STA-363-Lab-5

Phoebe Yan, Jean Ji & Rickey huang

3/10/2021

Question 1

Based on this, how many models do you think we fit in Stage 1 (Step 1) of BSS? Hint: Think carefully here. Some of the variables are categorical with more than two levels. Each level counts as a possible X.

We think when we fit the models in the stage 1, step 1 of BSS, we would fit 11 models with 15 possible X's. Each numeric variables can be served as a possible X to fit a model. There are 3 numeric variables among all variables. For the categorical variables with only 2 levels, we can use one level between the two to represent the variable and set it to 1 when the individual is in this level and 0 otherwise. There are 4 categorical variables in this data set. In addition, the variable *genhlth* is a categorical variable with 5 levels, which are *poor*, *fair*, *excellent*, *good*, and *very good*, and 4 of them can be served as separate variable choices with two levels for the model. Hence, we have $3 + 4 + 4 = 11$ models in total we can fit during the first step of the stage 1 of BSS.

Question 2

Once we have fit all these models in Stage 1 (Step 2) of BSS, what do you think we do? Hint: Same as Stage 1 (Step 1).

After fitting all these models in the stage 1, step 2, we would compute and compare the R^2 of these models we get during the step 2, and the LSLR model with the highest R^2 will be stored.

Question 3

How many β terms are in this full model? In other words, we end with Stage 1 (Step what) ? Hint: If you are stuck, just fit the model in R and look. You can do this by putting in all the predictors manually, or using ModelFull <- lm(weight ~ . , data = cdc).

```
##
## Call:
## lm(formula = weight ~ ., data = cdc)
##
## Coefficients:
##      (Intercept)  genhlthvery good       genhlthgood       genhlthfair
##       -119.22976           2.24914           3.92441           6.03632
##      genhlthpoor           exerany          hlthplan          smoke100
##          4.45523          -1.45066          -0.13040          -1.71940
##           height          wtdesire               age           genderf
##          4.16376           1.11618           0.08201         -21.95735
```

From the result above, in the full model, including the intercept term, we would have 12 β terms.

Question 4

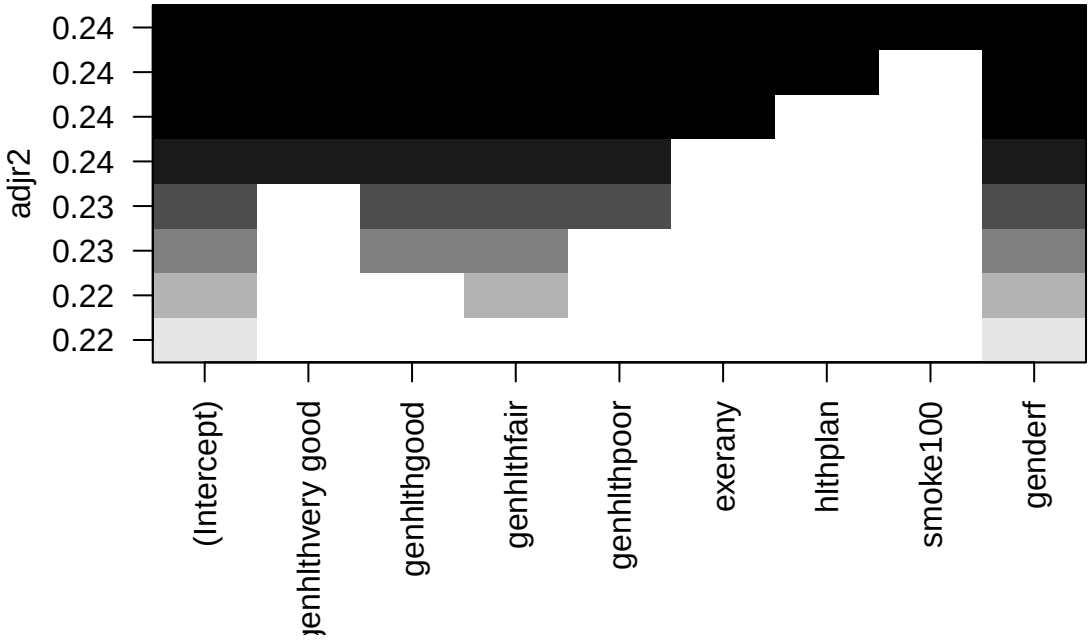
Look at only the categorical features in the data. Using these categorical features only, run the first stage of BSS and call the output BSScat. Remember to change the nvmax part of the code!!! You will notice that nothing seems to happen, as the output has been stored. Let's look at the R2adj value of each of these models by using the code summary(BSScat)$adjr2. What is the R2adj of the model fit with one X? (This is the first value). With two Xs? (This is the second value).

```
## [1] 0.2202324 0.2246977 0.2296282 0.2330062 0.2376003 0.2386057 0.2396686
## [8] 0.2397217
```

See from the computation result for adjusted R^2 s above, the R^2_{adj} of the model fitted with one X is 0.2202324. The R_{adj} for the model with two Xs is 0.2246977.

Question 5

Create a plot to help us see the values of the R2adj for all our models from Stage 1. To do this, you can use code plot(BSScat, scale = "adjr2"). Note: You can also use plot(BSScat, scale = "Cp") (which does AIC) or plot(BSScat, scale = "bic"), if you are wanting to use other metrics in the future.



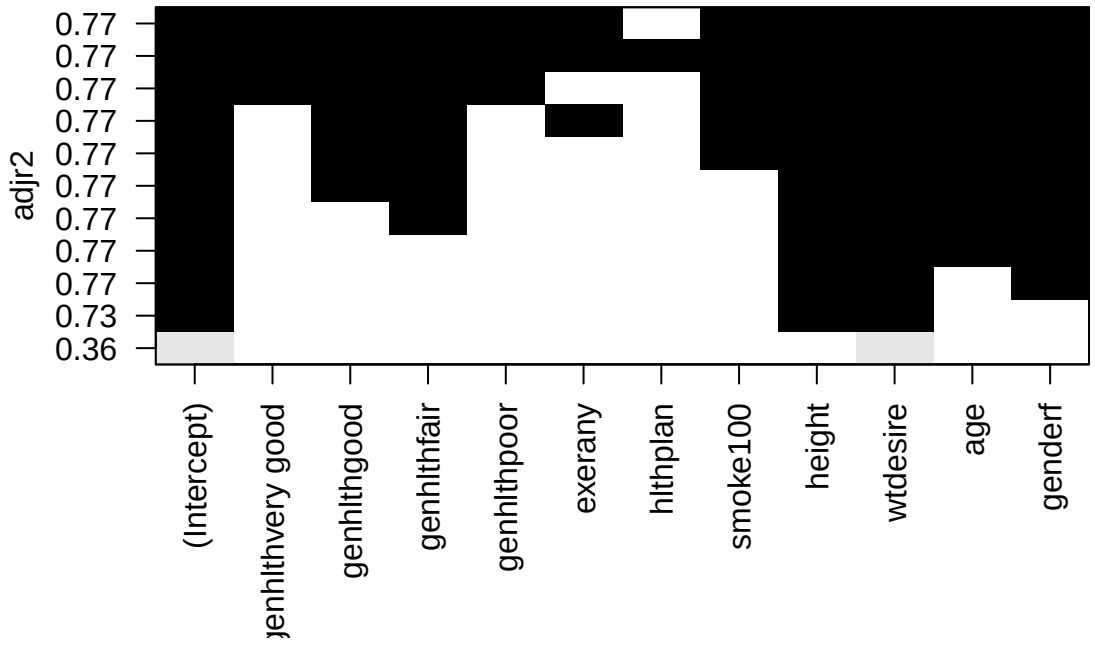
Question 6

What is the R2adj of the model with health Good, health Fair, gender female and the intercept?

From the plot made from the previous question, we can see that the model with health Good, health Fair, gender female and the intercept has an R^2_{adj} of 0.23

Question 7

Use all the possible feature variables (categorical and numeric) and run the first stage of BSS and call the output BSSall. Then, use the code plot(BSSall, scale = "adjr2") to plot the results.



Question 8

Which features are used in the model with the lowest value of R2adj, and what is the value of R2adj for that model?

In the lowest R^2_{adj} (0.36) model, the feature used are the desired weight in pounds and the intercept. The model with the highest R^2_{adj} (0.77) uses features of health Very Good, health Good, health Fair, exerany, health plan, height, desired weight in pounds, age, gender female and the intercept.

Question 9

Which features are used in the model with a R2adj of .73?

Based on the plot we made above, in the model with a R^2_{adj} of 0.73, the features used are the height, the desired weight in pounds and the intercept.

Question 10

Based on the results, which features would you choose to use? Explain. There is more than one correct answer here, so make sure you justify your reasoning.

```
##  [1] 0.3620242 0.7314531 0.7695811 0.7713019 0.7720435 0.7727094 0.7730423
##  [8] 0.7733458 0.7737070 0.7739298 0.7739196
```

Based on the results, we would like to choose the model with features health Good, Health Fair, smoke100, height, weight desired, age, gender female, and the intercept. because compare with others, this model has a adjusted R^2 of 0.7730423. The models with less features have R^2_{adj} much smaller than the model we choose, while the R^2_{adj} for the models with more feature than our model only increase at 0.0001 level. Thus, since we don't want to pursue a small increase in the R^2_{adj} which would trade off with the manipuility since we would have more variables in our model instead.