mjon238_assignment2

mjon238

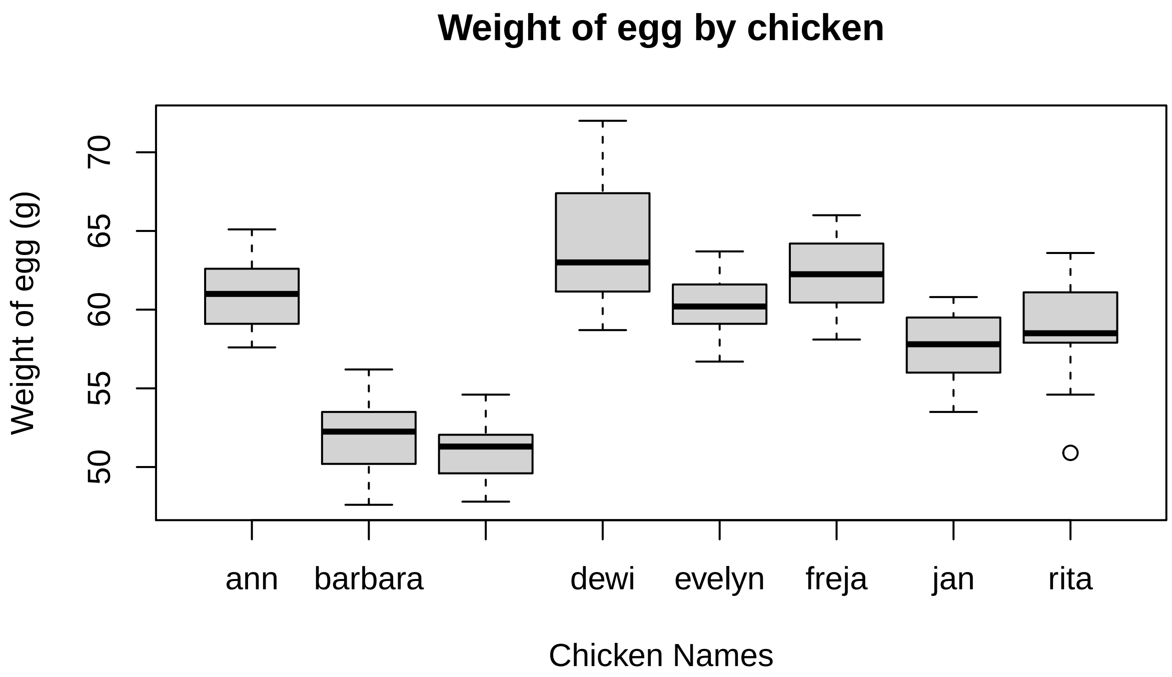
13/01/2022

Question 1)

```
#Creating a new variable called rita and adding it to the dataframe
rita <- ifelse(eggsdf$name == 'rita', 1, 0)
eggsdf <- cbind(rita, eggsdf)
```

a) Creating the box plots

```
boxplot(weight ~ name,
        data = eggsdf,
        main="Weight of egg by chicken",
        xlab="Chicken Names",
        ylab="Weight of egg (g)")
```



Rita's egg has most similar weight to Jan's eggs. It also has slightly similar weight to both Evelyn and Ann's, albeit theirs are slightly heavier. Dewi and Freja have heavier eggs than Rita. Barabara and Brook have much lighter eggs than Rita.

b) Weight Analysis

I would assume weight has no effect on whether the probability the egg is Rita's or not. This is because an increasing weight would more likely be Dewi or Frejas, and a decreasing weight would more likely be Barbara or Brook. A non-linear relationship could exist though.

I would assume weight and probability an egg is Rita's is unrelated.

c) Model Fitting

To determine whether this model is a good one we will first examine p-values, then compare AIC with the null model and compare deviance with the null model

```
fitWeight <- glm(rita ~ weight, family = 'binomial', data = eggsdf)
summary(fitWeight)
```

```
##
## Call:
## glm(formula = rita ~ weight, family = "binomial", data = eggsdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7107  -0.5669  -0.5277  -0.4612   2.1495
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.51813    2.79928  -1.614    0.107
## weight       0.04543    0.04748   0.957    0.339
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.55  on 156  degrees of freedom
## Residual deviance: 122.62  on 155  degrees of freedom
## AIC: 126.62
##
## Number of Fisher Scoring iterations: 4
```

The corresponding p-value for the weight coefficient is quite large and given this is the only variable this suggests the model does a bad job of determining odds which eggs are Ritas.

We will now fit the null model and make comparisons.

```
fitNull <- glm(rita ~ 1, family = 'binomial', data = eggsdf)

AIC(fitNull, fitWeight)
```

```
##           df      AIC
## fitNull    1 125.5491
## fitWeight  2 126.6171
```

```
AIC(fitWeight) - AIC(fitNull)
```

```
## [1] 1.068022
```

AIC of the null model is approximately the same as the AIC of the fitted model (difference of less than 2). This tells us both models are similarly supported by the data and the fitted model is NOT better than the null. This reinforces the idea the model is inappropriate.

I will run a deviance comparison to see if the null model (which is a submodel) is more appropriate than the fitted model.

We have a null hypothesis (H0) that the submodel is appropriate.

```
anova(fitNull, fitWeight, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: rita ~ 1
## Model 2: rita ~ weight
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       156     123.55
## 2       155     122.62  1  0.93198   0.3343
```

Given the high p-value, the H0 is not rejected and therefore the null model is appropriate. This suggests that weight on its own is an unnecessary variable and we have an inappropriate model.

Indicators such as P-Values, AIC and deviance all agree that the weight model is inappropriate.

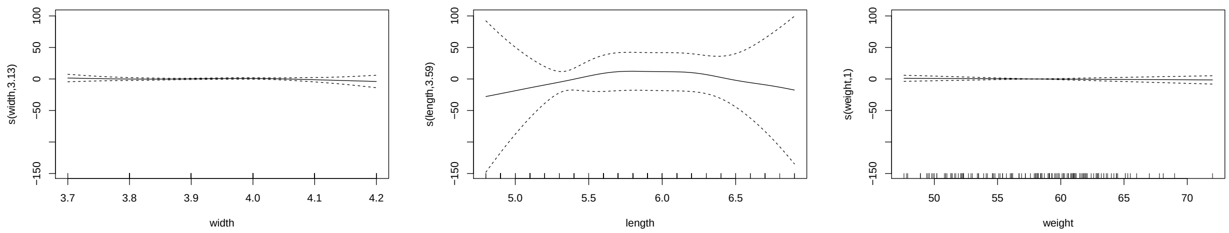
Question 2)

a) Model Searching

First thing I am going to do is look to see if there are non-linear relationships between my numeric variables. I will do this by fitting GAM plots.

```
gam.fit2 <- gam(rita ~ s(width, k=6) + s(length) + s(weight),
                data = eggsdf,
                family = 'binomial')

plot(gam.fit2)
```



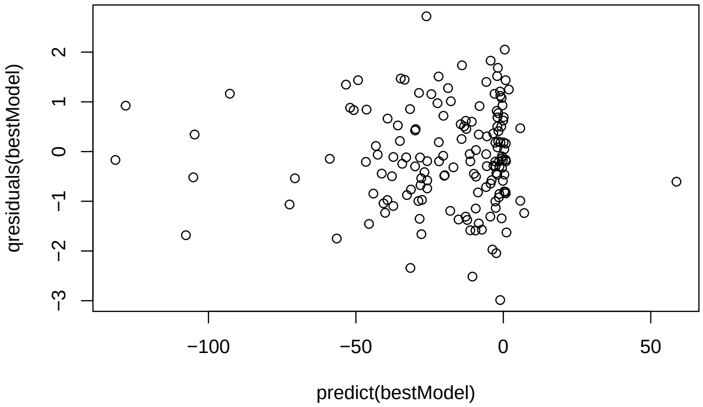
There are no non-linear relationships evident as we can draw a straight line through all the dotted lines.

Next I fitted a very complicated model, not with the intention of using it as my final model, but dredging it. Originally, dredging failed, so I cut down the number of interactions in the model using the following formula $rita = weight * width * length + colour + calcium.spots$. I then dredged again, with success. But my AIC was still relatively high. I then decided to do backwards stepwise selection instead.

This led to a lower AIC and a best model with approximately 87 AIC. I have selected this as my best model and will check the randomized quantile residuals. (Note I've hidden the stepwise selection code as it was too long.)

```
complexModel <- glm(rita ~ weight * width * length * colour * calcium.spots,
                    data = eggsdf,
                    family = 'binomial')
```

```
plot(predict(bestModel), qresiduals(bestModel))
```



The quantile residuals plot looks roughly appropriate. In general there is band around the mean with constant variance. However, there may be some overdispersion. So I will run a beta-binomial regression and check the rho value.

```
model2 <- vglm(bestModel$call$formula,
               family = 'betabinomial',
               eggsdf)
plogis(coef(model2)[2])
```

```
## (Intercept):2
##    0.01052996
```

The rho value is approximately 0.01, this indicates there is no overdispersion and my model is suitable.

A summary of the model is below

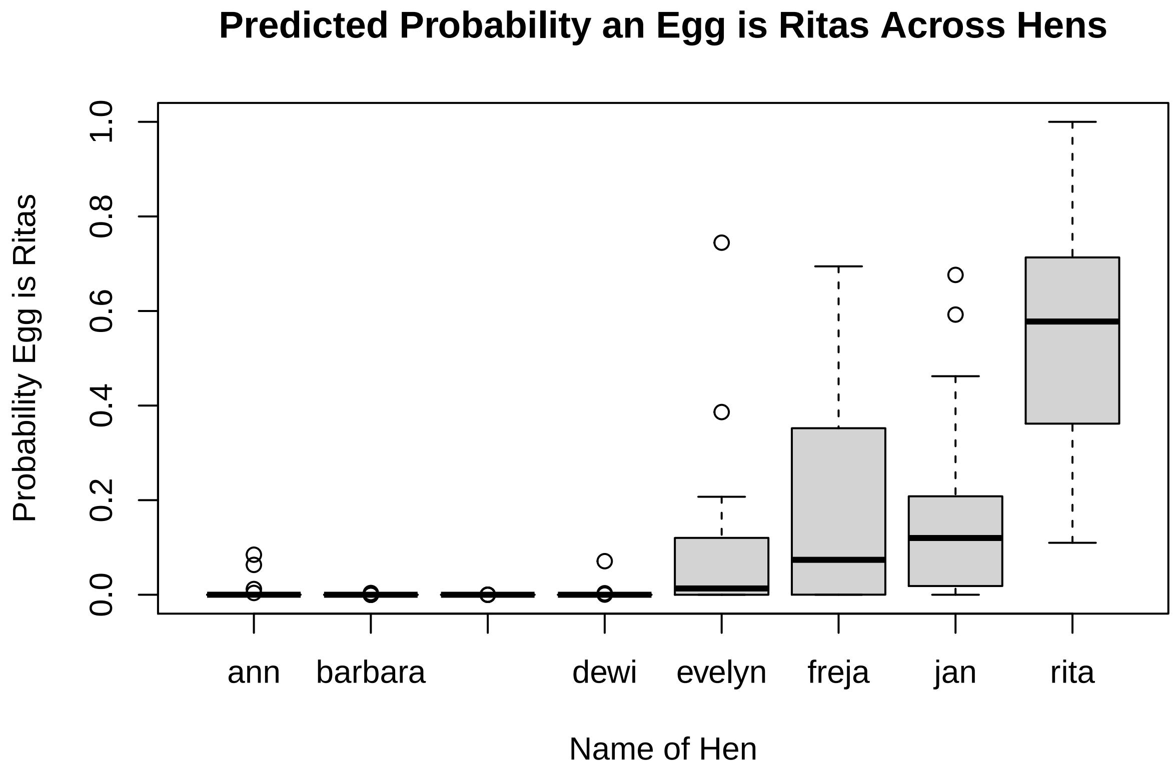
```
summary(bestModel)
```

```
##
## Call:
## glm(formula = rita ~ weight + width + length + colour + calcium.spots +
##     weight:width + weight:length + width:length + weight:colour +
##     length:colour + width:calcium.spots + length:calcium.spots +
##     weight:width:length, family = "binomial", data = eggsdf)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.65195  -0.15571  -0.00003   0.00000   2.10181
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -74499.225  36867.143  -2.021   0.0433 *
## weight                      1302.813    633.481   2.057   0.0397 *
## width                      18544.265   9266.235   2.001   0.0454 *
## length                     12752.243   6295.355   2.026   0.0428 *
## colourlight                  128.009     94.530   1.354   0.1757
## colourmedium                 157.248     92.921   1.692   0.0906 .
## calcium.spotsyes            -777.908   1038.509  -0.749   0.4538
## weight:width                -326.834    159.712  -2.046   0.0407 *
## weight:length               -223.069    108.172  -2.062   0.0392 *
## width:length               -3155.023   1579.648  -1.997   0.0458 *
## weight:colourlight             6.863      4.890   1.403   0.1605
## weight:colourmedium            6.457      4.864   1.328   0.1843
## length:colourlight           -93.630     51.505  -1.818   0.0691 .
## length:colourmedium          -94.403     50.021  -1.887   0.0591 .
## width:calcium.spotsyes       104.282    194.791   0.535   0.5924
## length:calcium.spotsyes       61.148     98.376   0.622   0.5342
## weight:width:length           55.645     27.227   2.044   0.0410 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.549  on 156  degrees of freedom
## Residual deviance:  52.723  on 140  degrees of freedom
## AIC: 86.723
##
## Number of Fisher Scoring iterations: 12
```

b) Predictions

```r
predictedValues <- predict(bestModel, type = "response")
eggsPredicted <- cbind(predictedValues, eggsdf)

boxplot(predictedValues ~ name, eggsPredicted,
        ylab = "Probability Egg is Ritas",
        xlab = "Name of Hen",
        main = "Predicted Probability an Egg is Ritas Across Hens")
```



The model does a very good job of distinguishing Ann, Barbara, Brook and Dewi's from Rita's.

It does a slightly worse job of distinguishing between Evelyn, Freja and Jans eggs with Ritas. But still does a good job overall. The average probability that that the model gives these hens eggs to be Ritas is all less than 20%. However, there are some concerning outliers, all visible in the boxplot.

Question 3)

a) Prediction with New Data

```
decisionProb <- predict(bestModel, newdata = eggsNewdf, type = 'response')

decision.df <- data.frame(decisionProb, eggsNewdf)%>%
  format(scientific = F)%>%
  mutate(decision = rank(desc(decisionProb)))%>%
  mutate(decision = replace(decision, decision>16, NA))%>%
  select(-decisionProb)

decision.df%>%na.omit%>%arrange(decision)
```

```
##    width length weight colour calcium.spots decision
## 43   3.8    6.1   58.4   dark            no        1
## 15   3.9    5.9   57.7 medium            no        2
## 32   4.1    5.8   62.9  light           yes        3
## 23   3.9    5.7   60.0 medium           no        4
## 38   3.9    5.8   60.0  light           no        5
## 13   4.0    5.8   61.0  light           no        6
## 20   3.9    5.7   57.1  light           no        7
## 37   4.0    5.7   60.2 medium           no        8
## 27   4.0    5.7   59.3  light           no        9
## 12   3.9    5.9   60.0  light           no       10
## 47   3.9    6.2   62.1  light          yes       11
## 30   3.9    5.9   60.8 medium           no       12
## 22   4.0    5.8   63.1  light           no       13
## 1    3.8    5.7   54.9 medium           no       14
## 29   3.9    5.9   61.5 medium           no       15
## 5    3.9    6.0   61.4  light           no       16
```

```
write.csv(decision.df, file = "mjon238.csv", row.names = FALSE)
```