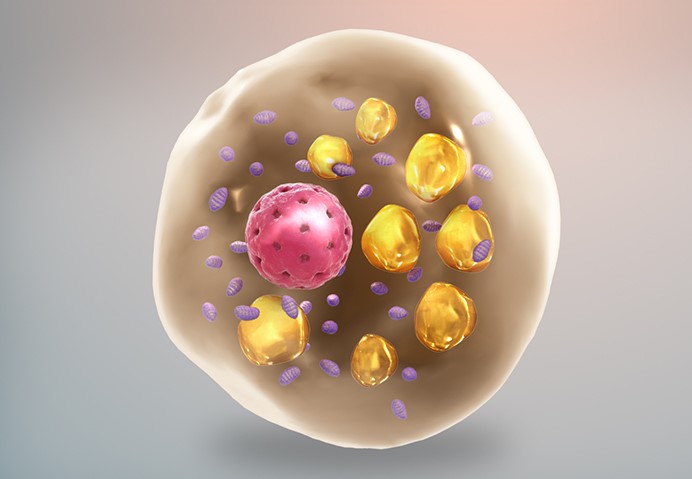
**Case Study: Identifying Determinants of the Presence and Volume of Brown Fat in Human**

Group 13: Lingfen Chen, Yayu Chen, Dhruv Goyal, Hyun Mok Kim, Kenneth Kim

****

**Table of Contents**

**1. Introduction**

**2. Objective**

**3. Data**

**4. Investigating the relationship between covariates**

**5. Building the model**

**6. Possible Limitations**

**7. Conclusion**

**8. References**

**Introduction**

Brown adipose tissue (BAT), also commonly known as brown fat, is one of the two primary adipose tissues. The most common adipose tissue is white fat which accumulates through large lipid droplets regulating body temperature by providing insulation for the body's organs (Marcin, A. 2020). On the other hand, brown fat helps to regulate body temperature by generating body heat and burns excess energy (Fox, K., & Smith, B. 2018). Brown fat was thought to be only found in babies and small/hibernating mammals. Newborns have a lot of brown adipose tissue since they have not developed the shiver mechanic to create heat. The common consensus was that adults lose brown fat but develop the shiver mechanic to respond to cold body temperature. This occurs through a process called thermogenesis (Marcin, A. 2020). Thermogenesis is explained by the body producing a hormone in response to the temperature stimuli, and the receptors in the brown adipose tissue picking up the hormone and then begins to create heat and burn fat (Marcin, A. 2020). In the process, scientists believe that burning brown fat also burns calories. Thus, scientists and the medical community have postulated that the activation of brown fat could act as a possible therapeutic treatment for obesity and other metabolic diseases. Brown fat activation is linked with lower body weight, and better glucose and lipid regulation (Statistical Society of Canada, 2011). This is incredibly important because obesity is the fifth leading cause of deaths with an estimated 2.8 million adults each year (EASO, 2020) and 86.3% of adult Americans are projected to become obese by 2030(All Answers Ltd. , November 2018). It is imperative that we find possible treatments so that we can save people’s lives.

**Objective**

The objective of this analysis is to identify the factors determining the existence and the volume of brown fat in a large cohort of cancer patients. We aim to investigate the relationship between many explanatory variables and the presence of brown fat. From our analysis regarding the most significant covariates, we will attempt to build a model in order to estimate the probability of having brown fat.

**Data**

In order to complete our objective outlined earlier, the Molecular Imaging Center at The University of Sherbrooke provides cancer patient data. This data is classified as ungrouped data as it is not sorted into categories and grouped manners. In this data set, there is a response variable pertaining to the presence of brown fat and there are many covariate variables.

Some of the variables in the data set include:

\* Sex: sex of the patient (Female=1, Male=2).

\* Diabetes: (No=0, Yes=1).

\* Age: Age of the patient in years.

\* Ext\_Temp: External Temperature.

\* Weight: in Kgs

\* Size: in cms.

\* BMI: Body Mass index.

And more for a total of 23 variables, for 4842 patients.

**Exploration of the DataSet**

We explore the dataset by taking a look at the number of male and female patients with brown fat and the corresponding average of brown fat volume in each gender:

|  |  |
| --- | --- |
|  |  |

There is a presence of brown fat in 245 female and 83 male patients from the dataset. We observe that the presence of brown fat is 3 times more like in female patients than male.

The female patients with presence of brown fat have a mean volume of 46.14 and the males have around 26. The female patients had greater brown volume fat than male.

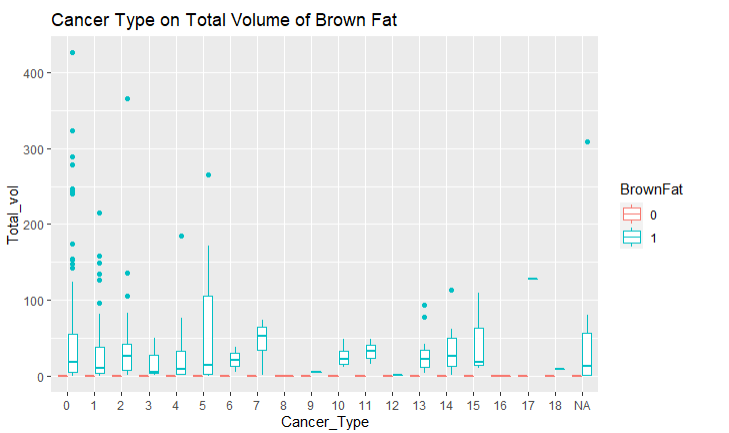


There were a total of 4,842 patients with 2,370 female patients and 2,472 male patients. The sample’s observed ratio of females having brown fat was 245 / 2,370 = 0.1033755 (10.34%) and males was 83 / 2,472 = 0.03357605 (3.36%).

The probability of a female patient having brown fat was approximately 7% more likely than a male.

Based on these results, we can focus on testing whether sex is a significant predictor of the prescense of brown fat.

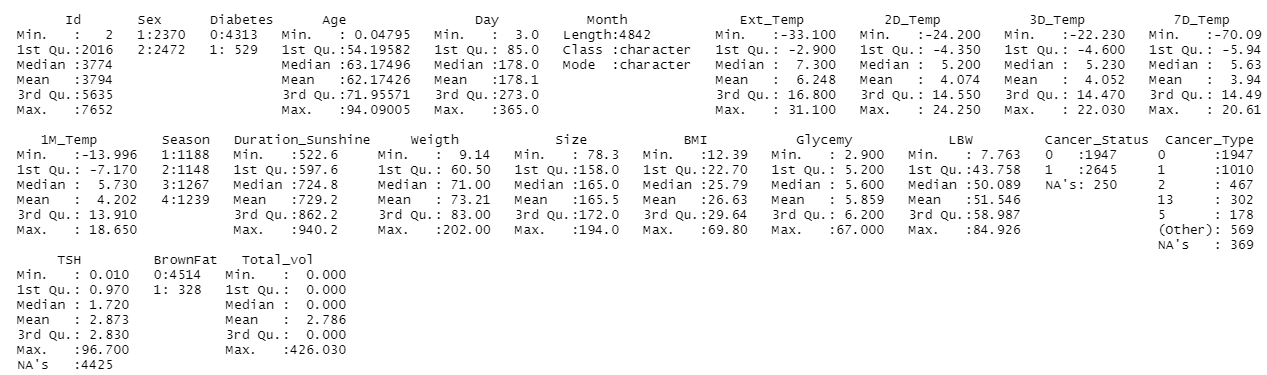
Since this data set is on cancer patients, we explore the data to see if different types of cancer can be a good predictor for the response variable.

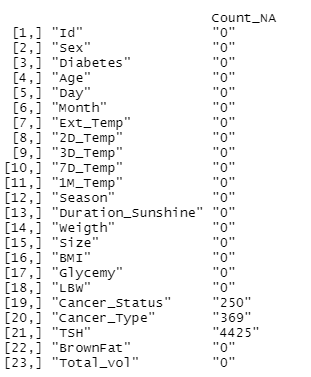


From this plot, we can see that some types of cancer may be a significant predictor for the presence of brown fat. We can also focus on cancer during the model selection process.

**Investigating the relationship between covariates**

Before we begin, we take a glimpse of the dataset:





From the summary we observed NA values in the column “Cancer\_Status”, “Cancer\_Type”, “TSH”. The NA values could represent patients who have not been tested. In the “TSH” column, there are 4,425 patients with NA values out of 4,842 patients (More than 90% of total patients) which means there is not enough information about “TSH” to test the relationship between other factors. Other research shows patients with BAT will tend to have higher TSH level(thyroid-stimulating hormone).(Lapa C, et al. 2015). The data set records TSH, but we chose not to include TSH in the model. Since 90% of TSH are missing, we thus made the decision to remove the “TSH” level.

To adjust NA values for “Cancer\_Status” and “Cancer\_Type”, it is appropriate to exclude the NA values of “Cancer\_Status” and “Cancer\_Type” as well to study the relationship between each factor as we still have enough information to analyze the relationship. Removing 250 NA values would represent only about 5.1% of the values, and R also uses **na.omit** and **na.exclude** to deal with NA values. This process removes the data of the patients who only have NA values for either “Cancer\_Status” or “Cancer\_Type”.

We can also look at possible effects of multicollinearity whether the variables with NA are highly correlated with another variable, we can drop the variables with NA values.

Interpolation is another possible solution, however since in the TSH column we only have a limited amount of values (approx. 9%), we believe that interpolation can possibly lead to over/under stating the effect on the response variables and thus produce inaccurate results.

We take a look at the variance covariance matrix to identify possible multicollinearity between variables. We plot the matrix for quantitative predictors:

|  |
| --- |
|  |
|  |
|  |

We can see from the variance covariance matrices, that there is a strong correlation between Weight and BMI, Sex and Size, Sex and Lean Body Weight, Size and LBW, Size and Weight. We can also see that the presence of Brown Fat and Total Brown Fat volume are correlated.

There is a strong correlation between the External temperature and all other temperature measurements, and with the Duration of Sunshine.

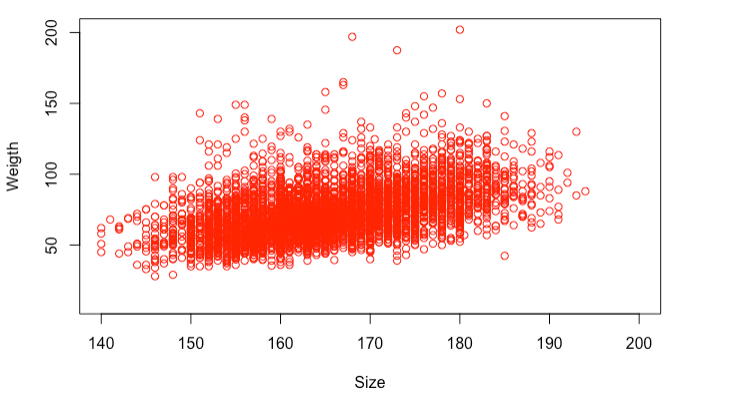
We then take a look at the categorical predictors and we plot the correlation of Categorical variables with BrownFat:

|  |
| --- |
|  |

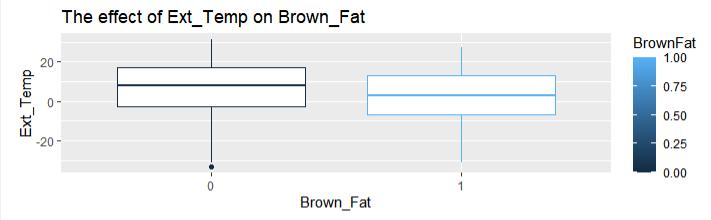
From these matrices, we can see that we do not need to examine all the covariates against the response variable, as examining one of the correlated predictors, we would see similar effects due to the principles of multicollinearity.

After removing variable “TSH”, we investigate which variables have significant effects on having brown fat. Since this model contains many variables, we observed there are some correlations between variables. For example, as identified earlier, size and weight, exemplifies multicollinearity.

**Correlation between Size and Weight**



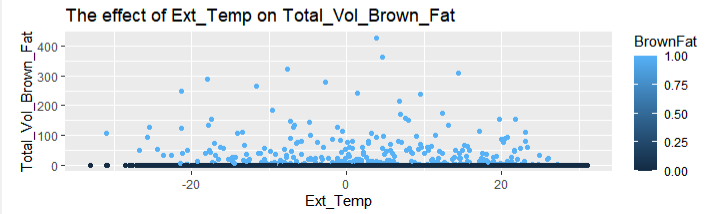
Based on our introduction we explore the most significant factor for the presence of brown fat which is low air temperature. This is because brown fat needs to be activated by a low ambient temperature to be detected. We then plot the external temperature against the response variable brown fat:



With this boxplot we look at the effects of external temperature on patients with and without brown fat. It is shown that external temperature among patients with brown fat have lower external temperatures on average.

We can also see that there is an outlier in external temperature for patients without brown fat. The mean external temperatures for the existence of brown fat is lower than the mean temperature for non-brown fat. However, there are large variabilities for both groups, evident by the long whiskers.

We then explore the relationship between external temperature and the total volume of brown fat:



From this scatter plot, we can see that the patients without brown fat have 0 volume of brown fat. The effect of external temperature on total volume of brown fat looks random which means the increase/decrease of the external temperature does not have a significant effect on total volume of brown fat.

Therefore, the relationship between external temperature and the total volume of brown fat may be independent, despite us identifying correlation between total brown fat volume and the presence of brown fat. This tells us that there may be other important predictors which will affect the presence and volume of brown fat.

Taking a look at the plot between other temperature measurements, *2D\_Temp* and *3D\_Temp* we can see that the plots look very similar which would confirm the multicollinearity effect that we observed in the matrix.

|  |  |
| --- | --- |
|  |  |

So we can say with some confidence that Ext\_Temp, 2D\_Temp, 3D\_Temp, 7D\_Temp, 1M\_Temp, Season, Duration\_Of\_Sunshine are correlated together and have a significant relationship with the presence of brown fat.

We can also see that from our correlation matrix that the Weight and BMI are correlated

We plotted the effect of BMI and weight on the presence of brown fat and the similar result of both plots represent multicollinearity.

|  |  |
| --- | --- |
|  |  |

From these plots, we see that patients with high brown fat volume are on the lower range of both weight and BMI, and these two variables could be a significant predictor. Thus the other highly correlated variables might also have a significant effect on brown fat due to multicollinearity.

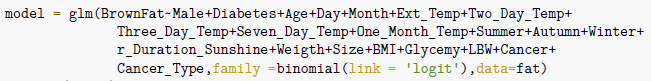
In order to deal with the possible identified multicollinearity we could drop one of the variables that is correlated less with the response variable or during the model selection process, see if the correlated variables become insignificant and then drop out.  
We aim to use a stepwise regression model selection process first, then see if any of the remaining significant predictors are highly correlated.

From there we can use VIF or fit multiple models, to see which predictors are the most significant in fitting the data.

**Building Model for Prediction**

To study the effect of explanatory variables against the response variable brown fat, we fit the model using logistic regression method as the response variable is a binary variable which only has 2 outcomes of having / not having a brown fat.

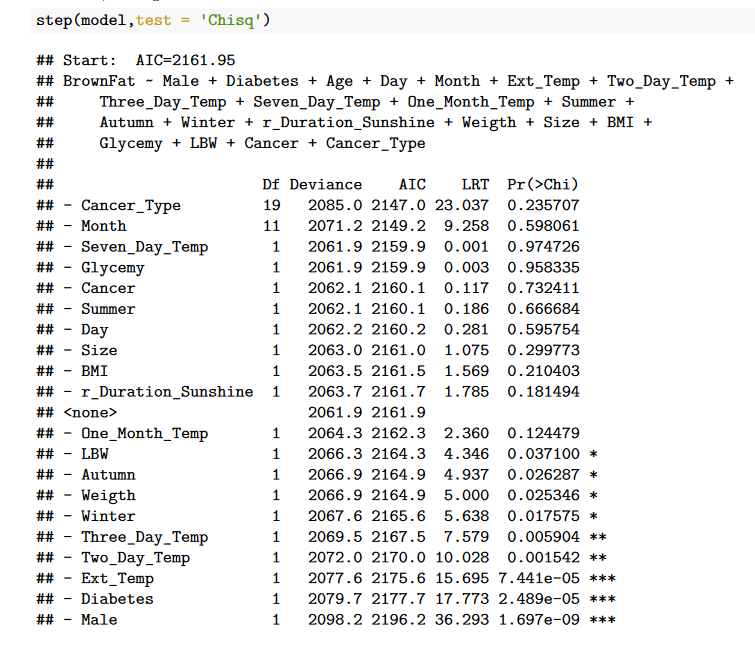
We first fit:

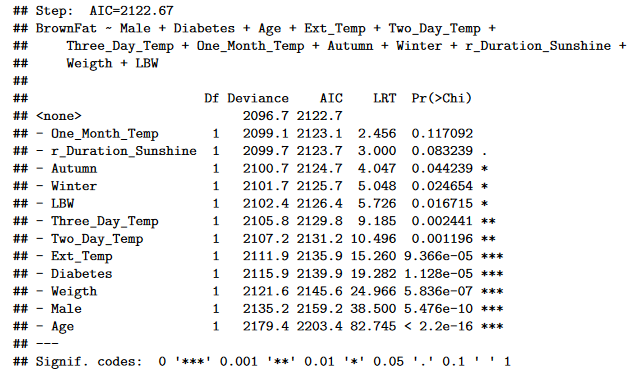


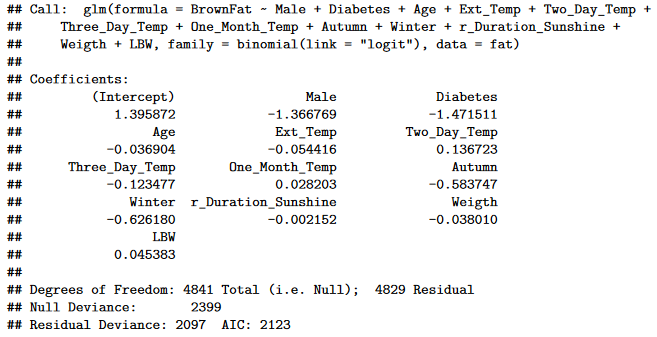
We decided to use a backward elimination method and started with the complex model including all the predictors. We have not included the interactions as we had many (23) explanatory variables. By using a backward elimination, we identify the least significant variables in the complex model. Then we refit the model with significant variables found from the previous part.

The backward elimination method automatically drops insignificant values until only significant values are remaining in the complex model.

We use R function step() so that the process is automatic:



After many iterations we arrive at the final model:



The resulting model will have the following predictor: ***Sex, Diabetes, Age, Ext\_temp,***

***Two\_day \_temp, Three\_day\_temp, One\_month\_temp, Season, Duration\_of\_sunshine, Weight, and LBW.***

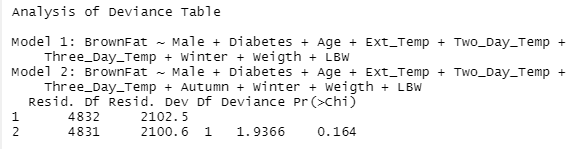
After Minimizing AIC, we can see that the model has some insignificant variables, so we fit and test if a simpler model is better.

We fit a second model that has the following predictors:

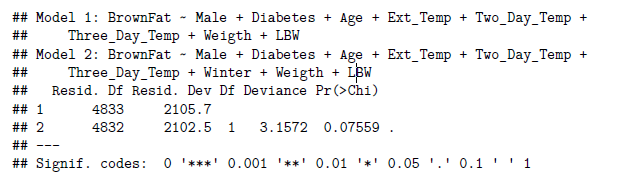
***Sex, Diabetes, Age, Ext\_temp, Two\_day \_temp, Three\_day\_temp, Season, Weight,***

***and LBW.***

We can use deviance of two models for comparison, thus we perform an analysis of deviance to test the null hypothesis that complex model fits better than the simpler model:



The p-value = 0.164 > 0.05 tells us that we fail to reject the null hypothesis. Therefore, we conclude that the simpler model fits as well as the complex model. We do this process again:

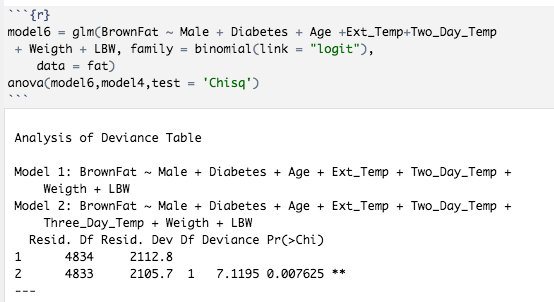


We again fail to reject the null hypothesis, and select the model with the following predictors:

***Sex, Diabetes, Age, Ext\_temp, Two\_day \_temp, Three\_day\_temp, Weight,***

***and LBW.***

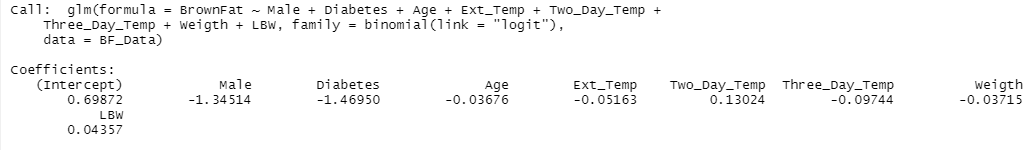
From our multicollinearity analysis we saw that some predictors may be correlated. We were able to drop out a lot of correlated predictors through backwards elimination as well as some predictors that we thought could be significant in our data exploration such as cancer.

In the latest model, we see ***Ext\_temp, Two\_day \_temp, Three\_day\_temp.*** We attempt a final test to see if we can remove some of the correlated predictors and see if it performs just as well in fitting the data:

From this analysis, we reject the null hypothesis and are not able to move forward with the simpler model.

Both two\_day\_temp and three\_day\_temp have significant effects on the presence of brown fat and are required by the model despite the two variables being highly correlated. This could possibly be explained by the importance of temperature in the activation of brown fat that we discussed in the introduction.

Our final model is:



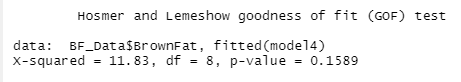
Based on the final model, we conclude that 1.the odds of having brown fat for women is 3.84 times the odds of having brown fat for men. 2. The odds of having brown fat for non-Diabetes people is 4.35 times the odds of having brown fat for people with Diabetes. 3. The odds for having brown fat is 0.96 times for every increase in age.

Next we do multiple goodness-of-fit tests on the final model to validate model selection.

We use the Hosmer-Lemeshow test, ROC curve, and classification table+predictive power.

The deviance given in the model cannot be used to test goodness-of-fit since the data is ungrouped data.

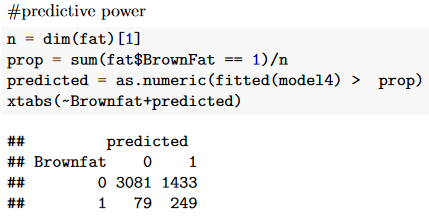
Appling the Hosmer and Lemeshow goodness of fit test on g > #of predictors + 1:

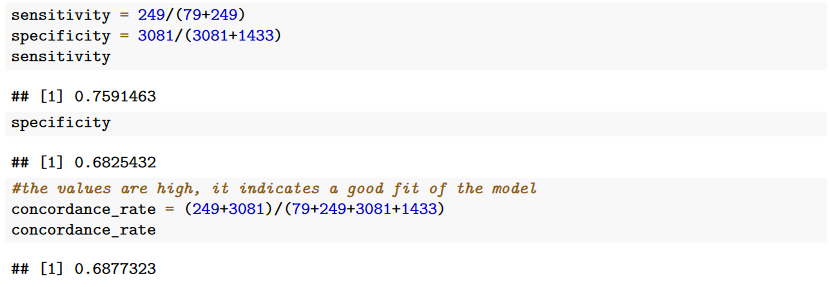


The null hypothesis is that the observed frequencies equal the expected frequencies, and since the output p-value is greater than 𝝰 =0.05, we fail to reject the null hypothesis. That is to say that the current model fits the data well.

The result of Hosmer and Lemeshow’s goodness of fit test for logistic regression inform us how well the data fits the model. The Hosmer and Lemeshow’s test is only effective when observing binary response variables. The reason to use this test if this test takes consideration of whether the observed values match the expected values of equal sized population groups.

To check the goodness-of-fit of the final model, we create a classification table and find predictive power of the logistic regression model to evaluate sensitivity and specificity of the final model. Higher sensitivity and specificity indicate a better fit of the model.

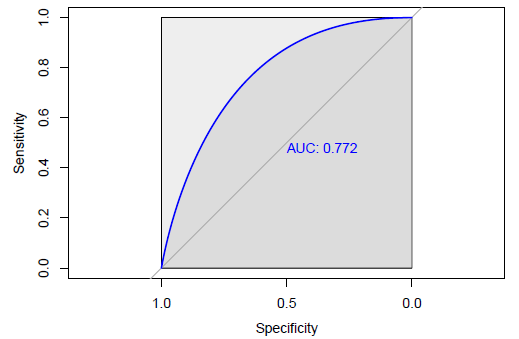




To check the goodness-of-fit, we first found that the cutoff value of the final model is 0.0677406 (prop). As both sensitivity and specificity are relatively high, the classification table and predictive power tell us that the final model is a good fit of the data.

For our final goodness of fit test, we compute the ROC curve, which is a plot of sensitivity as a function of (1 - specificity) for the possible cutoffs. The area under the ROC curve is called the concordance index, which tells us the probability that the predictions and the outcomes are agreeable. The larger the concordance index, the better. For example a concordance index of 0.5 indicates guesses at random.

Computing the ROC curve for our model results in:



The obtained concordance index for our ROC curve is 0.772 which is much higher than random guessing. Also the goodness-of-fit is explained by the high concordance rate. Therefore, we conclude that this logistic regression model is a good predictor for the presence of brown fat.

**Limitations of this Study**

Previously we mentioned that research shows patients with BAT will tend to have higher TSH level (Lapa C, et al. 2015). The data set records TSH, but we chose not to include TSH in the model since 90% of TSH are missing, we thus made the decision to remove the “TSH” level. Therefore, a possible strong predictor was removed due the lack of measurements of TSH.

Another possible limitation for this study is the possibility of overfitting the data. This occurs when the model that we fit perfectly follows the dataset. This would mean that the regression model perfectly encapsulates the variation between values in this data set (Javapoint). If over-fitting has occurred in this model, it would make our prediction worthless, as the model would only be useful for the known values of this particular dataset. Overfitting means the regression model has failed to pick up the relationship between predictors and the response, and memorized patterns within one dataset, making any predictions of new values untrustworthy and incorrect. A solution to this would be to create an evaluation dataset from a subsection of the original dataset, which the regression model does not train on and then compare how it performs on the two dataset.

Lastly, another limitation is the fact the model could be limited by variables in the dataset. There could be other strong predictors that model the presence of brown fat better.“Beth Israel Deaconess Medical Center (BIDMC) now indicate that brown fat can also act to regulate skeletal muscle function. … We knew that *muscles could regulate brown fat – exercising increases brown fat*”. (G. S, 2018) Thus, we could be measuring confounding effects of some predictors. Such as exercising can lead to decrease in LBW, and weight. Increase in muscle, muscle is dense and heavy, which increases LBW, size and weight. Since we’re not sure about the confounding effects, this is one of the major limitations to our analysis.

**Conclusion**

Brown adipose tissue (BAT), or brown fat, helps to regulate body temperature by generating body heat in response to cold temperatures. Through this the medical community has postulated that the activation of brown fat treatment for obesity and other metabolic diseases. After initial exploration of the dataset, we had an idea of some predictors that could be useful. We handled NA values, and then calculated the correlation matrix to see possible multicollinearity between covariates. So we can say with some confidence that Ext\_Temp, 2D\_Temp, 3D\_Temp, 7D\_Temp, 1M\_Temp, Season, Duration\_Of\_sunshine have a significant relationship with the presence of brown fat. We fit a most complex model and using a backward elimination method, we dropped out correlated variables.

Our final model, has the following significant predictors:

***Sex + Diabetes + Age + Ext\_Temp +Two\_Day\_Temp + Three\_Day\_Temp + Weight + LBW***

Some of these predictors are correlated, but dropping them hurt our model. We conclude that these predictors must be very important, given the background information regarding brown fat.

After multiple goodness of fit tests, Hosmer-Lemeshow, ROC Curve, etc. we are confident that this model fits the data well

As mentioned in the previous slide, while we are confident in the model we do recognize the possible limitations. We conclude this study with a quote from George Box that “All models are wrong, but some are useful”

**References**

1. Determinants of the Presence and Volume of Brown Fat in Human \| Statistical Society of Canada. (2011). Statistical Society of Canada. <https://ssc.ca/en/case-study/determinants-presence-and-volume-brown-fat-human>
2. Fox, K., & Smith, B. (2018, May 21). Everything You Need to Know About Brown Fat. Men's Journal. <https://www.mensjournal.com/health-fitness/everything-you-need-know-about-brown-fat/>
3. Glen, S. (2016, August 28). Hosmer-Lemeshow Test: Definition. Statistics How To. https://www.statisticshowto.com/hosmer-lemeshow-test/
4. All Answers Ltd. (November 2018). Heavy Cost of Obesity in Healthcare. Retrieved fromhttps://nursinganswers.net/essays/heavy-cost-of-obesity-in-healthcare.php?
5. Overfitting and Underfitting in Machine Learning - Javatpoint. (n.d.). Retrieved from https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning
6. Lapa, Constantin et al. “Activation of brown adipose tissue in hypothyroidism.” *Annals of medicine* vol. 47,7 (2015): 538-45. doi:10.3109/07853890.2015.1085126
7. Marcin, A. (2020, March 17). Brown Fat: What You Should Know. Healthline. <https://www.healthline.com/health/brown-fat> Wikipedia contributors. (2021, March 23). Brown adipose tissue. Wikipedia. <https://en.wikipedia.org/wiki/Brown\_adipose\_tissue>
8. Mikstas, C. (2020, April 20). What is Brown Fat? Retrieved April 11, 2021, from <https://www.webmd.com/diet/brown-fat-what-you-need-to-know#1>
9. Nedergaard, J., & Cannon, B. (2017). Brown Adipose Tissue - an overview \| ScienceDirect Topics. ScienceDirect. <https://www.sciencedirect.com/topics/neuroscience/brown-adipose-tissue>
10. Writer, G. S. (2018, October 31). Brown Fat Flexes Its Muscle. GEN - Genetic Engineering and Biotechnology News. <https://www.genengnews.com/topics/drug-discovery/brown-fat-flexes-its-muscle/>
11. Statistics. (2020, June 08). Retrieved from https://easo.org/media-portal/statistics/