mjon238_Assignment1

mjon238

09/01/2022

Question 1)

a) For each row in the data set, calculate the proportion of lobsters that survived.Print the proportions out, and add them as a new column to the data set.

```
#Print proportion of lobsters that survived for each row
`colnames<-`(lobsterdf[,3]/lobsterdf[,2], 'propotion_survived')
```

```
##    propotion_survived
## 1           0.0000000
## 2           0.1000000
## 3           0.1363636
## 4           0.3333333
## 5           0.5454545
## 6           0.5862069
## 7           0.7222222
## 8           0.7058824
## 9           0.8750000
## 10          1.0000000
## 11          1.0000000
```

```
#Add that to the dataframe
lobsterdf <- lobsterdf%>%
  transmute(size,
            n,
            survived,
            'proportion_survived' = survived/n)

#Show the dataFrame
lobsterdf
```

```
## # A tibble: 11 x 4
##     size     n survived proportion_survived
##    <dbl> <dbl>    <dbl>               <dbl>
##  1    27     5        0               0
##  2    30    10        1               0.1
##  3    33    22        3               0.136
##  4    36    21        7               0.333
##  5    39    22       12               0.545
##  6    42    29       17               0.586
```

```
##  7    45    18       13              0.722
##  8    48    17       12              0.706
##  9    51     8        7              0.875
## 10    54     6        6              1
## 11    57     1        1              1
```

b) Fit the same model that the authors did. Show the code you used to fit the model, and print the summary() output.

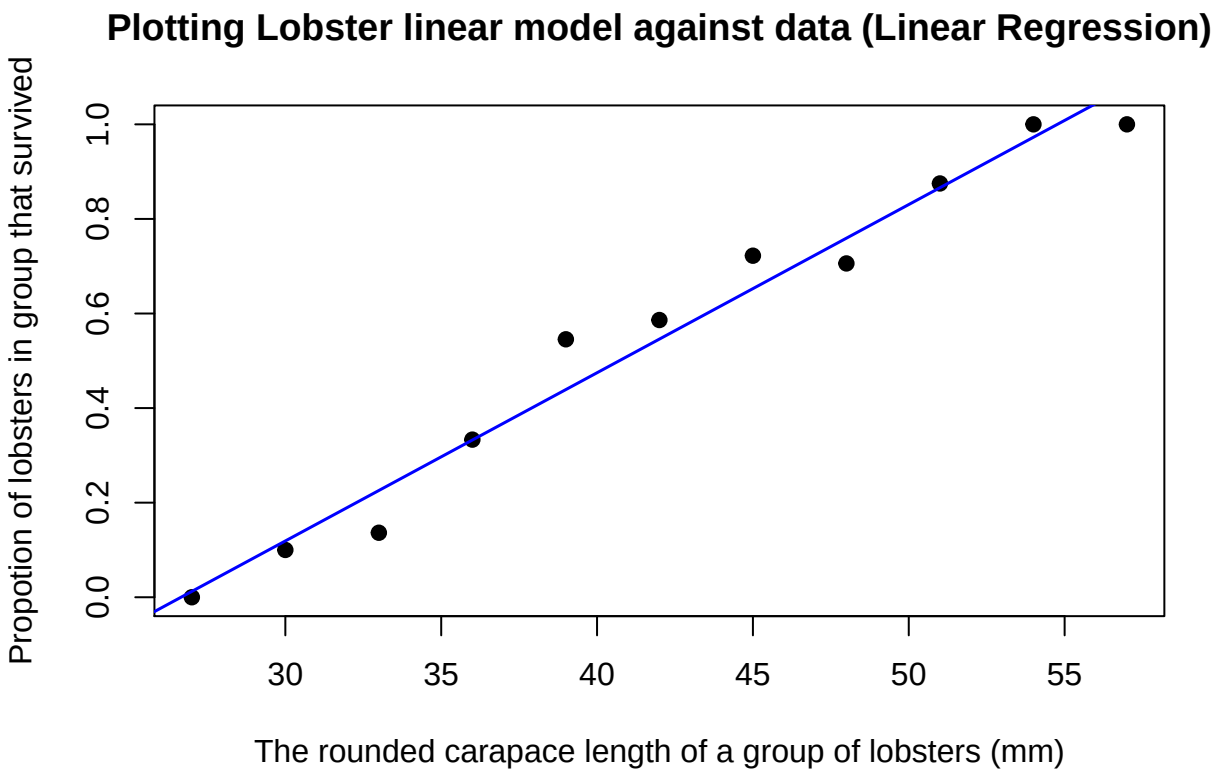
We are trying to determine how "a juvenile lobster's size was related to its vulnerability to predation". When using linear regression we will use proportion_survived as our response variable and size as an explanatory variable.

```
lobster.fit <- lm(proportion_survived ~ size,  data = lobsterdf)
summary(lobster.fit)
```

```
##
## Call:
## lm(formula = proportion_survived ~ size, data = lobsterdf)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.089376 -0.036212  0.000887  0.033829  0.106301
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.948038   0.086867  -10.91 1.72e-06 ***
## size         0.035569   0.002017   17.63 2.75e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06348 on 9 degrees of freedom
## Multiple R-squared:  0.9719, Adjusted R-squared:  0.9687
## F-statistic: 310.8 on 1 and 9 DF,  p-value: 2.752e-08
```

c) Create a scatter plot of size against the proportion of surviving lobsters.

```
#Plotting our linear regression model
plot(lobsterdf$size,
     lobsterdf$proportion_survived,
     main = "Plotting Lobster linear model against data (Linear Regression)",
     xlab = 'The rounded carapace length of a group of lobsters (mm)',
     ylab = 'Propotion of lobsters in group that survived',
     pch = 19)
abline(lobster.fit, col = 'blue', lwd = 1.5)
```



d) Interpret the effects estimated by your model.

We estimate that for every 1 millimeter increase in the rounded carapace length of a group of lobsters, the expected proportion of lobsters in that group that survive increases by approximately 0.035569.

e) List the reasons why by explaining how assumptions required by the model do not hold.

In linear regression the ith observations response ($proportion\_survived_i$) comes from a normal distribution. This indicates that for observation size = 27, the model suggests it would be likely that there is a negative of proportion of lobsters in that group that survived. This is clearly implausible and therefore the assumption is violated.

This becomes a clear issue when we either make predictions outside of the data range or when predicting proportion survived for the largest group (57mm).

```
#if size = 20
predict.lm(lobster.fit, newdata = data.frame(size = 20))
```

```
##         1
## -0.236658
```

```
#If size = 57
predict.lm(lobster.fit, newdata = data.frame(size = 57))
```

```
##        1
## 1.079395
```

Both results are either greater than 1 or less than 0, which is clearly nonsense.

Linear regression also assumes a constant variance. This is unlikely because lobster groups where the proportion survived = 0.5 will have a greater variation than the very large lobsters, which probably all survived, or the very small lobsters, which probably all died.

f) Select a model you think is appropriate for these data. Show the code you used to fit the model, and print the summary() output.

We will fit a logistic regression, with binomial distribution, and predict the odds of survival.

```
lobster.fit2 <- glm(cbind(survived, n - survived) ~ size, lobsterdf, family = "binomial")
summary(lobster.fit2)
```

```
##
## Call:
## glm(formula = cbind(survived, n - survived) ~ size, family = "binomial",
##     data = lobsterdf)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.12729  -0.43534   0.04841   0.29938   1.02995
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.89597    1.38501  -5.701 1.19e-08 ***
## size         0.19586    0.03415   5.735 9.77e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 52.1054  on 10  degrees of freedom
## Residual deviance:  4.5623  on  9  degrees of freedom
## AIC: 32.24
##
## Number of Fisher Scoring iterations: 4
```

g) Interpret the effects estimated by your model.

```
#Log odds
coef(lobster.fit2)
```

```
## (Intercept)        size
##  -7.8959697   0.1958579
```

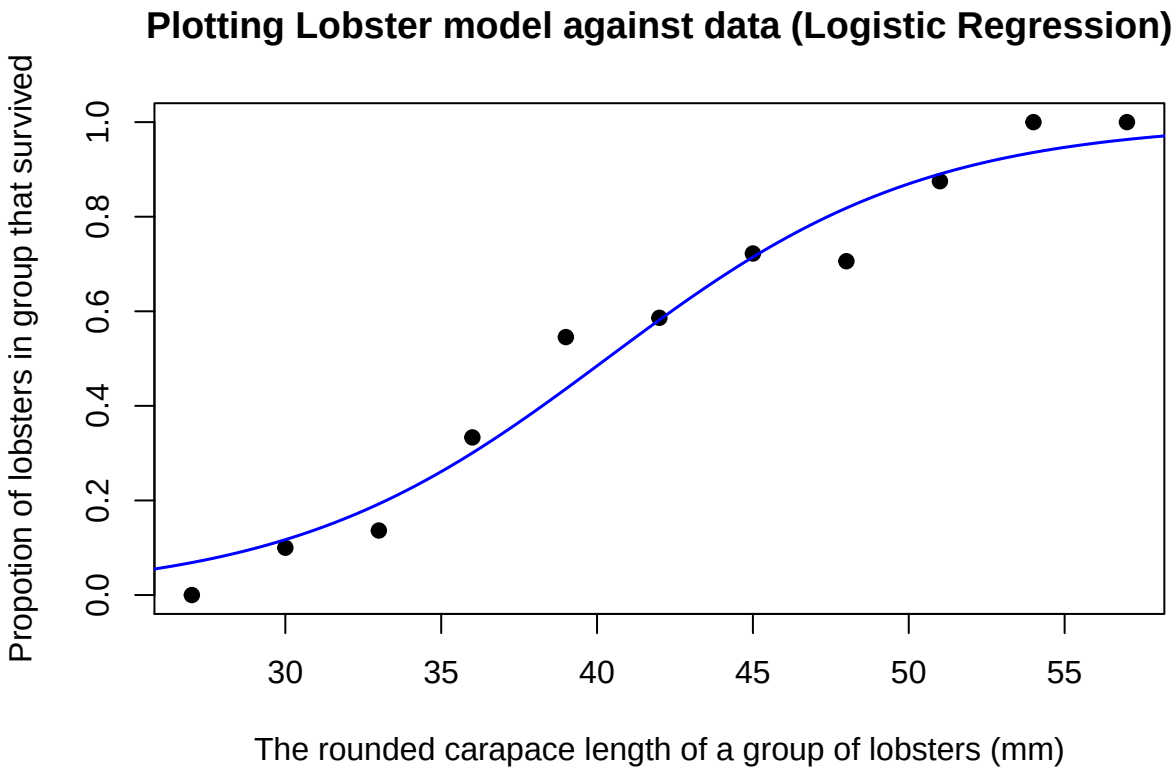
```
#Odds
exp(coef(lobster.fit2))
```

```
##  (Intercept)         size
## 0.0003722407 1.2163540760
```

We estimate that for every 1mm increase in the rounded carapace length of a lobster, the odds of survival are multiplied by 1.216 times.

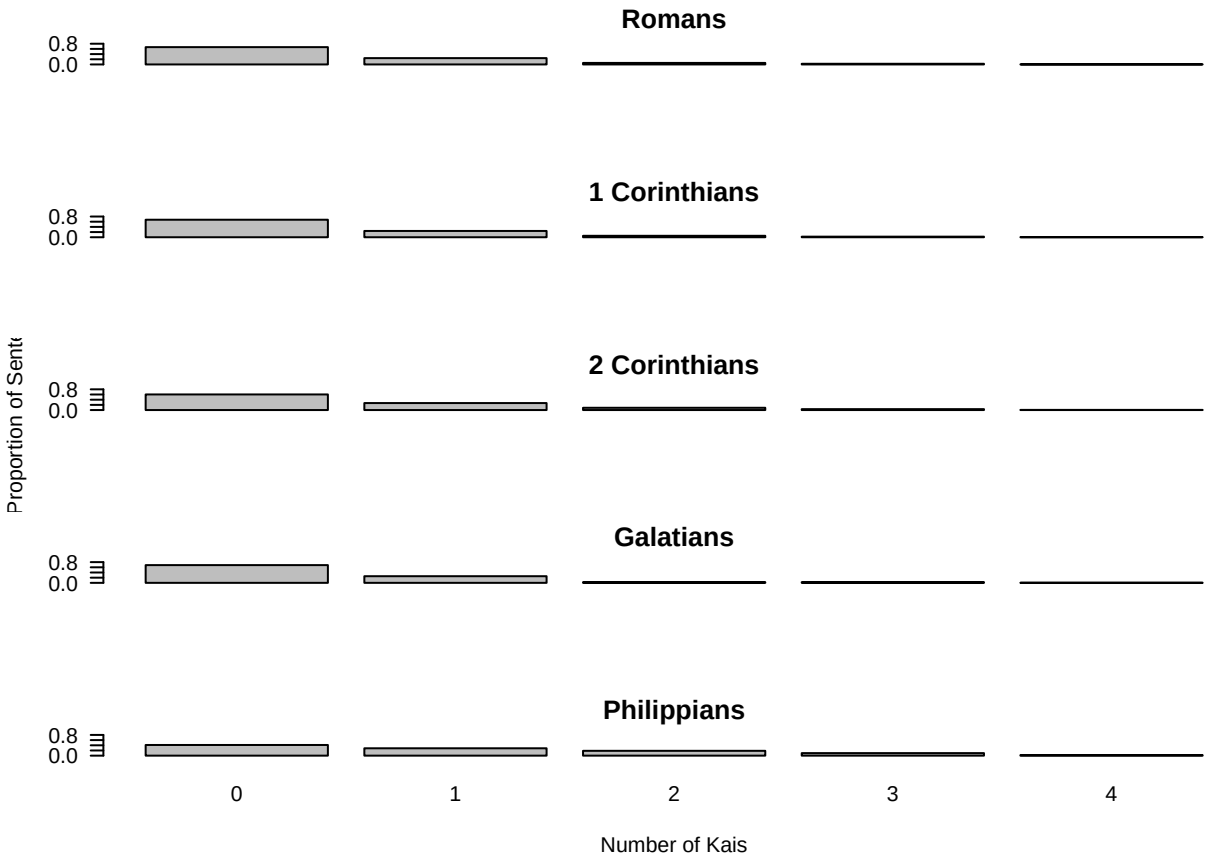
h) Recreate your plot

```
#define new data frame that contains predictor variable
newdata <- data.frame(size=seq(20, 60,len=500))

#use fitted model to predict values of proportion survived
newdata$proportion_survived <- predict(lobster.fit2, newdata, type="response")

#plot logistic regression curve
plot(proportion_survived ~ size, data = lobsterdf,
     main = "Plotting Lobster model against data (Logistic Regression)",
     xlab = 'The rounded carapace length of a group of lobsters (mm)',
     ylab = 'Propotion of lobsters in group that survived',
      pch = 19)
lines(proportion_survived ~ size, newdata, lwd=1.5, col = 'blue')
```



Question 2) a) Create this plot, and comment briefly.

```
## To make five plots stack on top of each other and to minimise margins between plots.
par(mfcol = c(5, 1), mar = c(4, 4, 2, 0.2), las = 1)
## A call to barplot() to create each individual plot.
barplot(prop.table(table(stPauldf$kais[stPauldf$work == "Romans"])),
main = "Romans", names.arg = rep("", 5), ylim = c(0, 0.8))
barplot(prop.table(table(stPauldf$kais[stPauldf$work == "Corinthians1"])),
main = "1 Corinthians", names.arg = rep("", 5), ylim = c(0, 0.8))
barplot(prop.table(table(stPauldf$kais[stPauldf$work == "Corinthians2"])),
main = "2 Corinthians", ylab = "Proportion of Sentences",
names.arg = rep("", 5), ylim = c(0, 0.8))
barplot(prop.table(table(stPauldf$kais[stPauldf$work == "Galatians"])),
main = "Galatians", names.arg = rep("", 5), ylim = c(0, 0.8))
barplot(prop.table(table(stPauldf$kais[stPauldf$work == "Philippians"])),
main = "Philippians", xlab = "Number of Kais", ylim = c(0, 0.8))
```



In each epistle kai is more often not used than used.

Philippians is the only epistle with clearly a different spread of kais. The proportion of sentences with 0 kais is lower than compared to other epistles. It also has proportionally the most sentences with 2 or 3 kais.

The other epistles have a similar spread. However Galatians and Romans have a proportionally higher number of sentences with 3 kais than 1 Corinthians and 2 Corinthians.

2 Corinthians has a proportionally higher number of sentences with 2 kais relative to Romans, Galatians and 1 Corinthians.

Based on the plots, I would assume that Philippians has a unique author. It is not clear if the other epistles have a unique author.

We will need to run a model.

b) Fit a model to compare how the average number of times the word "kai" appears in a sentence varies between the five works.

```
#We are looking at counts, so we will use a Poisson distribution

stPaul.fit1 <- glm(kais ~ work, family = 'poisson', data = stPauldf)
summary(stPaul.fit1)
```

```
##
## Call:
## glm(formula = kais ~ work, family = "poisson", data = stPauldf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4349  -0.9521  -0.9315   0.6993   3.3185
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.83514    0.06063 -13.774  < 2e-16 ***
## workCorinthians2  0.27944    0.09561   2.923  0.00347 **
## workGalatians    -0.05216    0.12909  -0.404  0.68617
## workPhilippians   0.86413    0.11489   7.521 5.43e-14 ***
## workRomans        0.04392    0.08656   0.507  0.61193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2059.1  on 1812  degrees of freedom
## Residual deviance: 2000.7  on 1808  degrees of freedom
## AIC: 3432.9
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(stPaul.fit1))
```

```
##      (Intercept) workCorinthians2    workGalatians  workPhilippians
##        0.4338118        1.3223884        0.9491782        2.3729455
##       workRomans
##        1.0448936
```

c) Fit a model to compare how the average number of times the word "kai" appears in a sentence varies between the five works.

$exp(\beta_0) = 0.4338118$.

If the epistle is '1 Corinthians', we estimate that the expected average number of kais per sentence to be multiplied by 0.4338.

The Romans coefficient is statistically insignificant. This means the expected number of kais per sentence for Romans is statistically equal to 1 Corinthians.

d) Conduct an appropriate hypothesis test and write a brief summary.

We can define H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

I.e Average number of kais per sentence are the same for all epistles, and therefore all works have the same author.

We will use Anova testing to test the null hypothesis

```
anova(stPaul.fit1, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: kais
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  1812     2059.1
## work  4   58.348      1808     2000.7 6.449e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Anova testing suggests some factor variables are significant
```

We reject the null (p-value approximately 0), there is strong evidence that there is at least one epistle with a different average number of kais per sentence. Suggesting there is more than one author.

e) For which works does there appear to be evidence of a difference from 1 Corinthians, in terms of the expected number of occurrences of "kai" per sentence?

Lets examine the summary again.

```
summary(stPaul.fit1)
```

```
##
## Call:
## glm(formula = kais ~ work, family = "poisson", data = stPauldf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4349  -0.9521  -0.9315   0.6993   3.3185
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.83514    0.06063 -13.774  < 2e-16 ***
## workCorinthians2  0.27944    0.09561   2.923  0.00347 **
## workGalatians    -0.05216    0.12909  -0.404  0.68617
## workPhilippians   0.86413    0.11489   7.521 5.43e-14 ***
```

```
## workRomans        0.04392    0.08656   0.507  0.61193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2059.1  on 1812  degrees of freedom
## Residual deviance: 2000.7  on 1808  degrees of freedom
## AIC: 3432.9
##
## Number of Fisher Scoring iterations: 6
```

Our base level is 1 Corinthians.

The P-Value for the Philippians epistle is very small (and clearly statistically significant). Therefore, it is highly likely that the Philippians epistle has a different number of average kais per sentence than 1 Corinthians, suggesting Philippians has a different author than 1 Corinthians.

The P-Value for 2 Corinthians is larger, but still statistically significant. This suggests there is some evidence that 2 Corinthians has a different number of average kais per sentence to 1 Corinthians and consequently a different author.

f) For which works does there appear to be evidence of a difference from Philippians, in terms of the number of the expected number of occurrences of "kai" per sentence?

We first change the baseline to Philippians, then reexamine a summary.

```
unique(stPauldf$work)
```

```
## [1] "Romans"       "Corinthians1" "Corinthians2" "Galatians"    "Philippians"
```

```
stPauldf$work <- factor(stPauldf$work, c('Philippians', "Romans",
                                         "Corinthians1", "Corinthians2",
                                         "Galatians"))

stPaul.fit2 <- glm(kais ~ work, family = 'poisson', stPauldf)
summary(stPaul.fit2)
```

```
##
## Call:
## glm(formula = kais ~ work, family = "poisson", data = stPauldf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4349  -0.9521  -0.9315   0.6993   3.3185
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.02899    0.09759   0.297    0.766
## workRomans       -0.82022    0.11550  -7.101 1.24e-12 ***
## workCorinthians1 -0.86413    0.11489  -7.521 5.43e-14 ***
## workCorinthians2 -0.58469    0.12243  -4.776 1.79e-06 ***
## workGalatians    -0.91629    0.15004  -6.107 1.01e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2059.1  on 1812  degrees of freedom
## Residual deviance: 2000.7  on 1808  degrees of freedom
## AIC: 3432.9
##
## Number of Fisher Scoring iterations: 6
```

We have set Philippians to our base, now all coefficients corresponding to different works are statistically significant. This suggests that average number of kais per sentence is statistically different to all other works.

I.e. Philippians has a unique author.

```
print('End Of Assignment 1')
```

```
## [1] "End Of Assignment 1"
```