Predictive Biomarkers for Obesity

Taylor Keating, Zichen Liu, Niki Petrakos

December 17, 2021

Scientific Background

Obesity is a major public health problem and increases risk of many health conditions such as coronary heart disease, stroke, certain types of cancers, and all-cause mortality[1]. While obesity is of concern due to the aforementioned reasons, the definition of obesity, which is based solely on body mass index (BMI)[2], is flawed[3]. This lends some difficulty for clinicians to appropriately treat their patients. For example, a patient with a BMI greater than or equal to 30 would be classified as obese, and hence at an increased risk of coronary heart disease, stroke, and cancer. However, some individuals with BMI greater than or equal to 30 are actually at a healthy weight and are not at an increased risk for other diseases[4].

Some research has shown that some physiological pathways are pathways for obesity, such as the insulin and insulin-like growth factor (IGF) axis and chronic inflammation[1]. This evidence, coupled with the flaws surrounding the current working definition of obesity, has sparked interest in the scientific community to create a better working definition of obesity — one that involves a patient's biomarker measurements. This new definition could also be useful for the purposes of personalized preventative medicine, where each patient is treated based on their own individual needs, rather than broad rules of thumb (such as the BMI rule of thumb for classifying obesity).

In a previously-published paper by Nimptsch, Konigorski, and Pischon titled "Diagnosis of obesity and use of obesity biomarkers in science and clinical medicine" (2019), the authors point out the relationship between certain adipokines, including leptin, adiponectin, and resistin, and obesity-associated health outcomes. Moreover, the authors stress the importance of further obesity biomarker research in order to give more insights into obesity-related disease etiology, which could help clinicians better identify which of their patients are at a higher risk for developing certain diseases.

Specific Aims

There are two main aims of this project:

1. Screen for biomarkers that are associated with obesity.
2. Feature-select for biomarkers associated with obesity using prediction-based regression methods (including LASSO linear, Ridge linear, and LASSO polynomial).

Note that the aim of this project is not to create a completely new definition of obesity, one devoid of BMI altogether. The aim of this project is to supplement the current working definition of obesity, using BMI, through the inclusion of biomarkers. Hence, to achieve these aims, obesity will be defined using BMI; however, the findings in this report can inform a predictive model, which would build upon the current definition of obesity by incorporating biomarkers in addition to a patient's BMI.

Data Description

The analysis is based on the utilization of the National Health and Nutrition Examination Survey (NHANES) dataset, which is a program of studies within the Centers for Disease Control and Prevention (CDC) designed to assess the health and nutritional status of adults and children in the United States. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the U.S.,15 of which are sampled from each year. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. In particular, the 2015 to 2016 dataset will be used for the purposes of this analysis.

The 2015 to 2016 NHANES data are available to the public on the CDC website. Prior to performing the analysis, the datasets in the form of XPT files must be downloaded, then read into dataframes using the haven library in R. Data from across multiple questionnaires and modules must be merged on participant ID. After subsetting the data to contain only columns relevant to the aims, columns with biomarker measurements will be renamed for convenience.

Additionally, some variables must be re-coded to fit modeling needs — for example, the binary indicator for obesity will be derived from a continuous BMI variable. Finally, the data will be filtered to only include complete cases. The cleaned dataset will be ready for the analysis methods described below.

Variables

To conduct this analysis, first the outcome of obesity has to be defined. Obesity has been previously classified as having a BMI of 30 or greater. Therefore, the BMI measurement from the data will be dichotomized with a cut-off of greater than or equal to 30 to create the outcome variable. Next, 37 biomarker lab variables will be collected in the NHANES 2015 to 2016 dataset, which could be potential biomarkers for obesity. However, many of these variables are repeats of the same biomarker using a different measurement (for example g/dL and g/L). Therefore, any duplicate measurement variables will be filtered out, thus resulting in 24 unique biomarkers. These variables include measurements on proteins (albumin), enzymes (alkaline phosphatase, aspartate aminotransferase, alanine aminotransferase), metabolism byproducts (bicarbonate, creatinine), as well as other types of biomarkers (such as glucose).

Methods

Screening Approach: To find biomarkers that are associated with obesity, a screening-based approach will be used first. This approach involves pairwise Spearman correlations between the binary outcome of obesity and each of the 24 biomarkers that are included in the lab variables section of the study. First, the dataset will be split 50:50 into training and validation sets. Next, the Spearman correlation between obesity and each biomarker will be calculated on the training data. Next, the biomarkers will be ranked by the absolute value of the correlation obtained. This ranking of biomarkers will be used to sequentially fit linear regression models to the training data using the biomarkers with the top N correlations in absolute value (this is done for $N = 1$ through $N = 24$). For example, the first model will be *obesity ~ biomarker1* and the last model will be

obesity ~ biomarker1 + biomarker2 + … + biomarker24. Finally, the mean squared error (MSE) will be evaluated from the predictions of these linear models on the validation data. The set of biomarkers included in the model that resulted in the lowest MSE when predicting the validation data will be reported as the biomarkers most associated with obesity from this screening-based approach.

Prediction Approach: To build a predictive model for obesity, three models will be fit to the data using 10-fold cross-validation using the package *glmnet* to evaluate which biomarkers together are most predictive of the outcome of obesity. Similar to in the screening-based approach, the outcome will be obesity and the predictors will be the 24 biomarkers. The three models to be fit are a LASSO regression model, a ridge regression model, and a LASSO regression model that includes every biomarker up to its 5th polynomial (to allow for more flexibility in capturing the relationship between each biomarker and obesity).

LASSO and ridge regression are analysis methods that combine feature selection with regularization. They can be particularly useful in situations where there are many candidate features, with some collinearity between features. Regularization in these regression methods is performed by minimizing the least squares and a penalty term. In LASSO regression, this penalty term includes a tuning parameter and $\sum_{i=1}^{p} |\beta_p|$, where β_p are the parameter coefficients. This penalty term in LASSO regression forces some of the parameter coefficients to be zero. In ridge regression, the penalty term includes a tuning parameter and $\sum_{i=1}^{p} (\beta_p)^2$, where β_p are the parameter coefficients. The resulting models from the 10-fold cross-validation for each of the three methods, after deriving the respective tuning parameters that minimize the MSE, will be reported.

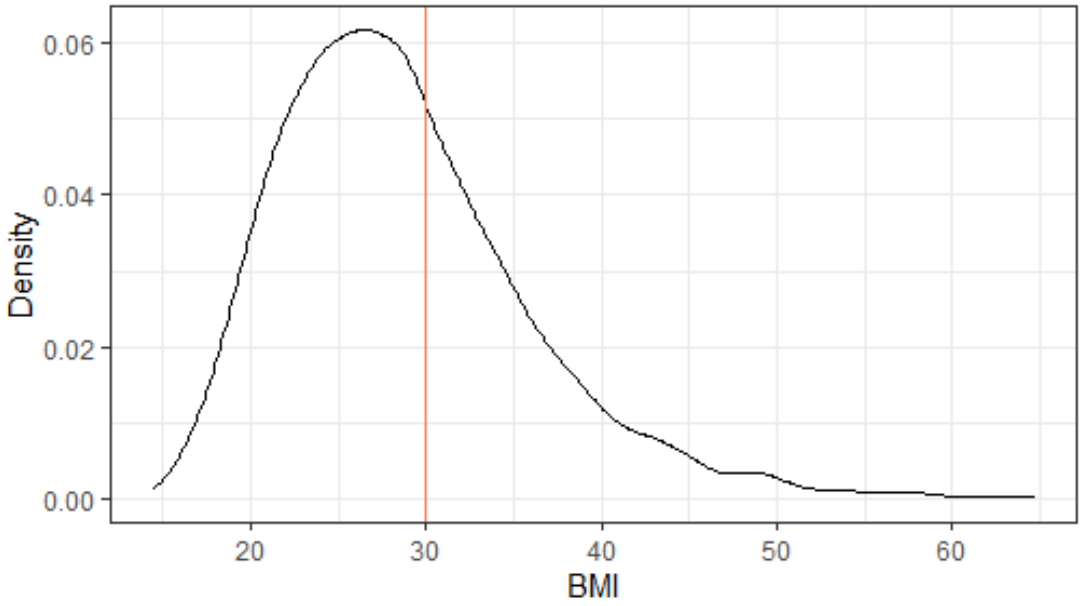
Finally, the specific biomarkers feature-selected for, in each of the four resulting methods (one screening-based and three prediction-based), will be compared.

Results

The analysis performed on the NHANES 2015 to 2016 dataset included a total of 6,744 observations with data on both BMI and biomarker measurements. After filtering to include complete-cases only, there were 6,169 observations used in the analysis.

Figure 1 below shows the distribution of BMI from the dataset, overlaid with a vertical line indicating a BMI of 30, the cut-off point defining the presence of obesity. There were 36% of individuals in the sample classified as being obese based on having a BMI of 30 or greater.

Figure 1. Distribution of BMI in NHANES 2015 to 2016 dataset



Furthermore, **Table 1** below shows the breakdown of the population by weight status defined by BMI cut-offs.

Table 1. Distribution of individuals in NHANES 2015 to 2016 by weight status

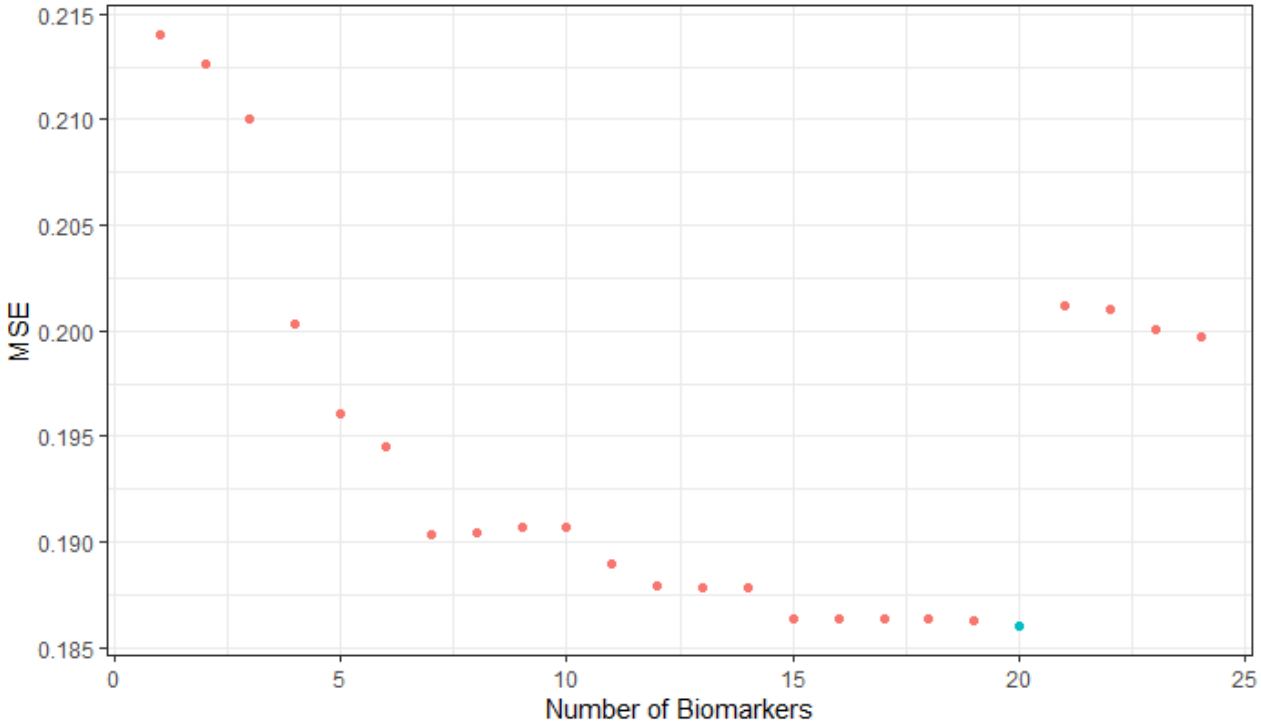
Weight Status	BMI Range	Number of Individuals (%)
Under	Less than 18.5	210 (3.4%)

Healthy	18.5 to 25	1850 (30%)
Over	25 to 30	1867 (30%)
Obese	30 to 40	1788 (29%)
Severe	40 and above	454 (7.3%)

The data are ideal for the intended analysis due to it containing a good representation of all types of weight statuses, and especially due to it featuring a large proportion of individuals with obesity or severe obesity, the population with the outcome of interest.

Screening Approach: First, presented are the results from the screening-based approach using ranking of pairwise Spearman correlations. **Figure 2** below shows the resulting MSE of the models fit on the training data with the top N biomarker correlations, evaluated on the validation set:

Figure 2. MSE of linear models with top N biomarkers ranked on correlation with obesity



From this figure, it is shown that the number of biomarkers included in the linear model that minimizes the MSE when predicting the validation data is 20. The top 20 most correlated biomarkers selected by this method are reported below in **Table 3**.

Prediction Approach: Next, **Table 2** below lists the tuning parameters derived for each of the 3 prediction-based methods after 10-fold cross-validations of 100 potential parameters. Each method resulted in a different tuning parameter that minimized the MSE.

Table 2. Tuning parameters derived from 10-fold cross-validation of 100 λ's

Method	Tuning parameter λ
Ridge linear	0.07
LASSO linear	0.01
LASSO with ≤ 5°polynomials	0.007

Using these tuning parameters, the various biomarkers feature-selected to be most predictive of obesity are listed in **Table 3** below.

Table 3. Selected biomarkers by method

	Did this method select for the biomarker?			
Biomarker Name	**Screening**	**Ridge Linear**	**LASSO Linear**	**LASSO Polynomial (Degrees)**
Albumin refrigerated serum	Y	Y	Y	Y (2°, 3°)
Alkaline Phosphatase		Y	Y	Y (1°, 5°)
Aspartate Aminotransferase		Y	Y	Y (1°, 3°, 4°, 5°)
Alanine Aminotransferase	Y	Y	Y	Y (1°, 3°)

Blood Urea Nitrogen		Y	Y	Y (1°)
Bicarbonate	Y	Y	Y	Y (1°)
Total Calcium	Y	Y		
Cholesterol refrigerated serum	Y	Y		Y (5°)
Creatine Phosphokinase	Y	Y		
Chloride	Y	Y		
Creatinine refrigerated serum	Y	Y	Y	Y (1°)
Globulin	Y	Y	Y	Y (1°)
Glucose refrigerated serum	Y	Y	Y	Y (1°, 4°)
Gamma Glutamyl Transferase	Y	Y		Y (5°)
Iron refrigerated serum	Y	Y	Y	Y (1°)
Potassium		Y		Y (1°)
Lactate Dehydrogenase	Y	Y	Y	Y (1°)
Sodium	Y	Y		
Osmolality	Y	Y		
Phosphorus	Y	Y	Y	Y (1°)
Total Bilirubin	Y	Y	Y	Y (1°)
Total Protein	Y	Y		
Triglycerides refrig serum	Y	Y	Y	Y (1°, 2°)
Uric acid	Y	Y	Y	Y (1°)

Discussion

As shown in **Table 3**, the screening-based method selected 20 biomarkers with the highest correlation with obesity. Using ridge regression selected all 24 biomarkers as being predictive of obesity. In contrast, using LASSO regression selected 15 biomarkers, and implementing polynomials up to 5 degrees to LASSO regression selected 18 unique biomarkers (to at least some polynomial degree) but 26 variables in total (including linear and higher power terms).

Comparing the biomarkers selected by each of the different methods, the following 12 biomarkers were selected to be included in all models:

1. Albumin refrigerated serum
2. Alanine Aminotransferase
3. Bicarbonate
4. Creatinine refrigerated serum
5. Globulin
6. Glucose refrigerated serum
7. Iron refrigerated serum
8. Lactate Dehydrogenase
9. Phosphorus
10. Total Bilirubin
11. Triglycerides refrig serum
12. Uric acid

These are the biomarkers included in the NHANES study that were unanimously determined to be predictive of obesity by all four methods and thus could be of interest to clinicians attempting to develop a personalized prevention method for obesity. Note that these results are in concordance with the literature, to some extent, as there are biomarkers that can be used to predict obesity for different individuals. However, the biomarkers chosen in the analysis of this project differ from the biomarkers chosen in the Nimptsch et al. (2019) paper.

For future work, it would be meaningful to repeat this analysis with a different starting definition of obesity. Rather than using BMI, it would be interesting to see how the results may or may not differ using a definition of obesity that is for example, based on waist size or other scientifically-sensible forms of identifying patients as being obese or not. This may also be of scientific interest since it has already been established that basing obesity on BMI is flawed.

References:

1. Nimptsch, K., Konigorski, S., & Pischon, T. (2019). Diagnosis of obesity and use of obesity biomarkers in science and clinical medicine. Metabolism, 92, 61–70. https://doi.org/https://doi.org/10.1016/j.metabol.2018.12.006
2. Centers for Disease Control and Prevention. (2021, June 7). Defining adult overweight & obesity. Centers for Disease Control and Prevention. Retrieved December 11, 2021, from https://www.cdc.gov/obesity/adult/defining.html.
3. Devlin, K. (2009, July 4). Top 10 reasons why the BMI is bogus. NPR. Retrieved December 11, 2021, from https://www.npr.org/templates/story/story.php?storyId=106268439.
4. Labos, C. (2019, February 8). Can you be obese but still be healthy? Office for Science and Society. Retrieved December 11, 2021, from https://www.mcgill.ca/oss/article/health-you-asked/can-you-be-obese-still-be-healthy.

Code Appendix

Taylor Keating, Zichen Liu, Niki Petrakos

December 17, 2021

Prepare the data

Read in the NHANES 2015-2016 biomarker and outcome data from .XPT files.

```
biomarkers <- read_xpt(paste0(getwd(), "/BIOPRO_I.XPT"))
outcome <- read_xpt(paste0(getwd(), "/BMX_I.XPT"))
```

Subset the outcome to only include ID and obesity variables, then create indicator variable for obesity.

```
outcome <- outcome %>% select("SEQN", "BMXBMI")
outcome$Obesity <- ifelse(outcome$BMXBMI >= 30, 1, 0)
```

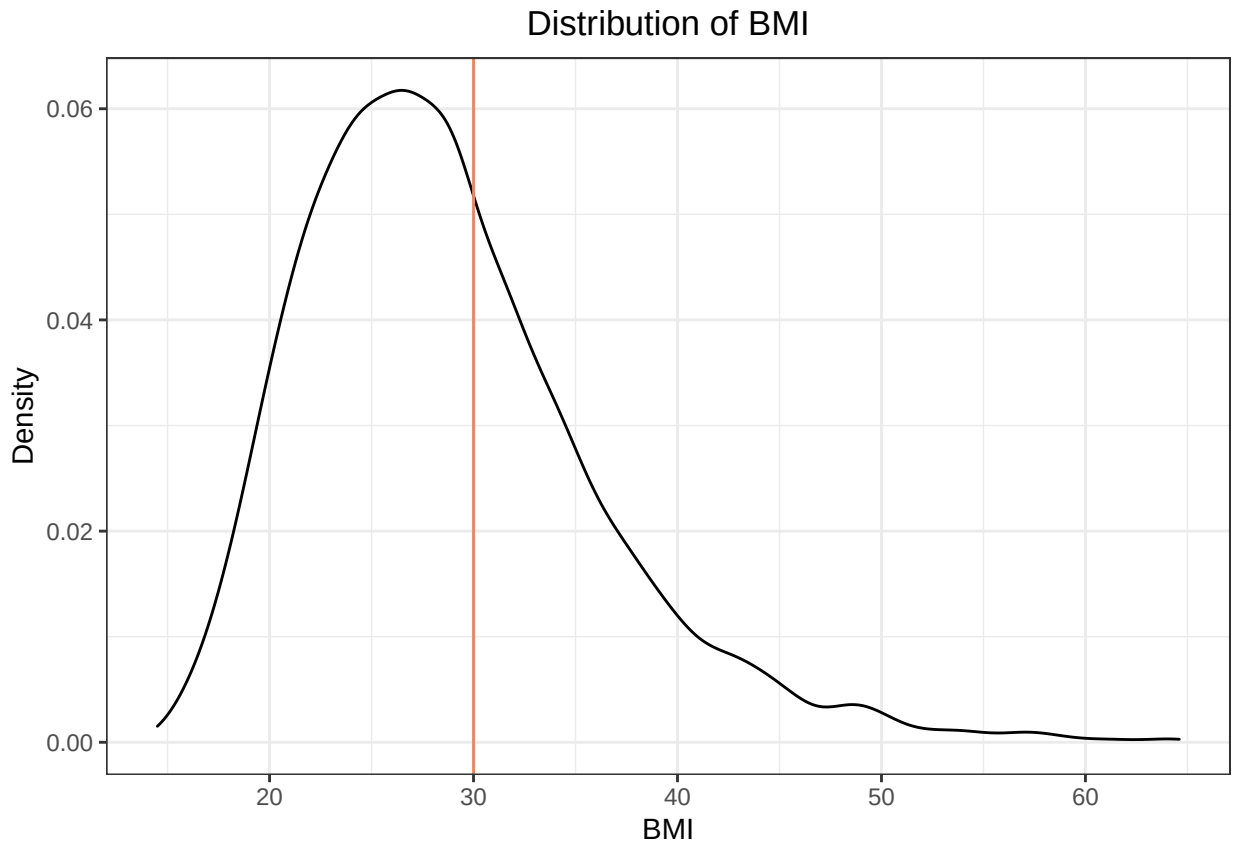
Merge outcome data and biomarker data, then exclude incomplete cases.

```
data <- inner_join(outcome, biomarkers, by = "SEQN")
data <- data[complete.cases(data),]
```

Exploratory analysis

Plot the distribution of BMI in our data.

```
ggplot(data = data, aes(x = BMXBMI)) + geom_density() +
  geom_vline(xintercept = 30, col = "coral") +
  labs(x = "BMI", y = "Density", title = "Distribution of BMI") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```




```
prop_obese <- sum(data$Obesity == 1) / nrow(data)
```

In our dataset, 0.36 of the population are obese based on BMI. We can view a detailed breakdown of the population by weight status, determined by BMI cut-offs.

```
data <- data %>% mutate("Weight_Status" = case_when(BMXBMI < 18.5 ~ "Under",
                                                    BMXBMI >= 18.5 & BMXBMI < 25 ~ "Healthy",
                                                    BMXBMI >= 25 & BMXBMI < 30 ~ "Over",
                                                    BMXBMI >= 30 & BMXBMI < 40 ~ "Obese",
                                                    BMXBMI >= 40 ~ "Severe",
                                                    TRUE ~ "NA"))
data %>% group_by(Weight_Status) %>% summarize(N = n()) %>% kable()
```

Weight_Status	N
Healthy	1850
Obese	1788
Over	1867
Severe	454
Under	210

Screening-based method

Further subset the data to only columns required for implementing the screening-based method.

```
data_clean <- data %>% select(-c("SEQN", "BMXBMI", "Weight_Status"))
data_clean <- data_clean %>% select(Obesity, everything())
```

Rename all biomarker columns to the proper labels.

```
labels <- get_label(biomarkers)
labels <- c(Obesity = "Obesity", labels)
colnames(data_clean) <- lapply(1:ncol(data_clean), function(x) labels[names(data_clean)[x]][[1]])
```

Some biomarker columns are repeats but with different units in parentheses; remove these repeated columns.

```
colnames(data_clean) <- str_remove(colnames(data_clean), "\\s\\((.*)\\)")
data_clean <- data_clean[, !duplicated(colnames(data_clean))]
```

Additionally, the commas and spaces in the biomarker names must be removed for the linear models to be fit without errors.

```
colnames(data_clean) <- colnames(data_clean) %>%
  str_remove(",") %>%
  str_replace_all(" ", "_")
```

Next, split the dataset 50/50 into training and validation sets.

```
set.seed(1)
n <- nrow(data_clean)
train <- sample(n, size = floor(0.5*n), replace = FALSE)
data_train <- data_clean[train, ]
data_test <- data_clean[-train, ]
```

Define a function `calc_corr` to calculate the spearman correlation between obesity and one biomarker.

```
calc_corr <- function(biomarker, data){
  return(cor(data[biomarker+1], data[1], method="spearman"))
}
```

Apply the function to calculate pairwise-correlations between obesity and all biomarkers in the training set. Then, rank the biomarkers from largest to smallest correlations.

```
corrs <- do.call(rbind, lapply(1:(ncol(data_train)-1), calc_corr, data = data_train))
colnames(corrs)[1] <- "Correlations"
corrs_sort <- data.frame(correlation = corrs[order(-abs(corrs)),])
```

Define a function `calc_mse` that fits a linear model with the N top ranked biomarkers on the training set, then calculates the mean squared error when predicting the validation set.

```
calc_mse <- function(n, data_train, data_test){
  biomarkers <- paste0(rownames(corrs_sort)[1:n], collapse = "+")
  model <- lm(eval(parse(text = paste0("Obesity~", biomarkers))), data = data_train)
  pred <- predict(model, data_test)
  errors <- data_test$Obesity - pred
  return(mean(errors^2))
}
```

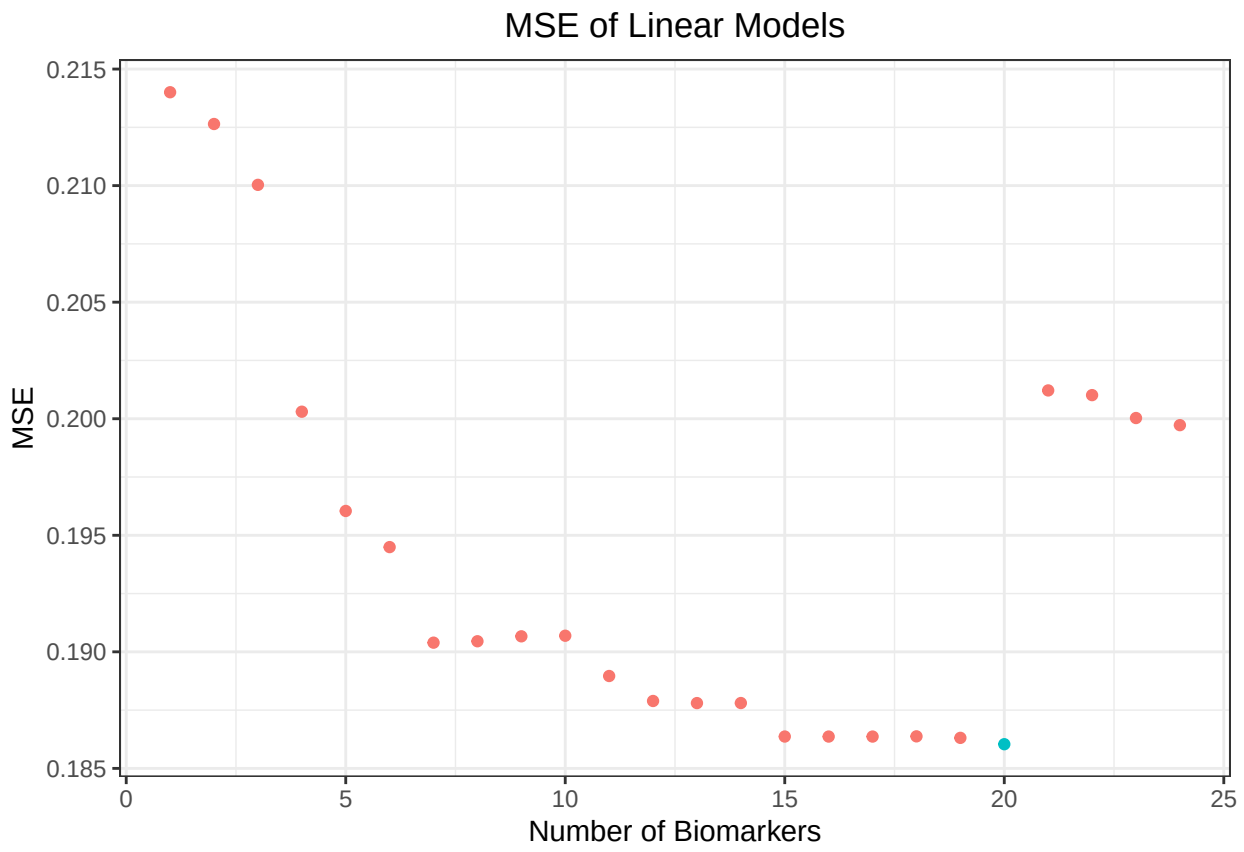
Apply the function to calculate the MSE for each fitted model from one biomarker to all biomarkers included.

```
mses <- do.call(rbind, lapply(1:(ncol(data_test) - 1), calc_mse,
                              data_train = data_train, data_test = data_test))
```

Display the calculated MSE's from each linear model graphically.

```
mse_df <- data.frame("Param" = 1:nrow(mses), "MSE" = mses)
mse_min <- which.min(mse_df$MSE)
mse_df$Min <- ifelse(mse_df$Param == mse_min, 1, 0)

ggplot(data = mse_df, aes(x = Param, y = MSE, col = as.factor(Min))) + geom_point() +
  labs(x = "Number of Biomarkers", y = "MSE", title = "MSE of Linear Models") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```



The number of biomarkers that minimizes the MSE is 20.

```
final_biomarkers_screen <- paste(rownames(corrs_sort)[1:mse_min], collapse = ", ")
```

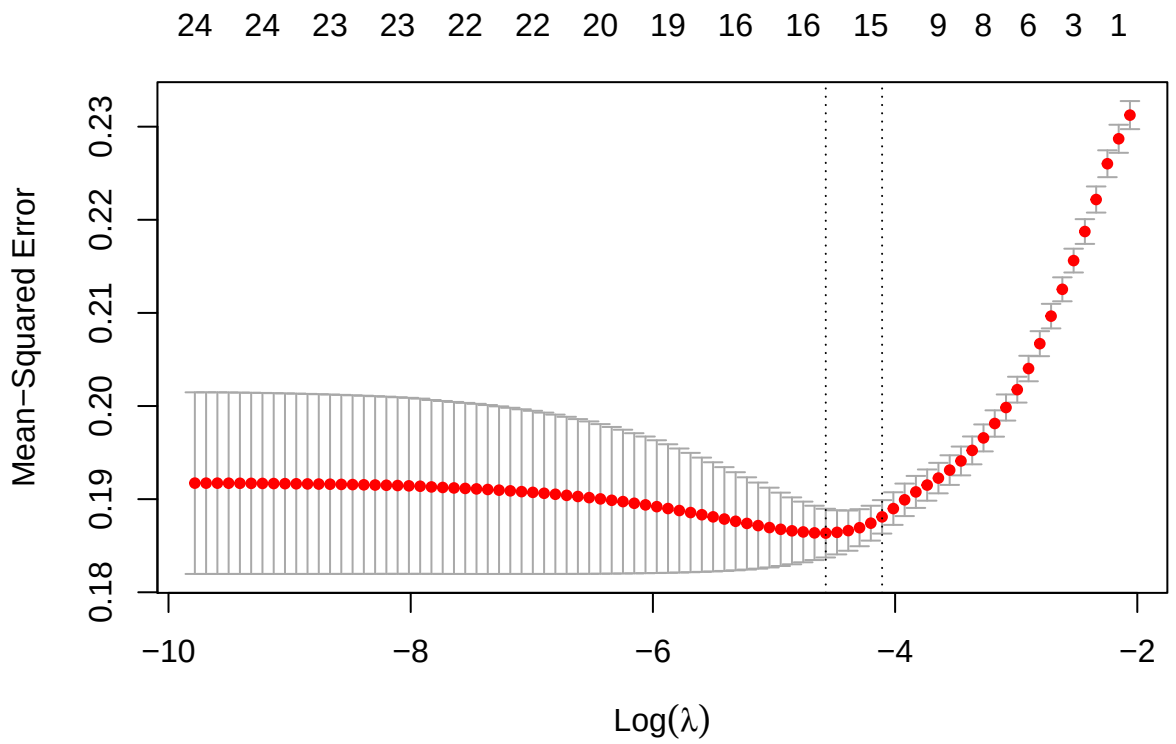
The 20 biomarkers that were chosen from the screening-based method to be most correlated with the outcome of obesity are Albumin_refrigerated_serum, Gamma_Glutamyl_Transferase, Triglycerides_refrig_serum, Uric_acid, Alanine_Aminotransferase, Glucose_refrigerated_serum, Iron_refrigerated_serum, Globulin, Lactate_Dehydrogenase, Total_Calcium, Phosphorus, Total_Bilirubin, Bicarbonate, Cholesterol_refrigerated_serum, Creatinine_refrigerated_serum, Osmolality, Creatine_Phosphokinase, Total_Protein, Chloride, Sodium.

Prediction-based methods

Lasso regression with linear models

We use `glmnet` for 10-fold cross-validation on 100 λ values to determine the tuning parameter that minimizes the MSE.

```
set.seed(456)
nlambda <- 100
fit_lasso <- cv.glmnet(x = as.matrix(data_clean)[,-1], y = as.matrix(data_clean)[,1],
                       alpha = 1, nlambda = nlambda)
plot(fit_lasso)
```



Get the coefficients produced by this λ to see which biomarkers were selected for.

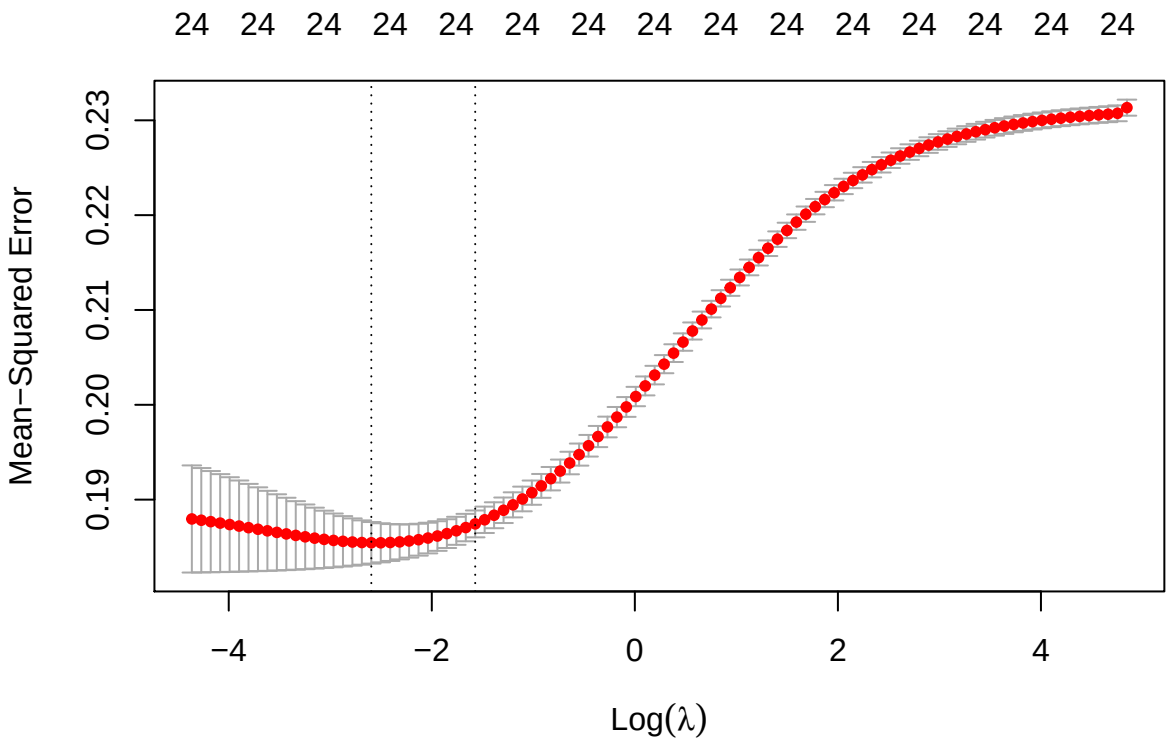
```
coef_lasso <- summary(coef(fit_lasso, s = "lambda.min"))
results_lasso <- lapply(list(coef_lasso$i), function(x) names(data_clean)[x])[[1]]
final_biomarkers_lasso <- paste(results_lasso[2:length(results_lasso)], collapse = ", ")
```

The 15 biomarkers that were chosen from using lasso regression with linear models to be most correlated with the outcome of obesity are Albumin_refrigerated_serum, Alkaline_Phosphatase, Aspartate_Aminotransferase, Alanine_Aminotransferase, Blood_Urea_Nitrogen, Bicarbonate, Creatinine_refrigerated_serum, Globulin, Glucose_refrigerated_serum, Iron_refrigerated_serum, Lactate_Dehydrogenase, Phosphorus, Total_Bilirubin, Triglycerides_refrig_serum, Uric_acid.

Ridge regression with linear models

We use `glmnet` for 10-fold cross-validation on 100 λ values to determine the tuning parameter that minimizes the MSE.

```
set.seed(789)
nlambda <- 100
fit_ridge <- cv.glmnet(x = as.matrix(data_clean)[,-1], y = as.matrix(data_clean)[,1],
                       alpha = 0, nlambda = nlambda)
plot(fit_ridge)
```



Get the coefficients produced by this λ to see which biomarkers were selected for.

```
coef_ridge <- summary(coef(fit_ridge, s = "lambda.min"))
results_ridge <- lapply(list(coef_ridge$i), function(x) names(data_clean)[x])[[1]]
final_biomarkers_ridge <- paste(results_ridge[2:length(results_ridge)], collapse = ", ")
```

The 24 biomarkers that were chosen from using ridge regression with linear models to be most correlated with the outcome of obesity are Albumin_refrigerated_serum, Alkaline_Phosphatase, Aspartate_Aminotransferase, Alanine_Aminotransferase, Blood_Urea_Nitrogen, Bicarbonate, Total_Calcium, Cholesterol_refrigerated_serum, Creatine_Phosphokinase, Chloride, Creatinine_refrigerated_serum, Globulin, Glucose_refrigerated_serum, Gamma_Glutamyl_Transferase, Iron_refrigerated_serum, Potassium, Lactate_Dehydrogenase, Sodium, Osmolality, Phosphorus, Total_Bilirubin, Total_Protein, Triglycerides_refrig_serum, Uric_acid.

Lasso regression with polynomial models

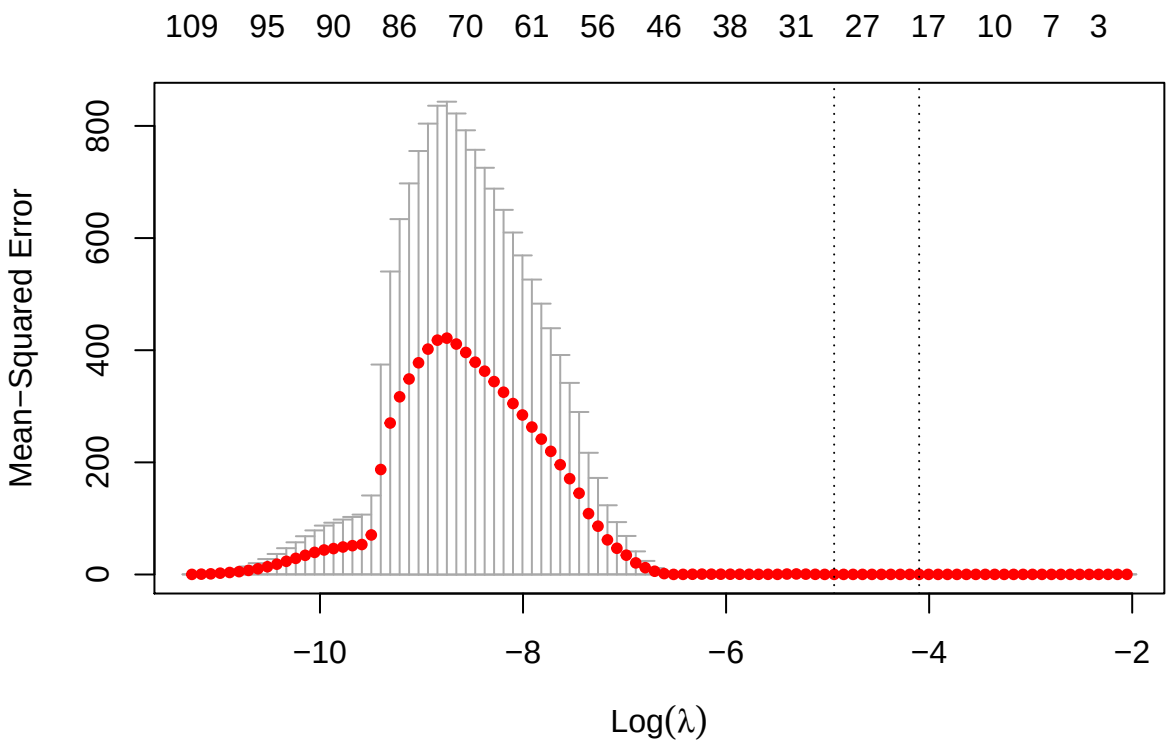
Add polynomial terms up to 5 degree for each biomarker.

```
max_degree <- 5
selected_matrix <- cbind(data_clean$Obesity,
                         matrix(apply(data_clean[, -1], 2,
                                      FUN = poly, degree = max_degree, raw = T),
                                nrow = nrow(data_clean),
                                byrow = F))

colnames(selected_matrix)<- c("Obesity", sapply(names(data_clean[, -1]),
                                                FUN = paste, 1:max_degree, sep = "_poly_"))
```

We use `glmnet` for 10-fold cross-validation on 100 λ values to determine the tuning parameter that minimizes the MSE.

```
set.seed(123)
nlambda <- 100
fit_poly <- cv.glmnet(x = selected_matrix[, -1], y = selected_matrix[, 1],
                      alpha = 1, nlambda = nlambda)
plot(fit_poly)
```



Get the coefficients produced by this λ to see which biomarkers were selected for.

```
coef_poly <- summary(coef(fit_poly, s = "lambda.min"))
results_poly <- lapply(list(coef_poly$i), function(x) colnames(selected_matrix)[x])[[1]]
final_biomarkers_poly <- paste(results_poly[2:length(results_poly)], collapse = ", ")
```

The 26 biomarkers and their polynomials that were chosen from using lasso regression with polyomials models to be most correlated with the outcome of obesity are Albumin_refrigerated_serum_poly_2, Albumin_refrigerated_serum_poly_3, Alkaline_Phosphatase_poly_1, Alkaline_Phosphatase_poly_5, Aspartate_Aminotransferase_poly_1, Aspartate_Aminotransferase_poly_3, Aspartate_Aminotransferase_poly_4, Aspartate_Aminotransferase_poly_5, Alanine_Aminotransferase_poly_1, Alanine_Aminotransferase_poly_3, Blood_Urea_Nitrogen_poly_1, Bicarbonate_poly_1, Cholesterol_refrigerated_serum_poly_5, Creatinine_refrigerated_serum_poly_1, Globulin_poly_1, Glucose_refrigerated_serum_poly_1, Glucose_refrigerated_serum_poly_4, Gamma_Glutamyl_Transferase_poly_5, Iron_refrigerated_serum_poly_1, Potassium_poly_1, Lactate_Dehydrogenase_poly_1, Phosphorus_poly_1, Total_Bilirubin_poly_1, Triglycerides_refrig_serum_poly_1, Triglycerides_refrig_serum_poly_2, Uric_acid_poly_1.