Problem Set 8 Solution Attempt

Lars L. Ankile

April 30, 2020

Problem 1.

a) The model equation is $\log(\text{odds of failure}) = 15 - 0.2322 \cdot temp$, where the coefficients are gotten from the below code.

```
# Load the data
library(mcsm)
data("challenger")

# Get smallest temp in data
min(challenger$temp)
```

```
## [1] 53
```

```
# Fit the model
challenger_model <- glm(oring ~ temp, data = challenger,
                        family = binomial(link = "logit"))
summary(challenger_model)
```

```
## 
## Call:
## glm(formula = oring ~ temp, family = binomial(link = "logit"), 
##     data = challenger)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.0611  -0.7613  -0.3783   0.4524   2.2175  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  15.0429     7.3786   2.039   0.0415 *
## temp         -0.2322     0.1082  -2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
## 
## Number of Fisher Scoring iterations: 5
```

```
# Save coefficients for later
b0 <- summary(challenger_model)$coef[1, 1]
b1 <- summary(challenger_model)$coef[2, 1]
```

b) The smallest temperature in the data set is 53 degrees Fahrenheit, so no, the intercept does not have a meaningful interpretation.

c) The slope coefficient is -0.2322, which means that the log of the odds for failure decreases by negative 0.23 per degree increase in temperature. According to the model above, the p-value of the slope is 0.032, which is less than 0.05 and can be considered significant given significance level of 0.05.

d) The odds of O-ring failure is 0.3, which means that failure is less likely than failure not occurring.

```r
# Use r as a calculator
exp(b0 + 70 * b1)
```

```
## [1] 0.2986478
```

e) the odds ratio of the odds of O-ring failure with 60 F to 75 F is 32.5.

```r
# Use r as a calculator
exp(b1 * (60 - 75))
```

```
## [1] 32.53906
```

f) The probability of O-ring failure when it's 30.9 F during launch is very close to 1 (0.9996).

```r
# Use r as a calculator
odds <- exp(b0 + 30.9 * b1)

odds / (odds + 1)
```

```
## [1] 0.9996178
```

g) According to the model, there is a significant relationship between temperature and odds of failure. Furthermore, for 30.9 F the probability is almost 1, which suggest that the temperature was the reason for failure. However, the model isn't necessarily valid for temperatures lower than 53 F, so it's not that open and shut. Still, the evidence is compelling.

Problem 2.

a)

i. There are 54 females and 66 males in the population. The median age is 17.1 years, while the mean age is 17.71 years, so the age data is slightly right-skewed. The lowest and highest age is 12.09 and 23.85 years, respectively.

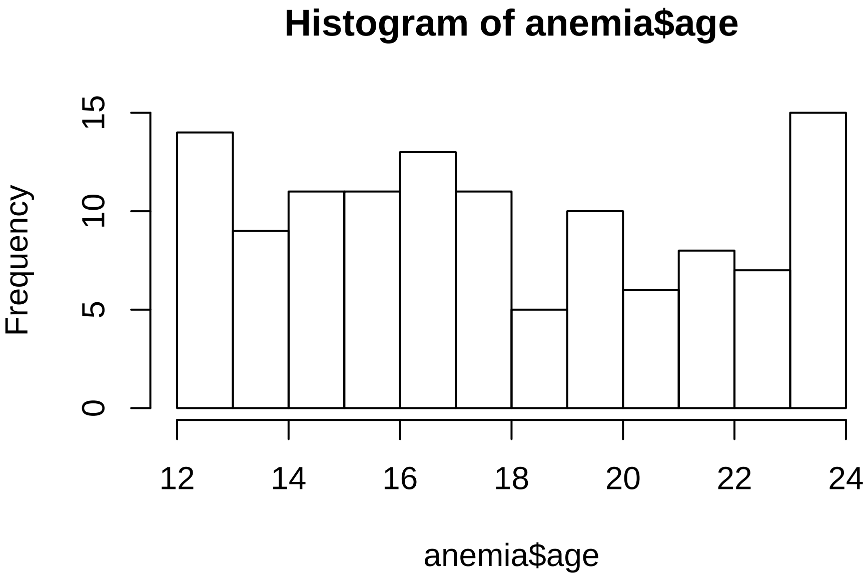
```
# Load the data
load('datasets/anemia.Rdata')
summary(anemia$sex)
```

```
## female   male
##     54     66
```

```
summary(anemia$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.09   14.62   17.10   17.71   21.00   23.85
```

```
hist(anemia$age, breaks = 10)
```



ii. 17.5% of the children in this data is considered substantially underweight.

```
underweight <- nrow(anemia[anemia$whz < qnorm(0.05), ])

underweight / nrow(anemia)
```

```
## [1] 0.175
```

iii. From the below calculation, we see that 63% of the children in the study was iron deficient.

```
num_deficient <- nrow(anemia[anemia$iron < 0, ])

num_deficient / nrow(anemia)
```

```
## [1] 0.6333333
```

iv. 67.5% of children in this study are anemic, compared to roughly 15% in the US, so the difference is huge!

```r
num_anem <- nrow(anemia[anemia$hb < 10.5, ])
num_anem / nrow(anemia)
```

```
## [1] 0.675
```

b) Below I've fitted a log(odds) model to the data for the binary anemia variable as the response and sex as the explanatory variable. The equation came out to be log(odds of anemia) = 0.5306 + 0.3751 * (child is male). This would suggest that the prevalence of anemia is higher with boys, but the coefficient was not significant at a $\alpha = 0.05$ significance level, so the relationship appears to be weak, if it exists.

```r
anemia$anemia <- anemia$hb < 10.5
anem_model1 <- glm(anemia ~ sex, data = anemia, family = binomial(link = "logit"))
summary(anem_model1)
```

```
## 
## Call:
## glm(formula = anemia ~ sex, family = binomial(link = "logit"), 
##     data = anemia)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.5781  -1.4094   0.8240   0.9619   0.9619  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept)   0.5306     0.2818   1.883   0.0597 .
## sexmale       0.3751     0.3916   0.958   0.3381  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 151.34  on 119  degrees of freedom
## Residual deviance: 150.42  on 118  degrees of freedom
## AIC: 154.42
## 
## Number of Fisher Scoring iterations: 4
```

```r
b0 <- anem_model1$coefficients[[1]]
b1 <- anem_model1$coefficients[[2]]
```

c) The odds of a female in the sample having anemia is 1.7, which is greater than 1, so it's more likely than not than a female is diagnosed with anemia.

```r
exp(b0)
```

```
## [1] 1.7
```

d) From the below model we see that the relationship between presence of anemia and age is log(odds of anemia) = 2.23 - 0.084 * (age of child). This means that the older a child is, the less likely it is to be anemic. Again, though, the relationship seems to be weak.

```r
anem_model2 <- glm(anemia ~ age, data = anemia, family = binomial(link = "logit"))
summary(anem_model2)
```

```
##
## Call:
## glm(formula = anemia ~ age, family = binomial(link = "logit"),
##     data = anemia)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7172  -1.3492   0.7888   0.8798   1.0807
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.22983    1.00314   2.223   0.0262 *
## age         -0.08377    0.05442  -1.539   0.1237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 151.34  on 119  degrees of freedom
## Residual deviance: 148.94  on 118  degrees of freedom
## AIC: 152.94
##
## Number of Fisher Scoring iterations: 4
```

e) From the below, it seems to be a difference between how age is associated per sex. The age coefficient is close to zero (-0.0053), while the coefficient for age:sexmale is a lot larger (-0.16). This suggests that the age effect for anemia (lower prevalence with higher age) is larger for boys than for girls.

```r
anem_model3 <- glm(anemia ~ age + sex + sex*age,
                   data = anemia, family = binomial(link = "logit"))
summary(anem_model3)
```

```
##
## Call:
## glm(formula = anemia ~ age + sex + sex * age, family = binomial(link = "logit"),
##     data = anemia)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.0078  -1.3959   0.7309   0.9577   1.2073
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.625542   1.381323   0.453    0.651
## age         -0.005299   0.075459  -0.070    0.944
## sexmale      3.254487   2.054421   1.584    0.113
## age:sexmale -0.160325   0.111053  -1.444    0.149
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 151.34  on 119  degrees of freedom
## Residual deviance: 146.02  on 116  degrees of freedom
## AIC: 154.02
##
## Number of Fisher Scoring iterations: 4
```

f)

i. the odds of anemia for the iron deficient children is 1.92, while it is 1.53 for the non deficient children, i.e. it's definitively higher for the iron dificient children.

```
anem_model4 <- glm(anemia ~ iron, data = anemia, family = binomial(link = "logit"))
b0 <- anem_model4$coefficients[[1]]
b1 <- anem_model4$coefficients[[2]]

# Odds of anemia for child with iron deficiency
exp(b0 + b1 * (-1.48))
```

```
## [1] 1.920628
```

```
# Odds of anemia for child without iron deficiency
exp(b0 + b1 * (0.18))
```

```
## [1] 1.527817
```

ii. We can see from the below model summary that the p-value of the slope is much smaller than 0.05 (> 0.002), which would suggest a significant relationship at significance level $\alpha = 0.05$.

```
summary(anem_model1)
```

```
##
## Call:
## glm(formula = anemia ~ sex, family = binomial(link = "logit"),
##     data = anemia)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5781  -1.4094   0.8240   0.9619   0.9619
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5306     0.2818   1.883   0.0597 .
## sexmale       0.3751     0.3916   0.958   0.3381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 151.34  on 119  degrees of freedom
## Residual deviance: 150.42  on 118  degrees of freedom
## AIC: 154.42
##
## Number of Fisher Scoring iterations: 4
```

iii. From the below model summary, we can see that after adjusting for the confounding factors sex, age, weight, family wealth, and recent diarrhea episodes, the relationship between anemia and iron levels still persists. Furthermore, it seems to have been strengthened, because the coefficient is larger in magnitude and the associated p-value is even smaller.

```r
anemia_model5 <- glm(
  anemia ~ iron + sex + age + whz + wealth + diarrhea,
  data = anemia,
  family = binomial(link = "logit")
)

summary(anemia_model5)
```

```
##
## Call:
## glm(formula = anemia ~ iron + sex + age + whz + wealth + diarrhea,
##     family = binomial(link = "logit"), data = anemia)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2220  -1.0230   0.5370   0.8216   1.5301
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.73836    1.23097   1.412 0.157893
## iron        -0.17602    0.05099  -3.452 0.000556 ***
## sexmale      0.40888    0.43995   0.929 0.352690
## age         -0.15722    0.06597  -2.383 0.017166 *
## whz         -0.38727    0.23191  -1.670 0.094928 .
## wealth       0.28728    0.15868   1.810 0.070226 .
## diarrheaYes  0.49961    0.56526   0.884 0.376775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 151.34  on 119  degrees of freedom
## Residual deviance: 128.68  on 113  degrees of freedom
## AIC: 142.68
##
## Number of Fisher Scoring iterations: 4
```

g)

i. After thorough analysis that takes into consideration many possible factors, there is strong evidence that increased level of iron levels decrease a child's odds of having anemia. Right now, roughly 65% of children in this district suffer from anemia, so the intervention is expected to have a huge impact on children's wellbeing in the district.

ii. Since we haven't done a controlled and randomized study, it's hard to conclude too definitively that introducing iron supplements will decrease the rate of anemia. It is possible that, for some other reason, children with low iron levels just happens to also have a high rate of anemia. However, the relationship is strong, so it seems reasonable to at least test the treatment and monitor the situation.

Problem 3.

a) After fitting the model to the different variables, I find that the coefficients that are significant at a $\alpha = 0.1$ level are wing length and total brightness. Total brightness is negatively correlated with nestling fate, while wing length is positively correlated. This makes sense as longer wings probably makes the bird more adept at flying away, while brighter colors makes it more visible to predators.

```r
# Load the data
load('datasets/rubythroats.Rdata')

# Fit the model
model <- glm(
  nestling.fate ~ carotenoid.chroma + bib.area + total.brightness
    + weight + wing.length + tarsus.length,
  data = rubythroats,
  family = binomial(link = "logit")
)

summary(model)
```

```
##
## Call:
## glm(formula = nestling.fate ~ carotenoid.chroma + bib.area +
##     total.brightness + weight + wing.length + tarsus.length,
##     family = binomial(link = "logit"), data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8550  -0.8988  -0.4402   1.0220   1.7890
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -24.057818  13.046850  -1.844  0.06519 .
## carotenoid.chroma  -4.774799   3.302978  -1.446  0.14829
## bib.area           -0.001272   0.002833  -0.449  0.65341
## total.brightness   -0.130880   0.043669  -2.997  0.00273 **
## weight             -0.358164   0.419582  -0.854  0.39332
## wing.length         0.521821   0.234826   2.222  0.02627 *
## tarsus.length       0.476708   0.484340   0.984  0.32500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.736  on 70  degrees of freedom
## Residual deviance: 79.655  on 64  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 93.655
##
## Number of Fisher Scoring iterations: 4
```

b)

i. According to the below model, there's a significant relationship between if the bird laid a second clutch and nestling fate.

```r
# Fit a model
model <- glm(
  second.clutch ~ nestling.fate,
  data = rubythroats,
  family = binomial(link = "logit")
)

summary(model)
```

```
##
## Call:
## glm(formula = second.clutch ~ nestling.fate, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0302  -0.8262  -0.2144  -0.2144   2.7511
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3.761      1.012  -3.718 0.000201 ***
## nestling.fateFledged     3.405      1.070   3.182 0.001462 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 76.370  on 77  degrees of freedom
## Residual deviance: 55.615  on 76  degrees of freedom
##   (7 observations deleted due to missingness)
## AIC: 59.615
##
## Number of Fisher Scoring iterations: 6
```

ii. In this model, the two predictors that are the most statistically significant for whether the bird lays a second clutch that year is nestling fate and total brightness.

```r
# Fit a model
model <- glm(
  second.clutch ~ nestling.fate + carotenoid.chroma
    + bib.area + total.brightness,
  data = rubythroats,
  family = binomial(link = "logit")
)

summary(model)
```

```
##
## Call:
## glm(formula = second.clutch ~ nestling.fate + carotenoid.chroma +
##     bib.area + total.brightness, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.87779  -0.25235  -0.09075  -0.02136   2.10799
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -21.397205   7.593148  -2.818  0.00483 **
## nestling.fateFledged   5.527158   1.740419   3.176  0.00149 **
## carotenoid.chroma     11.585799   6.085778   1.904  0.05694 .
## bib.area               0.007019   0.003801   1.847  0.06480 .
## total.brightness       0.149085   0.068533   2.175  0.02960 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.503  on 75  degrees of freedom
## Residual deviance: 39.076  on 71  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 49.076
##
## Number of Fisher Scoring iterations: 7
```

iii. From the below model we see that the birds are more likely to have a second clutch both if the first clutch survived and if they have brighter colors. However, if the first clutch survived, the brightness is suddenly negatively correlated. This could mean that brighter birds are more likely to find a mate, but their offspring is more likely to be predated.

```r
# Fit a model
model <- glm(
  second.clutch ~ nestling.fate + total.brightness
    + nestling.fate * total.brightness,
  data = rubythroats,
  family = binomial(link = "logit")
)
```

```
summary(model)
```

```
##
## Call:
## glm(formula = second.clutch ~ nestling.fate + total.brightness +
##     nestling.fate * total.brightness, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.17919  -0.84586  -0.07953  -0.02768   2.14776
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -10.8693     6.7704  -1.605    0.108
## nestling.fateFledged                    9.7318     6.8536   1.420    0.156
## total.brightness                        0.2033     0.1594   1.276    0.202
## nestling.fateFledged:total.brightness  -0.1671     0.1659  -1.007    0.314
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.940  on 76  degrees of freedom
## Residual deviance: 52.024  on 73  degrees of freedom
##   (8 observations deleted due to missingness)
## AIC: 60.024
##
## Number of Fisher Scoring iterations: 8
```

c)

i. For the below model, the variables weight, wing length, first clutch size, and having a second clutch were positively correlated with survival.

```
# Fit a model
model1 <- glm(
  survival ~ carotenoid.chroma + bib.area + total.brightness
    + weight + wing.length + tarsus.length + first.clutch.size
    + second.clutch,
  data = rubythroats,
  family = binomial(link = "logit")
)

summary(model1)
```

```
##
## Call:
## glm(formula = survival ~ carotenoid.chroma + bib.area + total.brightness +
##     weight + wing.length + tarsus.length + first.clutch.size +
##     second.clutch, family = binomial(link = "logit"), data = rubythroats)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9501  -0.5706  -0.2036   0.4403   2.5038
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -24.322666  22.759433  -1.069   0.2852
## carotenoid.chroma -18.517610   6.804872  -2.721   0.0065 **
## bib.area           -0.002710   0.004754  -0.570   0.5686
## total.brightness   -0.108595   0.063858  -1.701   0.0890 .
## weight              0.605506   0.648570   0.934   0.3505
## wing.length         0.842892   0.430007   1.960   0.0500 *
## tarsus.length      -0.886169   0.837672  -1.058   0.2901
## first.clutch.size   2.620596   1.132561   2.314   0.0207 *
## second.clutchYes    1.664275   1.075925   1.547   0.1219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.438  on 48  degrees of freedom
## Residual deviance: 34.764  on 40  degrees of freedom
##   (36 observations deleted due to missingness)
## AIC: 52.764
##
## Number of Fisher Scoring iterations: 6
```

ii. This second model turned out to have a higher AIC-level (64.389 to 52.764), so it’s less parsimonious, and we can regard the first model as the better one in this case.

```
# Fit a model
model2 <- glm(
  survival ~ carotenoid.chroma + total.brightness
    + wing.length + first.clutch.size,
  data = rubythroats,
  family = binomial(link = "logit")
)

summary(model2)
```

```
##
## Call:
## glm(formula = survival ~ carotenoid.chroma + total.brightness +
##     wing.length + first.clutch.size, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0203  -0.7386  -0.4338   0.8047   2.4505
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -20.15330   10.45426  -1.928   0.0539 .
## carotenoid.chroma  -9.49612    3.89478  -2.438   0.0148 *
## total.brightness   -0.07515    0.04582  -1.640   0.1010
## wing.length         0.46054    0.22795   2.020   0.0433 *
## first.clutch.size   1.56415    0.70098   2.231   0.0257 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 72.997  on 55  degrees of freedom
## Residual deviance: 54.389  on 51  degrees of freedom
##   (29 observations deleted due to missingness)
## AIC: 64.389
##
## Number of Fisher Scoring iterations: 5
```

iii. Since I don't know what normal physical attributes are for these birds, but they are identical, I set all other variables to 0. I got that the bird that had 5 eggs the first time had odds of survival of $7 \cdot 10^{-5}$, while the bird that only had 3 had odds of survival of $3.7 \cdot 10^{-7}$, i.e. much lower, even though it's very low for both birds.

```r
coef <- model1$coefficients
first_num <- coef[['first.clutch.size']]
second_clutch <- coef[['second.clutchYes']]
b0 <- coef[['(Intercept)']]

# Odds for bird that laid 5 eggs
exp(b0 + second_clutch + 5 * first_num)
```

```
## [1] 7.081697e-05
```

```r
# Odds for bird that laid 3 eggs
exp(b0 + second_clutch + 3 * first_num)
```

```
## [1] 3.749012e-07
```

iv Both of these birds have significantly better odds of survival than the two birds, at odds of survival of 0.261 and 0.199 for bird A and B, respectively. Bird A also has a higher odds of survival than bird B.

```r
bib <- coef[['bib.area']]
brightness <- coef[['total.brightness']]
carotenoid <- coef[['carotenoid.chroma']]
tarsus <- coef[['tarsus.length']]
wing <- coef[['wing.length']]
weight <- coef[['weight']]
```

```
# Female A
exp(b0 + 350 * bib + 35 * brightness + 0.9 * carotenoid
    + 19.5 * tarsus + 51 * wing + 10.8 * weight
    + 4 * first_num + second_clutch)
```

```
## [1] 0.2606475
```

```
# Female B
exp(b0 + 300 * bib + 20 * brightness + 0.85 * carotenoid
    + 19.0 * tarsus + 50 * wing + 10.9 * weight
    + 3 * first_num + second_clutch)
```

```
## [1] 0.1990654
```

d) In the last analysis we see, slightly counter-intuitively, that the bird with more ornamentation (bird A) also has the higher odds of survival. This is counter-intuitive since carotenoid chroma, overall brightness, bib area, and tarsus length all are negatively correlated with survival in the model. What seems to be working in bird A's favor is that it had more eggs in the first clutch. The ability to have many eggs in the first clutch is probably an indication of fitness. There might also be some effect that males are probably more attracted to more ornamented females, so the better ornamented females attract better males and will in turn have children with "better" genes.

Problem 4.

Written Report

Problem 4 Appendix

```
#load the data
```

Problem 5.

Written Report

Problem 5 Appendix.

```
#load the data
```