Tracheal tube", "Laryngeal mask"))

plot(jitter(Y,.2) ~ D, pch=2-T, data=sorethroat, ylab="Pr(SoreThroat)",xlab="Duration", main="Interacti
curve(predict(m1, data.frame(D=x,T=1), type="response"), lty=1, add=T)
curve(predict(m2, data.frame(D=x,T=0), type="response"), lty=2, add=T)
legend("bottomright", inset=.05, pch=1:2, lty=1:2,legend=c("Tracheal tube", "Laryngeal mask"))
```



These two graphs appear to be relatively similar. There does not appear to be any significant differences in duration of sore throat predicted in the interaction vs the main effects model.