STAT 426 Assignment 6

Due Wednesday, October 13, 11:59 pm.

Submit through Moodle.

Name: Brianna Diaz

Netid: bdiaz22

Submit your work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown. Be sure to show your work.

Problem 1. (8 pts)

The following data are from a retrospective study of 2659 patients. The data show estimated daily numbers of cigarettes smoked and counts of Cancer patients and Control patients for each level. The variable `clevel` is an ordinal score for the level of smoking.

```
dfs
```

```
##   dailycig clevel Cancer Control
## 1        0      0      7       6
## 2      < 5      1     55     129
## 3     5-14      2    489     570
## 4    15-24      3    475     431
## 5    25-49      4    293     154
## 6      50+      5     38      12
```

a) (1 pt) Here are the estimated coefficients from fitting the logistic regression of Cancer/Control on `dailycig`.

```
coefficients(glm(cbind(Cancer, Control) ~ dailycig,
                 family=binomial, data=dfs))
```

```
##   (Intercept)     dailycig0 dailycig15-24 dailycig25-49  dailycig5-14
##    -0.8524792     1.0066299     0.9496859     1.4956992     0.6992053
##   dailycig50+
##     2.0051587
```

What is the reference category for smoking here, in other words, which smoking level does the intercept correspond to?

Answer: The reference category is dailycig<5

b) (1 pt) An equivalent model is to treat `clevel` as a factor (categorical) variable. Here are the coefficients for that model:

```
coefficients(glm(cbind(Cancer, Control) ~ factor(clevel),
                 family=binomial, data=dfs))
```

```
##     (Intercept) factor(clevel)1 factor(clevel)2 factor(clevel)3 factor(clevel)4
##      0.15415068     -1.00662990     -0.30742455     -0.05694397      0.48906933
## factor(clevel)5
##      0.99852883
```

What is the reference level of smoking in this case, i.e., which level does the intercept correspond to?

Answer: Reference level is factor(clevel)0.

c) (2 pts) The models in a) and b) are equivalent, and both correspond to the saturated model. Explain why, or demonstrate with computations that these models are saturated.

```
daily <- (glm(cbind(Cancer, Control) ~ dailycig,
                 family=binomial, data=dfs))
summary(daily)
```

```
##
## Call:
## glm(formula = cbind(Cancer, Control) ~ dailycig, family = binomial,
##     data = dfs)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.8525     0.1610  -5.294 1.20e-07 ***
## dailycig0       1.0066     0.5792   1.738   0.0822 .
```

```
## dailycig15-24   0.9497     0.1742   5.450 5.02e-08 ***
## dailycig25-49   1.4957     0.1893   7.901 2.78e-15 ***
## dailycig5-14    0.6992     0.1724   4.055 5.01e-05 ***
## dailycig50+     2.0052     0.3682   5.446 5.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  9.6054e+01  on 5  degrees of freedom
## Residual deviance: -1.5166e-13  on 0  degrees of freedom
## AIC: 45.73
##
## Number of Fisher Scoring iterations: 3
```

```
factor <- (glm(cbind(Cancer, Control) ~ factor(clevel),
                family=binomial, data=dfs))
summary(factor)
```

```
##
## Call:
## glm(formula = cbind(Cancer, Control) ~ factor(clevel), family = binomial,
##     data = dfs)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.15415    0.55635   0.277   0.7817
## factor(clevel)1 -1.00663    0.57919  -1.738   0.0822 .
## factor(clevel)2 -0.30742    0.55975  -0.549   0.5829
## factor(clevel)3 -0.05694    0.56031  -0.102   0.9191
## factor(clevel)4  0.48907    0.56518   0.865   0.3869
## factor(clevel)5  0.99853    0.64744   1.542   0.1230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.6054e+01  on 5  degrees of freedom
## Residual deviance: 3.0642e-14  on 0  degrees of freedom
## AIC: 45.73
##
## Number of Fisher Scoring iterations: 3
```

Answer: Based off the results of the deviance residuals, is shows that it is a one to one model and therefore saturated.

d) (2 pts) The models in a) and b) both have the form,

$$\text{logit}(\pi_i) = \alpha + \beta_i, \quad i = 2, \ldots, 6.$$

Perform a likelihood ratio test of

$$H_0 : \beta_2 = \cdots = \beta_6 = 0$$
$$H_a : \text{at least one } \beta_i \neq 0$$

What do you conclude from the test result?

```
drop1(daily, test="Chisq")
```

```
## Single term deletions
##
## Model:
## cbind(Cancer, Control) ~ dailycig
##          Df Deviance    AIC    LRT  Pr(>Chi)
## <none>         0.000  45.73
## dailycig  5   96.054 131.78 96.054 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Based off the results I conclude that we reject the null for the p-value = 2.2e-16.

e) (2 pt) Using the likelihood ratio approach (G^2 deviance test), test the adequacy of a simplified logit model that treats `clevel` as a quantitative variable rather than as a categorical factor variable. What do you conclude from this test?

```
factor2 <- (glm(cbind(Cancer, Control) ~ clevel,
                family=binomial, data=dfs))
drop1(factor2, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## cbind(Cancer, Control) ~ clevel
##        Df Deviance     AIC    LRT  Pr(>Chi)
## <none>      9.887  47.617
## clevel  1   96.054 131.784 86.167 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Based off the test I conclude the model is not a good fit because its p-value = 2.2e-16; therefore, rejecting the null.

Problem 2. (12 pts)

The following data were reported on the FDA website from a randomized, prospective clinical trial of a vaccine for Covid-19. The outcome is onset of covid-19 after at least 14 days (Covid), or no onset (NoCovid). The data below are for the age groups 18-64 and 65+.

```
dfv
```

```
##     Age Treatment Covid NoCovid
## 1 18-64   Placebo   441   15111
## 2 18-64   Vaccine   157   15395
## 3   65+   Placebo    68    3924
## 4   65+   Vaccine    16    3954
```

a) (2 pts) Fit the saturated logistic regression model (the full model with interaction) using the (Covid, Nocovid) frequencies as the response. Display the model summary. Is the Wald coefficient test of the interaction statistically significant at level $\alpha = 0.05$?

```
slr <- glm(cbind(Covid, NoCovid) ~ Age*Treatment, family=binomial, data=dfv )

summary(slr)
```

```
##
## Call:
## glm(formula = cbind(Covid, NoCovid) ~ Age * Treatment, family = binomial,
##     data = dfv)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.53413    0.04831 -73.157  < 2e-16 ***
## Age65+                    -0.52123    0.13151  -3.963 7.39e-05 ***
## TreatmentVaccine          -1.05142    0.09364 -11.229  < 2e-16 ***
## Age65+:TreatmentVaccine -0.40312    0.29408  -1.371     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  2.0907e+02  on 3  degrees of freedom
## Residual deviance: -1.7435e-12  on 0  degrees of freedom
## AIC: 33.443
```

```
##
## Number of Fisher Scoring iterations: 3
```

Answer: We do not reject the Null because the p-value for the intersection is 0.17 which is greater than .05.

b) (2 pts) Let θ_{18-64} and θ_{65+} denote the conditional Covid/NoCovid odds ratios for Placebo versus Vaccine for the two age groups. Express the null hypothesis of no interaction in the logit model as a hypothesis about θ_{18-64} and θ_{65+}.

Answer:

$$H_0 : \beta_1 8 - 64 = \beta_6 5 + = 1$$

c) (2 pts) Fit the additive model (dropping the interaction term) and display the model summary.

Answer:

```
slr2 <- glm(cbind(Covid, NoCovid) ~ Age + Treatment, family=binomial, data=dfv )

summary(slr2)
```

```
##
## Call:
## glm(formula = cbind(Covid, NoCovid) ~ Age + Treatment, family = binomial,
##     data = dfv)
##
## Deviance Residuals:
##       1        2        3        4
## -0.2459   0.4143   0.6411  -1.1650
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.52228    0.04727 -74.512  < 2e-16 ***
## Age65+            -0.61250    0.11737  -5.218 1.81e-07 ***
## TreatmentVaccine -1.09669    0.08862 -12.375  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.0730  on 3  degrees of freedom
## Residual deviance:   2.0004  on 1  degrees of freedom
## AIC: 33.443
```

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

d) (2pts) Test for homogeneous association by performing a likelihood ratio test of the additive model versus the saturated model. What do you conclude?

```
anova(slr2, slr, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Covid, NoCovid) ~ Age + Treatment
## Model 2: cbind(Covid, NoCovid) ~ Age * Treatment
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         1     2.0004
## 2         0     0.0000  1   2.0004   0.1573
```

Answer: I can conclude that this is association between the two because the P-value is 0.1573 which is greater than 0.05, so we do not reject the Null.

e) (2 pts) Compute a profile likelihood confidence interval for coefficients, and translate into a profile likelihood confidence interval for the ratios of odds of Covid onset for vaccine versus placebo and for Age 65+ versus Age 18-64.

Answer:

```
confint(slr2)
```

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

##                       2.5 %     97.5 %
## (Intercept)      -3.6162341 -3.4308974
## Age65+           -0.8493105 -0.3886271
## TreatmentVaccine -1.2729165 -0.9253019
```

```
exp(confint(slr2))
```

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

##                       2.5 %     97.5 %
## (Intercept)      0.02688373 0.03235789
## Age65+           0.42770974 0.67798704
## TreatmentVaccine 0.28001376 0.39641173
```

f) (2 pts) Discuss how to interpret the confidence interval results. Is the vaccine effective, and if so how effective? What possible explanations might there be for a different odds of onset for the 65+ group versus the 18-64 group?

Answer: The vaccine is effective because the probability intervals are smaller than one. Those who are 65+ might have weaker immune system so the vaccine does not work as well as someone who is between 18-64.