Project

Xuqing Luo, Peixi Jiang (Section D)

2018/4/24

Abstract

In our project, we use data about Starbucks drinks. The goal of our modeling is to find out that among all the data provided, which variables actually affect the calories of drinks in Starbucks. Through exploratory data analysis step, we eliminate variables dietary fiber, vitamin A, vitamin C, caffeine, and saturated fat; using the Vision Rule 10%, we eliminate categorical beverage and create two indicators (size and milk) for beverage prep after using F-test; we decide to transform two predictors (iron and calcium) into polynomial of third power after comparing r-squared and standard error; after checking correlation, we add "sugar*carbohydrate-cholesterol" to our model- RoughModel. Then, we use best subset selection and find out that our BSSModel is good with a high adjusted r square. After conditions check, we get to our final model:$\widehat{Calories} = 2.8075+10.1957(Somthiesyes)+0.4482(TotalFat)+10.1478(TransFat)+4.0403(Sugar)+6.4179(Protein) - 0.5029(Sodium) - 15.6606(nonfatmilkyes) + 12.6142(soymilkyes) - 4.1077(sizeVenti)$. We are comfortable with this model because it explains more than 90% of the variance in the response variable.

Data

Tatman, R. (Ed.). (2017, July 20). Nutrition facts for Starbucks Menu. Retrieved May 1, 2018, from https://www.kaggle.com/starbucks/starbucks-menu

Our data is about different Starbucks drinks and their contents. Each row in the dataset corresponds to the contents of one particular type of drink. We decide to use calories as our response variable, and the potential predictors might be sugar, total fat, protein, trans fat, cholesterol, and carbohydrate. There are 242 rows in our dataset with no missing data.

EDA

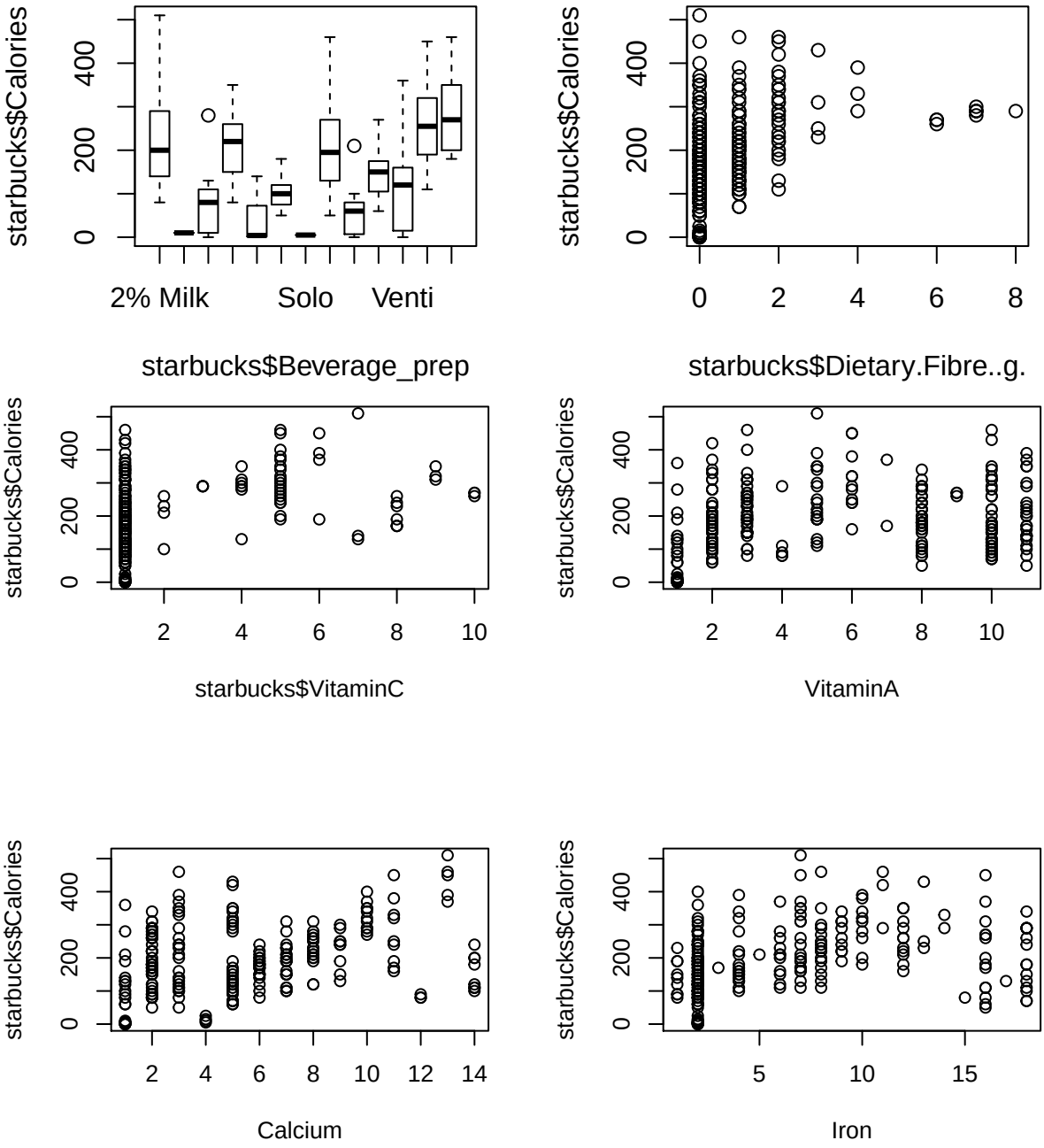
There are 11 possible explanatory variables for our response variable- calories. In the first step, through visualizing the scatterplots, we eliminate dietary.Fibre because the scatterplot is in fan shape, which means that there's no significant relationship between dietary fiber and calories; we also eliminate Vitamin A and caffeine because the relationships are possibly linear but nearly horizontal; for saturated fat and Vitamin C, we eliminate them because the relationship between dots are horizontal and there are just a few dots on the plots; for sodium, we consider eliminating it because the linear relationship is really weak. We do have outliers within the data of each variable.

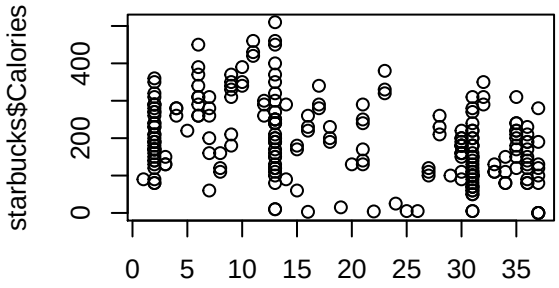
Then, using the Vision Rule 10%, we eliminate categorical beverage because it is the names of different drinks; we also decide to combine different categories into size (Tall, Grande, Venti) and milk (0.02, nonfat, whole, soy) to create indicators for Beverage prep because there is too little data in each group. After using F-test for both indicators size and milk, with small p-values, we are confident to reject null models and state that size and milk type are statistically significant to predict calories.

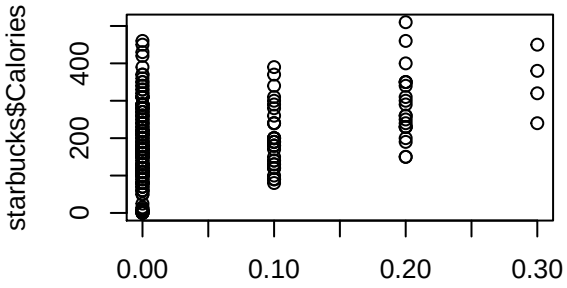
For next step, we decide to transform calsium into polynomial of third power because the scatterplot has two curve shapes, and we also decide to transform iron into plolynomial of second power because the scatterplot has one curve shape. We decide to pick the models with transformation for both predictors because they have smaller standard error and larger r-squared.

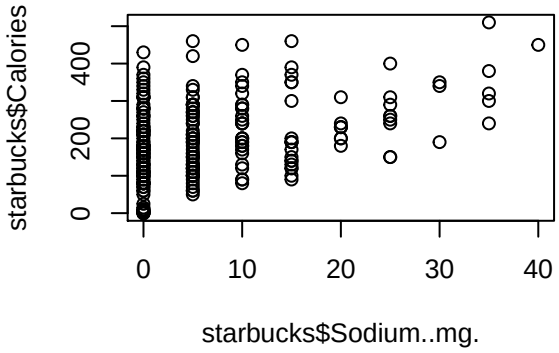
We check correlation in our final step in EDA. For sugar and cholesterol, the correlation is so high as 0.98 that we can delete one of them and only take sugar into account; because the correlations between Cholesterol and Carbohydrate, sugar and Carbohydrate are rather high around 0.75, we decide to add "sugar*carbohydrate-cholesterol" to our previous model.

Finally, we get to our EDA rough model with predictors: Beverage_category, Total_Fat, Trans.Fat, Total.Carbohydrates, Cholesterol, Sugars, Protein, poly(Iron, 2), Sodium, poly(Calcium, 3) , milk_0.02, milk_nonfat, milk_soy, milk_whole, size_Grande, size_Tall , size_Venti and sugar*cholesterol-cholesterol.









create indicator for Beverage_prep about size

```
size_Tall <- ifelse(starbucks$Beverage_prep == "Tall" | starbucks$Beverage_prep=="Tall Nonfat
                    Milk","yes","no")
size_Grande <- ifelse(starbucks$Beverage_prep == "Grande" | starbucks$Beverage_prep == "Grande
                     Nonfat Milk","yes","no")
size_Venti <- ifelse(starbucks$Beverage_prep == "Venti" | starbucks$Beverage_prep == "Venti Nonfat
                    Milk","yes","no")
```

create indicator for Beverage_prep about milk

```
milk_0.02 <- ifelse(starbucks$Beverage_prep == "2% Milk" ,"yes","no")
milk_nonfat <- ifelse(starbucks$Beverage_prep == "Grande Nonfat Milk" | starbucks$Beverage_prep ==
                        "Short Nonfat Milk" | starbucks$Beverage_prep == "Tall Nonfat Milk" |
                        starbucks$Beverage_prep == "Venti Nonfat Milk","yes","no")
milk_whole <- ifelse(starbucks$Beverage_prep == "Whole Milk","yes","no")
milk_soy <- ifelse(starbucks$Beverage_prep == "Soymilk","yes","no")
```

test whether size is influencial using F-test

```
NullModelSize <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.
                    +Cholesterol..mg.+Sugars..g.+Protein..g.+Iron+Sodium..mg.+Calcium+milk_0.02
                    +milk_nonfat+milk_soy+milk_whole, data = starbucks)
AltModelSize <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.
                   +Cholesterol..mg.+Sugars..g.+Protein..g.+Iron+Sodium..mg.+Calcium+milk_0.02
                   +milk_nonfat+milk_soy+milk_whole+size_Grande+size_Tall+size_Venti,
                   data = starbucks)
#anova(NullModelSize,AltModelSize)
```

test whether milk type is influencial using F-test

```
NullModelMilk <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.+Choleste
AltModelMilk <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.+Cholestero
#anova(NullModelMilk,AltModelMilk)
```

Transform Iron into polynomia of third power

```
IronPolyModel2 <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.+Choleste
#summary(IronPolyModel2)
#summary(AltModelMilk)
```

Transform Calcium into polynomia of third power

```
CalPolyModel3 <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.+Choleste
#summary(CalPolyModel3)
#summary(IronPolyModel2)
```

Check Correlation

```
##                          Total_Fat Total.Carbohydrates..g. Cholesterol..mg.
## Total_Fat                1.0000000               0.3443939        0.2536465
## Total.Carbohydrates..g.  0.3443939               1.0000000        0.7666536
## Cholesterol..mg.         0.2536465               0.7666536        1.0000000
## Sugars..g.               0.2047457               0.7714065        0.9841958
## Protein..g.              0.4245402               0.4106293        0.3604488
## VitaminC                 0.1441490               0.2403555        0.3347756
## VitaminA                 0.1227884               0.3633081        0.2314017
## Calcium                  0.3335017               0.2182064        0.1183339
## Iron                     0.2091519               0.2749934        0.3437644
## Sodium..mg.              0.4185155               0.2902948        0.1994770
##                          Sugars..g. Protein..g.  VitaminC    VitaminA
## Total_Fat                 0.2047457  0.42454021 0.1441490  0.12278840
## Total.Carbohydrates..g.   0.7714065  0.41062928 0.2403555  0.36330807
## Cholesterol..mg.          0.9841958  0.36044876 0.3347756  0.23140167
## Sugars..g.                1.0000000  0.26306079 0.3034039  0.23133660
## Protein..g.               0.2630608  1.00000000 0.3684670 -0.04275548
## VitaminC                  0.3034039  0.36846698 1.0000000  0.08527310
## VitaminA                  0.2313366 -0.04275548 0.0852731  1.00000000
## Calcium                   0.1096013  0.56568575 0.1099517 -0.15028197
## Iron                      0.2879397  0.18295432 0.2111544  0.31158942
## Sodium..mg.               0.2059687  0.49623281 0.3022343 -0.05902922
##                              Calcium         Iron Sodium..mg.
## Total_Fat                 0.333501734  0.209151909  0.41851549
## Total.Carbohydrates..g.   0.218206428  0.274993354  0.29029477
## Cholesterol..mg.          0.118333943  0.343764418  0.19947701
## Sugars..g.                0.109601279  0.287939742  0.20596872
## Protein..g.               0.565685745  0.182954319  0.49623281
## VitaminC                  0.109951693  0.211154405  0.30223433
## VitaminA                 -0.150281968  0.311589416 -0.05902922
## Calcium                   1.000000000  0.005871216  0.35148840
## Iron                      0.005871216  1.000000000 -0.11409445
## Sodium..mg.               0.351488396 -0.114094449  1.00000000
```

Our rough model

```
RoughModel <- lm(Calories~Beverage_category+Total_Fat+Trans.Fat..g.+Total.Carbohydrates..g.+Sugars..g.+
```

Model Selection

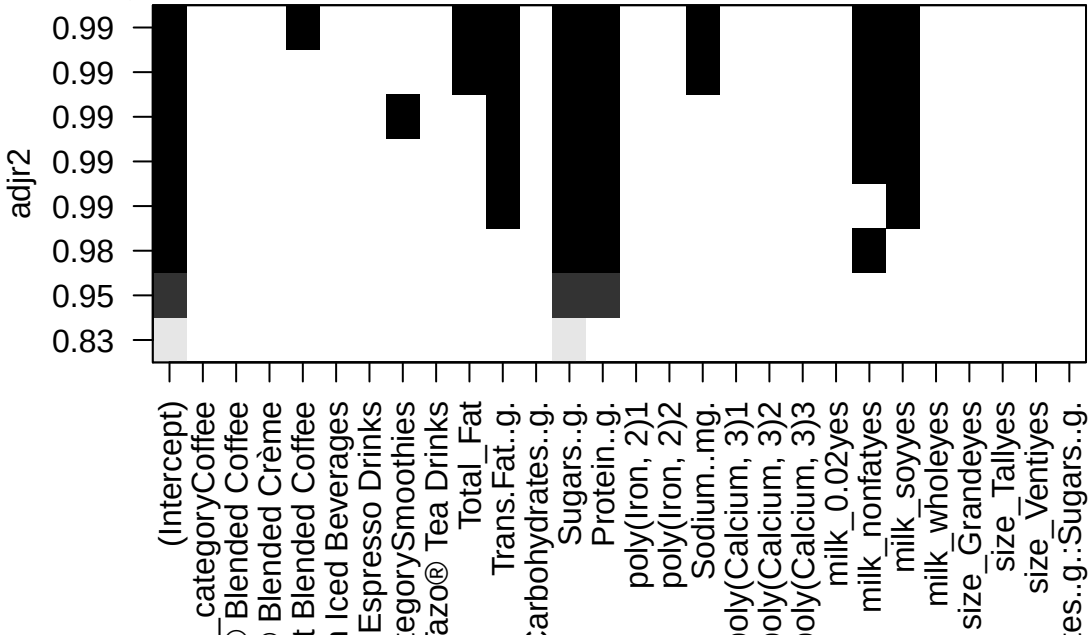
First of all, by looking at its summary, we can see what can be removed from rough model to form the refined model.

After comparison, the p-value for the following predictors are too big (bigger than alpha = 0.5) that we remove them in our refined model: poly(Cal,2 and 3), milk_0.02, milk_soy, milk_whole, and size_Tall.

Then, we create a model for the refined model, named it RefinedModel. And we run anova to see whether RoughModel or RefinedModel is better.

With a p-value = 0.01018, we reject null and continue using our rough model

After that, we continue with a best set selection to see which are the best combinations of predictors.

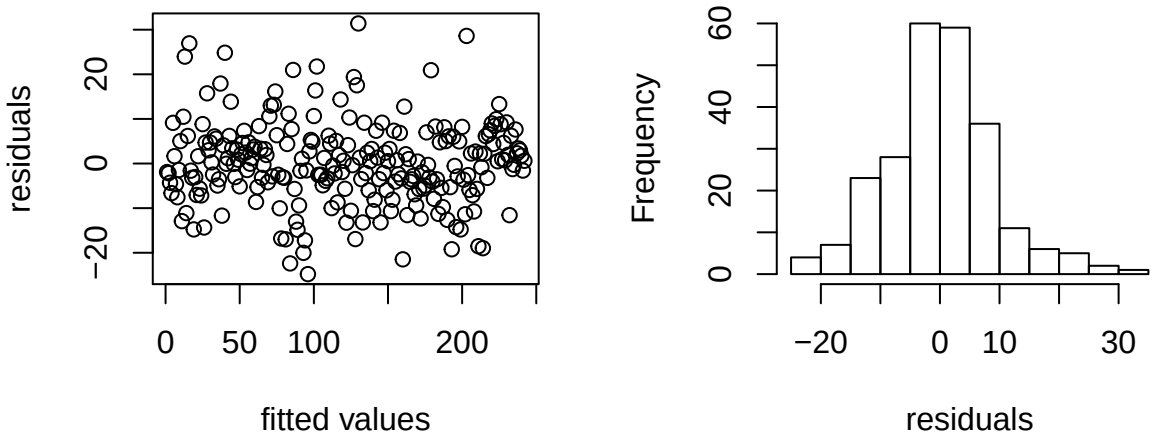


The graph above shows the adjustive r square for different combinations of predictors. Since the adjustive r square differs for so little, we take all the predictors appear in the graph as our BSSModel. Before creating the model, we need to create indicator for the categorical variables that gets in.

After having the BSSModel, we run anova of BSSModel and RoughModel to see which is better

The p-value is very small (2.157e-05), which means we should use RoughModel. However, the adjustive r square for BSSModel(0.991706) and RoughModel(0.9929377) differs so little that it is not convincing enough for us to add so many predictors. So we use the BSSModel as our final model.

Conditions for inference



1. Linearity: after polyomial, the plots shows linear relationships between the numeric variables and calories
2. Zero mean: according to the residual plot above, there are random scatter points above and below the 0 line
3. Constant variability: according to the residual plot above, variability of points around the 0 line is roughly constant.
4. Independence: there is no obvious reason that independence is violated
5. Random: the sample was selected randomly
6. Normality: according to the histagram of residuals above, the distribution of residual is unimodel and symmetric

Final Model

$\widehat{Calories} = 2.8075 + 10.1957(Somthiesyes) + 0.4482(TotalFat) + 10.1478(TransFat) + 4.0403(Sugar) + 6.4179(Protein) - 0.5029(Sodium) - 15.6606(nonfatmilkyes) + 12.6142(soymilkyes) - 4.1077(sizeVenti)$

R-square = 0.9917 Residual standard error = 9.368

Analysis

The model shows that calories of starbucks drinks are positively affected by its total fat, trans fat, sugar, and protein. Also, The drinks that are smothies are expected to have 10.1957 more calories than others. The drinks that contains soymilk are expected to have 12.6142 more calories. Plus, calories have negative relationships with sodium. Also, if nonfat milk is used, on average, the drinks will have 15.6606 less calories. Surprisingly, the drinks that are venti tend to have less calories also.

Conclusion

In conclusion, we are quite confident with our model since over 90% of the variance in the response variables are explained by our model. Indeed, there are some surprising part in our final model, but that's the result of our model selection, so we choose to trust technology.