1. Calories

*Summary*

In this calories context, our goal is to obtain the “best” model by which we can get the most accurate prediction for a given value of X. Thus, I divide my report into 3 parts: data preprocessing, model fitting and model prediction.

For data preprocessing, I deal with some problems including: multicollinearity, checking model assumptions, influential observations and outliers.

For model fitting, I choose the model under some constraints of the form (1) and conclude that the best model is like the form (2).





For prediction, I gain a new set of data and predict the calories of it.

* 1. Data Preprocessing

1.1.1 Data Structure and Analysis

We have the dataset about “common house food“, which includes the ingredients for each food items. So we need to first analyze which variable should be included into the model.

It includes 22 columns and 962 observations. Since there is one row is missing and I remove it from the dataset at first.

After that, we should get a general idea of the dataset, I classify the data into the following shape:

|  |  |
| --- | --- |
| Food Items |  |
| Weight (in grams) |  |
| Water (in grams) |  |
| KCal |  |
| Protein (in grams) |  |
| Cholesterol (in mg) |  |
| Carbohydrates (in grams) |  |
| Fats (in grams) | Fat  SatFat  MonoUnSatFat  PolyUnSatFat |
| Minerals (in mg) | Ca  P  Fe  K  Na |
| Vitamins | VitaA (IU)  VitaA (RE)  Thiamin (in mg)  Riboflavin (in mg)  Niacin (in mg)  VitaC (in mg) |

Table 1.1 Data Structure on Common House Food Dataset

Based on the above data classification table, I decide to include the following 8 variables in the initial model and each variable may have one or more items:

Water, Weight, Protein, Carbohydrates, Cholesterol, Vitamins, Minerals, Fats.

1.1.2 Data Diagnostics

1.1.2.1 Multicollinearity

(1) Detecting Multicollinearity Using Variance Inflation Factors

The idea scenario is that all the variables are uncorrelated. So for explanatory variables, we should do diagnostics for multicollinearily.

The full model is in the form:



Since there are 20 variables in the model, I choose to use variance inflation factor to do diagnostics for multicollinearily.

Variance for beta(k) is defined as follows:



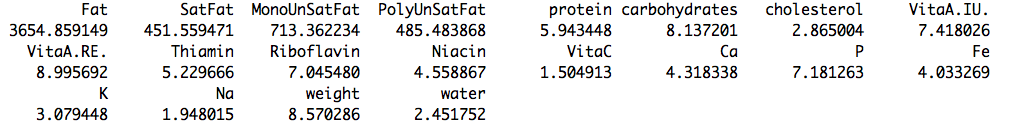


Table 1.2 VIF for explanatory variables for full model

We find that 4 of them (all regarding fats items ) are even greater than 10. So, we can conclude that the multicollinearily among these 4 items will have a large impact on the inference.

Also, the mean of all the VIF values is 269.4273, which is greater than 1. So, there may be serious multicollinearily problems and we should do some transformation for the data.

(2) Transformation for Reducing Data-based Multicollinearity

In order to reduce the multicollinearily, we could opt to remove some insignificant variables out of the predictors from the model.

Alternatively, if we have a good scientific reason for needing both of the predictors to remain in the model, we could go out and collect more data and then will reduce the multicollinearily among the predictors.

For the context in this problem, I choose to use the first method and by using the stepwise function in R and with BIC+Both criterion since it better for result of model fitting. The result is below in the table:

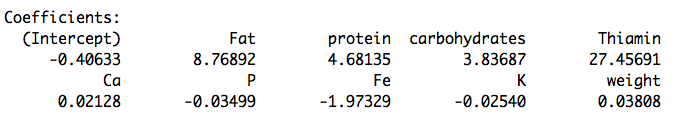


Table 1.3 Coefficients for the Reduced Model

Then we remove 11 insignificant variables (VitaA.IU., VitaA.RE., Niacin, water, PolyUnSatFat, SatFat, MonoUnSatFat, cholesterol, Na, Riboflavin, VitaC) from the original full model, and the reduced model is the following form:



Again, I use variance inflation factor to do diagnostics for multicollinearily. Then the result is in the following table.

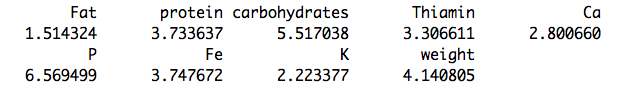


Table 1.4 VIF for Explanatory Variables for Reduced Model

Then, we find that all of the item fat is smaller than 10. Also the mean of all the VIF values is 3.72818, which is smaller than the full model. The result is much better than the original data. So we can conclude that the effect of multicollinearity on the model inference is much smaller than before.

1.1.2.2 Regression Assumptions

(1) Diagnostic Plots

Based on the model selected in the I take studentized residuals to check the assumptions. If the studentized residuals are large, it means that the observations have outliers and we need to do transformation on the data. The results can be seen from the following figures:

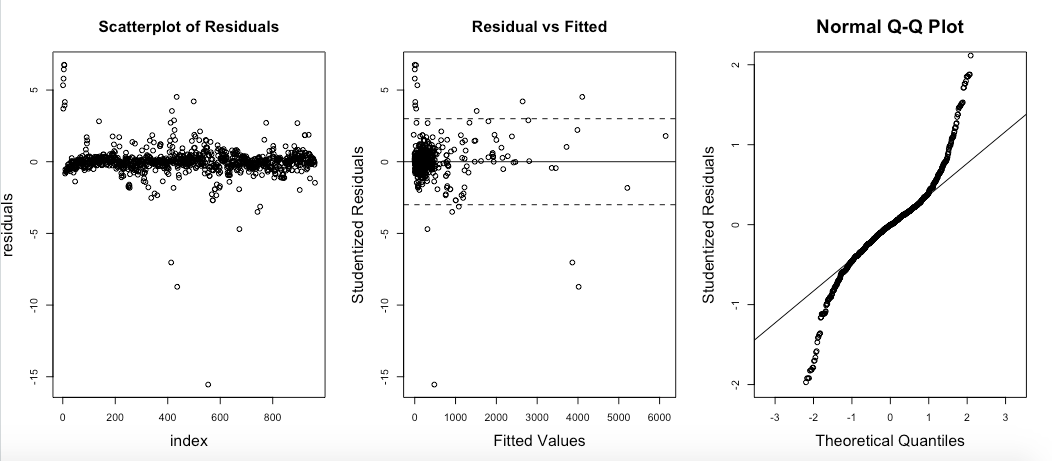


Figure 1.5 Checking Regression Assumptions on Original dataset

Assumption 1: Independence



Based on the scatterplot of residuals, we can conclude that the assumption of independence is satisfied. So we do not need to do any transformation about the data.

Assumption 2: Linearity



From the data pattern, we could easily find that the model is linear regression.

Assumption 3: Equal Variance



Based on the plot of residuals v.s. fitted, we find that the variance of residuals do not have distinct structures and then we conclude that the assumption of equal variance is nearly satisfied well, so we do not need to do some transformation.

Assumption 4: Normal Distribution



Based on the normal QQ plot, we conclude that the data is not fully from normal distribution, so we may need to do some kinds transformation about the data.

(2) Boxcox Checking

I choose to use boxcox to judge if I indeed need to do some transformation on the dataset.

By doing boxcox on the reduced model, we can easily find based on the following figure that lambda is 1 and it means that we do not need to do any transformation of powerful number. I think maybe it’s because the model is not so sensitive to the assumption of normal distribution that we do not need to completely access this assumption.

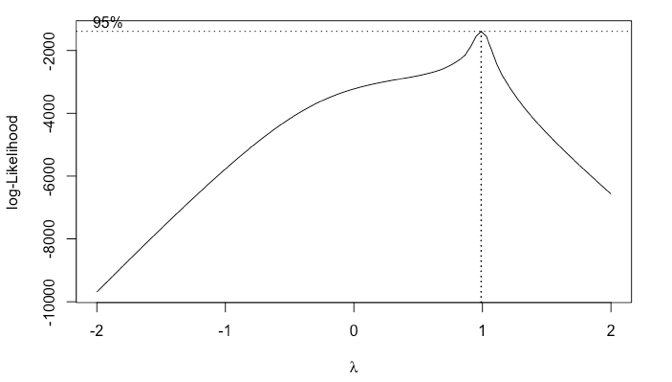


Figure 1.6 Boxcox of for Reduced Model

1.1.2.3 Measures of influence

I take studentized residuals to check the influence. If the studentized residuals are large, it means that the observations are influential points. I use the following four methods to access it: Leverage, DFFITS, Cook’s distance, DFBETAS. We can all four plots shows that there may be some influential observations in the data. They may not all be outliers, so we need to check further.

601%20final%20figure/Rplot08.pdf

Figure 1.7 Identification of Influential Observations

1.1.2.4 Outliers

In order to find the outliers, we plot the reduced model and we can identify from the figure below that the observation of 413, 436, 554 are three outliers. So we need to drop these three data point as the first step.

Rplot.pdf

Figure 1.8 Plot of Reduced Model

After dropping these three outliers, we fit the new reduced model with new dataset and plot the model again. I find that this time the outlier is not so extinct. So we stop and use these data to continue the analysis.

Rplot03.pdf

Figure 1.9 Plot of Reduced Model without Outliers

* 1. Model Fitting

We consider a model in the following term:



And then, we need to come up with a procedure to obtain the best model.

I choose to use the best subsets method with various model selection criteria. A best subsets procedure identifies a group of subset models that gives the best values of a given criterion. Since the number of explanatory variables in this context is large, all possible best best subsets may not be feasible. In this case, a stepwise selection procedure offers a feasible approach.

I use “myregsub.R” to as a procedure to choose the best model under the constraint of the model with seven explanatory variables.

Then, based on the result, I find that the best model which fits the requirements includes the following seven variables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of variables | Weight  Protein  Fat  Chol  Carb  K  Thiamin | rsq | rss | adjr2 | cp | bic |
| 7 | 0.9990175 | 278026.0 | 0.9990103 | 92.18697 | -6600.344 |

Table 1.10 Best Subsets Procedure

The final model now is like this way:



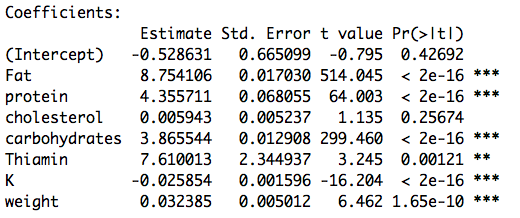


Table 1.11 R Summary of Final Model

And also we can find that multiple R-squared is 0.9993, which means that this model can perfectly explain 99.93% of the variables, which will lead to a highly accurate prediction for a given value of X.

* 1. Prediction

In order to test my model’s prediction power, I need to use the new data to predict and get it’s performance.

I gain the following new dataset:

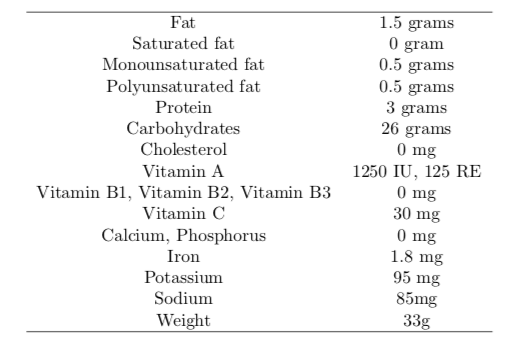


Table X value for Prediction

Since my final model only includes 7 variables: Weight, Protein, Fat, Chol, Carb, K, Thiamin. I just put these 7 into the final model.

Then, I get the result as follows:

(1) Predictive Value



(2) Predictive Interval



