STA363_Lab_5

Zishan Shao

2022-09-29

```r
# load the data
source("http://www.openintro.org/stat/data/cdc.R")
```

Question 1:

We have a client who wants to build a model for Y = how much weight a person wants to lose/gain. This is important to study as it can reflect current societal health trends. Perceptions of a "target" weight can also lead to mental health struggles, life style choices, changes in exercise or other life habits, and so on. Is this a regression or classification task?

ANS: The response variable in this case is numerical, so this model performs a regression task.

```r
# we don't have a particular variable, but we do have the things we need to build it
cdc$wtchange<- cdc$weight -cdc$wtdesire
```

Question 2:

Create a box plot of the response variable. Does it seem like anything is unusual? If so, explain what and how you will choose to handle it before analysis.

ANS:

```r
# create the boxplot
library(ggplot2)
ggplot(cdc,aes(x ="wtchange", y = wtchange)) + geom_boxplot(fill='gold', col = 'black') + labs(title="F
```

Figure 1:

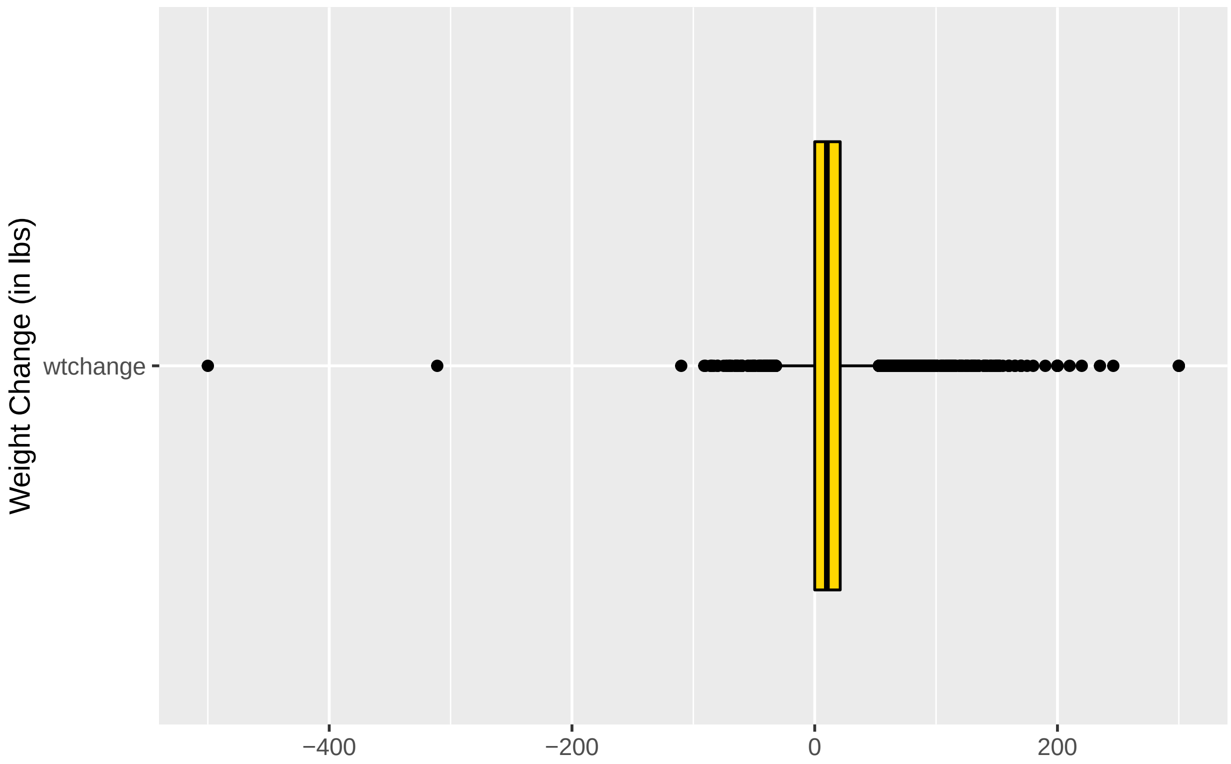
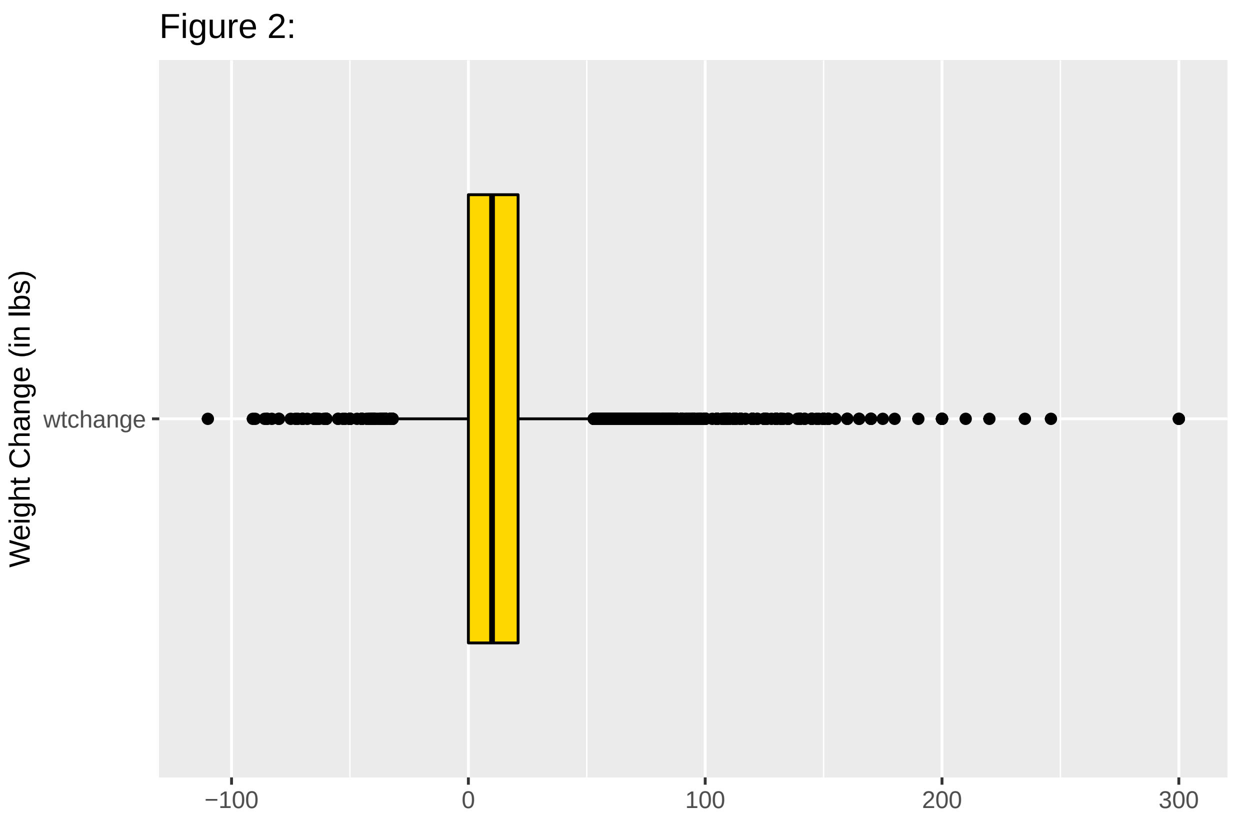


Figure 1 indicates that there are some unreasonable responses in this data. The positive number indicates the pounds a person expected to lose and negative number indicates the pounds a person expected to gain. However, for points on the left, it is unreasonable to gain over 500 lbs or 300 lbs, which is not a common case, so we suspect that these data are errors.

```
# get rid of the data unreasonable
cdc <- cdc[cdc$wtchange > -200,]
ggplot(cdc,aes(x ="wtchange", y = wtchange)) + geom_boxplot(fill='gold', col = 'black') + labs(title="Fi
```



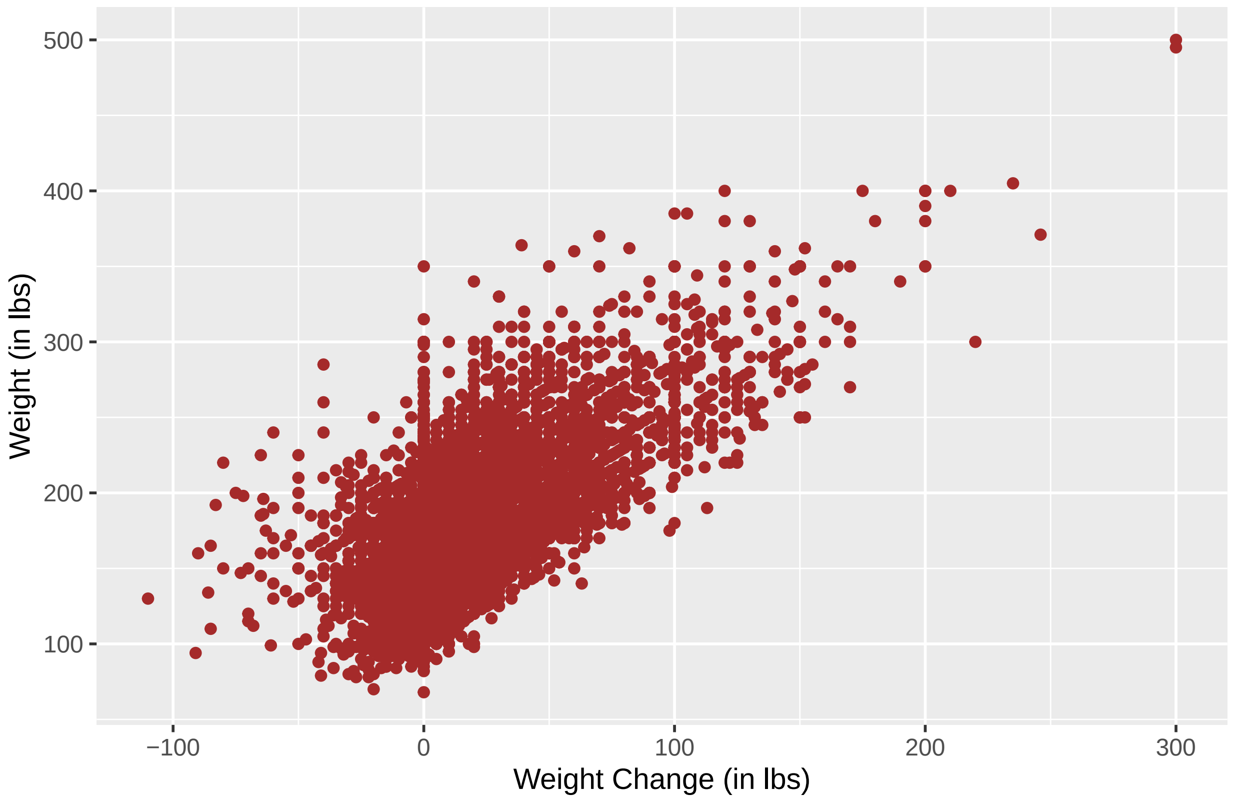
Question 3:

Create an appropriate plot to visualize the relationship between weight and desired weight change. Label your plot Figure 2. Based on shape alone, does LSLR seem like a reasonable choice?

ANS:

```
# create a scatterplot to illustrate the relationship
ggplot(cdc, aes(x = wtchange, y = weight)) + geom_point(color = "brown") + labs(title="Figure 3:", x =
```

Figure 3:



From the scatterplot, we could see that the weight change has a moderate linear relationship with weight and has no obvious outliers, so LSLR seem like a reasonable choice.

Question 4:

Write down the population form of the LSLR regression model (so include the error term, no hats) in matrix form in the white space in your Markdown file. This is to help you practice writing matrix notation. Remember that to start an equation you use $ in the white space, and that the R Code help site for our course has examples. Example: $Y = \beta_0 + \beta_1 X + \epsilon$.

ANS:

$$\boldsymbol{Y} = \begin{bmatrix} 1 & Weight_1 \\ 1 & Weight_2 \\ \vdots & \vdots \\ 1 & Weight_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X_D} \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Question 5:

Create the needed design matrix for an LSLR model for Y= desired weight change and X = weight and store the result as XD. Show the first few rows of your matrix (head(XD)).

ANS:

```
Y <- cbind(cdc$wtchange)  # create the matrix of Y
XD <- cbind(1,cdc$weight)  # create the designed matrix
knitr::kable(head(XD), caption = "Header of the XD", col.names = c("Intercept", "Weight"))
```

Table 1: Header of the XD

Intercept	Weight
1	175
1	125
1	105
1	132
1	150
1	114

Question 6:

State the dimensions of Y and the dimensions of XD.

ANS: The dimension of Y is [19998, 1], and the dimension of XD is [19998, 2]

Question 7:

**

ANS:

$$\hat{\beta} = (\boldsymbol{X_D^T X_D})^{-1}\boldsymbol{X_D^T Y}$$

Question 8:

From the matrix representation from the previous question, use R to estimate slope and intercept for the LSLR regression line. Show the code you use to do this and show your results. Note: You may not use lm for this question.

ANS:

```
# compute the slope based on previous equation
Beta <- solve(t(XD) %*% XD) %*% t(XD) %*%Y   # solve is the inverse
Beta
```

```
##            [,1]
## [1,] -46.882520
## [2,]   0.362535
```

```
# prove if with lm, the result are same
gg <- lm(wtchange ~ weight, cdc)
gg
```

```
##
## Call:
## lm(formula = wtchange ~ weight, data = cdc)
##
## Coefficients:
## (Intercept)       weight
##    -46.8825       0.3625
```

Therefore, the coefficient matrix β should be [-46.8825, 0.3625]

Question 9:

Based on your LSLR line, find the estimated desired weight change for the 5th individual in the data.

ANS:

```
# given the model
w <- cdc$weight[5]
y <- -46.8825 + 0.3625*w
sprintf("The estimated desired weight change of 5th individual: %s lbs", y)
```

```
## [1] "The estimated desired weight change of 5th individual: 7.4925 lbs"
```

Based on the model, the 5th individual in the data should change (by losing) 7.4925 lbs.

Question 10:

Why do we need to remove wtdesire from the data before running BSS? Make sure you remove that column from the data before proceeding.

ANS: This is because the wtchange is derived from the wtdesire and the weight, which means the wtchange is strongly related with these two variables. What's more, the weight and the desired weight are also strongly related (for instance, if one person is overly obese, then he/she/they is expected to lose more weight). Therefore, it is no need to add two strongly correlated variables in a model.

Question 11:

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.

ANS: There are totally 8 variables as features for training, so we should create at least 8 models. Notifying that the general health (genhlth) has 5 different categories, which, if in the model, this means 4 different new variables with one category as base level. Therefore, there should be 8 - 1 + 4 = 11 models in BSS stage 1 (minus 1 is counted as one of 4 categories of genhlth).

Question 12:

Once we have fit all the models in with two features, how do you think we choose one? Hint: Same as Stage 1 (Step 1).

ANS: We could compare the Rˆ2 of these models and choose the one with highest Rˆ2. This step is feasible because all models have two features and thus is comparable in Rˆ2.

Question 13:

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 features manually, or using

ANS:

There should be 12 β_i in this full model. This is because there are total 11 terms to consider and we also need one intercept. So there will be 12 β_i in the full model.

Question 14:

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 feature? (This is the first value). With two features? (This is the second value).

ANS:

```
# select the best subset of the model
library(leaps)
# nvmax will tell the R the maximum number of beta terms we want to allow in out model
# BSSout <- regsubsets( Y ~ X1 + X2 + X3, data = cdc, nvmax = 3, method="exhaustive")

set.seed(114514)

# find out model by looking only the categorical features
BSScat <- regsubsets(wtchange ~ genhlth + exerany + hlthplan + gender + smoke100, data = cdc, nvmax = 8

# the order indicates the number of variables in the model
knitr::kable(summary(BSScat)$adjr2, col.names = "Adjusted r^2", caption = "BSS Stage 1 (Part 1)")
```

Table 2: BSS Stage 1 (Part 1)

Adjusted r^2
0.0240895
0.0324532
0.0392157
0.0462087
0.0519596
0.0552613
0.0558168
0.0557699

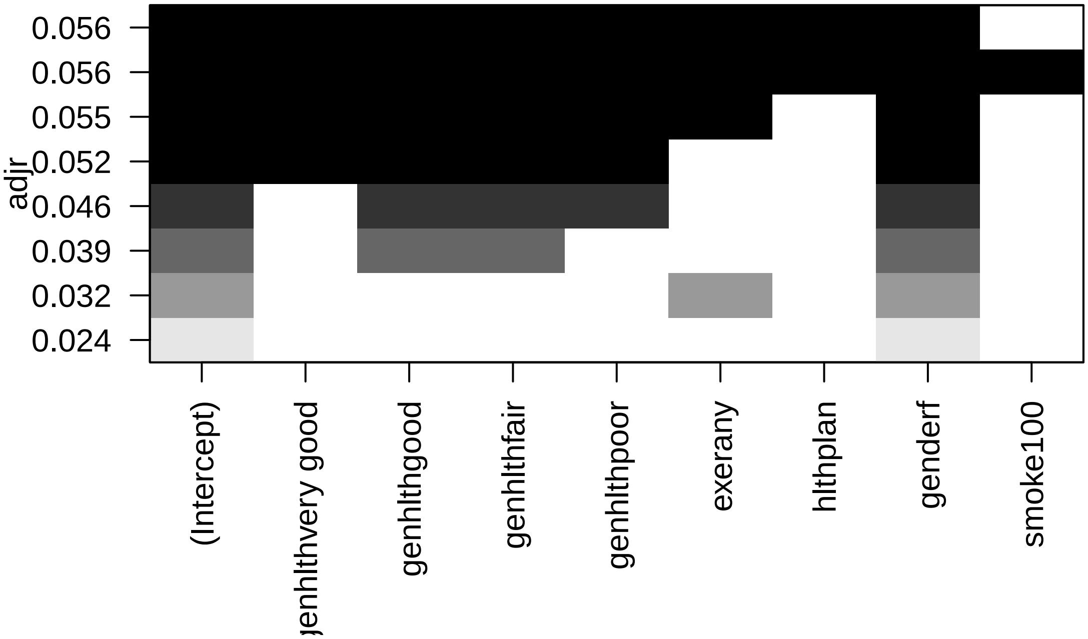
The R^2_{adj} is about 0.0241 with one feature and 0.0325 with two features.

Question 15:

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

ANS:

```
plot(BSScat, scale = "adjr")  # , title("Figure 4: Adjusted R-squared of BSS models")
```



Question 16:

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

ANS: The R2adj is 0.039 for the model with health Good, health Fair, gender female and the intercept.

Question 17:

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.

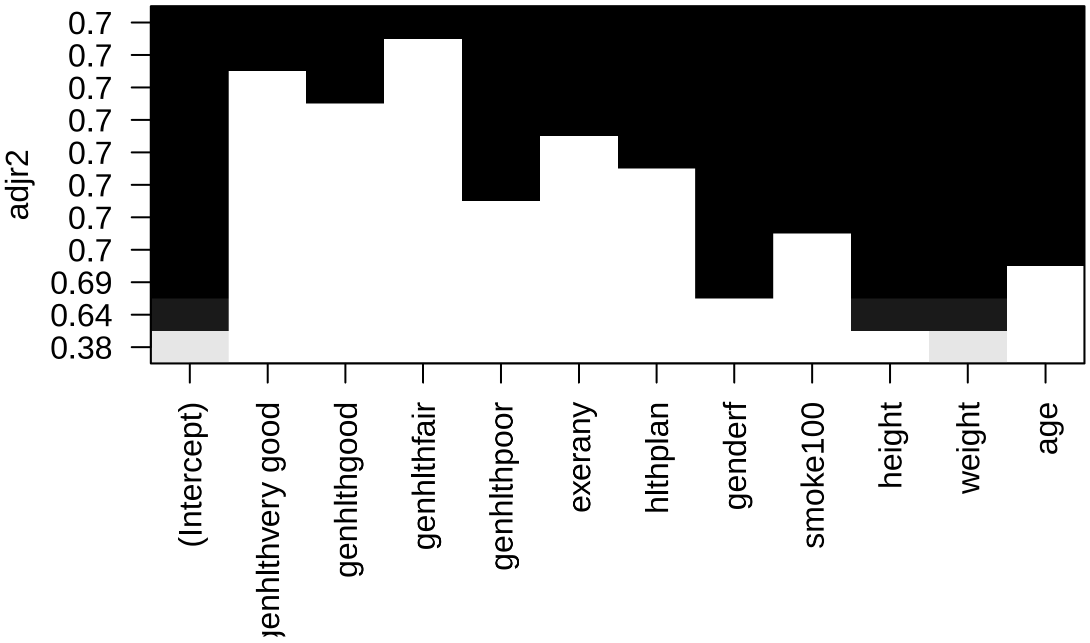
ANS:

```
# select the best subset of the model
library(leaps)
# nvmax will tell the R the maximum number of beta terms we want to allow in out model
# BSSout <- regsubsets( Y ~ X1 + X2 + X3, data = cdc, nvmax = 3, method="exhaustive")

set.seed(114514)

# find out model by looking only the categorical features
BSSall <- regsubsets(wtchange ~ genhlth + exerany + hlthplan + gender + smoke100 + height + weight + ag

plot(BSSall, scale = "adjr2")
```



```
#, row.names = c("M1","M2","M3","M4","M5","M6","M7","M8","M9","M10","M11")
```

Question 18:

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

ANS:

The model with lowest R2adj value used the weight as the feature, which gives the R2adj of 0.38.

Question 19:

Which features are used in the model with a R2adj of 0.6955846? To figure this out, look at the values for each model. Figure out which of the models (1,2,3, etc) has the desired value. Then, run summary(BSSall)$which[HERE,], where the HERE is replaced with the number of the model you want.

ANS:

```
# the R2adj = 0.6955846 is for the model 4, which uses 4 features
knitr::kable(summary(BSSall)$adjr2,  col.names = "Adjusted r^2", caption = "BSS of Full Model")
```

Table 3: BSS of Full Model

Adjusted r^2
0.3770674
0.6376873
0.6946867
0.6955846
0.6960346
0.6961822
0.6962403
0.6963053
0.6963221
0.6963236
0.6963376

```
summary(BSSall)$which[4,]
```

```
##      (Intercept) genhlthvery good      genhlthgood      genhlthfair
##             TRUE            FALSE            FALSE            FALSE
##      genhlthpoor          exerany         hlthplan          genderf
##            FALSE            FALSE            FALSE             TRUE
##         smoke100           height           weight              age
##            FALSE             TRUE             TRUE             TRUE
```

Based on the result from the summary, the model should includes the intercept, gender female, height, weight and age.

Question 20:

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.

ANS:

Based on the previous result, we could see that the model 4 has the least variables with comparatively large adjusted R-square values. Therefore, model 4 should be the best choice and we should use variables: intercept, gender female, height, weight and age.