Brown Fat Case Study

Group Number: Project 5
Zidi Gao (1004945919)
Tongfei Li (1004759460)
Mengjiao Liu (1005491353)
HaoyangYu (1007570433)
STAC51H3 S 20221
Due Date: April 8, 2022

Background and Significance

We are no strangers to fat, but brown fat initially, research thought that brown fat was only found in newborns; however, with the advancement of technology, brown fat has gradually been found in adults. The function of brown fat is to allow mammals to adapt to cold environments, and this fat is activated only at low temperatures. The factors that determine the presence and volume of brown fat in adults are still to be studied, so the main objective of this study is to examine the relationship between numerous factors and the presence of brown fat in a large number of cancer patients.

Exploratory data analysis

Response variable:
BrownFat: The existence of brown fat. (No=0, Yes=1)

Quantitative Explanatory variables:
Age: Age of the patient in years.
Day: Day of the year.
Ext_Temp: External Temperature.
2D_Temp: Average temperature of the last 2 days.
3D_Temp: Average temperature of the last 3 days.
7D_Temp: Average temperature of the last 7 days.
1M_Temp: Average temperature of last month.
Duration_Sunshine: Sunshine duration.
Weight: in Kgs
Size: in cms.
BMI: Body Mass index.
Glycemia.
Lean Body Weight.
TSH
Total_Vol: Total volume of Brown Fat.

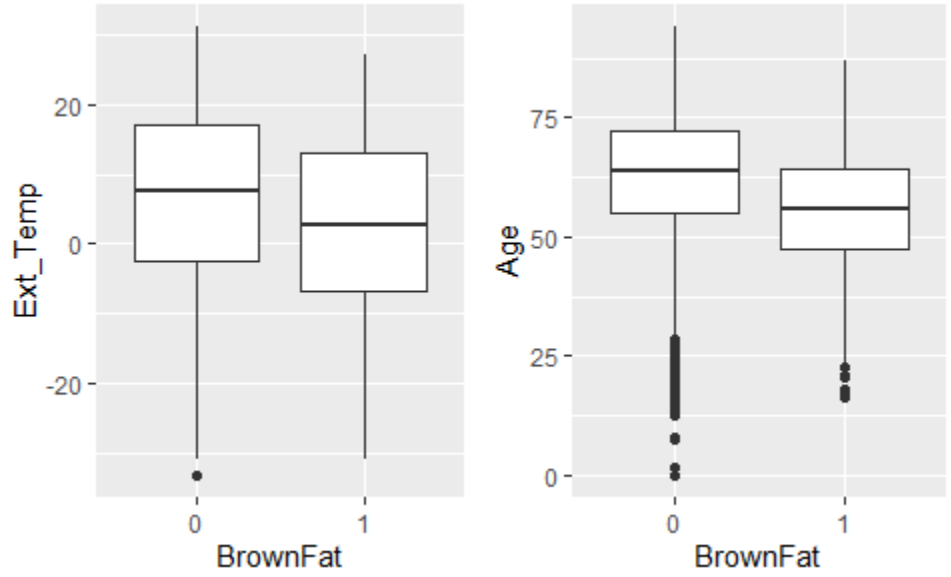
Categorical explanatory variables:
Sex: sex of the patient (Female=1, Male=2).
Diabetes: (No=0, Yes=1).
Month: Month of the exam.
Season: Spring=1, Summer=2, Autumn=3, Winter=4.
Cancer_Status: (No=0, Yes=1).
Cancer_Type: (No=0, lung=1, digestive=2, Oto-Rhino-Laryngology=3, breast=4, gynecological
(female)=5, 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).
BrownFat: (No=0, Yes=1).

Variable Selection Before Model Fitting:

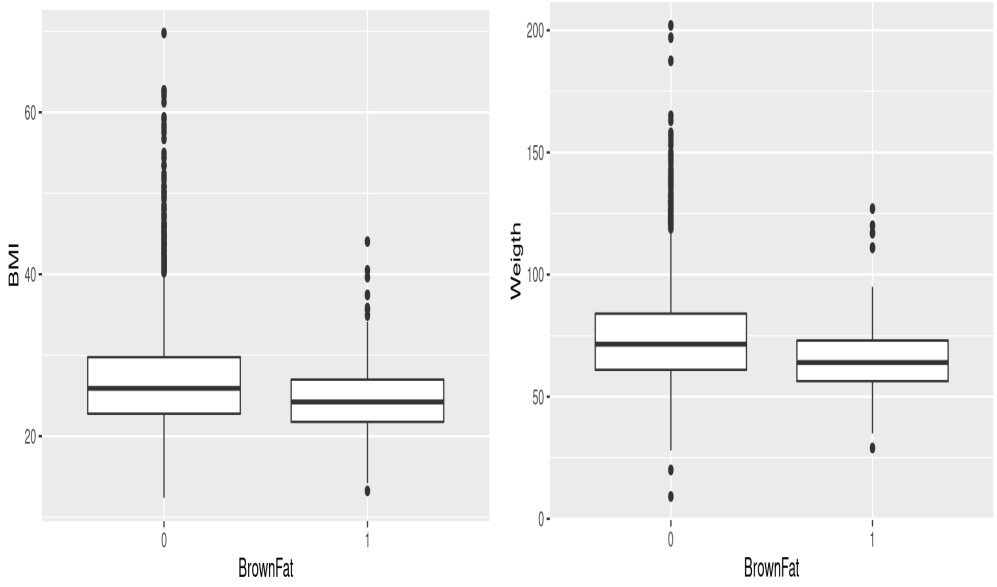
ID: There is no relationship between ID number and the existence of brown fat, so we can remove this variable.
TSH: There are approximately 91.39% of responses in TSH are missing, which may cause a selection bias if we use TSH as an explanatory variable. Thus, we can remove TSH from the explanatory variables.
Total_vol: Since our goal is to predict the existence of brown fat, the volume of brown fat is highly correlated with our response variable (existence of brown fat). We can remove this variable from explanatory variables before fitting the model.

(There are only approximately 6.77% of the participants have brown fat.)

Plots:



According to the boxplot above, we can clearly see that temperature and age do have effects on the existence of brown fat.



And we can find that BMI and Weight have an effect on the existence of brown fat。The graph shows that BMI and weight, both mean and maximum, were higher in the absence of brown fat than in the presence of brown fat in the body.

Model

Since we have a binary response variable on the existence of brown fat, we decide to fit a logistic model which is more popular in health sciences because coefficients can be interpreted in terms of odds ratios.

1. Only Main Effect Model
   Variables: Sex, Diabetes, Age, Day, Month, Ext_Temp, 2D_Temp, 3D_Temp, 7D_Temp, 1M_Temp, Season, Duration_Sunshine, Weight, Size, BMI, Glycemy, LBW, Cancer_Status, Cancer_Type
   There are 19 explanatory variables with 51 coefficients in total.

2. Reduced Main Effect Model
   Then we can reduce the main effect model with the two-sided step function:

```
Step:  AIC=2123.89
BrownFat ~ Sex + Diabetes + Age + Ext_Temp + `2D_Temp` + `3D_Temp`
+
    `1M_Temp` + Season + Duration_Sunshine + Weigth + LBW
```

   check if we can further simplify the model

```
`1M_Temp`          1   2099.1 2125.1  3.237  0.072009 .
Season             3   2102.1 2124.1  6.219  0.101418
Duration_Sunshine  1   2099.7 2125.7  3.781  0.051825 .
```

   further, remove Season, Duration_Sunshine and 1M_Temp
   Variables: Sex, Diabetes, Age, Ext_Temp, 2D_Temp, 3D_Temp, Weight, LBW
   There are 8 explanatory variables with 9 coefficients in total.

3. Interaction Effect Model
   Variables: Sex,Diabetes,Age,Ext_Temp,`2D_Temp`,`3D_Temp`,Weigth,LBW and their interaction terms
   There are 36 variables (including interaction terms) with 37 coefficients in the model.

4. Reduced Interaction Effect Model
   Then we can reduce the interaction model with the two-sided step function:

```
Step:  AIC=2108.23
BrownFat ~ Sex + Diabetes + Age + Ext_Temp + `2D_Temp` + `3D_Temp`
+
    Weigth + LBW + Sex:Diabetes + Sex:Age + Diabetes:Weigth +
    Diabetes:LBW + Age:Weigth + Ext_Temp:`2D_Temp` +
`3D_Temp`:Weigth +
    Weigth:LBW + Ext_Temp:Weigth
```

   Again, check if we can further simplify the model:

```
Age:Weigth         1   2075.0 2109.0 2.7625 0.096496 .
```

   further, remove the interaction term between Age and Weight for p-value=0.096

Variables: Sex, Diabetes, Age, Ext_Temp, 2D_Temp, 3D_Temp, Weigth, LBW, Sex:Diabetes, Sex:Age, Diabetes:Weigth, Diabetes:LBW, Ext_Temp:2D_Temp, 3D_Temp:Weigth, Weigth:LBW, Ext_Temp:Weigth
There are 16 variables (including interaction terms) with 17 coefficients.

Discussion/Conclusion

(Convey your findings to a broader audience. Reference everything including the source of images and sources of information.)

1. Homogeneous association between variables (compare main effect model with interaction model) – decision: include the interaction terms

```
Analysis of Deviance Table

Model 1: BrownFat ~ Sex + Diabetes + Age + Ext_Temp + `2D_Temp` + `3D_Temp` +
    Weigth + LBW + Sex:Diabetes + Sex:Age + Diabetes:Weigth +
    Diabetes:LBW + Ext_Temp:`2D_Temp` + `3D_Temp`:Weigth + Weigth:LBW +
    Ext_Temp:Weigth
Model 2: BrownFat ~ Sex + Diabetes + Age + Ext_Temp + `2D_Temp` + `3D_Temp` +
    Weigth + LBW
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4825     2075.0
2      4833     2105.7 -8  -30.653 0.000162 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here in our ANOVA table, the deviance of comparing the main effect model and model with interaction terms is -30.653. The corresponding P_value is 0.000162, a lot lower than alfa=0.05, which is very significant, so we cannot assume the homogeneous association here, the complex model with interaction is better.
The homogeneous association suggests the same odds ratio of one variable given a second variable.
There are 16 interaction terms in our model, which are: “`Sex:Diabetes + Sex:Age + Diabetes:Weigth + Diabetes:LBW + Ext_Temp:`2D_Temp` + `3D_Temp`:Weigth + Weigth:LBW + Ext_Temp:Weigth`”.
Take “Sex:Age” for example, there is significant evidence that the effect of Age on BrownFat is dependent on Sex, and vice versa. Take all else constant, when age changes, the change of expected brown fats of Female and Male are different.

2. GOF test (Hosmer-Lemeshow test, with g > p + 1, LRT test)
   Here p (number of covariates) is 16, choose g=18 > 17

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod_inter_rr$y, fitted(mod_inter_rr)
X-squared = 19.651, df = 16, p-value = 0.2363
```

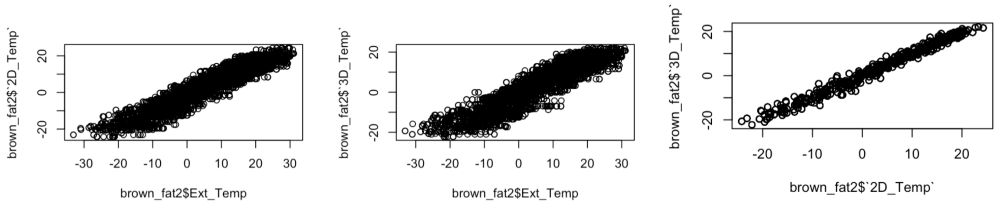
the p-value is large, and fails to reject H0, we have enough evidence to show that observed frequencies equal to expected frequencies. So this model fits the data well.

3. Interpretation and conjecture of some of the significant terms beta

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.4072353  1.3043144  -1.079  0.28063
Sex2                0.2532918  0.4977098   0.509  0.61081
Diabetes1          -8.7650997  4.6126461  -1.900  0.05740 .
Age                -0.0290306  0.0049092  -5.913 3.35e-09 ***
Ext_Temp           -0.1817817  0.0584146  -3.112  0.00186 **
`2D_Temp`           0.1217679  0.0414336   2.939  0.00329 **
`3D_Temp`           0.0730166  0.0713934   1.023  0.30643
Weigth             -0.0049811  0.0185015  -0.269  0.78776
LBW                 0.0814400  0.0314411   2.590  0.00959 **
Sex2:Diabetes1     -4.1129188  2.4234841  -1.697  0.08968 .
Sex2:Age           -0.0268812  0.0084498  -3.181  0.00147 **
Diabetes1:Weigth   -0.1272248  0.0877788  -1.449  0.14723
Diabetes1:LBW       0.3512024  0.1994582   1.761  0.07828 .
Ext_Temp:`2D_Temp` -0.0009005  0.0004208  -2.140  0.03236 *
`3D_Temp`:Weigth   -0.0024676  0.0009348  -2.640  0.00830 **
Weigth:LBW         -0.0006716  0.0003531  -1.902  0.05721 .
Ext_Temp:Weigth     0.0020364  0.0008508   2.393  0.01669 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For main effect terms:
e^beta_diabetes = exp(-8.7650997) = 0.0001560866: Keep all else constant, estimated odds that diabetes patients have brown fats are 0.000156 times estimated odds for non-diabetes.
e^beta_sex = exp(0.2532918) = 1.288259: Keep all else constant, estimated odds that male patients have brown fats are 28.8% higher than estimated odds for females.
e^beta_age = exp(-0.0290306) = 0.9713867 : Keep all else constant, one year increase in age, odds of having brown fats decreased by 2.86%.
From above, diabetes patients are very less likely to have brown fats, males are more likely to have brown fats, and odds of having brown fats decrease by age, which is consistent with the statement that brown fats present more likely in newborns but they also appear in adults.

Moreover, e^beta_Exttemp=0.8337833 < 1 < e^beta_3D=1.075748 < e^beta_2D=1.129492, it is very interesting that increase in 2-day and 3-day temperature lead to increase in odds of having brown fats, but increase in external temperature leads to decrease in odds.
Here I have two conjectures about this: one is that when 2 days and 3 days temperatures are higher, but there is a sudden drop in temperature(external temperature lower), the probability of having brown fats against no brown fats is higher. Another is that there should be multicollinearity among these temperature variables because the change of one variable must lead to a change of another, so the value of a single beta does not mean a lot of things.
After careful consideration, we suggest the second conjunction, since there is a high correlation between these three variables, and multicollinearity is very high.



Attempts to reduce multicollinearity:
First, we tried to delete one of the temperature variables, when we delete 3D_Temp, we have to delete the interaction of `3D_Temp`:Weigth, but reducing this significant term increases AIC by 2117.5-2109=8.5, and it makes both the other two variables not significant. Tries of reducing the other two variables also cause similar situations, so this attempt fails.

```
Ext_Temp           -4.133e-02  2.624e-02  -1.575  0.11525
`2D_Temp`           2.690e-02  1.534e-02   1.754  0.07950 .  AIC: 2117.5
```

Second, we tried to center the temperature terms, but it only works for quadratic terms, this didn't change AIC or correlations between variables, and this attempt fails. Therefore we decided to keep all three terms since we need to follow the hierarchy order of deleting terms.

For interaction terms:
e^beta_Sex:Age = exp(-0.0268812) = 0.9734769: On one year increase in age, the change(decrease) in odds of male with brown fats is 2.65% more than odds for females.

Estimated odds ratio for Diabetes effect given Gender is:
Female (sex=0) exp(-8.765)=0.000156, Male (sex=1) exp(-8.765-4.1129)=2.55e-06 . The difference between diabetes and non-diabetes with brown fats are a lot bigger among males than females.

Estimated odds ratio for Gender effect given Diabetes is:
(Diabete=0): exp(0.2532918) = 1.288259, (Diabete=1): exp(0.2532918-4.1129188)=0.02107586. Among non-diabetes, Males are more likely to have brown fats than females. One impressive thing is that, among diabetes, males are very less likely to have brown fats than females!

4. Predictive power – classification tables

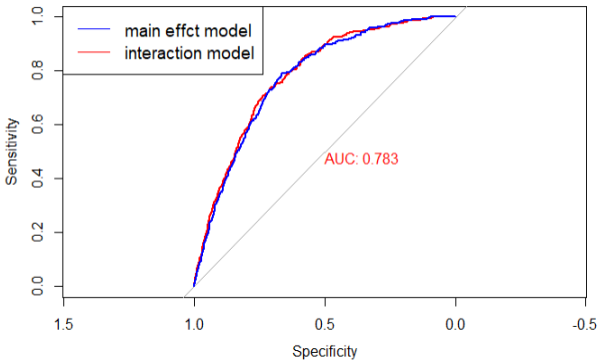
y	predicted 0	1
0	2946	1568
1	74	254

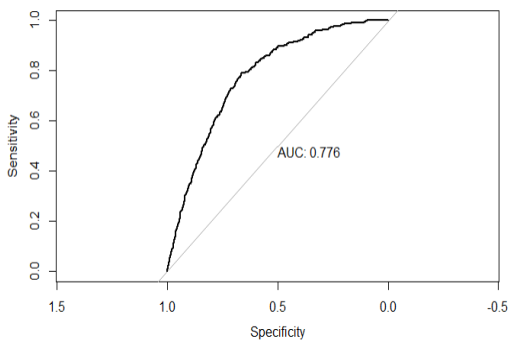
Sensitivity = 77.4% Specificity = 65.3% Concordance rate =66.1 %

From the two-by-two table of data, we get sensitivity and specificity, For sensitivity(77.4%), it is larger than 50%, which means if we use this model to predict who has Brown Fat then the success rate is really high. For specificity(65.3%)the rate is higher than 50%, which indicates a better fit for the model. If we use the interaction model to predict who does not have Brown Fat, there is a good chance that our prediction is correct. And finally, the concordance rate(66.1 %) is over 50% which tells us the overall success rate is high and if

we use the model to test whether there is brown fat the result is likely to be correct. The interaction model is good for being used to fit the data.

5. Comparing Main Effect Model with InteractionModel by ROC curve:





Here are plots of ROC curves. The first plot shows the comparison of the main effect model and interaction model, and AUC (Area under the ROC Curve) for the interaction model is 0.783. However, in the first plot, the main effect curve and interaction curve are quite similar to each other, therefore we draw a separate ROC curve for the main effect model and the AUC is 0.776 which is smaller than the AUC of the interaction model's 'curve. The comparison proves once again that the interaction model is better suited to the data. In conclusion, the interaction model has a 78.3% success rate to distinguish the positive class values from the negative class values of Brown Fat data.

6. Conclusion

Our goal for the Brown Fat case study is to build a model to estimate the probability of having brown fat. After cleaning and analyzing the data, we build the best model which depends on 8 main effect variables and 8 interactions. The model has an accuracy of about 79% and fits the data well.

By deep-learning the model, we find some variables significantly affect the appearance of Brown Fat. Firstly male patients have a 28.8% higher estimated odds than female patients to have brown fat. Secondly, with increasing age the probability of brown fat's occurrence is on a decreasing trend, we suppose this is the reason why brown fat is present more likely in newborns and adults rarely have brown fat. Thirdly, for patients who have diabetes women are much more likely to have brown fat than men, this ratio is exactly the opposite of non-diabetes patients, but this is hard to say diabetes influences females to have brown fat much easier. Last but not least, we find 3 temperature variables (External Temp, 2 days average Temp, 3 days average Temp) are strong correlate to each other, which confuses us about the relationship between temperature and brown fat appearance. Therefore for future research, we suggest more detailed record temperature, in particular, the range of temperature changes. This may be helpful for finding out how temperature influences Brown Fat.

Assigned job description

Zidi Gao Assigned Job:

1: Do the predictive power and calculate Sensitivity, Specificity, Concordance rate then interpret them.

2: Compare ROC curve of main effect model and Interaction model

3:Conclude the whole report and give suggestions.

Tongfei Li Assigned Job:

1. Test homogeneous association between variables (compare main effect model with interaction model), perform Hosmer-Lemeshow test to see goodness of fit of interaction model, and do all of the interpretation of beta values, explore effects of several terms to brown fats.

2. Write part 1, 2, 3 in report-discussion.

Mengjiao Liu Assigned Job:

fit, reduce and choose the model

HaoyangYu Assigned Job:

fit exploratory data analysis and introduction.

Reference

Lakhal-Chaieb, L. (Ed.). (2011). *Statistical Society of Canada*. Determinants of the Presence and Volume of Brown Fat in Human | Statistical Society of Canada. Retrieved March 31, 2022, from https://ssc.ca/en/case-study/determinants-presence-and-volume-brown-fat-human