**Nkamnebe Chukwudi Nnaemeka**

**921-31-1827**

**DS560**

**Homework 4 Solutions.**

**Question 1.**

# (a)  
  
In this dataset, kyphosis = 1 means kyphosis is present; kyphosis = 0 means kyphosis is absent

kyphosis <- read.csv("~/WIU Graduate Program/Spring 2022/DS 560 - Categorical Data Analysis/kyphosis.csv")  
# View(kyphosis)  
logreg <- glm(kyphosis$Kyphosis ~ kyphosis$Age\_months, data = kyphosis, family = binomial(link = logit))  
summary(logreg)

##   
## Call:  
## glm(formula = kyphosis$Kyphosis ~ kyphosis$Age\_months, family = binomial(link = logit),   
## data = kyphosis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3126 -1.0907 -0.9482 1.2170 1.4052   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.572693 0.602395 -0.951 0.342  
## kyphosis$Age\_months 0.004296 0.005849 0.734 0.463  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 55.051 on 39 degrees of freedom  
## Residual deviance: 54.504 on 38 degrees of freedom  
## AIC: 58.504  
##   
## Number of Fisher Scoring iterations: 4

Notice that the p-value for the estimate of Age\_months is 0.46, consequently, for this model, we conclude that Age has no significant effect  
  
# (b)

library(ggplot2)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# Showing difference in dispersion in the mean ages  
ggplot(kyphosis, aes(x = Kyphosis, y = mean(Age\_months))) +  
 geom\_col() +  
 labs(title = "Mean Age for Kyphosis category",  
 x = "Kyphosis, 0 = Absent, 1 = Present",  
 y = "Mean Age in Months") +  
 theme(plot.title = element\_text(hjust = 0.5))

Chart, bar chart

Description automatically generated

# Showing difference in dispersion in the std ages  
ggplot(kyphosis, aes(x = Kyphosis, y = sd(Age\_months))) +  
 geom\_col() +  
 labs(title = "Standard Deviation of Age for Kyphosis category",  
 x = "Kyphosis, 0 = Absent, 1 = Present",  
 y = "St.d Age in Months") +  
 theme(plot.title = element\_text(hjust = 0.5))

Chart, bar chart

Description automatically generated Notice that for both the mean and standard deviation of ages in two Kyphosis categories, the category in which Kyphosis was absent has a hugher value for the mean and standard deviation  
# (c)

kyphosis2 <- cbind(kyphosis, kyphosis[, 1] ^ 2)  
colnames(kyphosis2) <- c("Age\_months", "Kyphosis", "Age\_months\_Squared")  
View(kyphosis2)  
  
logitReg <- glm(kyphosis2$Kyphosis ~ kyphosis2$Age\_months + kyphosis2$Age\_months\_Squared, family = binomial, data = kyphosis2)  
summary(logitReg)

##   
## Call:  
## glm(formula = kyphosis2$Kyphosis ~ kyphosis2$Age\_months + kyphosis2$Age\_months\_Squared,   
## family = binomial, data = kyphosis2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.482 -1.009 -0.507 1.012 1.788   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0462547 0.9943478 -2.058 0.0396 \*  
## kyphosis2$Age\_months 0.0600398 0.0267808 2.242 0.0250 \*  
## kyphosis2$Age\_months\_Squared -0.0003279 0.0001564 -2.097 0.0360 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 55.051 on 39 degrees of freedom  
## Residual deviance: 48.228 on 37 degrees of freedom  
## AIC: 54.228  
##   
## Number of Fisher Scoring iterations: 4

Notice that the squared age term is significant at 5% significance level when using the logit of the probability of Present (1). This means that for every one unit increase in the square of age (in months) the odds of Kyphosis been present reduces by 0.03%  
.

ggplot(kyphosis2, aes(fitted(logitReg))) +  
 geom\_histogram() +  
 labs(title = "Histogram of fitted values",  
 x = "Fitted Logit",  
 y = "Frequency") +  
 theme(plot.title = element\_text(hjust = 0.5))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Chart, histogram

Description automatically generated

# Solution to Problem 2

# (a)  
  
Firstly, I imported the dataset into R after rearranging the dataset

MBPT <- read.csv("~/WIU Graduate Program/Spring 2022/DS 560 - Categorical Data Analysis/Myers-Briggs Personality Types Samples.csv")  
View(MBPT)  
  
MBPT\_fit <- glm(Yes / (Yes + No) ~ E\_I + S\_N + T\_F + J\_P, weights = Yes + No, family = binomial, data = MBPT)  
summary(MBPT\_fit)

##   
## Call:  
## glm(formula = Yes/(Yes + No) ~ E\_I + S\_N + T\_F + J\_P, family = binomial,   
## data = MBPT, weights = Yes + No)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2712 -0.8062 -0.1063 0.1124 1.5807   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.1140 0.2715 -7.788 6.82e-15 \*\*\*  
## E\_IIntroversion -0.5550 0.2170 -2.558 0.01053 \*   
## S\_NSensing -0.4292 0.2340 -1.834 0.06664 .   
## T\_FThinking 0.6873 0.2206 3.116 0.00184 \*\*   
## J\_PPerceiving 0.2022 0.2266 0.893 0.37209   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 30.488 on 15 degrees of freedom  
## Residual deviance: 11.149 on 11 degrees of freedom  
## AIC: 73.99  
##   
## Number of Fisher Scoring iterations: 4

The prediction equation is given as;  
  
(Y = 1) = - 2.1140 - 0.5550(E\_I) - 0.4292(S\_N) + 0.6873(T\_F) + 0.2022(J\_P)  
where;  
  
The indicator variables are set up such that;  
  
E\_I = 1, when its Introversion and E\_I = 0, when its Extroversion  
S\_N = 1, when its Sensing and S\_N = 0, when its iNtuitive  
T\_F = 1, when its Thinking and T\_F = 0, when its Feeling  
J\_P = 1, when its Perceiving and J\_P = 0, when its Judging  
  
 # (b)  
  
From (Y = 1) = - 2.1140 - 0.5550(E\_I) - 0.4292(S\_N) + 0.6873(T\_F) + 0.2022(J\_P)  
  
E\_I = 0; S\_N = 1; T\_F = 1; and J\_P = 0

exp(- 2.1140 - 0.5550\*(0) - 0.4292\*(1) + 0.6873\*(1) + 0.2022\*(0))

## [1] 0.1563122

 # (c)

max(fitted(MBPT\_fit))

## [1] 0.2271486

Based on the model parameter estimates, the personality type with the highest probability of drinking alcohol frequently is ENTP because of the concept of odds; The estimates are analysed as follows;  
  
 (i): Notice from the prediction equation, that E\_I has a -ve sign, and it is for when a person has a personality that includes Introversion. For E\_I, the estimated odds that a person who is introverted drinks alcohol frequently is exp(-0.5550) = 0.574 times the odds that a person who is extroverted, which means extroverts are predicted to drink alcohol more frequently than introverts,  
  
 (ii): The estimate for S\_N is -0.4292, which means that the estimated odds that a person who is sensing drinks alcohol frequently is exp(-0.4292) = 0.6510 times the odds that a person who is iNtuitive. That is, iNtuitive persons are predicted to drink alcohol more frequently than a person who is sensing  
  
 (iii): The estimate for T\_F is 0.6873, which means that the estimated odds that a person who is Thinking drinks alcohol frequently is exp(0.6878) = 1.989 times the odds that a person who is feeling. This means that Thinking persons are predicted to drink alcohol more frequently than a person who is feeling  
  
 (iv): The estimate for J\_P is 0.2022, which means that the estimated odds that a person who is Perceiving drinks alcohol frequently is exp(0.2022) = 1.22 times the odds that a person who is Judging. This means that a Perceiving person is more likely to drink alcohol than a person who is Judging  
  
This is why the personality ENTP has the highest probability of drinking alcohol frequently

# Solution to Problem 3

 # (a)

crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat", head = TRUE)  
  
Q3a <- glm(y ~ weight + factor(color), family = binomial, data = crabs)  
summary(Q3a)

##   
## Call:  
## glm(formula = y ~ weight + factor(color), family = binomial,   
## data = crabs)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1908 -1.0144 0.5101 0.8683 2.0751   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.2572 1.1985 -2.718 0.00657 \*\*   
## weight 1.6928 0.3888 4.354 1.34e-05 \*\*\*  
## factor(color)2 0.1448 0.7365 0.197 0.84410   
## factor(color)3 -0.1861 0.7750 -0.240 0.81019   
## factor(color)4 -1.2694 0.8488 -1.495 0.13479   
## ---  
## 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: 188.54 on 168 degrees of freedom  
## AIC: 198.54  
##   
## Number of Fisher Scoring iterations: 4

The model is given as;  
logit[P(Y = 1)] = -3.2572 + 1.6928*weight + 0.1448*c2 - 0.1861*c3 - 1.2694*c4  
  
where;  
c2 = 1 for color = medium, 0 otherwise,  
c3 = 1 for color = medium dark, 0 otherwise,  
c4 = 1 for color = dark, 0 otherwise  
  
The interpretation of the parameter estimates are as follows;  
  
 (i): The Intercept Estimate - The value of Logit function when weight is 0 and color is none of the colors mentioned in the model  
  
 (ii): The estimate of the weight parameter - 1 unit increase in weight increases the odds of Y = 1 by exp(1.6928) assuming other factors remains unchanged  
  
 (iii): For medium-light colored crabs (category 1), c2 = c3 = c4 = 0, and the prediction equation is logit[P(Y = 1)] = −3.2572 + 1.6928*weight*  
  
*(iv) Similarly, for dark crabs, c2 = c3 = 0 and c4 = 1, so the prediction equation becomes logit[P(Y = 1)] = (−3.2572 - 1.2694) + 1.6928*weight; and so on  
  
 (v) This model assumes lack of interaction between weight and color, the effect of weight is the same for all colors  
  
  
 (b)

Q3b <- glm(y ~ weight + color, family = binomial, data = crabs)  
summary(Q3b)

##   
## Call:  
## glm(formula = y ~ weight + color, family = binomial, data = crabs)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1596 -0.9998 0.5237 0.8825 1.9109   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0316 1.1161 -1.820 0.0687 .   
## weight 1.6531 0.3825 4.322 1.55e-05 \*\*\*  
## color -0.5142 0.2234 -2.302 0.0213 \*   
## ---  
## 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: 190.27 on 170 degrees of freedom  
## AIC: 196.27  
##   
## Number of Fisher Scoring iterations: 4

 (i): The estimate of the weight parameter; 1 unit increase in weight increases the odds of Y = 1 by exp(1.6531) assuming color remains unchanged  
  
 (ii) At a given weight, for every one-category increase in color darkness, the estimated odds of a satellite multiplies by exp(-0.51