CW1

dizhen

4/26/2020

Contents

Import and clean the data 1

Visualization 4

Fit the multiple linear regression model 5

Diagnostic tests 5

Set directory

```
setwd('D:/git/DPH112-xjtlu/cw1')
```

Import and clean the data

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
HINTS <- read_csv("HINTS.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   personid = col_character(),
##   stratum = col_character(),
##   app_region = col_character(),
##   updatedate = col_character(),
##   caother_os = col_character(),
##   occupationstatus_os = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```r
HINTS.FOUR <- HINTS[,c("bmi","averagedailytvgames","genderc","generalhealth","weight")]
str(HINTS.FOUR)
```

```
## tibble [3,677 x 5] (S3: tbl_df/tbl/data.frame)
##  $ bmi                : num [1:3677] 22 34.6 19.2 31 31.4 32 26.4 30.8 30.6 -9 ...
##  $ averagedailytvgames: num [1:3677] 4 6 5 4 5 8 2 6 1 2 ...
##  $ genderc            : num [1:3677] 1 1 2 2 1 1 2 1 1 2 ...
##  $ generalhealth      : num [1:3677] 2 4 3 3 2 3 2 2 4 -9 ...
##  $ weight             : num [1:3677] 158 255 105 210 225 198 149 215 213 -9 ...
```

```r
head(HINTS.FOUR)
```

```
## # A tibble: 6 x 5
##     bmi averagedailytvgames genderc generalhealth weight
##   <dbl>               <dbl>   <dbl>         <dbl>  <dbl>
## 1  22                     4       1             2    158
## 2  34.6                   6       1             4    255
## 3  19.2                   5       2             3    105
## 4  31                     4       2             3    210
## 5  31.4                   5       1             2    225
## 6  32                     8       1             3    198
```

```r
# check independet variable: averagedailytvgames
table(HINTS.FOUR$averagedailytvgames)
```

```
## 
##  -9  -4   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 
## 101  39  85 502 844 684 496 294 225  39 152  16  94   4  37   1   8  14   8   2 
##  18  19  20  21  24 
##   7   1  19   1   4
```

```r
table(HINTS.FOUR$genderc)
```

```
## 
##   -9    1    2 
##   69 1424 2184
```

```r
table(HINTS.FOUR$generalhealth)
```

```
## 
##   -9   -5    1    2    3    4    5 
##  105   15  374 1199 1355  495  134
```

```
table(HINTS.FOUR$weight[which(HINTS.FOUR$weight< 0)])
```

```
##
##  -9  -4
## 148   1
```

```
table(HINTS.FOUR$bmi[which(HINTS.FOUR$bmi< 0)])
```

```
##
##  -9  -4
## 162  15
```

```
summary(HINTS.FOUR$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -9.0   140.0   170.0   169.5   200.0   442.0
```

```
HINTS.CLEAN <- HINTS.FOUR %>%
  filter(averagedailytvgames >= 0) %>%
  filter(genderc >0) %>%
  filter(generalhealth > 0) %>%
  filter(weight > 0) %>%
  filter(bmi >0) %>%
  mutate(genderc = factor(genderc)) %>%
  mutate(generalhealth = factor(generalhealth))

str(HINTS.CLEAN)
```

```
## tibble [3,331 x 5] (S3: tbl_df/tbl/data.frame)
##  $ bmi                : num [1:3331] 22 34.6 19.2 31 31.4 32 26.4 30.8 30.6 36.3 ...
##  $ averagedailytvgames: num [1:3331] 4 6 5 4 5 8 2 6 1 1 ...
##  $ genderc            : Factor w/ 2 levels "1","2": 1 1 2 2 1 1 2 1 1 1 ...
##  $ generalhealth      : Factor w/ 5 levels "1","2","3","4",..: 2 4 3 3 2 3 2 2 4 3 ...
##  $ weight             : num [1:3331] 158 255 105 210 225 198 149 215 213 260 ...
```

```
head(HINTS.CLEAN)
```

```
## # A tibble: 6 x 5
##     bmi averagedailytvgames genderc generalhealth weight
##   <dbl>               <dbl> <fct>   <fct>          <dbl>
## 1  22                     4 1       2                158
## 2  34.6                   6 1       4                255
## 3  19.2                   5 2       3                105
## 4  31                     4 2       3                210
## 5  31.4                   5 1       2                225
## 6  32                     8 1       3                198
```

The sample size is 3677 in total with no NA.

The independent variable `AverageDailyTVGames` is a categorical variable with values from 0 to 24 hours, plus -4 meaning unreadable or non-conforming numeric response, and -9 meaning missing data. There are 101 records of AverageDailyTVGames = -9 and 39 records of AverageDailyTVGames = -4. Here, we discard the those records.

genderc is a categorical variable with values -9(missing data), 1(Male) and 2(Female). There are 69 records with genderc = -9. Here, we discard those records.

generalhealth is a categorical variable with values from 1 to 5, plus -5 meaning multiple responses selected in error, and -9 meaning missing data. There are 105 records of generalhealth = -9 and 15 records of generalhealth = -5. Here, we discard the those records.

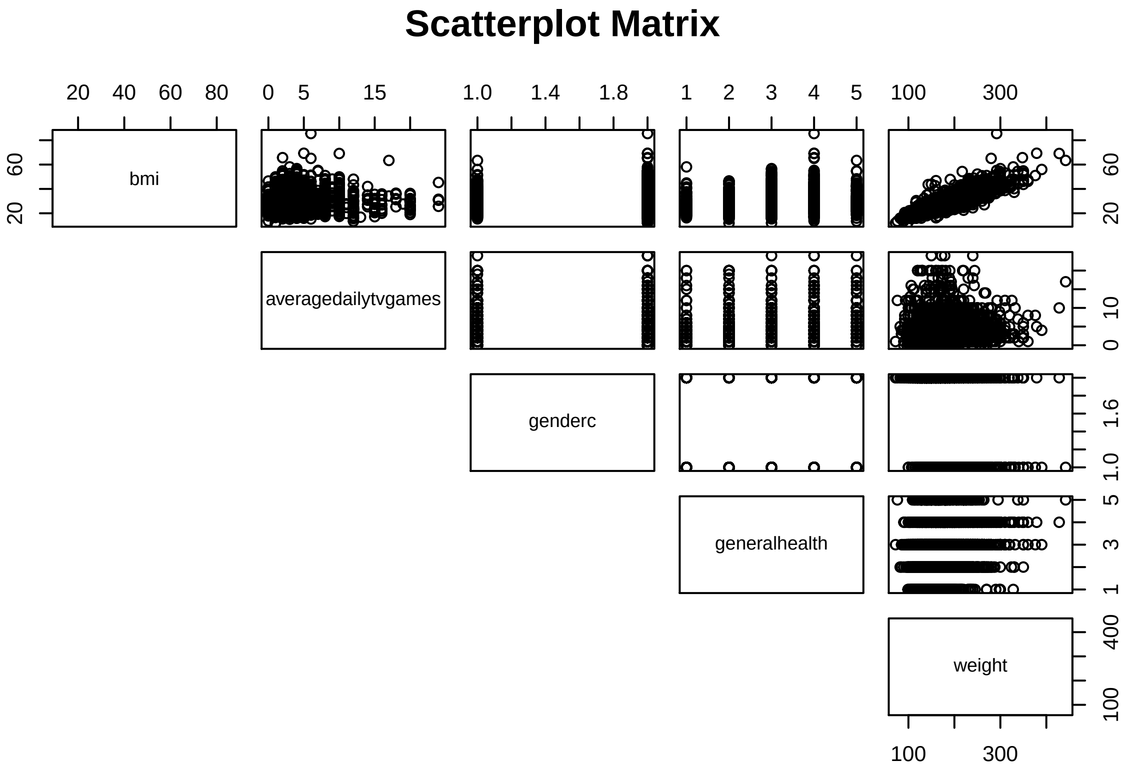
weight is a continuous variable with the maximum value of 442 and minimum value of 40 in pounds, plus -9 meaning missing data, and -4 meaning unreadable or non-conforming numeric response. There are 148 records of weight = -9 and 1 record of weight = -4. Here we discard them.

The dependent variable bmi is continuous with two invalid value -9 (missing data) and -4 (unreadable or nonconforming numeric response). There are 162 records with -9 and 15 records with -4. Here we discard those records.

After cleaning the data, now the sample size is 3331.

Visualization

```
pairs(~bmi + averagedailytvgames + genderc + generalhealth + weight, data = HINTS.CLEAN,
      lower.panel = NULL,
      main = "Scatterplot Matrix")
```



Fit the multiple linear regression model

```
HINTS.CLEAN.LM1 <- lm(bmi ~ averagedailytvgames + genderc + generalhealth + weight, data = HINTS.CLEAN)
summary(HINTS.CLEAN.LM1)
```

```
##
## Call:
## lm(formula = bmi ~ averagedailytvgames + genderc + generalhealth +
##     weight, data = HINTS.CLEAN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.822  -1.639  -0.132   1.454  38.709
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.237564   0.262918   0.904   0.3663
## averagedailytvgames 0.043933   0.016365   2.685   0.0073 **
## genderc2            4.252627   0.103966  40.904  < 2e-16 ***
## generalhealth2      0.384230   0.168077   2.286   0.0223 *
## generalhealth3      0.840329   0.168702   4.981 6.64e-07 ***
## generalhealth4      1.649914   0.203090   8.124 6.29e-16 ***
## generalhealth5      1.527405   0.292519   5.222 1.88e-07 ***
## weight              0.138312   0.001178 117.442  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.753 on 3323 degrees of freedom
## Multiple R-squared:  0.8229, Adjusted R-squared:  0.8225
## F-statistic:  2205 on 7 and 3323 DF,  p-value: < 2.2e-16
```

```
confint(HINTS.CLEAN.LM1, level = 0.95)
```

```
##                           2.5 %    97.5 %
## (Intercept)         -0.27793261 0.7530609
## averagedailytvgames  0.01184616 0.0760189
## genderc2             4.04878264 4.4564706
## generalhealth2       0.05468377 0.7137754
## generalhealth3       0.50955930 1.1710988
## generalhealth4       1.25172058 2.0481080
## generalhealth5       0.95386932 2.1009413
## weight               0.13600246 0.1406207
```

Diagnostic tests

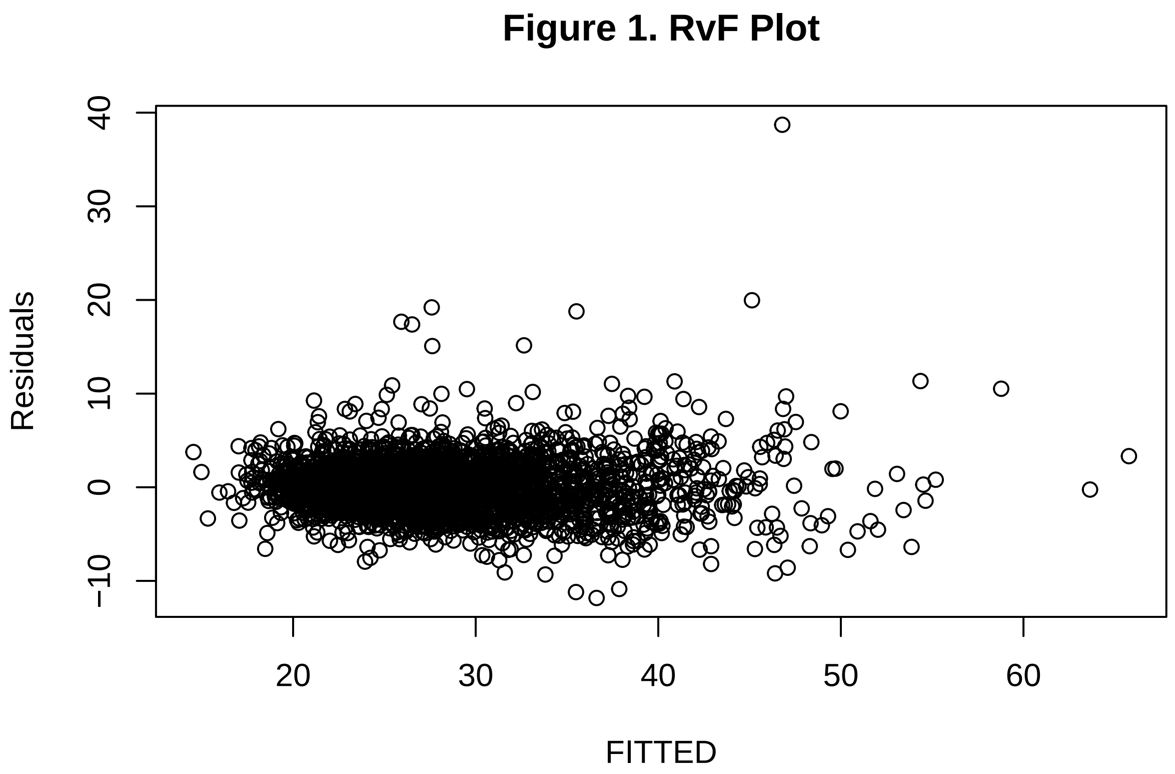
Assessing assumptions about the form of the model

1. Plot the residuals against the fitted values (RvF plot)

```
HINTS.CLEAN$FITTED <- predict(HINTS.CLEAN.LM1, type = "response")
HINTS.CLEAN$RESID <- resid(HINTS.CLEAN.LM1)
head(HINTS.CLEAN)
```

```
## # A tibble: 6 x 7
##     bmi averagedailytvgames genderc generalhealth weight FITTED  RESID
##   <dbl>               <dbl> <fct>   <fct>          <dbl>  <dbl>  <dbl>
## 1  22                     4 1       2                158   22.7 -0.651
## 2  34.6                   6 1       4                255   37.4 -2.82
## 3  19.2                   5 2       3                105   20.1 -0.873
## 4  31                     4 2       3                210   34.6 -3.55
## 5  31.4                   5 1       2                225   32.0 -0.562
## 6  32                     8 1       3                198   28.8  3.18
```

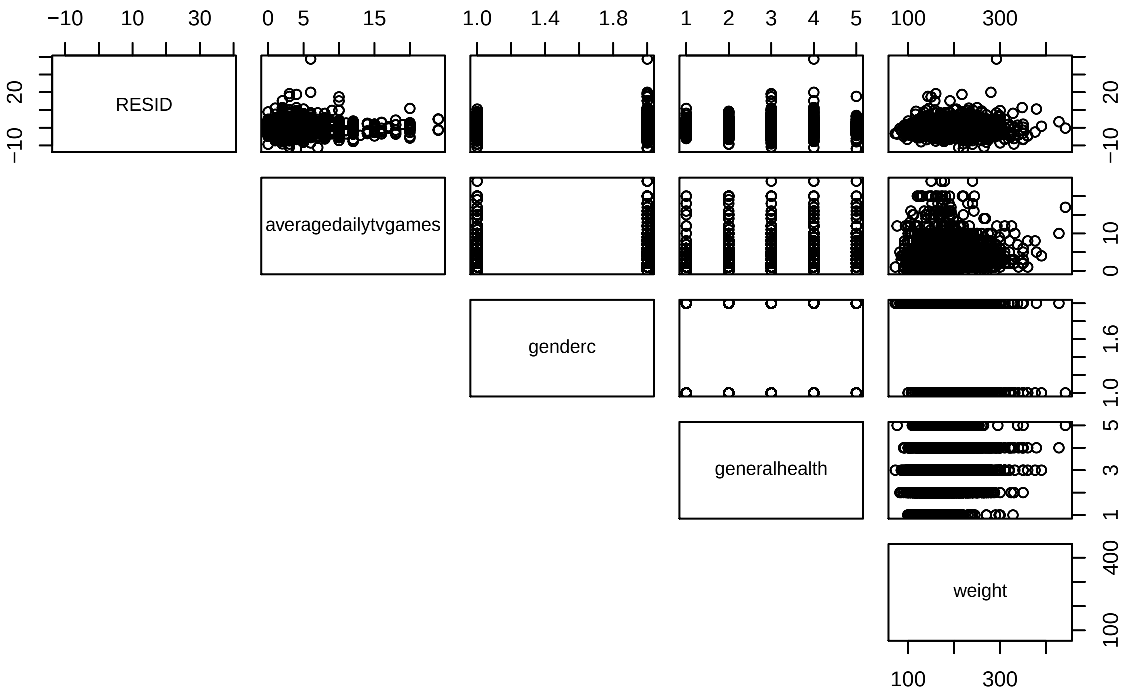
```
plot(RESID~FITTED, data = HINTS.CLEAN,
     main = "Figure 1. RvF Plot",
     ylab = "Residuals")
```



2. Plot the residuals against each of the predictors

```
pairs(~RESID + averagedailytvgames + genderc + generalhealth + weight, data = HINTS.CLEAN,
      lower.panel=NULL,
      main = "Figure 2. Scatterplot Matrix")
```

Figure 2. Scatterplot Matrix



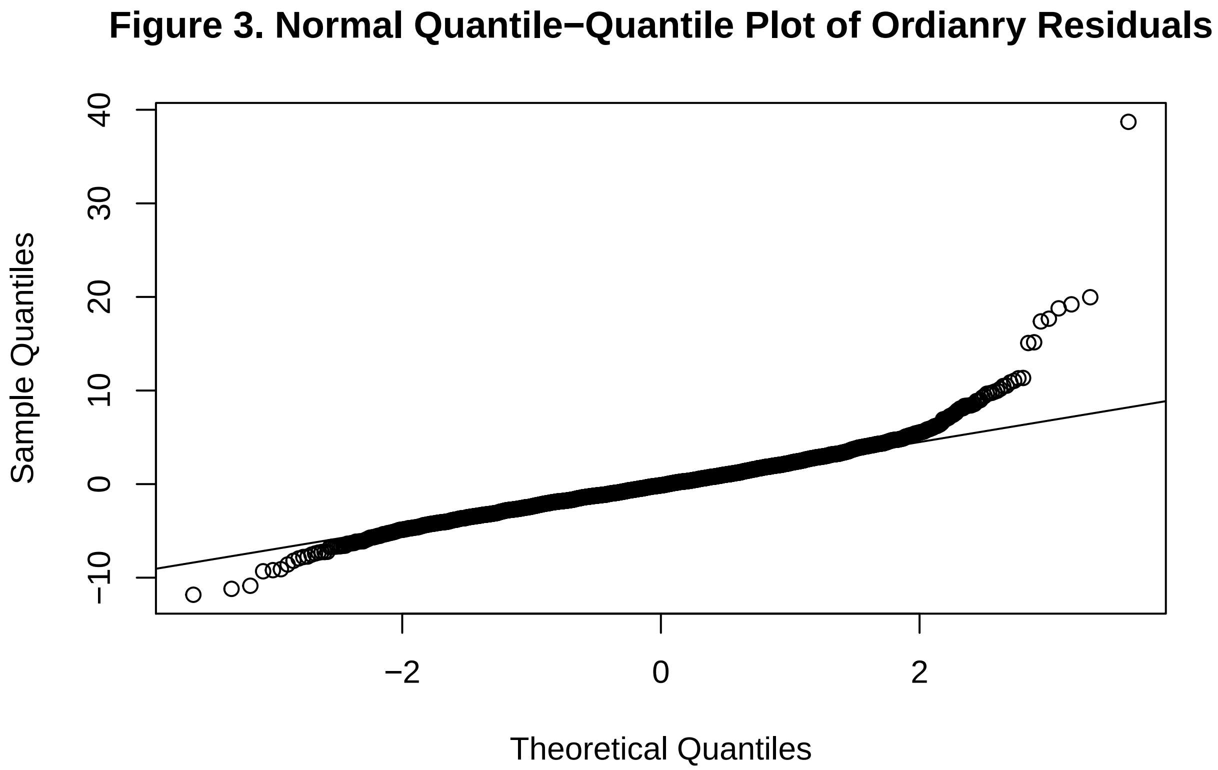
Conclusion:

In the RvF plot(Figure 1.), we notice some deviation. In the scatterplot matrix (Figure 2.), we are only interested in the first row. It shows some deviation. The assumption of linearity may be violated.

Assessing assumptions about the errors

Produce a quantile-quantile plot of the residuals.

```r
qqnorm(HINTS.CLEAN$RESID,
       main = "Figure 3. Normal Quantile-Quantile Plot of Ordianry Residuals")
qqline(HINTS.CLEAN$RESID)
```



Conclusion:

First, Figure 3 shows some deviation in the right tail and left tail. The assumption of normality may not be satisfied. Second, there is wedge-shaped pattern in Figure 2. The assumption of equal variance of residuals may not be satisfied. Third, Evidence about independence is available from the design of data.

Assessing assumptions about the predictors

Calculate variance inflation factor or VIF to detect the presence of multicollinearity between the predictors.

```
# The fitted model
HINTS.CLEAN.LM1 <- lm(bmi ~ averagedailytvgames + genderc + generalhealth + weight, data = HINTS.CLEAN)
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
vif(HINTS.CLEAN.LM1)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## averagedailytvgames 1.071067  1        1.034924
## genderc             1.136481  1        1.066058
## generalhealth       1.125911  4        1.014934
## weight              1.192721  1        1.092118
```

Conclusion:

There are three assumptions about the predictors. First, the independent variables are nonrandom, which is satisfied. Second, the independent variables are meansured without error, which is also satisfied. Third, the independent variables are linearly independent of each other which is tested by VIF. The VIF values for all predictors are all $<$ 10, indicating the absence of collienarity.

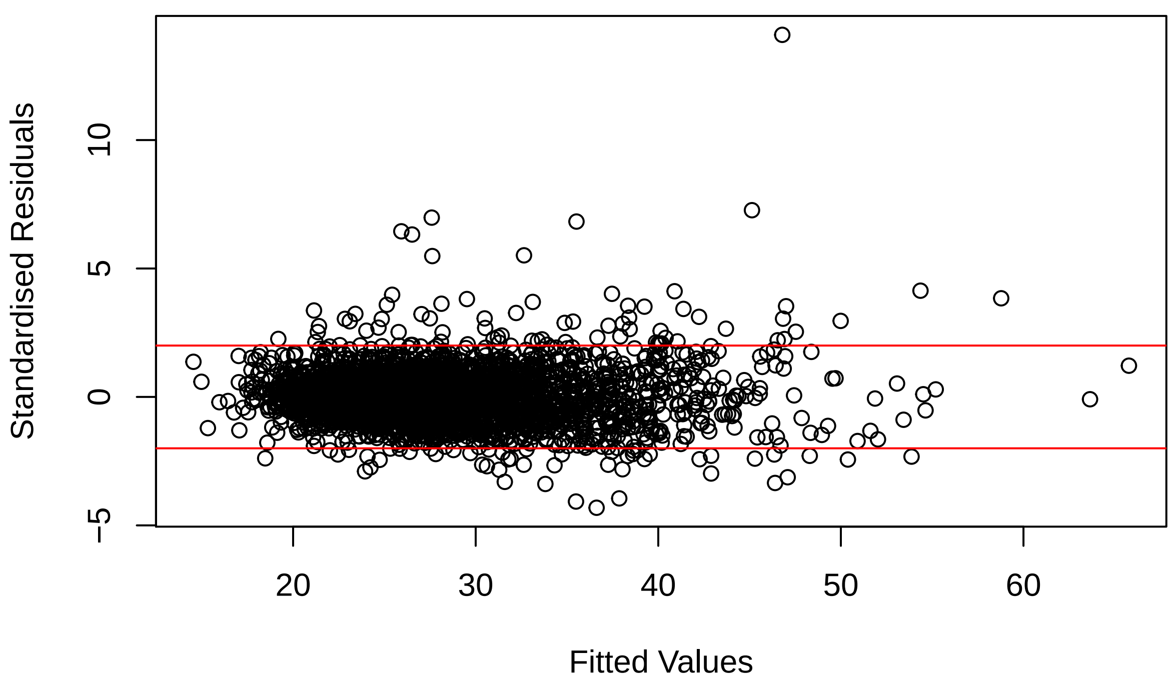
Assessing the assumption about the observations

1. Evaluate the presence of outliers in the dependent variable using standardised residuals.

There are 127 observations greater or less than two standard deviations larger than the mean. They are outliers of the dependent variable.

```r
HINTS.CLEAN$RSTAND <- rstandard(HINTS.CLEAN.LM1)
plot(RSTAND ~ FITTED, data = HINTS.CLEAN,
     ylab = "Standardised Residuals",
     xlab = "Fitted Values",
     main = "Figure 5. Standardised RvF Plot")
abline(h = c(-2, 2), col = "red")
```

Figure 5. Standardised RvF Plot



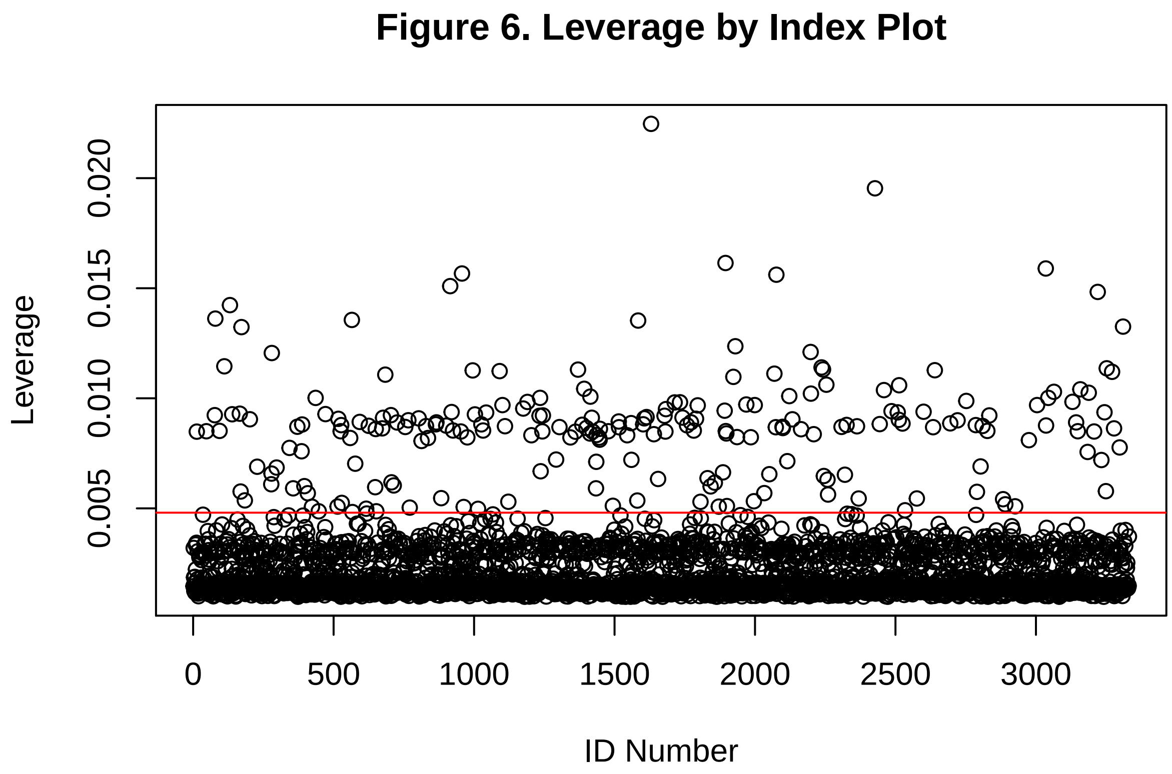
```r
HINTS.CLEAN[abs(HINTS.CLEAN$RSTAND) > 2,]
```

```
## # A tibble: 127 x 8
##      bmi averagedailytvgames genderc generalhealth weight FITTED RESID RSTAND
##    <dbl>               <dbl> <fct>   <fct>          <dbl>  <dbl> <dbl>  <dbl>
##  1  22                     8 2       3                158   27.5 -5.54  -2.01
##  2  38.5                   6 2       3                300   47.1 -8.59  -3.13
##  3  46.7                   5 2       5                247   40.4  6.30   2.30
##  4  30.3                   2 2       4                230   38.0 -7.74  -2.82
##  5  31.2                   8 1       2                160   23.1  8.10   2.94
##  6  27.1                   3 1       2                148   21.2  5.88   2.14
##  7  25.1                   4 2       3                190   31.8 -6.69  -2.43
##  8  39.8                   2 2       3                204   33.6  6.17   2.24
##  9  35.1                   4 1       2                198   28.2  6.92   2.51
## 10  32                     8 2       3                236   38.3 -6.32  -2.30
## # ... with 117 more rows
```

2. Evaluate the presence of outliers in the independent variables using leverage values.

The plot shows 221 observations are high-leverage points. They are outliers of the independent variable.

```r
HINTS.CLEAN$HAT <- hatvalues(HINTS.CLEAN.LM1)
HAT.CUT <- 2 * (7 + 1)/ nrow(HINTS.CLEAN)
ID <- seq(1,nrow(HINTS.CLEAN),by = 1)
plot(HAT ~ ID, data = HINTS.CLEAN,
     ylab = "Leverage",
     xlab = "ID Number",
     main = "Figure 6. Leverage by Index Plot")
abline(h = HAT.CUT, col = "red")
```



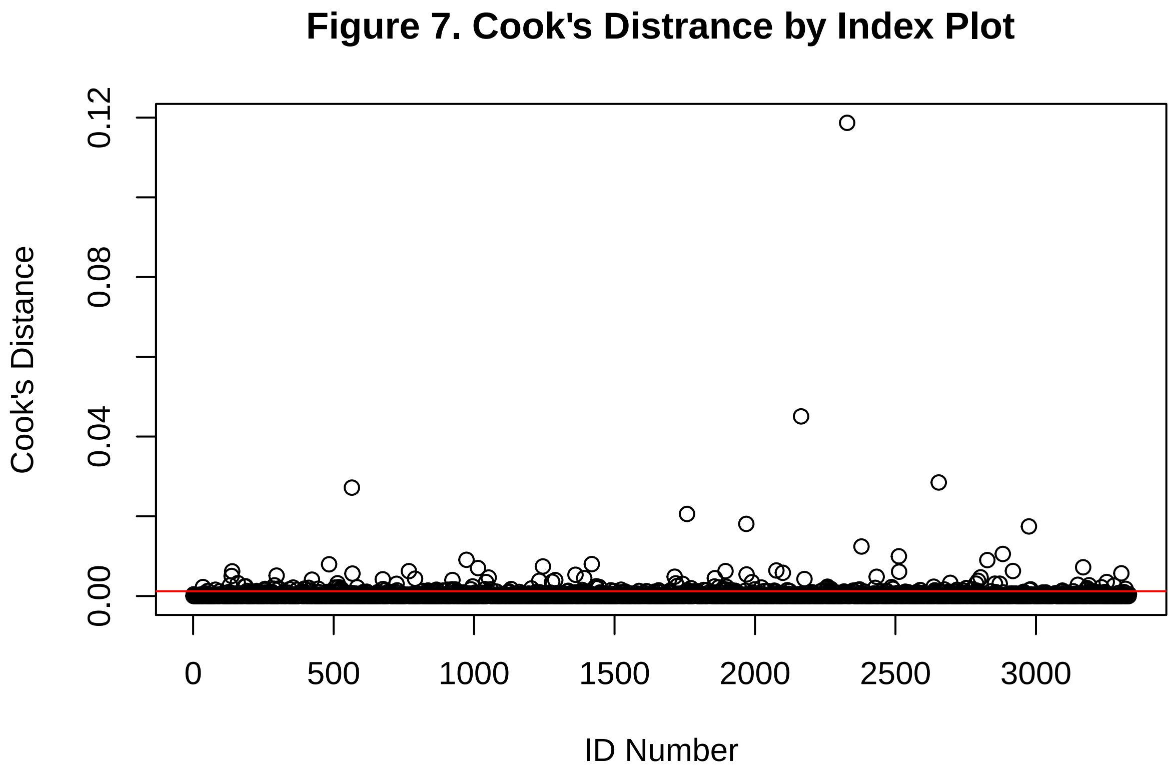
```
HINTS.CLEAN[HINTS.CLEAN$HAT > HAT.CUT,]
```

```
## # A tibble: 221 x 9
##      bmi averagedailytvg~ genderc generalhealth weight FITTED   RESID  RSTAND
##    <dbl>            <dbl> <fct>   <fct>          <dbl>  <dbl>   <dbl>   <dbl>
##  1  28.7                3 2       5                162   28.6  0.144   0.0526
##  2  29                  3 2       5                180   31.0 -2.05   -0.746
##  3  36.6                2 2       5                220   36.5  0.0660  0.0241
##  4  36                 18 2       5                230   38.6 -2.62   -0.958
##  5  38                  8 2       5                208   35.1  2.86    1.04
##  6  22.7               20 2       4                120   23.6 -0.916  -0.335
##  7  69.1               10 2       4                428   65.8  3.32    1.22
##  8  46.7                5 2       5                247   40.4  6.30    2.30
##  9  21.8               16 2       1                135   23.9 -2.07   -0.754
## 10  24.4               15 2       3                138   25.1 -0.677  -0.246
## # ... with 211 more rows, and 1 more variable: HAT <dbl>
```

3. Evaluate the presence of influential observations using Cook's distance.

The plot shows there are 173 observations with high Cook's distance values.

```
HINTS.CLEAN$COOK <- cooks.distance((HINTS.CLEAN.LM1))
COOK.CUT <- 4/length(ID)
plot(COOK ~ ID, data = HINTS.CLEAN,
     ylab = "Cook's Distance",
     xlab = "ID Number",
     main = "Figure 7. Cook's Distrance by Index Plot")
abline(h = COOK.CUT, col = "red")
```



```
HINTS.CLEAN[HINTS.CLEAN$COOK > COOK.CUT,]
```

```
## # A tibble: 173 x 10
##      bmi averagedailytvg~ genderc generalhealth weight FITTED RESID RSTAND
##    <dbl>            <dbl> <fct>   <fct>          <dbl>  <dbl> <dbl>  <dbl>
##  1  31.9                6 2       1                235   37.3 -5.36 -1.95
##  2  21.4                1 1       1                121   17.0  4.38  1.60
##  3  36                 18 2       5                230   38.6 -2.62 -0.958
##  4  69.1               10 2       4                428   65.8  3.32  1.22
##  5  38.5                6 2       3                300   47.1 -8.59 -3.13
##  6  46.7                5 2       5                247   40.4  6.30  2.30
##  7  30.3                2 2       4                230   38.0 -7.74 -2.82
##  8  41.5                3 1       4                323   46.7 -5.19 -1.89
##  9  31.2                8 1       2                160   23.1  8.10  2.94
## 10  51.3                6 2       3                299   46.9  4.35  1.58
## # ... with 163 more rows, and 2 more variables: HAT <dbl>, COOK <dbl>
```

THE END