BTRY 6020 Homework IV

NAME: ANDRES CASTANO

NETID: student AC986

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

Question 1.

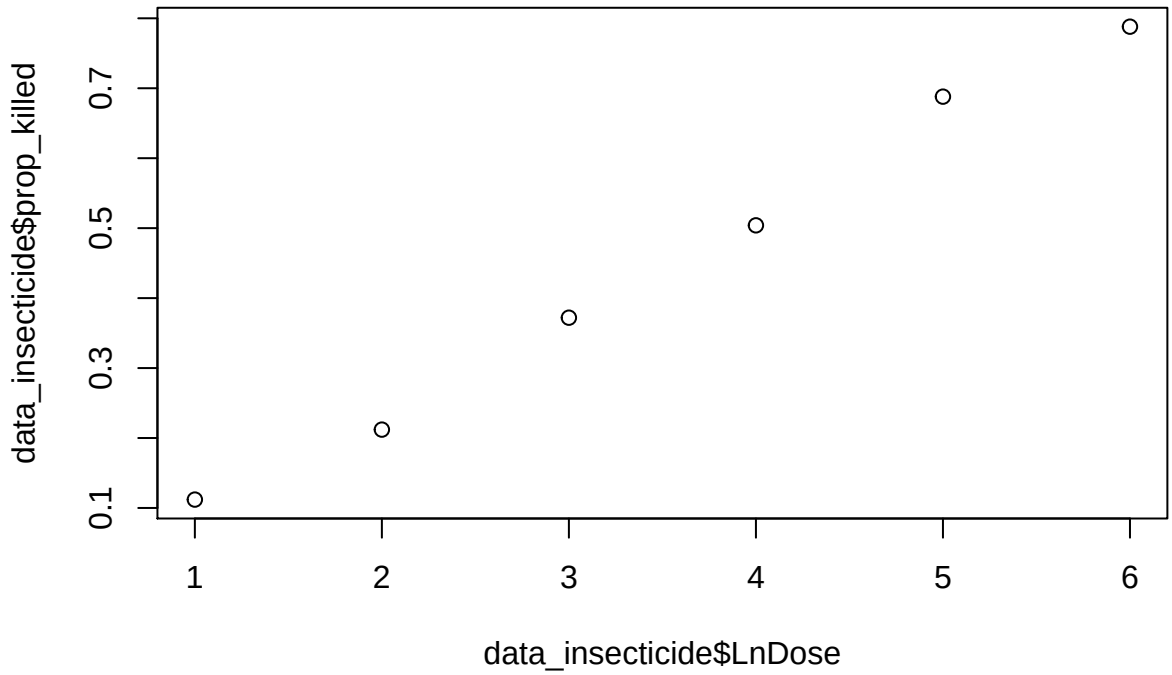
In an experiment with a newly developed insecticide, 1500 experimental insects were divided at random into six groups of 250 each. Each insect in a given group was exposed to a certain dose of the insecticide. A day later, the number of insects out of 250 that had died was recorded. Data appear in Hwk4Q1DatSp17. (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)
data_insecticide = read_excel("Hwk4Q1DatSp17.xlsx")
head(data_insecticide)
```

```
##   LnDose Num NumDied
## 1      1 250      28
## 2      2 250      53
## 3      3 250      93
## 4      4 250     126
## 5      5 250     172
## 6      6 250     197
```

```
data_insecticide$prop_killed=(data_insecticide$NumDied)/(data_insecticide$Num)
plot(data_insecticide$LnDose,data_insecticide$prop_killed)
```




```
# calculate number of insects and proportion that not died
data_insecticide$NumNoDied=data_insecticide$Num-data_insecticide$NumDied
data_insecticide$prop_nokilled=(data_insecticide$Num - data_insecticide$NumDied)/(data_insecticide$Num)
```

The data seems to fit the logistic distribution well, we see that the probability that the insects got kill increase with the amount of dose but is delimited in the interval $[0, 1]$.

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

```
# model fit
insecticide.glm=glm(cbind(NumDied,NumNoDied)~LnDose,family='binomial',data=data_insecticide)
summary(insecticide.glm)
```

```
##
## Call:
## glm(formula = cbind(NumDied, NumNoDied) ~ LnDose, family = "binomial",
##     data = data_insecticide)
##
## 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 maximum likelihood estimators for β_0 and β_1 are -2.64367 and 0.67399, respectively. On the other hand, the fitted response function is:

$$logit(\pi) = log_e(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = -2.64367 + 0.67399 * LnDose_i$$

Where $\hat{\pi}_i$ is $\hat{p}(y_i = 1|LnDose)$ which represent the estimated probability that the insects not died whithin one day given the insecticide. This estimated probability is equal to $\hat{\pi}_i = \frac{1}{1+exp(-(-2.64367+0.67399LnDose_i))}$.

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

The 90(1-α)% confidence interval for $exp(\beta_1)$ is defined as:

$$exp(\hat{\beta}_1 \pm z_{\frac{0.10}{2}} * SE(\hat{\beta}_1))$$

In R we can get this interval as follows:

```
b1=0.67399
se_b1=0.03911
z=-qnorm(0.05)
lower=b1 - (z)*se_b1
upper=b1 + (z)*se_b1
c(exp(lower), exp(upper))
```

```
## [1] 1.839805 2.092418
```

With the following orders, we can verified our calculations:

```
library(MASS)
# using standard errors
exp(confint.default(insecticide.glm, level = 0.90))
```

```
##                    5 %      95 %
## (Intercept) 0.05499875 0.0919137
## LnDose      1.83981322 2.0924203
```

```
# using profiled log-likelihood
exp(confint(insecticide.glm, level = 0.90))
```

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

```
##                    5 %       95 %
## (Intercept) 0.05477389 0.09155635
## LnDose      1.84143267 2.09436198
```

The intervals are calculated using the two formulations do not difer to much. The professor used the definition exp(confint(model, level = 0.90)) during class, so this should be the correct calculation. We are 95% confident that the odds of died increase between 100(1.84143267-1)=84.4% and 100(2.09436198-1)=109.4% for each unit of increase in the dose.

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)).

The probability that insects exposed to a ln(dose)=3.5 die is equal to:

$$\hat{p}(y_i = 1|Ln(Dose) = 3.5) = \frac{1}{1 + exp(-(-2.64367 + 0.67399 * 3.5))} = 0.429$$

In R we get this estimated probability as follows:

```
predic_dose35 = predict(insecticide.glm, data.frame(LnDose=3.5), type="response", se.fit=TRUE)
predic_dose35
```

```
## $fit
##         1
## 0.4293018
##
## $se.fit
##          1
## 0.01468008
##
## $residual.scale
## [1] 1
```

```
#95% C.I for the estimated probabilty
lower= predic_dose35$fit - (-qnorm(0.025))*predic_dose35$se.fit
upper= predic_dose35$fit + (-qnorm(0.025))*predic_dose35$se.fit
c(lower,upper)
```

```
##         1         1
## 0.4005294 0.4580742
```

The probability that insects die is 42.9%. On the other hand, we are 95% confident that for a ln(Dose) of 3.5, the probability that insects die is between 40% and 45.8%.

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.

The value of median lethat dose at which 50% of the insects are expected to die is equal to $-\frac{\hat{\beta}_0}{\hat{\beta}_1}$

```
# median lethal dose (mld)
coef=coefficients(insecticide.glm)
coef
```

```
## (Intercept)      LnDose
##  -2.6436750   0.6739928
```

```
mld=-(coef[1]/coef[2])
mld
```

```
## (Intercept)
##    3.922409
```

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 question is ¿Does the emotional stability increase the likelihood (probability) of perform a task?

We know that the probability of perform a difficult task given the emotional score is not lineary related with the parameters β_0 and β_1 and can be defined as:

$$p(TaskComp_i = 1|EmotScore) = \frac{1}{1 + exp(-(-\beta_0 + \beta_1 * EmotScore_i))}$$

After obtain the maximum likelihood estimates for β_0 and β_1 we need to test whether $\beta_1 > 0$ or not. β_1 give us the direction in which the emotional stability score affect the likelihhod of perform a difficult task. When $\beta_1 > 0$ means that the emotial stability affects positively the likelihood of perform a difficult task. Then, we can define a one side test for β_1 as follows:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 > 0$$

Then we can define the wald test statistic:

$$TS = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

This test statistics is distributed standard normal. So we use the standard normal distribution to make the one side test ($\beta_1 > 0$). Finally, based on the p-value of the test we should decide.

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.

The parameter of interest is $\hat{\beta}_1$ (which is a random variable). we assume that β_1 is consistent and asymptotically normal distributed, another key assumptions are:

1) Observations are independent.
2) X fixed (if X is random then we assume that X is independent of the error)
3) No outliers driving the results

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.

```
library(readxl)
data_task = read_excel("Hwk4Q2DatSp17.xlsx")
head(data_task)
```

```
##   TaskComp EmotScore
## 1        0       474
## 2        0       432
## 3        0       453
## 4        1       481
## 5        1       619
## 6        0       584
```

```
#Model
task.glm=glm(TaskComp~EmotScore, family = binomial, data=data_task)
#Summary
summary(task.glm)
```

```
##
## Call:
## glm(formula = TaskComp ~ EmotScore, family = binomial, data = data_task)
##
## 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
```

After obtain the maximum likelihood estimates for β_0 and β_1 , we need to test whether $\beta_1 > 0$ or not. β_1 give us the direction in which the emotional stability score affect the likelihhod of perform a difficult task. When $\beta_1 > 0$ means that the emotial stability affects positively the likelihood of perform a difficult task. Then, we can define a one side test for β_1 as follows:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 > 0$$

Then we can define the wald test statistic:

$$TS = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

This test statistics is distributed standard normal. So we use the standard normal distribution to make the one side test (for $\beta_1 > 0$).

```
#test statistics
b1= 0.018920
se_b1 = 0.007877
```

```
ts=b1 /se_b1
ts
```

```
## [1] 2.40193
```

```
pvalue = pnorm(ts, lower.tail=FALSE)
pvalue
```

```
## [1] 0.008154422
```

P-value<0.05, then we have evidence to say that the emotional stability increases the likelihood of performing a task. The coefficient $\hat{\beta}_1 = 0.018920$ means that for each unit of increase in the emotional stability score, the odd in favor of perform the task increase by 100(exp(0.018920)-1)=1.91%.

100(1-0.05)% confidence interval for $\hat{\beta}_1$ is:

```
library(MASS)
coeff_ci=confint.default(task.glm, level = 0.95)
coeff_ci
```

```
##                       2.5 %      97.5 %
## (Intercept) -18.887682442 -1.73016791
## EmotScore     0.003481266  0.03435839
```

```
#95% CI interval in expo
lower=coeff_ci[2,1]
lower
```

```
## [1] 0.003481266
```

```
upper=coeff_ci[2,2]
upper
```

```
## [1] 0.03435839
```

```
c(lower,upper)
```

```
## [1] 0.003481266 0.034358393
```

```
# expressed in exponential form
exp(c(lower,upper))
```

```
## [1] 1.003487 1.034955
```

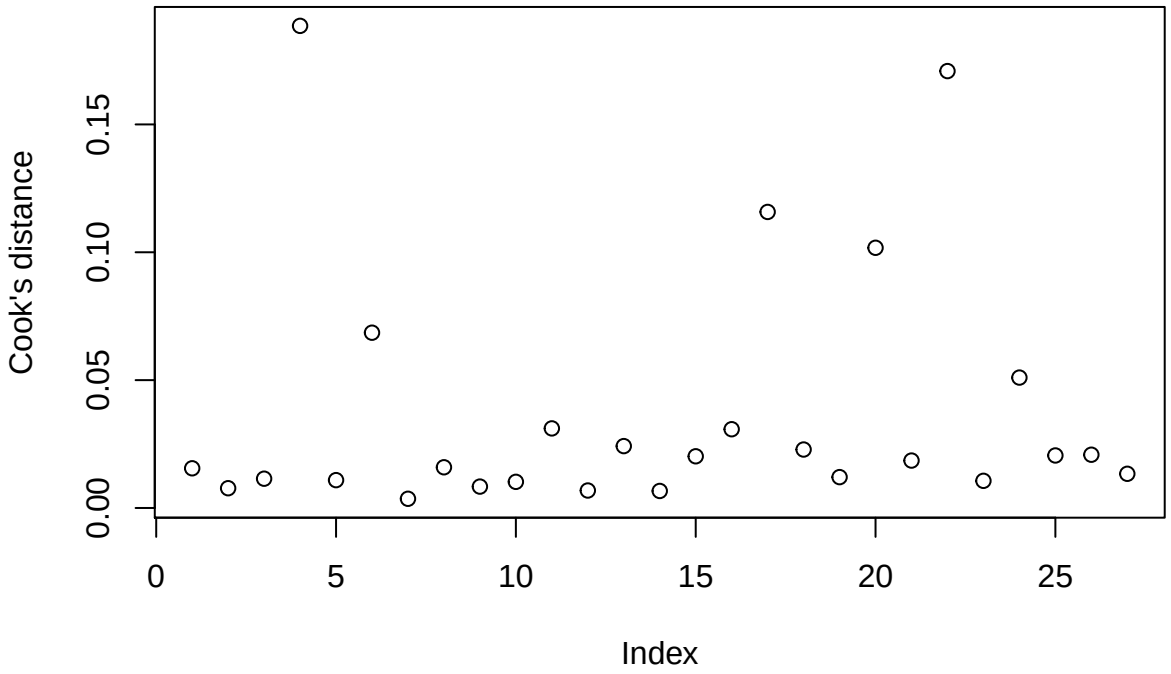
D) Discuss whether the assumptions stated in Part B above are met sufficiently for the validity of the statistical inferences; use graphs and other tools where applicable.

- The asummpation of independence in the observations is met by the fact that the data collection process in the only participant company was a random sample.
- The assumption of exogeinity between the X and the error could be in jeopardy (we normally have not used the the logit formulation from discrete choice modelling in econometrics that includes the error in the modelling process). for example, the skill of the workers measured by their academic degree (or another skill measure) is not include in the explanatory variables and it is possible that the level of skill of the workers may affect the ability of the workers to deal with their frustractions and therefore might affect their emotional inteligent scores.
- The results only have internal validity (the company sampled). The purpose of the psychologist is to find an answer that not only have internal validity (the company used in the sample) but also external validity, e.g, that its conclusiones can be extended to the rest of the firms. In our case, it is very naive to think that this results have external validity for at least two reasons:

1. We do not know what is the sector that represent the company used and it is possible to think that the ability to perform difficult task may vary from companies in different sectors. A difficult task may require a different set of skills in different sectors (for instance, a difficult task for a software engineer compared to a difficult task for rental manager).
2. The sample could not be representative of the companies in general, then it will not represent the hetegoneity in skills that the different firms could have.

- The assumption of not outliers driving conclusions can be tested by extending the cooks distance measure to the logistic regression case:

```
# cook's distance
cook_tasks = cooks.distance(task.glm)
plot(cook_tasks , ylab = "Cook's distance")
```



There is not highly influential points (cooks distance>1).

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.

I consider that the samply scheme is only valid to reach conclusions that are internally valid, e.g, only valid to the sampled firm. However, as far as I can understand the question, it seems to me that the purpose of the psychologist was to find an answer that not only have internal validity, but also external validity, e.g, that its conclusiones can be extended to the rest of the firms. In our case, I consider that we can reasonably claim that one firm can not be representative to all firms for different sectors, therefore the sampling scheme is not sufficient to meet the objective of the study.

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.

The statistical results support the ideat that the emotional intelligence has a positive effect in the likelihood to perform a difficult task. The results of the cooks distance measure also show that the outliers are no driving the conclusions. However, since the sample scheme its no representative of the population of firms

and we have not control for another potential explanatory variables, the conclusions are only valid internally and should not be extended to the rest of the companies. Therefore, the objective of the study is not met.