Effects on the Probability of Having Brown Fat

Viacheslav Petruniak, 1002881045, (Graphs, Modeling, formatting, Background and Significance)
Hanli Fu, 1004717586, (Graphs, Modeling, formatting, Exploratory data analysis)
Yihan Chen, 1004745993, (Graphs, Modeling, formatting, Model)
Leo Jia, 1005294398, (Graphs, Modeling, formatting, Exploratory data analysis)

2022-04-09

Background and Significance

Our Goals:

- Background to understand the goal and content of the report
- Clear explanation for why this work is important and relevant
- Present the goal of the study

Introduction

The data set for this report is Brown Fat data set. Originally it came in .xml format but it also was converted to .csv

Brown fat is the type of fat that allows humans to survive in cold environments. It is said that it can be detected in newborns or after exposure to low temperatures. (BrownFat, 1) The data also contains a large cohort of cancer patients which may lead to some interesting inquiries if cancer affects the presence of brown fat.

The question of interest is how certain conditions affect the presence of brown fat in the human body. The conditions include multiple types of cancer so it can be potentially used to warn about predisposition to cancer in a patient.

Generalized linear model, goodness of fit test based on deviance, step functions and likelihood ratio test were used to draw conclusions about the data and what affects the presence of brown fat which includes multiple cancer types.

Exploratory data analysis

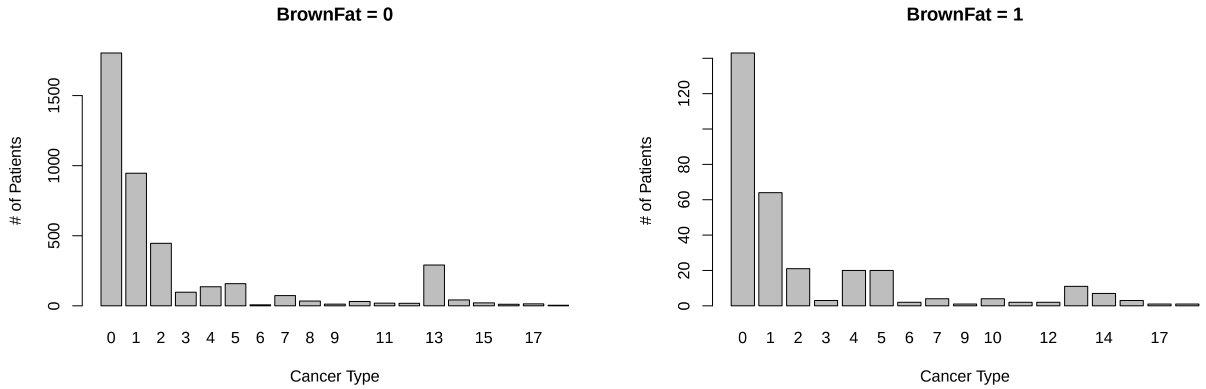
For exploratory data analysis the number of cases with brown fat and the volume of brown fat are chosen as the response variable. The explanatory variables are one of the other 20 variables included in the .xml. These include the following:

- Sex: sex of the patient (Female=1, Male=2).
- Diabetes: (No=0, Yes=1).
- Age: Age of the patient in years.
- Day: Day of the year.
- Month: Month of the exam.
- Ext_Temp: External Temperature.

- 2D_Temp: Average temperature of last 2 days.
- 3D_Temp: Average temperature of last 3 days.
- 7D_Temp: Average temperature of last 7 days.
- 1M_Temp: Average temperature of last month.
- Season: Spring=1, Summer=2, Autumn=3, Winter=4.
- Duration_Sunshine: Sunshine duration.
- Weight: in Kgs
- Size: in cms.
- BMI: Body Mass index.
- Glycemia.
- Lean Body Weight.
- Cancer_Status: (No=0, Yes=1).
- Cancer_Type: (No=0, lung=1, digestive=2, Oto-Rhino-Laryngology=3, breast=4, gynaecological (female)=5, genital (male)=6, urothelial=7, kidney=8, brain=9, skin=10, thyroid=11, prostate=12, non-Hodgkin lymphoma=13, Hodgkin=14, Kaposi=15, Myeloma=16, Leukemia=17, other=18).
- TSH

Firstly, it is good to start by getting a visualization of the presence of brown fat in different cancer types. Originally, ggplot2 was used to display the data, however, the graphs were hard to understand and sometimes meaningless. Because of this it was decided to use a more simple bar graph, which started to make more sense. A filter was applied on BrownFat to generate more graphs which presented the data better.

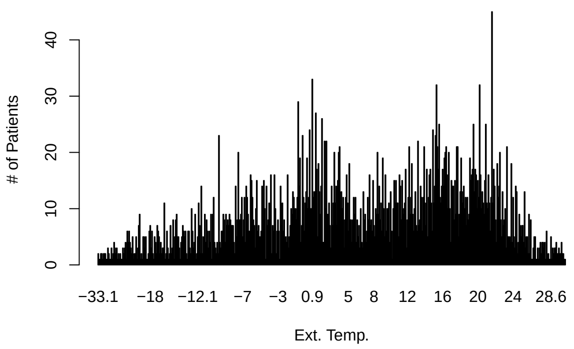
Using the filter and simple bar graph these two graphs were generated. It is the comparison of cancer type and the presence of brown fat.



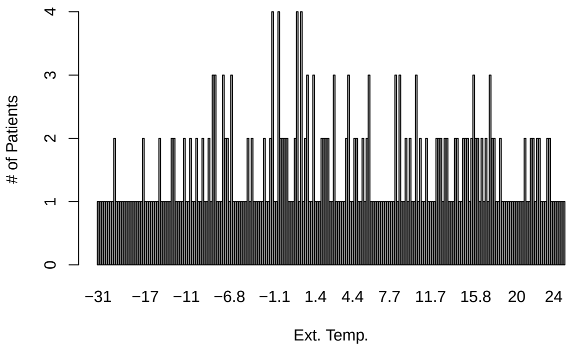
From the graph, it can be seen that cancer type 13 (non-Hodgkin lymphoma) may be significant to the presence of brown fat.

Later on in the report we concluded that external temperature, diabetes, weight, sex, age were the most significant terms using GLMs. We can generate bar graphs for these variables.

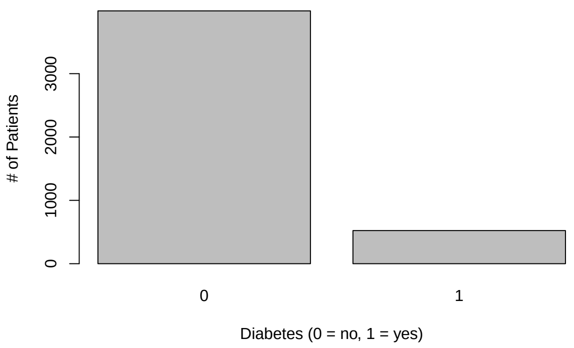
BrownFat = 0 for Ext. Temp.



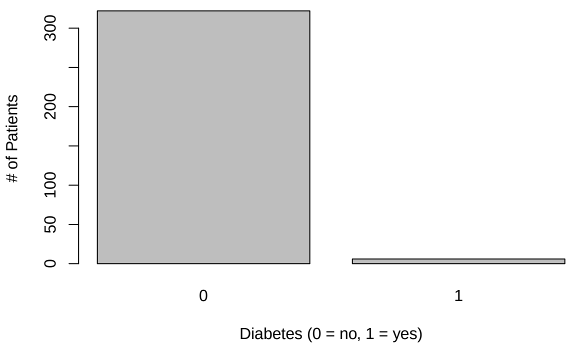
BrownFat = 1 for Ext. Temp.



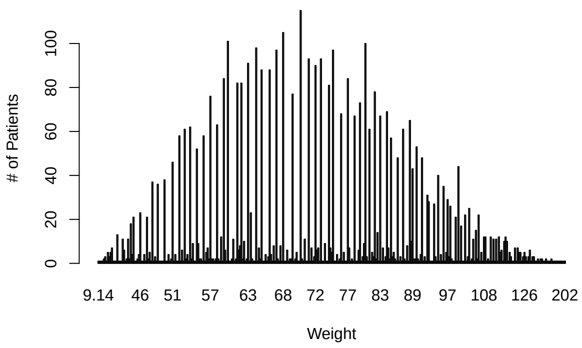
BrownFat = 0 for Diabetes



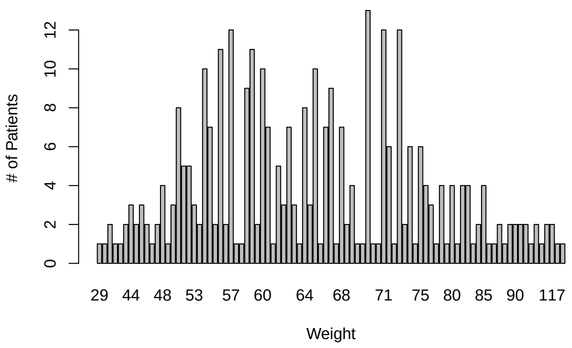
BrownFat = 1 for Diabetes

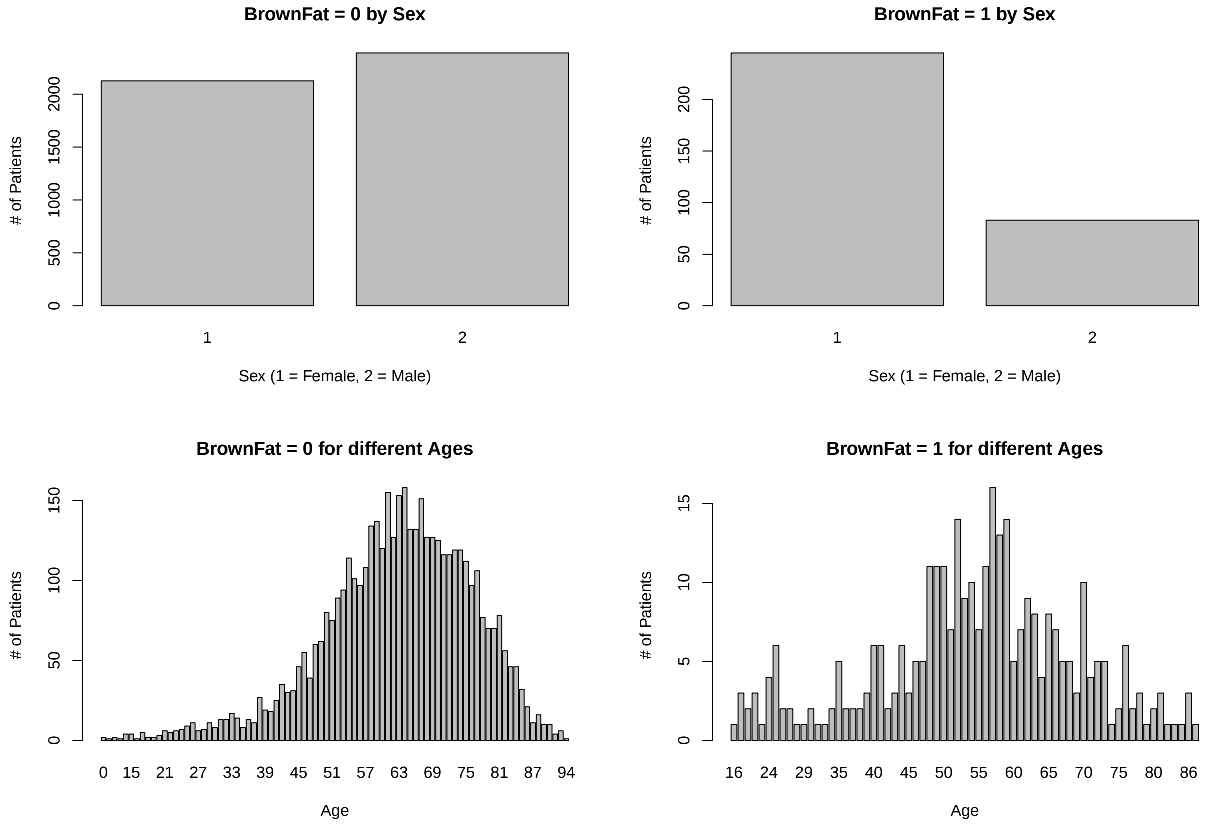


BrownFat = 0 for Weight



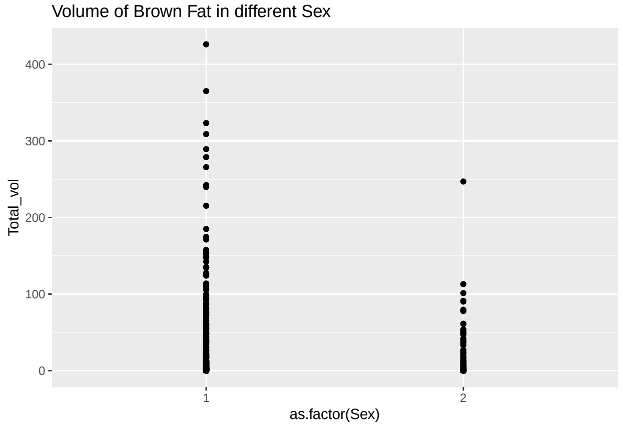
BrownFat = 1 for Weight

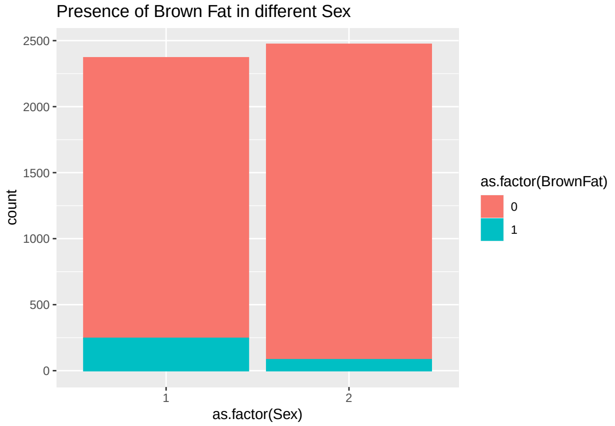




From these graphs we can loosely conclude a few things. External Temperature: We can see that those who have Brown Fat tend to have lower external temperatures. We can see that those without Brown Fat generally have higher external temperatures. Diabetes: There is not much we can confidently conclude but it does seem like those with Brown Fat have a lower amount of people with Diabetes. Weight: There is not much we can conclude here. Sex: It seems that if you have Brown Fat it is more likely that you are a male. Age: Brown Fat seems to be more dominant in those ages 48 to around 60.

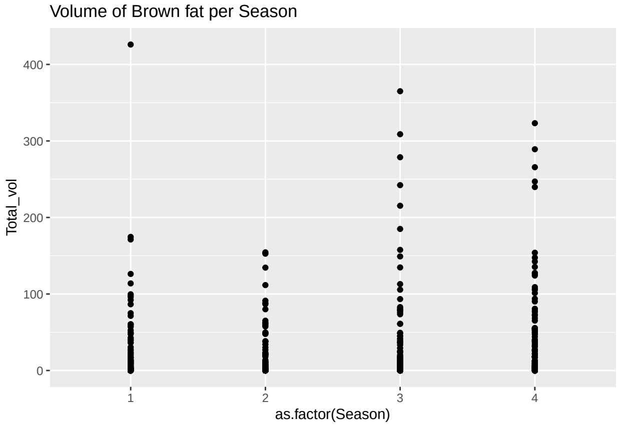
For other categorical variables, two kinds of graphs are used to make it easier to see if they are significant to the presence of brown fat. Here are two examples, respectively, the variable sex and the variable season. The first column of the graphs are point graphs with total volume of brown fat as y, and the second column, a bar graph to show whether the variable significantly affects the presence of brown fat.

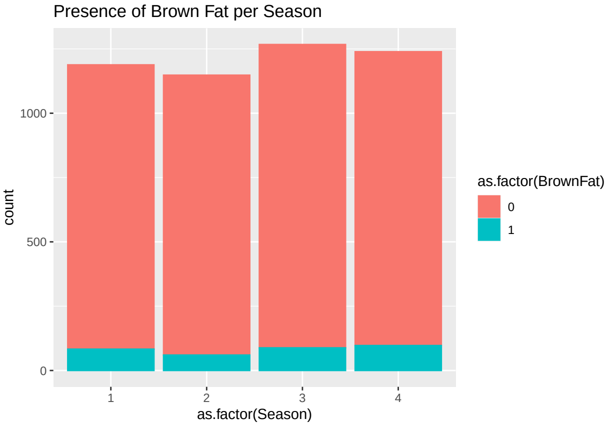




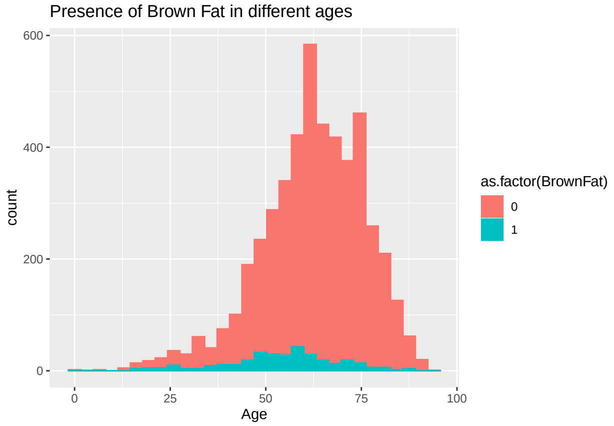
We can see the distribution of people who have BrownFat in the first graph, which uses the Total_vol as y.

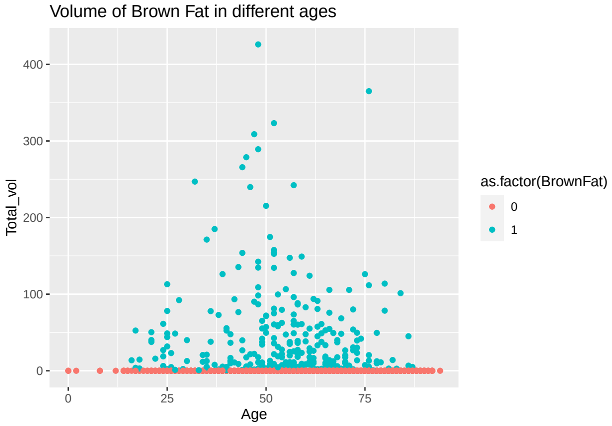
As it shown on the first graph, the total volume of brown fat that females have are likely to be higher than those males have. While in the second graph, we have the comparison of count in one bar graph, which shows that we have more females who have brown fat than males when the total amount of females is less then the total amount of males. Combining these two graphs together, we know that sex could have a significant effect on the presence of brown fat.



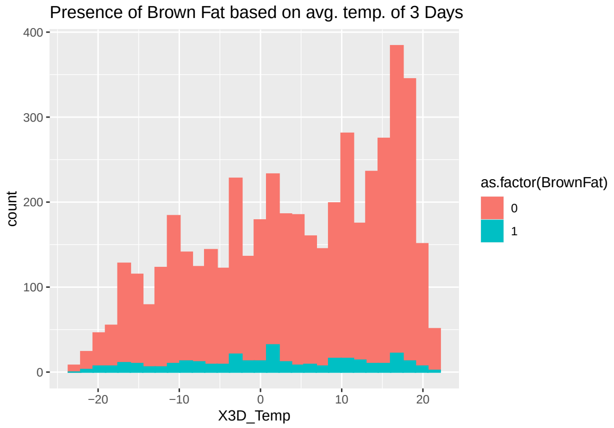


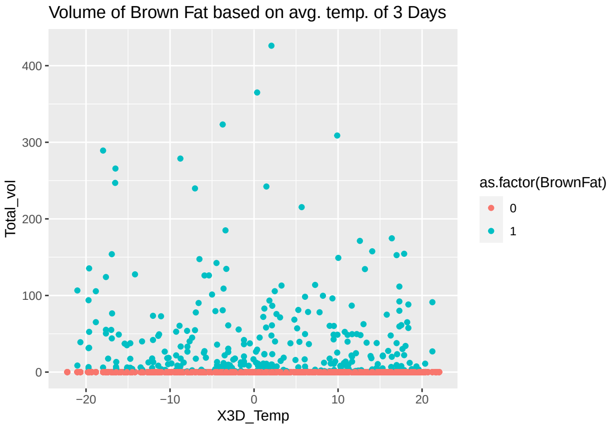
Next, the numerical variables. The first column is the histogram of variables. It can be seen, the distribution of people who do not have brown fat is more obvious than those who have brown fat. So a point graph was chosen to give a different view for people who have brown fat.





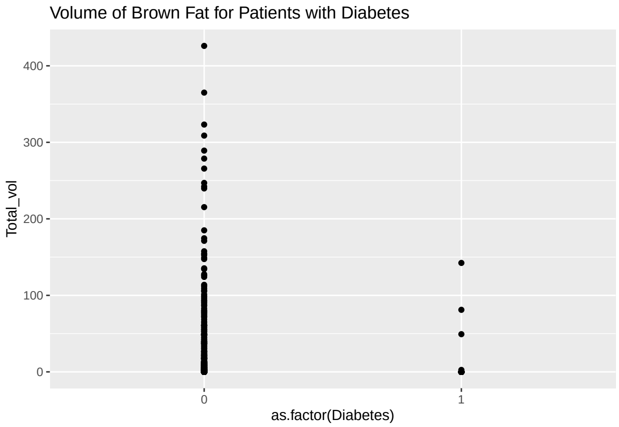
We can see the distribution of people who do not have BrownFat in the first graph, which is a histogram of age with respect to the presence of brown fat. As it shown on the first graph, the distribution of people who do not have brown fat is very clear – a right-skewed distribution, while the distribution of the people who have brown fat is not easy to see. So in the second graph, we choose total volume of brown fat as y to generate a point graph, which we can see the distribution of them in x scale and get a rough idea of the distribution of total volume. In this case, we have most of the points centered around the midpoint of age, and less as to the both sides of the midpoint, which means the distribution of people who have brown fat is more likely to be normal. Combining these two graphs together, we know that age have different distributions with respect to the presence of brown fat, so it could have a significant effect on the presence of brown fat.

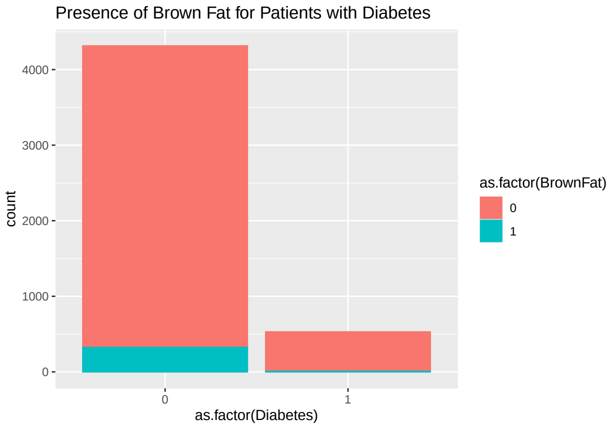




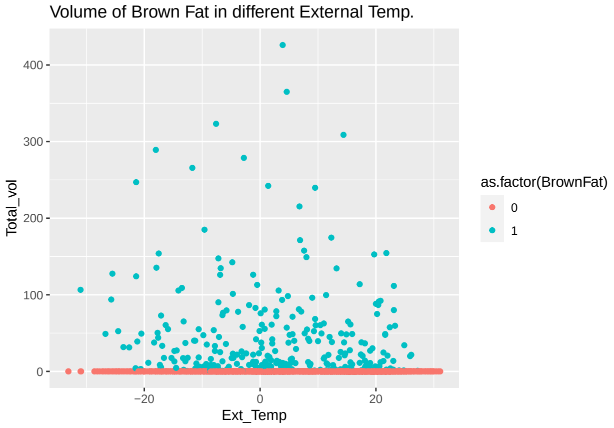
Additional examples

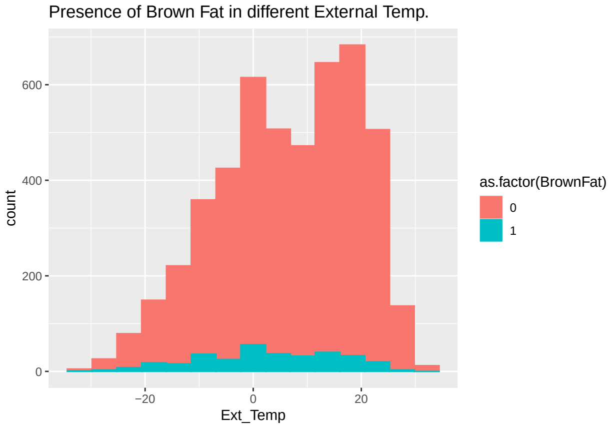
Graph of volume of brown fat and presence of brown fat vs presence of diabetes:





Graphs of brown fat vs external temperature:





Model

Steps for Model Building:

- Choose all variables with GLM (binomial, log-link)
- Step function to reduce AIC
- Choose significant terms from summary to find most significant interaction terms
- Add them to the last model
- Validate

Model Building using GLM

First, build a generalized linear model using binomial method and logit link function with all variables. Use a binomial method because Y is binary, yes or no for brown fat. From summary get AIC = 2161.95. Then use step function to reduce the AIC and get the model with AIC=2123.89

```
#Start:  AIC=2161.95
model_test = glm(data= dat,BrownFat~as.factor(Sex)+as.factor(Diabetes)+Age+Day+Month+
                     Ext_Temp+TwoD_temp+ThreeD_temp+SevenD_temp+oneM_temp+as.factor(Season)+
                     Duration_Sunshine+Weigth+Size+BMI+Glycemy+LBW+as.factor(Cancer_Status)+
                     as.factor(Cancer_Type)
                 , family =binomial(link = 'logit'))
summary(model_test)




#try step()
step(model_test, direction = 'backward', test = 'Chisq')
#with AIC 2123.89

#                         Df Deviance    AIC    LRT  Pr(>Chi)
# <none>                       2095.9 2123.9
# - as.factor(Season)      3   2102.1 2124.1  6.219  0.101418
# - oneM_temp              1   2099.1 2125.1  3.237  0.072009 .
# - Duration_Sunshine      1   2099.7 2125.7  3.781  0.051825 .
# - LBW                    1   2101.6 2127.6  5.664  0.017320 *
# - ThreeD_temp            1   2105.1 2131.1  9.179  0.002448 **
# - TwoD_temp              1   2106.6 2132.6 10.659  0.001096 **
# - Ext_Temp               1   2111.2 2137.2 15.340 8.980e-05 ***
# - as.factor(Diabetes)    1   2115.2 2141.2 19.289 1.124e-05 ***
# - Weigth                 1   2120.9 2146.9 25.006 5.715e-07 ***
# - as.factor(Sex)         1   2134.4 2160.4 38.496 5.486e-10 ***
# - Age                    1   2178.4 2204.4 82.469 < 2.2e-16 ***
# ---
```

```
after_step = glm(formula = BrownFat ~ as.factor(Sex) + as.factor(Diabetes) +
    Age + Ext_Temp + TwoD_temp + ThreeD_temp + oneM_temp + as.factor(Season) +
    Duration_Sunshine + Weigth + LBW, family = binomial(link = "logit"),
    data = dat)
summary(after_step)
```

```
##
## Call:
```

```
## glm(formula = BrownFat ~ as.factor(Sex) + as.factor(Diabetes) +
##     Age + Ext_Temp + TwoD_temp + ThreeD_temp + oneM_temp + as.factor(Season) +
##     Duration_Sunshine + Weigth + LBW, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3888  -0.4037  -0.2723  -0.1858   3.0223
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           3.004205   1.232758   2.437 0.014811 *
## as.factor(Sex)2      -1.364528   0.227291  -6.003 1.93e-09 ***
## as.factor(Diabetes)1 -1.471768   0.420498  -3.500 0.000465 ***
## Age                  -0.036875   0.004006  -9.205  < 2e-16 ***
## Ext_Temp             -0.054511   0.014021  -3.888 0.000101 ***
## TwoD_temp             0.137952   0.042334   3.259 0.001119 **
## ThreeD_temp          -0.123417   0.040866  -3.020 0.002527 **
## oneM_temp             0.036954   0.020569   1.797 0.072395 .
## as.factor(Season)2   -0.211946   0.239970  -0.883 0.377118
## as.factor(Season)3   -0.722405   0.329495  -2.192 0.028346 *
## as.factor(Season)4   -0.652931   0.279965  -2.332 0.019691 *
## Duration_Sunshine    -0.002716   0.001394  -1.948 0.051418 .
## Weigth               -0.037865   0.008738  -4.333 1.47e-05 ***
## LBW                   0.044979   0.019282   2.333 0.019663 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2399.3  on 4841  degrees of freedom
## Residual deviance: 2095.9  on 4828  degrees of freedom
## AIC: 2123.9
##
## Number of Fisher Scoring iterations: 7
```

In summary there are 5 terms that are significant with p-value < 0.001. Next, find significant interaction terms with 5 most significant variables with step function again.

```
#try to find significant interaction term through 5 most significant variable
model_test_interaction = glm(formula = BrownFat ~Ext_Temp* as.factor(Diabetes)
                      *Weigth*as.factor(Sex) * Age, family = binomial(link = "logit"), data = dat)

step(model_test_interaction)
#with AIC=2126
# glm(formula = BrownFat ~ Ext_Temp + as.factor(Diabetes) + Weigth + as.factor(Sex)+ Age +
# as.factor(Diabetes):as.factor(Sex) + Weigth:as.factor(Sex) + as.factor(Diabetes):Age +
# Weigth:Age + as.factor(Sex):Age + as.factor(Diabetes):as.factor(Sex):Age + Weigth:as.factor(Sex):Age,
# family = binomial(link = "logit"), data = dat)
```

Add those terms to the original model with all variables. With the new function, AIC = 2115.1. AIC decreased by 8 compared to the previous model.

```
#add interaction terms to the original model

model_test_interaction = glm(data= dat,BrownFat~as.factor(Sex)+
            as.factor(Diabetes)+Age+Day+Month+Ext_Temp+
            TwoD_temp+ThreeD_temp+SevenD_temp+oneM_temp+as.factor(Season)+
            Duration_Sunshine+Weigth+Size+BMI+Glycemy+
            LBW+as.factor(Cancer_Status)+as.factor(Cancer_Type)+
            as.factor(Diabetes):as.factor(Sex) + Weigth:as.factor(Sex) +
            as.factor(Diabetes):Age + Weigth:Age + as.factor(Sex):Age +
            as.factor(Diabetes):as.factor(Sex):Age +Weigth:as.factor(Sex):Age,
            family =binomial(link = 'logit'))


step(model_test_interaction)

# Step:  AIC=2115.1
# BrownFat ~ as.factor(Sex) + as.factor(Diabetes) + Age + Ext_Temp +
#     TwoD_temp + ThreeD_temp + oneM_temp + as.factor(Season) +
#     Duration_Sunshine + Weigth + LBW + as.factor(Sex):as.factor(Diabetes) +
#     as.factor(Sex):Weigth + as.factor(Diabetes):Age + Age:Weigth +
#     as.factor(Sex):Age + as.factor(Sex):as.factor(Diabetes):Age +
#     as.factor(Sex):Age:Weigth
```

Here’s the final model:

```
updated_model = glm(data=dat, BrownFat ~ as.factor(Sex) + as.factor(Diabetes) + Age + Ext_Temp +
    TwoD_temp + ThreeD_temp + oneM_temp + as.factor(Season) +
    Duration_Sunshine + Weigth + LBW + as.factor(Sex):as.factor(Diabetes) +
    as.factor(Sex):Weigth + as.factor(Diabetes):Age + Age:Weigth +
    as.factor(Sex):Age + as.factor(Sex):as.factor(Diabetes):Age +
    as.factor(Sex):Age:Weigth, family =binomial(link = 'logit'))
summary(updated_model)
```

Validation using goodness of fit test based on deviance:

Residual deviance: 2073.1 on 4821 degrees of freedom from summary.

```
pchisq(2073.1, df=4821, lower.tail=FALSE)
```

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

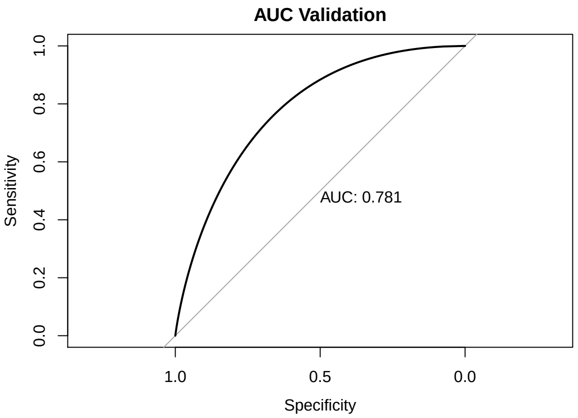
P-value shows that there is no evidence of lack of fit.

Validation with AUC

AUC tells how much the model is capable of distinguishing between classes. Let’s say we take two data points belonging to separate classes then there is 78% chance the model would be able to segregate them correctly.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



The final model has AUC=0.781 which means there is sufficient chance.

Discussion/Conclusion

Based on the model selected, brown fat depends on age and temperature which makes sense since until recently it was believed that this fat was present only in newborns and this fat needs to be activated by a low ambient temperature. Interestingly, cancer status does not appear in the final model, meaning there might not be a relationship between cancer and brown fat. That means for the future prediction of the cancer the brown fat might not be a good factor.

Limitations

One of the factors that could mean that the model is not the best fit is because of multicollinearity. There are 5 variables that depend on external temperature and it could affect the selection of preliminary model at first step function call.

There were some outliers that could affect the result

References

1. ggplot2 package - RDocumentation. (2021). RDocumentation. https://www.rdocumentation.org/packages/ggplot2/versions/3.3.5
2. R Markdown Cookbook (2021) https://bookdown.org/yihui/rmarkdown-cookbook/figures-side.html
3. BrownFat. (n.d.).