Determine the possible factors that activate brown fat as well as confirm whether external temperature affects the activation of brown fat.

Group 11: Md Shams Rahman, Manish Suresh, Senthooran Yogeswaran

28/03/2022

Jobs performed by each member

Md Shams Rahman

Job	Description
Introduction	Investigated the background information of the case study
Model Validation	Performed the validation on the model to make sure it is sound/effective
Conclusion	Contributed in giving final conclusive statement of the case study
Report Refining	Performed grammar/refining of the case study

Senthooran Yogeswara

Job	Description
Model building	Built the models used in the case study
Model Selection	Performed model selection process for the final model used in the case study
Conclusion	Contributed in giving final conclusive statement of the case study
Limitations	Discussed the limitations of the model we chose

Manish Suresh

Job	Description
Description	Briefle described what to expect from the case study
Data Cleaning	Performed the cleaning process of the data and documented the process a series of steps
Data Analysis	Performed the exploratory data analysis in visual format
Limitations	Discussed the limitations of the model we chose
References	Added the APA Citation at the end of the document

Introduction

Background and Significance

In this case study, the objective is to determine which biological factors impact the probability of a human having brown fat.

Brown fat, also known as brown adipose tissue, refers to a special type of body fat that allows organisms to adapt to cold temperatures; it produces heat to help maintain body temperature, with its greater amount of mitochondria than white fat. It was well known that this type of fat is present in small animals or newborn humans, presumably due to their smaller form factor.[1] However, it was investigated that brown fat may also be present in adults. The dataset that will be used for this case study collects observations of many possible factors that may impact the presence of brown fat, including: the sex of the patient, if they have diabetes, their body temperatures for certain time periods, and more.

Brief description of our analyses

We begin by the data cleaning process, which includes removing any unnecessary variables, rows with NA values, etc. We would then proceed with our exploratory data analysis, which includes plots of our variables and our analyses based on the results. Afterwards, we begin the model building and model selection processes, choosing our best model based on the criteria. Upon selecting said model, we will perform model validation and model diagnostics on all models and verify that they satisfy all the regression assumptions, and that our best model doesn't suffer from over-fitting with the original dataset. We will also discuss the potential limitations of our model, before finally coming to the conclusion of our case study.

Data Cleaning

To briefly summarize this section, first removed variables that are unnecessary. We would then remove rows that containing NA values. After this, we check for duplicate values, as well as corrupted data (e.g. data record of a person having no cancer but has a cancer status or vice versa, or a diabetes, sex, or season level that does not belong to levels 1,2, ...etc). Finally, we would convert our dataset into a usable format, which involves converting categorical variables into a factor format, to avoid making them quantitative.

1. Removed columns that are unnecessary.

```
brownfat.working.data <- brownfat.data %>% select(-c(Id, TSH, Total_vol))
```

Id is for organizational purposes; therefore, it provides no significant value to brown fat.
TSH contains too many NA values. Upon further research[2], it was shown to have an insignificant effect to having brown fat.
Total_vol doesn't factor into the probability of having brown fat, in fact its the other way around.

2. Checked if there is any missing data.

```
for (value in names(brownfat.working.data)){
  index <- brownfat.working.data[,value] == "NA"
  if(sum(index) > 0){
    brownfat.working.data <- brownfat.working.data[!index,]
  }
}
```

There were a few rows with missing data that were promptly removed.

3. Checked if there is any duplicate data.

```
duplicated.data.frame(brownfat.working.data) %>%
  enframe(.) %>%
  filter(value == TRUE) %>%
  nrow(.)
```

```
## [1] 0
```

There are no duplicate records, hence we can move forward with further cleaning of the data.

4. Checked if there is any corrupted data.

```
brownfat.working.data %>%
  filter(Cancer_Status != 0 & Cancer_Status != 1) %>%
  filter(Cancer_Type >18) %>%
  filter(Diabetes != 0 & Diabetes != 1) %>%
  nrow(.)
```

```
## [1] 0
```

The process involved checking whether the categorical variables are as identified as mentioned in the data description.

5. Checked if the correct dependency is kept between cancer_status and cancer_type.

```
brownfat.working.data %>%
  mutate(Cancer_Verification = ifelse(Cancer_Status == 1 & Cancer_Type > 1
                                    || Cancer_Status == 0 & Cancer_Type == 0, 1, 0)) %>%
  filter(Cancer_Verification == 0) %>%
  nrow(.)
```

```
## [1] 0
```

The process involved checking whether a cancer status of 1 was properly associated with one of the given 18 cancer types, while a cancer status of 0 has cancer type 0.

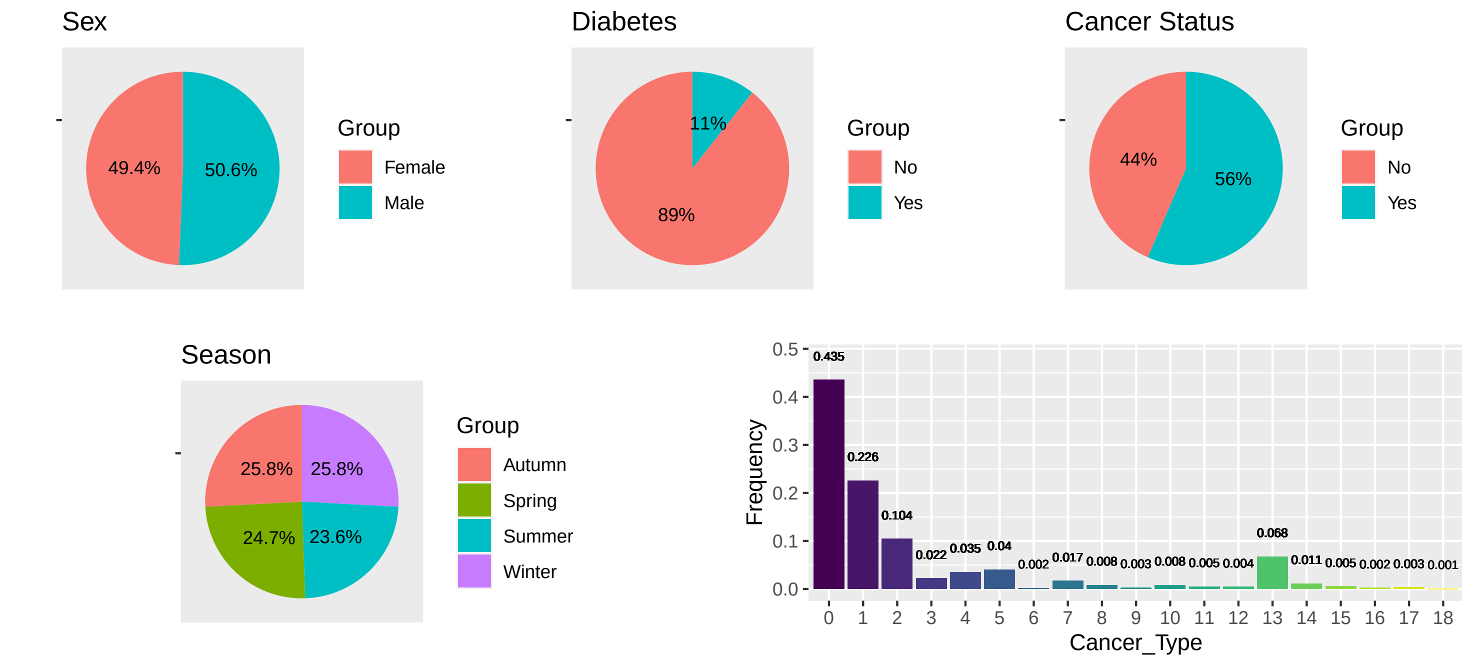
6. Finally convert the data into usable format.

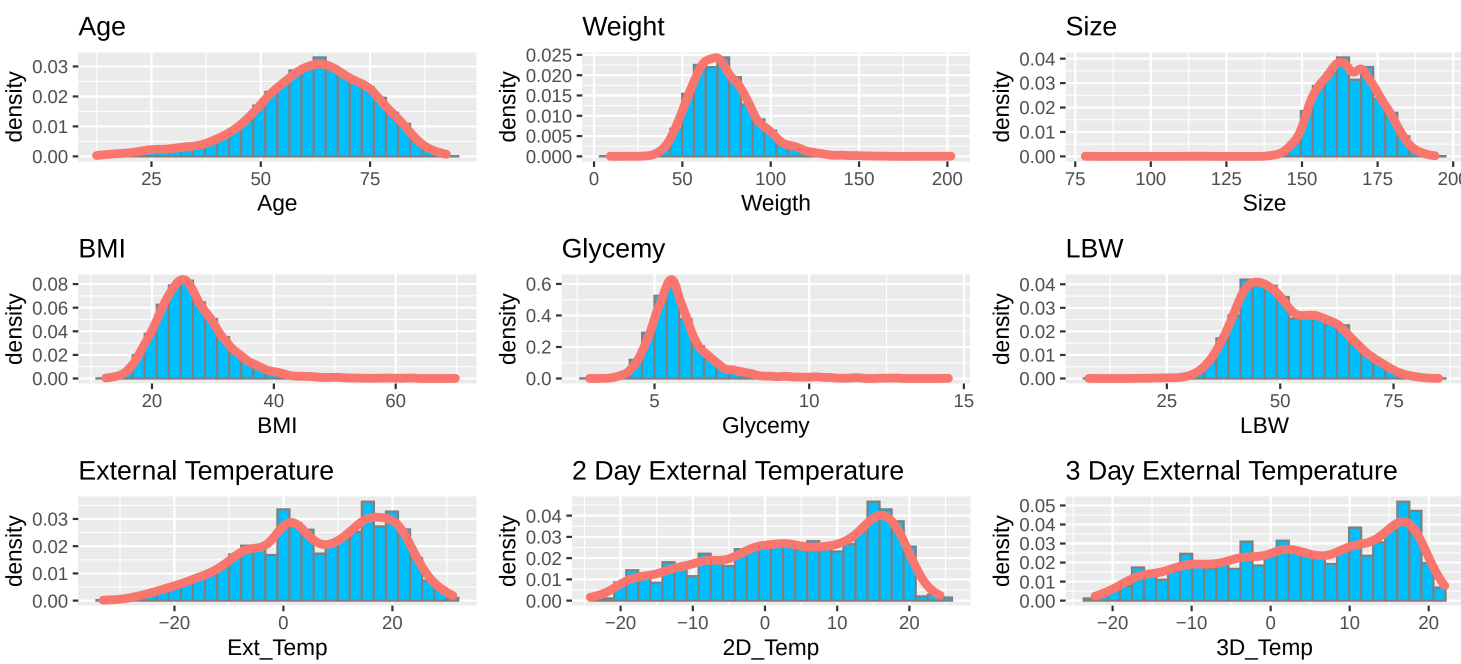
```
brownfat.working.data$Cancer_Status = as.numeric(brownfat.working.data$Cancer_Status)
brownfat.working.data$Cancer_Type = as.numeric(brownfat.working.data$Cancer_Type)
```

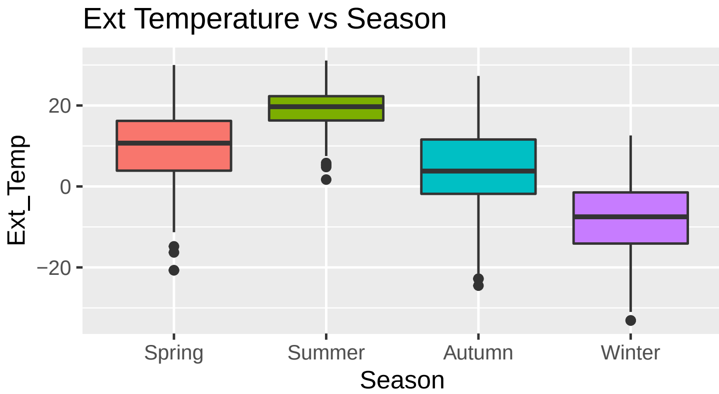
Converted the categorical variables into factors.

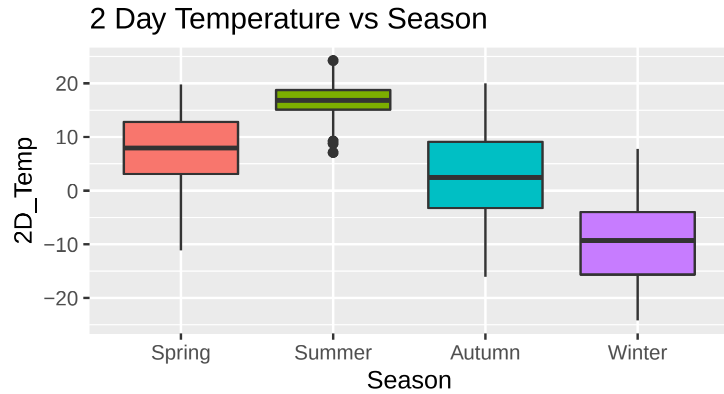
Data Analysis

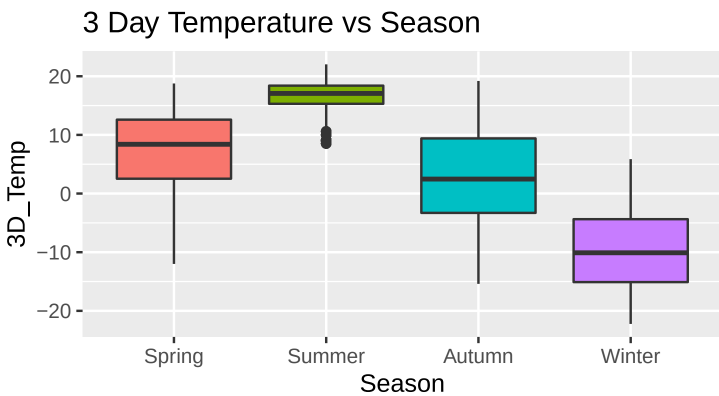
Descriptive statistics for the predictors

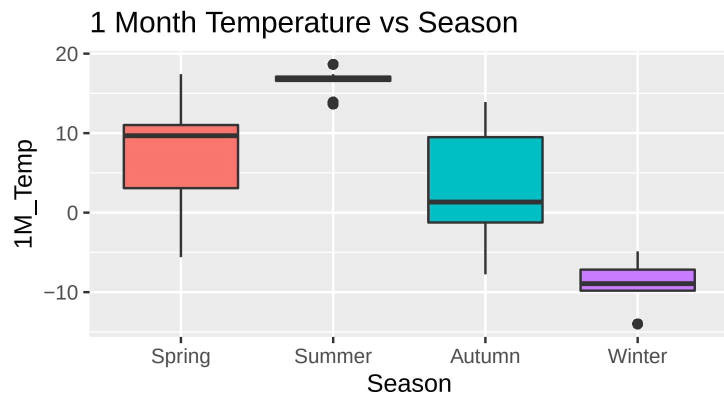






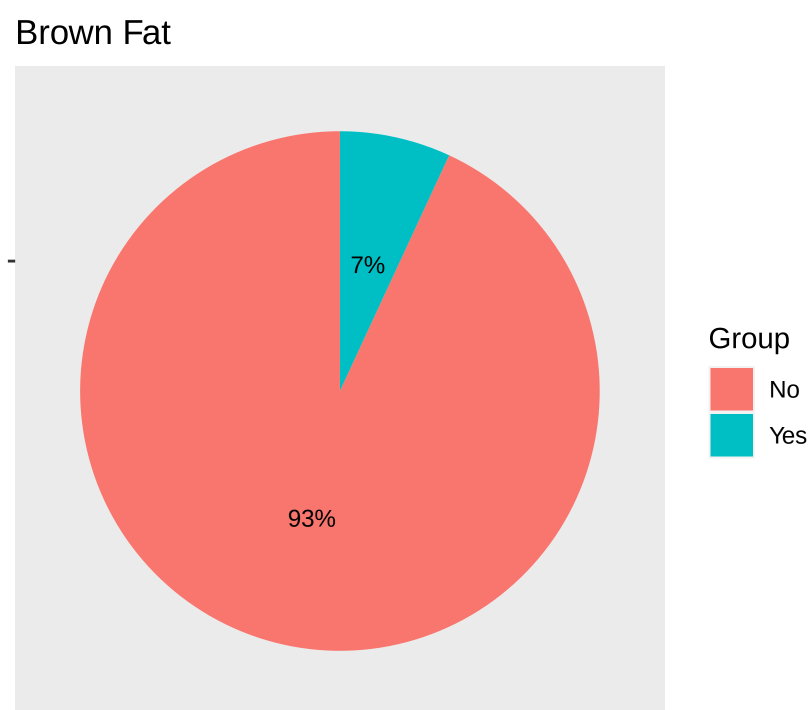




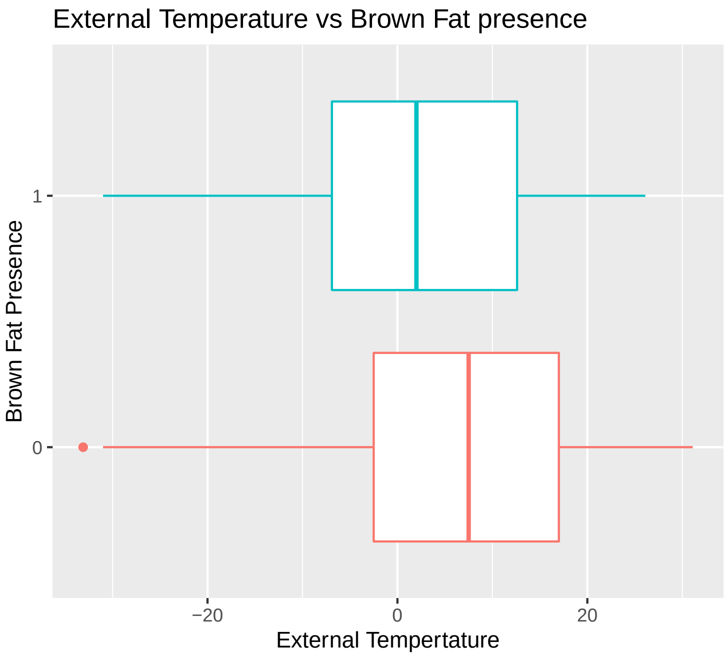


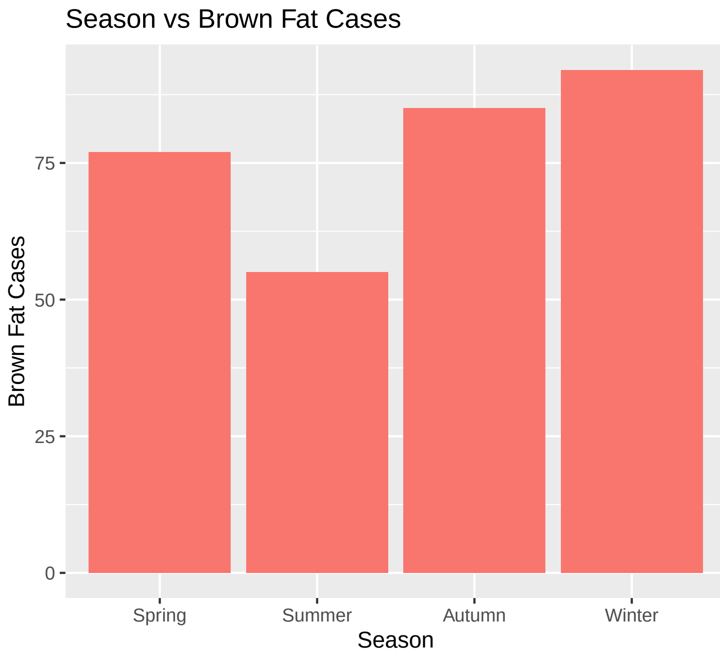
All variables are shown to be approximately normally distributed. In this case of the temperature, it would make sense to observe it in tandem with the month.

Distribution of the response variable



Relationship between the response variable and important predictors





Model Building

No Interactions model

As for building the model for this data, since the response variable is a categorical variable, we need to consider a generalized linear model. Since the response is a 0 or a 1, the family of the model is binomial distribution.

When building generalized linear models, we want to make sure which significant predictors we do not want to remove from the model based off of their P-values. This is due to several predictors often suffering from multicollinearity. As in, the correlation among predictors is meaningless since all variables are in the model; therefore, none of them are relevant to helping with our research. A variable may seem to have little effect because it overlaps considerably with other predictors in the model; in other words, said variable is predicted well by the other predictors.

We run backward, forward, and stepwise elimination, removing variables that don't have significant p-values.

The optimizations are as follows:

Selection.Process	AIC	Significant.Predictors
Backwards Step	1998.809	Ext_Temp, Sex, Diabetes, Age, Season, Weight, LBW
Forwards Step	2002.140	Age, Sex, BMI, Ext_Temp, Diabetes, Season
Stepwise	2002.140	Age, Sex, BMI, Ext_Temp, Diabetes, Season

The backwards elimination model has the least AIC, which makes it the best of our no interactions models. ### Two-Term Interactions model Once again, we run backward, forward, and stepwise elimination, removing variables that don't have significant p-values.

The optimizations are as follows:

Selection.Process	AIC	Significant.Predictors
Backwards Step	1992.204	Ext_Temp, Sex, Diabetes, Age, Season, Weigth, LBW, Ext_Temp:Season, Sex:Diabetes, Sex:LBW, Diabetes:Weigth, Season:LBW
Forwards Step	1988.415	Ext_Temp, Sex, Diabetes, Age, Season, Weigth, LBW, Age:Sex, Age:Diabetes, Weigth:LBW, Age:Weigth, Ext_Temp:Season
Stepwise	1988.415	Ext_Temp, Sex, Diabetes, Age, Season, Weigth, LBW, Age:Sex, Age:Diabetes, Weigth:LBW, Age:Weigth, Ext_Temp:Season

The forward/stepwise model has the least AIC, making it the best of the two-term interactions models.

Comparing the backwards elimination model with no interactions to the forward/stepwise model with two-term interactions, the latter model seems to be better since it has a lower AIC value. As such, we select this model to be our best one of them all.

Model Validation

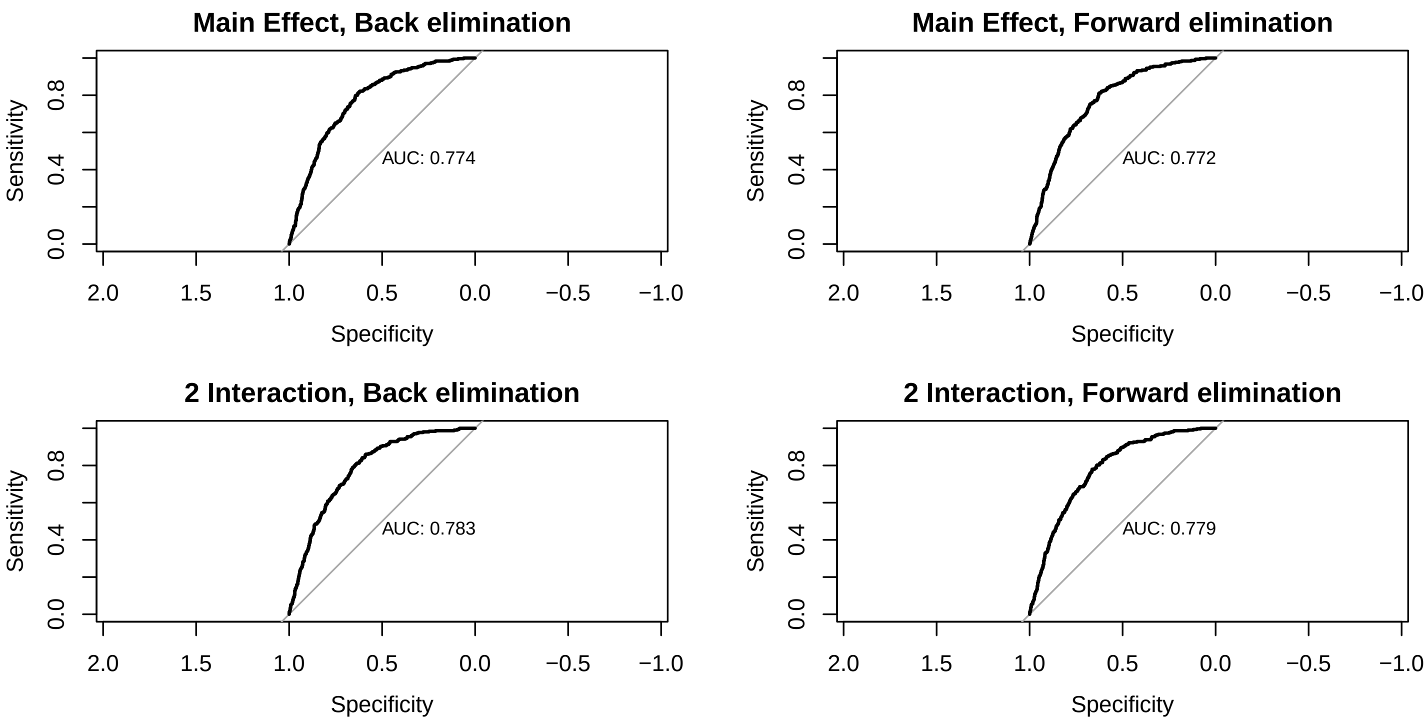
The penultimate step to complete our analysis is to verify whether our best model is valid or not; we perform model validation on said model, as well as our other models to check for whether our best model overfits with our dataset. We conduct 3 different methods of model validation: * Computing the sensitivity, specificity and concordance rates * Plotting the ROC curves * Plotting the pearson residuals against the fitted values

Sensitivity, Specificity, and Concordance Rates

Model	Sensitivity	Specificity	Concordance.Rate
Main Effects with Back Elimination	0.7411003	0.6762728	0.6807512
Main Effects with Forward Elimination	0.7508091	0.6757925	0.6809747
2 interactions with Back Elimination	0.7993528	0.6476945	0.6581712
2 interactions with Forward Elimination	0.7864078	0.6433718	0.6532529

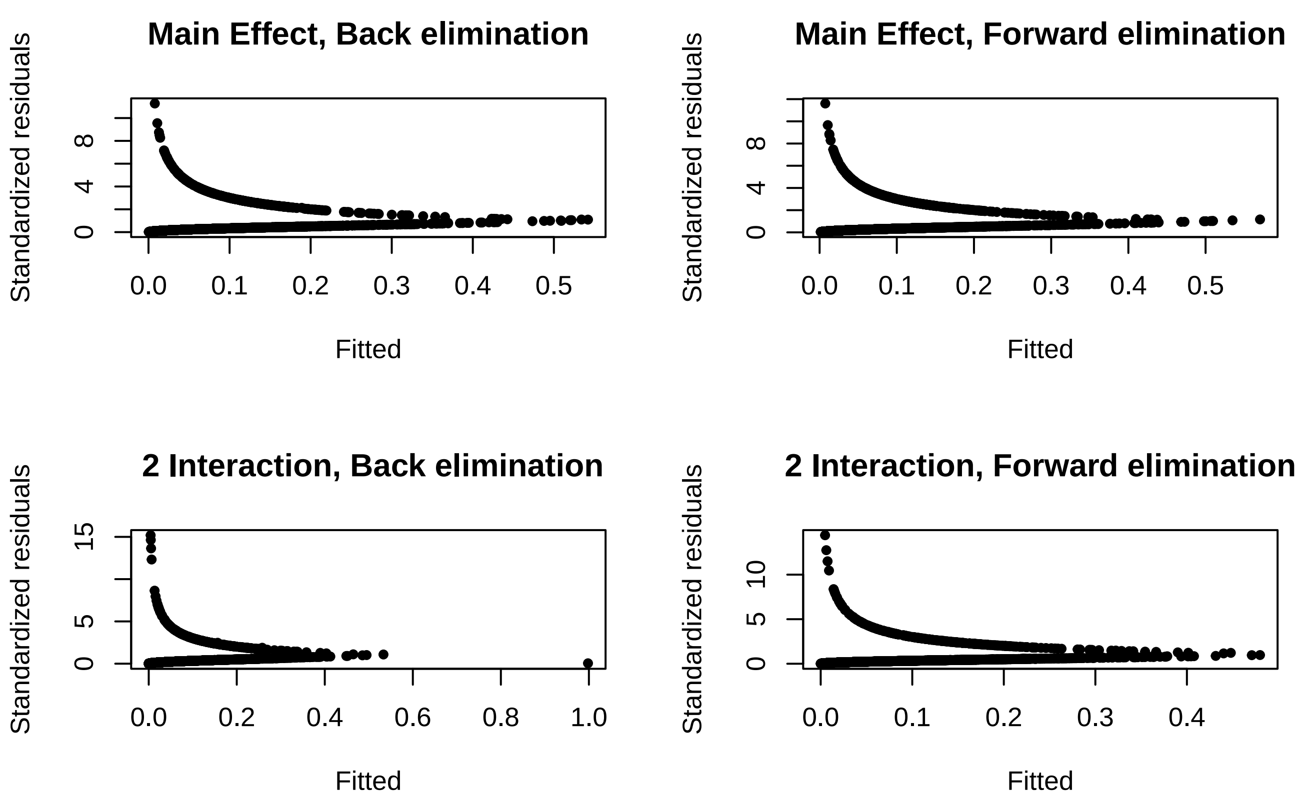
Upon computing every value for each model, we can verify that all of them are over 50%, which passes validation.

ROC Curve



The ROC curves for each model are all concave; furthermore, all concordance indices are fairly large as needed.

Pearson's Residual vs Fitted Values



The pearson residuals appear to be very high for every model, which means that every model does have room for error. That being said, every model still passes validation on the other two aspects, and these results don't show that our best model is any less valid than the other models.

Discussion and Conclusions

Summary

Throughout our case study, we have determined that the best model to represent the factors that affect the probability of having brown fat is the model with two-term interactions found by forward selection. These factors include the following main effects: age, sex, weight, external temperature, having diabetes, lean body weight, and season. There are also the following two-term effects involved: Age:Sex, Age:Diabetes, Weight:Lean Body Weight, Age:Weight, and External Temperature:Season.

This isn't enough, as some factors have a greater significance to the probability of having brown fat than others do, making them more relevant to this study. Upon looking at the factors with the smallest p-values with a confidence level of 90%, we determine the following main effects to be most significant to this case:

- External Temperature; this makes sense as external temperature affects body temperature, which brown fat regulates. With each unit increase of external temperature, the probability of having brown fat decreases by approximately 4.68%.
- Lean Body Weight; this is essentially a measure of body fat, which potentially correlates to having brown fat or not. With each unit increase of lean body weight, the probability of having brown fat increases by approximately 8.83%.
- Season (Winter); brown fat does regulate for extreme cold temperatures, so it would make sense that the coldest season affects the presence of brown fat. When the season is winter, the probability of having brown fat decreases by approximately 33.3%.

And for the two-term effects, the following are significant:

- Age:Sex; Interestingly, neither age nor sex are significant with their main effects, but their interaction suggests that if one was male, then age has a significant effect on the probability of having brown fat. If one was male, then with each unit increase of age, the probability of having brown fat decreases by approximately 2.15%.
- Age:Diabetes; Like before, neither age nor diabetes are significant on their own, but having diabetes suggests that age plays a significant role. If one did have diabetes, then with each unit increase of age, the probability of having brown fat decreases by approximately 6.02%.
- Weight:Lean Body Weight; Lean Body Weight is dependent on weight, so it's no surprise that they share some significance as an interaction term. With each unit increase of weight · lean body weight, the probability of having brown fat decreases by approximately 0.07%
- External Temperature:Season (Winter); both are significant terms, so it makes sense that their interaction is also significant; if the season was winter, then the external temperature would have a greater effect on the probability of having brown fat. When the season is winter, then with each unit increase of external temperature, the probability of having brown fat increases by approximately 3.93%.

Limitations

- Even though our research shows the confirmation of our hypothesis, external temperature affects the activation of brown fat; the inference on other variables can be a hit or miss. The reason is because the number of cases of brown fat is a mere 7, so our results are hardly the true to what happens in real life.

- Human biology is very complicated. Activating brown fat may be a result of many predictors that can go beyond our given dataset, although external temperature is one of the main effects. Therefore, we cannot generalize the findings and apply it outside of the context of the study due to the sample size.

References

[1] Donald Hensrud, M. D. (2020, November 5). Brown Fat: Why you want it. Mayo Clinic. Retrieved April 9, 2022, from https://www.mayoclinic.org/healthy-lifestyle/weight-loss/expert-answers/brown-fat/faq-20058388#:~:text=Brown%20fat%2C%20also%20call

[2] Zhang J, Wu H, Ma S, Gao L, Yu C, Jing F, Zhao J. TSH promotes adiposity by inhibiting the browning of white fat. Adipocyte. 2020 Dec;9(1):264-278. doi: 10.1080/21623945.2020.1783101. PMID: 32579056; PMCID: PMC7469524=