STAT 426 Assignment 7

Due Tuesday, October 19, 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 comma-separated data file 'surgery.csv' is included with this assignment. The data concern patients having surgery under general anesthesia. The variables are

Y = whether a patient experienced a sore throat on waking (0 = no, 1 = yes)

D = duration of surgery (in minutes)

T = type of device used to secure the airway (0 = laryngeal mask airway, 1 = tracheal tube)

a) (2 pts) Read the data into a data frame and display the first few rows of the data set. (Hint: check R help for 'read.csv')

```
surgery = read.csv("surgery.csv", header = TRUE)
head(surgery)
```

```
##   Patient  D T Y
## 1       1 45 0 0
## 2       2 15 0 0
## 3       3 40 0 1
## 4       4 83 1 1
## 5       5 90 1 1
## 6       6 25 1 1
```

b) (2 pts) Fit a logit model that includes `D`, `T`, and their interaction. Based on the model summary, is the interaction term significant at level $\alpha = 0.05$?

```
logit = glm(Y ~ D * T, family = binomial, data = surgery)

summary(logit)
```

```
##
## Call:
## glm(formula = Y ~ D * T, family = binomial, data = surgery)
##
## 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
## D            0.02848    0.03429   0.831   0.4062
## T           -4.47224    2.46707  -1.813   0.0699 .
## D:T          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
```

```
#p-value: 0.1966, interaction not significant.
```

c) (2 pts) Fit the additive model and display the model summary. Provide likelihood ratio confidence intervals for the coefficients and interpret the effects of the two variables.

```
add_mod = glm(Y ~ D + T, family = binomial, data = surgery)

summary(add_mod)
```

```
##
## Call:
```

```
## glm(formula = Y ~ D + T, family = binomial, data = surgery)
##
## 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
## D            0.06868    0.02641   2.600  0.00931 **
## T           -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
```

```
confint(add_mod)
```

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

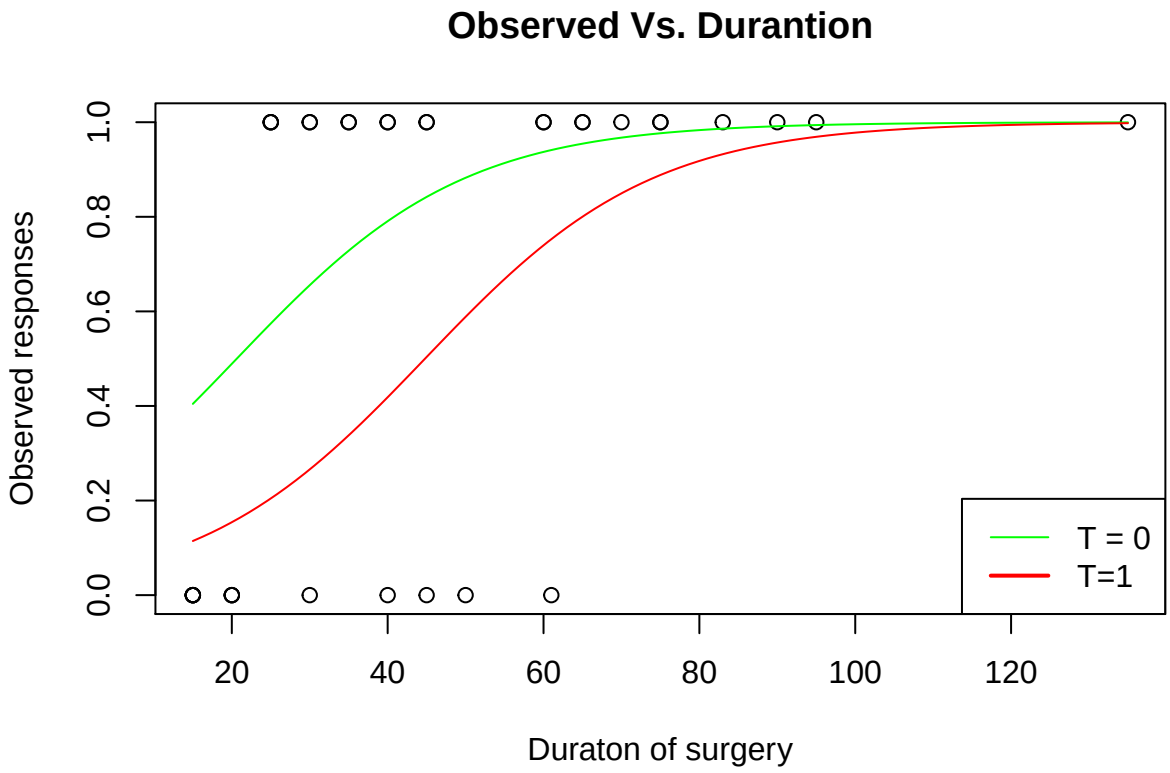
```
##                   2.5 %     97.5 %
## (Intercept) -3.80786158 0.61635531
## D            0.02547651 0.13216711
## T           -3.64627873 0.07434058
```

```
#The effect the two variables have according to the confidence interval D is significa
#This is because D is to the right of ) and T wraps around.
```

d) (2 pts) For the additive model, make a scatter plot of the observed responses versus duration of surgery. Indicate which device was used for each point using the plotting character. (One method is to use pch=T in the plot command). The add two curves to the plot, using different types of dashes or colors: the fitted probability response curve when T=0, and the fitted response curve when T=1. Also include a legend to indicate which curve is which.

```
plot(Y~D, data = surgery, main = "Observed Vs. Durantion", xlab = "Duraton of surgery",
curve(predict(add_mod, data.frame(T = 0, D=x), type="response"),col="Green",
add=TRUE)
```

```
curve(predict(add_mod, data.frame(T = 1, D=x), type="response"),col="Red",
add=TRUE)
legend("bottomright",
       c("T = 0", "T=1"),
       col = c("Green", "Red"), lwd = 1:2)
```



Problem 2. (12 pts)

Install faraway package: For this exercise, first make sure you have the `faraway` package installed. Check your R Studio list of packages. If the list includes `faraway`, you have it. If not, click the 'install' button in the packages window, and start typing `faraway` into the packages dialog box. It should auto-complete. Select 'faraway' and hit the **install** button.

Data: Aflatoxin B1, a type of mold that grows on peanuts and grains, was fed to lab animals at varying doses and the number responding with liver cancer were recorded. `dose` is the dosage in parts per billion, and, for each dose, `total` is the number of test animals and `tumor` is the number with lever cancer. The data are displayed below.

```
library(faraway)
aflatoxin
```

```
##   dose total tumor
## 1    0    18     0
```

```
## 2    1    22     2
## 3    5    22     1
## 4   15    21     4
## 5   50    25    20
## 6  100    28    28
```

a) (2 pts) Consider three link function models for these data:

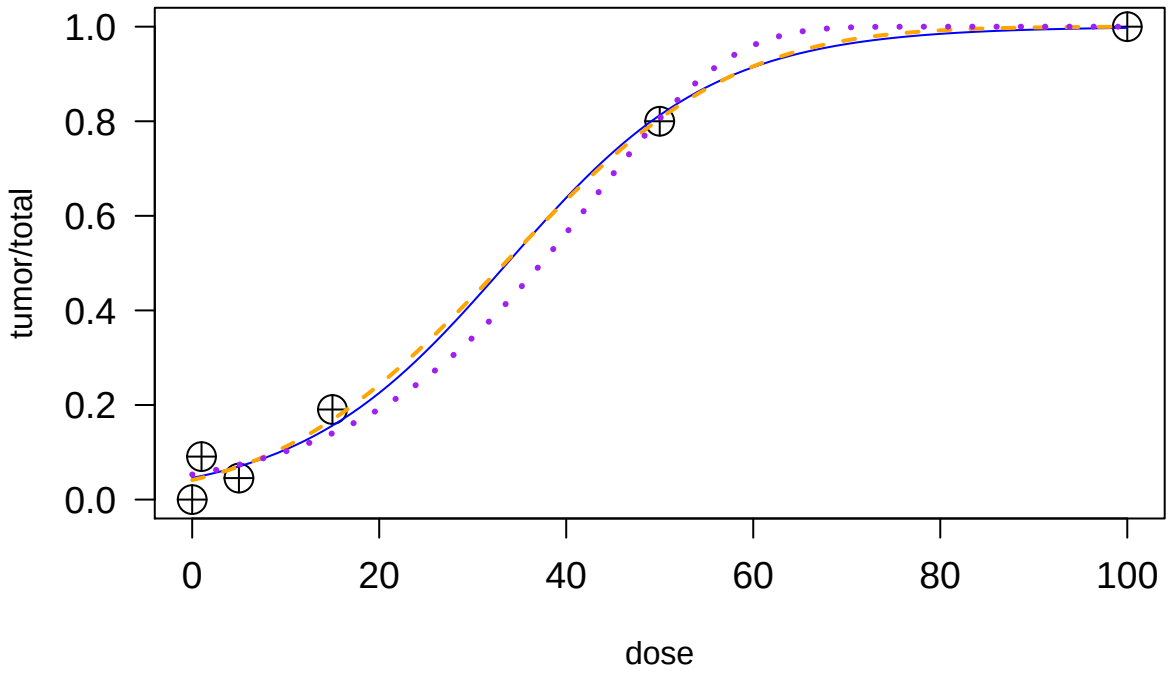
$$\begin{aligned} \text{Logit:} \quad & \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta\, dose \\ \text{Probit:} \quad & \Phi^{-1}(\pi) = \alpha + \beta\, dose \\ \text{Complementary log-log:} \quad & \log(-\log(1-\pi) = \alpha + \beta\, dose \end{aligned}$$

Make a scatter plot of the observed proportions with liver cancer versus dose. Add the fitted response curves for the three models to the graph.

```
mod.logit = glm(cbind(tumor,total-tumor) ~ dose,
family=binomial, data=aflatoxin)

mod.probit = glm(cbind(tumor,total-tumor) ~ dose,
family=binomial(link = probit), data=aflatoxin)

mod.cloglog = glm(cbind(tumor,total-tumor) ~ dose,
family=binomial(link = cloglog), data=aflatoxin)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
plot(tumor/total~ dose, data=aflatoxin, pch=10, cex=2, cex.axis=1.2, las=1)

curve(predict(mod.logit, data.frame(dose=x), type="response"), add=TRUE, lty=1, col="Blu

curve(predict(mod.probit, data.frame(dose=x), type="response"), add=TRUE, lty=2, col="Or

curve(predict(mod.cloglog, data.frame(dose=x),
type="response"), add=TRUE, lty=3, col="Purple", lwd=3)
```



b) (2 pts) Compare the deviances for the three models. Based on the curves and the deviances, which model appears to fit the best?

```
1-pchisq(deviance(mod.logit),df.residual(mod.logit))
```

```
## [1] 0.5752128
```

```
1-pchisq(deviance(mod.probit),df.residual(mod.probit))
```

```
## [1] 0.6225645
```

```
1-pchisq(deviance(mod.cloglog),df.residual(mod.cloglog))
```

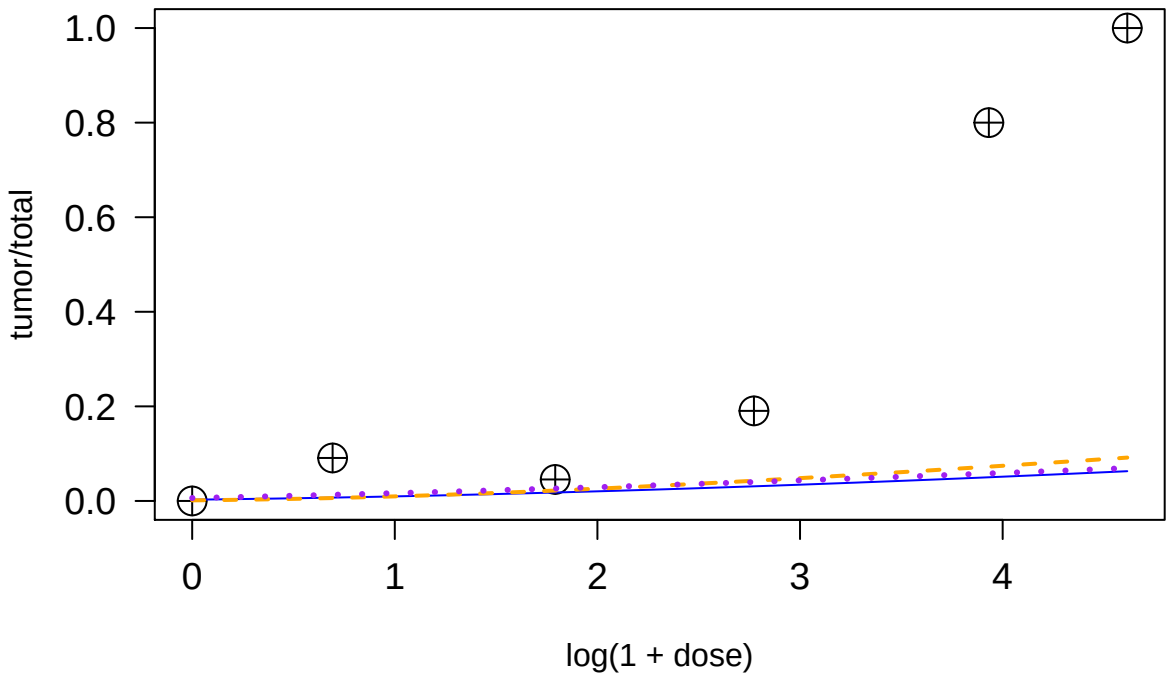
```
## [1] 0.5455295
```

```
#The model that appears the best is probit link
```

c) (2 pts) Make another scatter plot with three link function curves like the one in a), but this time use $\log(1 + dose)$ as the predictor variable, instead of *dose*.

```
mod.logit_log = glm(cbind(tumor,total-tumor) ~ log(1+dose),
family=binomial, data=aflatoxin)

mod.probi_logt = glm(cbind(tumor,total-tumor) ~ log(1+dose),
```

```
family=binomial(link = probit), data=aflatoxin)

mod.cloglog_log = glm(cbind(tumor,total-tumor) ~ log(1+dose),
family=binomial(link = cloglog), data=aflatoxin)

plot(tumor/total~ log(1+dose), data=aflatoxin, pch=10, cex=2, cex.axis=1.2, las=1)

curve(predict(mod.logit_log, data.frame(dose = x), type="response"), add=TRUE, lty=1, co

curve(predict(mod.probi_logt, data.frame(dose = x), type="response"), add=TRUE, lty=2, c

curve(predict(mod.cloglog_log, data.frame(dose = x),
type="response"), add=TRUE, lty=3, col="Purple", lwd=3)
```



d) (2 pts) Compare the deviances for the three $log(1 + dose)$ models. Based on the curves and deviances, which model appears to fit the best.

```
1 - pchisq(deviance(mod.logit_log), df.residual(mod.logit_log))
```

```
## [1] 0.02278691
```

```
1 - pchisq(deviance(mod.probi_logt), df.residual(mod.probi_logt))
```

```
## [1] 0.00726951
```

```r
1 - pchisq(deviance(mod.cloglog_log), df.residual(mod.cloglog_log))
```

```
## [1] 0.2187301
```

```r
#The best model is log-log link
```

e) (2 pts) Among the 6 different models that you fit, choose which one fits the best and use it to predict the proportion developing liver cancer at a dose of 20.

```r
#summary(mod.probit)
#sumary(mod.probit)$coefficients

mod_probit_pre = predict(mod.probit,newdata=data.frame(dose=20), interval="confidence",

mod_probit_pre
```

```
## $fit
##          1
## -0.6975651
##
## $se.fit
## [1] 0.1673653
##
## $residual.scale
## [1] 1
```

f) (2 pts) Provide a 95% confidence interval for your prediction in **e)**.

```r
beta = coefficients(mod.probit)
beta[1] + beta[2]*20
```

```
## (Intercept)
##  -0.6975651
```

```r
low = -0.6975651 - 1.96*0.1673653

high = -0.6975651 + 1.96*0.167365

c(low, high)
```

```
## [1] -1.0256011 -0.3695297
```