BTRY 6020 Homework IV Solution

NAME: student name

NETID: student NetID

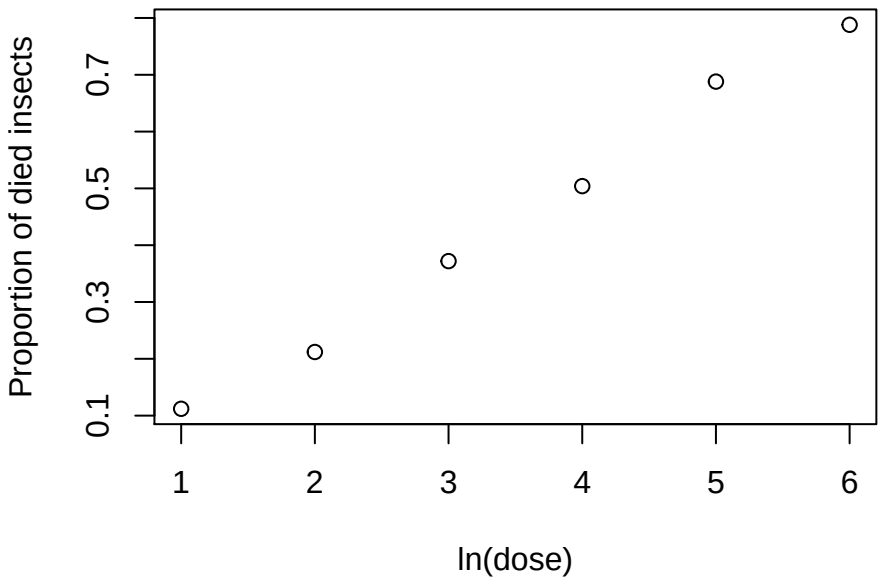
DUE DATE: March 13 2017, by 8:40 am

Question 1.

In an experiment with a newly developed insecticide, 1500 experimental insects were divided at random into six groups of 250 each. Each insect in a given group was exposed to a certain dose of the insecticide. A day later, the number of insects out of 250 that had died was recorded. Data appear in Hwk4Q1DatSp17. (Note this "grouped" data is in the format of CocaineTreatment2 of Lab 5, and must be handled accordingly)

A) For each dose level, calculate the proportion of insects that were killed within one day. Plot these proportions against the ln(dose) given in the data. Does a logistic model appear to fit the data?

```
library(readxl)
Q1DF <- read_excel("~/BTRY6020Sp17/Homework/Hwk4/Hwk4Q1DatSp17.xlsx")
Q1DF$PropDied <- Q1DF$NumDied/Q1DF$Num
plot(Q1DF$LnDose, Q1DF$PropDied, xlab="ln(dose)", ylab="Proportion of died insects")
```



The above plot shows valid curvature between the insect death probability and the predictor (ln(dose)) in a logistic regression. So a logistic model appears to be a good fit here.

B) Find the maximum likelihood estimates for β_o and β_1. State the fitted response function.

```
# Add a column of number of survivors
Q1DF$NumSurv <- Q1DF$Num - Q1DF$NumDied
Q1GLM <- glm(cbind(NumDied, NumSurv) ~ LnDose, data = Q1DF, family = binomial(logit))
summary(Q1GLM)
```

```
##
## Call:
## glm(formula = cbind(NumDied, NumSurv) ~ LnDose, family = binomial(logit),
##     data = Q1DF)
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -0.5092  -0.1115   0.7461  -0.2869   0.4744  -0.5599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.64367    0.15610  -16.93   <2e-16 ***
## LnDose       0.67399    0.03911   17.23   <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: 383.0695  on 5  degrees of freedom
## Residual deviance:   1.4491  on 4  degrees of freedom
```

```
## AIC: 39.358
##
## Number of Fisher Scoring iterations: 3
```

The MLE estimates for β_0 and β_1 are -2.64367 and 0.67399, respectively. The fitted model is

$$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.64367 + 0.67399 \times LnDose.$$

where $\pi = \mathbb{P}$(a death occurs).

C) Obtain and interpret a 90% confidence interval for $exp(\beta_1)$.

```
exp(confint(Q1GLM, parm = "LnDose", level = 0.90))
```

```
##      5 %     95 %
## 1.841433 2.094362
```

The 90% confidence interval for exp(β_1) is (1.841433, 2.094362). That is, we are 90% confident the odds of death are 1.84 to 2.09 times greater for every unit increase in LnDose.

D) Insects are exposed to a ln(dose) level of 3.5. What is the probability each will die? (Use an appropriate inferential procedure; sample R code: predict(GLMName, newdata, type="response", se.fit=T)).

```
# Construct 95% CI for P(Death|LnDose = 3.5)
PiHat <- predict(Q1GLM, newdata=data.frame(LnDose = 3.5), type="response", se.fit=T)
CInt <- c(PiHat$fit - 1.96*PiHat$se.fit, PiHat$fit + 1.96*PiHat$se.fit)
PiHat$fit
```

```
##         1
## 0.4293018
```

```
PiHat$se.fit
```

```
##          1
## 0.01468008
```

```
CInt
```

```
##         1         1
## 0.4005289 0.4580748
```

The estimated probability of a death of an insect is 0.4293018 when the ln(dose) level 3.5. And we are 95% confident that the probability of death at LnDose = 3.5 is between .401 and .458.

E) Give a point estimate for the median lethal dose (what entymologists refer to as the LD50)-the dose at which 50% of the insects are expected to die.

```
# At median lethal dose, log odds = 0
# Solve 0 = -2.64367 + 0.67399 LnDose for LnDose
LnDose <- 2.64367/0.67399
LnDose
```

```
## [1] 3.922417
```

```
exp(LnDose)
```

```
## [1] 50.52242
```

The estimated median lethal LnDose is 3.922417, so the median lethal dose is exp(3.922417) = 50.52242.

Question 2.

A psychologist conducted a study to determine if emotional stability is related to an employee's ability to complete a difficult and often frustrating task. Emotional stability was measured by the score on a written test commonly used to measure this. A random sample of 27 employees were selected from a single company that was willing to participate in the study. Data appears in Hwk4Q3DatSp17.

Does the likelihood of being able to do this task increase with emotional stability?

A) Formulation of the research question and choice of the appropriate statistical technique used to answer this question.

The research aims to determine how emotional stability affects an employee's ability to complete a task. We can fit a logistic regression with the emotional stability as the predictor and whether the task is completed (0-1) as the response to answer the question.

B) Notation for the random variable(s) and parameter(s) of interest; define these explicitly. Give the distributional assumptions for your random variable(s) and state all assumptions necessary for the statistical application you intend to use.

Let Y_i be 1 or 0 depending on whether the ith employee completes the task or not. Let X_i be the ith employee's emotional score. Assume $Y_i \sim ind\ Bernoulli\left(\frac{1}{1+e^{-\beta_0-X_i\beta_1}}\right)$ where $\frac{1}{1+e^{-\beta_0-X_i\beta_1}}$ stands for the probability that the ith employee completes the task.

There are two assumptions here: independence, and that the relationship between probability of being able to complete the task and emotional score follow a logistic relationship.

C) Calculations for the analysis. For hypothesis and significance tests, formulate the null and the alternative hypotheses, calculate the value of your test statistic, and then calculate your p-value. For confidence intervals, show and apply the appropriate formula. Use $\alpha = .05$ if not otherwise specified.

The null and alternative hypotheses are

$$H_0 : \beta_1 = 0 \quad vs \quad H_a : \beta_1 > 0.$$

```
library(readxl)
Q2DF <- read_excel("~/BTRY6020Sp17/Homework/Hwk4/Hwk4Q2DatSp17.xlsx")
Q2GLM <- glm(TaskComp ~EmotScore, data = Q2DF, family = binomial(logit))
summary(Q2GLM)
```

```
##
## Call:
## glm(formula = TaskComp ~ EmotScore, family = binomial(logit),
##     data = Q2DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7845  -0.8350   0.5065   0.8371   1.7145
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.308925   4.376997  -2.355   0.0185 *
## EmotScore     0.018920   0.007877   2.402   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 37.393  on 26  degrees of freedom
## Residual deviance: 29.242  on 25  degrees of freedom
## AIC: 33.242
##
## Number of Fisher Scoring iterations: 4
```

From the output, the test statistic is $z = 2.402$ and the p-value for this one-sided test is $\frac{0.0163}{2} = .00815$ which is less than $\alpha = 0.05$. So we reject H_0 in favor of H_a at $\alpha = 0.05$ and conclude that emotional stability increasess an employee's ability to complete a task.

```
confint(Q2GLM, parm = "EmotScore", level = 0.95)
```

```
##       2.5 %      97.5 %
## 0.005370759 0.037202638
```

D) Discuss whether the assumptions stated in Part B above are met sufficiently for the validity of the statistical inferences; use graphs and other tools where applicable.

There are only two assumptions here: independence and that the probability of task completion for a given emotional score follows a logistic function. Independence is guaranteed by a random sample of employees (and assuming appropriate testing conditions where they can't see each other's papers!). There is no way to evaluate whether or not the logistic relationship holds when there are not multiple repeat observations at each emotional score.

E) Discuss the sampling scheme and whether or not it is sufficient to meet the objective of the study. Be sure to include whether or not subjective inference is necessary and if so, defend whether or not you believe it is valid.

Since the data is randomly sampled from a single company, any inference from the study (and the model) can only reflect this one company. Therefore, the sampling scheme is not sufficient to meet the goal of the study if we want to extrapolate results to a bigger population. Perhaps if we knew more about the company and its employees we could subjectively infer to a greater population.

F) State the conclusions of the analysis. These should be practical conclusions from the context of the problem, but should also be backed up with statistical criteria (like a p-value, etc.). Include any considerations such as limitations of the sampling scheme, impact of outliers, etc., that you feel must be considered when you state your conclusions.

We conclude from the hypothesis testing that emotional stability does affect an employee's ability to complete a task for employees from the company in the study. However, to make a conclusion for a bigger population, we need a more diverse sample (one from more companies) or more information about the type of company and its employees to subjectively infer beyond this limited sampled population.