ExerciseB

Table of contents

Exercise B

We start by loading the data from the csv

```
dataWeightHeight = read_csv("C:/AppliedDataScienceAndStatistics/Applied-Data-Science-and-S
```

Initial exploratory data analysis.

Before we build our initial graph we will check if all the collums have the right data structure

```
str(dataWeightHeight)
```

```
spc_tbl_ [544 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ height: num [1:544] 152 140 137 157 145 ...
 $ weight: num [1:544] 47.8 36.5 31.9 53 41.3 ...
 $ age   : num [1:544] 63 63 65 41 51 35 32 27 19 54 ...
 $ male  : num [1:544] 1 0 0 1 0 1 0 1 0 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   height = col_double(),
  ..   weight = col_double(),
```

```
  ..   age = col_double(),
  ..   male = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

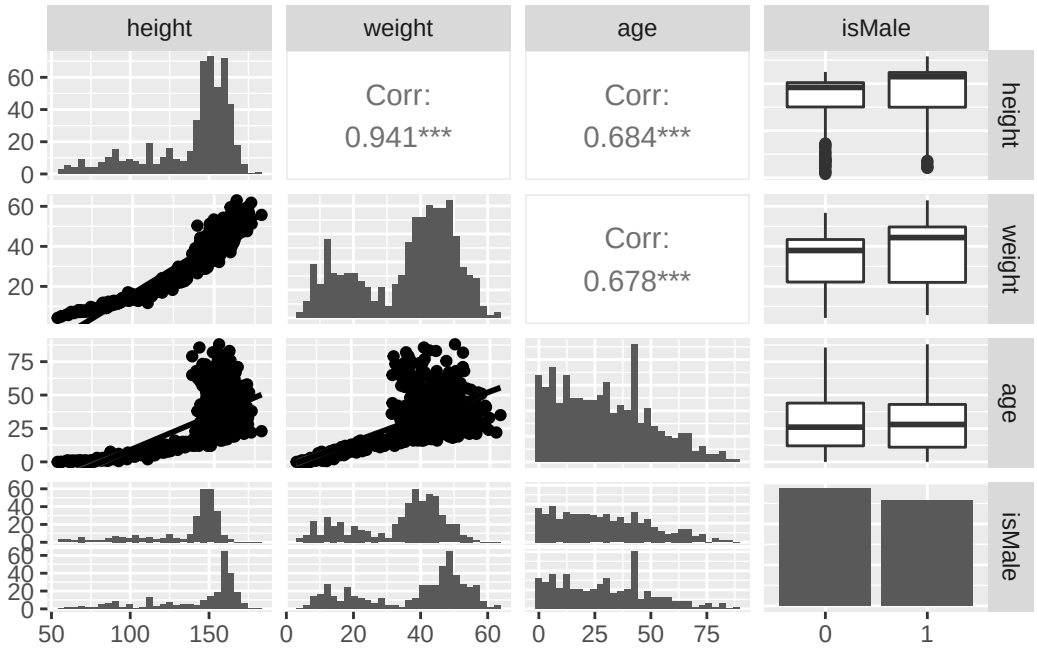
As suspected, they don't as the gender collum is loaded as a numerical, which does not make sense as a male is not bigger or smaller than a female.

```
dataWeightHeight$male = as.factor(dataWeightHeight$male)

# Change column name of the male column to become more explicit while
# we are at it
colnames(dataWeightHeight)[colnames(dataWeightHeight) == "male"] <- "isMale"

str(dataWeightHeight)
```

```
spc_tbl_ [544 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ height: num [1:544] 152 140 137 157 145 ...
 $ weight: num [1:544] 47.8 36.5 31.9 53 41.3 ...
 $ age   : num [1:544] 63 63 65 41 51 35 32 27 19 54 ...
 $ isMale: Factor w/ 2 levels "0","1": 2 1 1 2 1 2 1 2 1 2 ...
 - attr(*, "spec")=
  .. cols(
  ..   height = col_double(),
  ..   weight = col_double(),
  ..   age = col_double(),
  ..   male = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
ggpairs(dataWeightHeight, lower = list(continuous = "smooth"), diag = list(continuous = "b
    axisLabels = "show")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Looking at the height histogram it seems that height follows a normal distribution that is very heavily skewed to the left. The weight seem to follow a bimodal distribution, a distribution that is a mixture of two distinct distributions with same variance but different mean, of two normal distributions. Age seems to follow a slowly decreasing exponential distribution with a very high outlier in the middle. As for gender we can see there is slightly more female entries then male entries. More importantly comparing the other values with weight: Height follows a curvilinear increase that appears to increase almost exponentially. Age on the first half of the value range seems to follow a pretty linear relationship with weight till around age 30 where it has a pretty sharp increase in standard error compared to the average of the values. Furthermore, males follows the a very similar bimodal relationship to females but the males right distribution is more tilted to the right than the female one. Comparing the independent variables with weight, height seems to have a exponential relationship with weight as the increase was initially minimal but it gradually increases. Lastly, age after 30 years of age seems to a very big variance on its relationship with weight making is a much less reliable weight predictor.

Data separation prior to modeling

Before starting to compute our models we will first separate the data into training and testing data sets, using the caTools library

```r
# make this example reproducible
set.seed(1)

# use 80% of dataset as training set and 20% as test set
sample <- sample.split(dataWeightHeight$weight, SplitRatio = 0.8)
## note that the column selected above can be any column.
train <- dataWeightHeight[sample, ]
test <- dataWeightHeight[!sample, ]
```

Linear model

```
Call:
lm(formula = weight ~ height + age + isMale, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9177  -3.7974   0.0003   3.0598  13.4653

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.20361    1.38727 -23.214   <2e-16 ***
height        0.47893    0.01209  39.602   <2e-16 ***
age           0.03839    0.01571   2.445   0.0149 *
isMale1       0.74046    0.48475   1.527   0.1274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.958 on 431 degrees of freedom
Multiple R-squared:  0.8862,    Adjusted R-squared:  0.8854
F-statistic:  1119 on 3 and 431 DF,  p-value: < 2.2e-16


      RMSE        R2
1 4.849264 0.8967102
```

As we can see from the R-squared a simple linear model already explains approximately 89,6% of the variance that influence a person's weight. According to our RMSE this model will on average have a prediction error of 4.849264. Also from the p-values we can see that the gender is not as significant as the other values and therefore I will be removing it on the next model

Next we will try to improve the predictive power of our model by changing the relationship between height and weight from a simple linear regression to a Logarithmic Regression to

better represent the exponential increase in height to weight relationship we had observed in the previous graph.

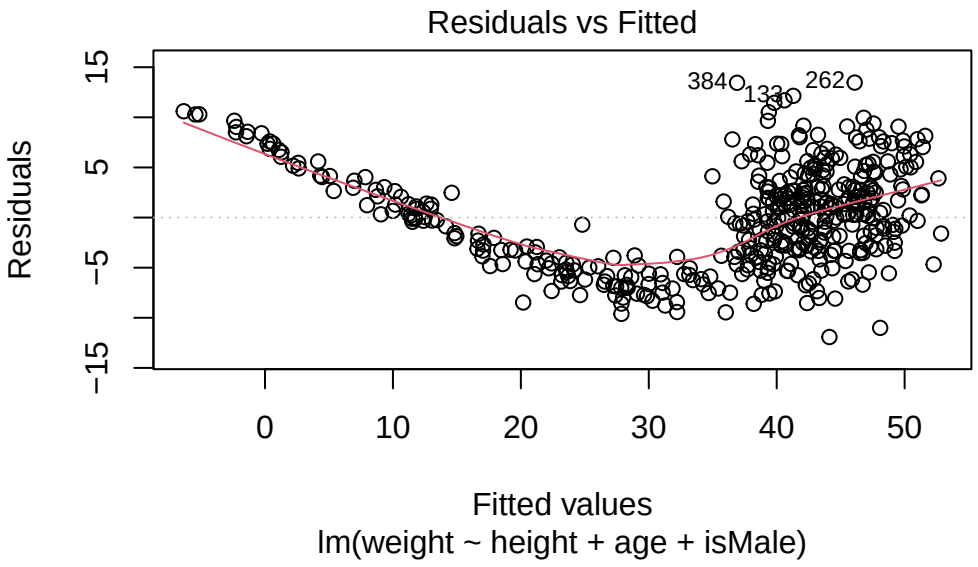
Improved model

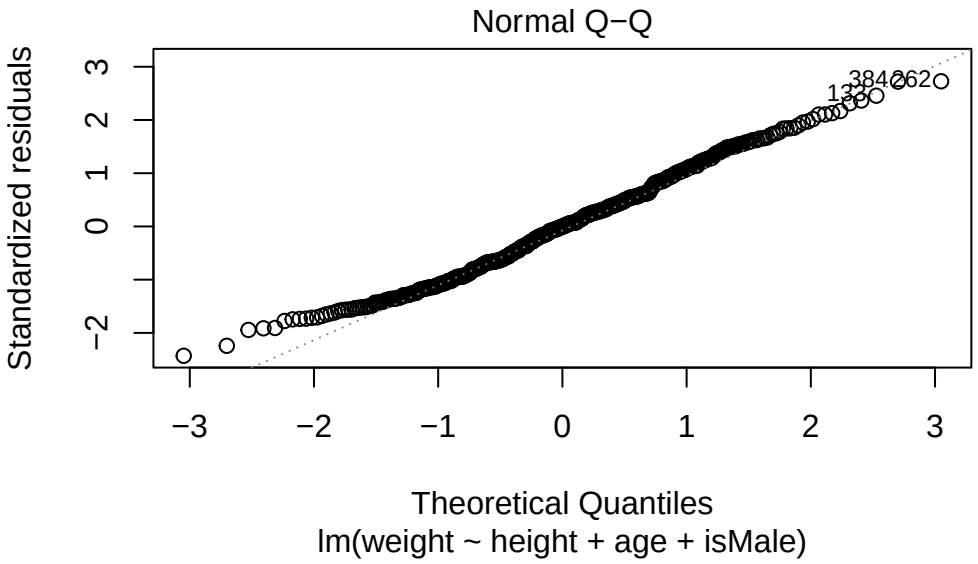
```
Call:
lm(formula = log(weight) ~ log(height) + age, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46070 -0.08547  0.00579  0.08466  0.38320

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.4347587  0.1556185 -47.776  < 2e-16 ***
log(height)  2.2056176  0.0332092  66.416  < 2e-16 ***
age          0.0019971  0.0003824   5.223 2.75e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1248 on 432 degrees of freedom
Multiple R-squared:  0.9531,    Adjusted R-squared:  0.9529
F-statistic:  4393 on 2 and 432 DF,  p-value: < 2.2e-16


       RMSE        R2
1 0.1086593 0.9654727
```

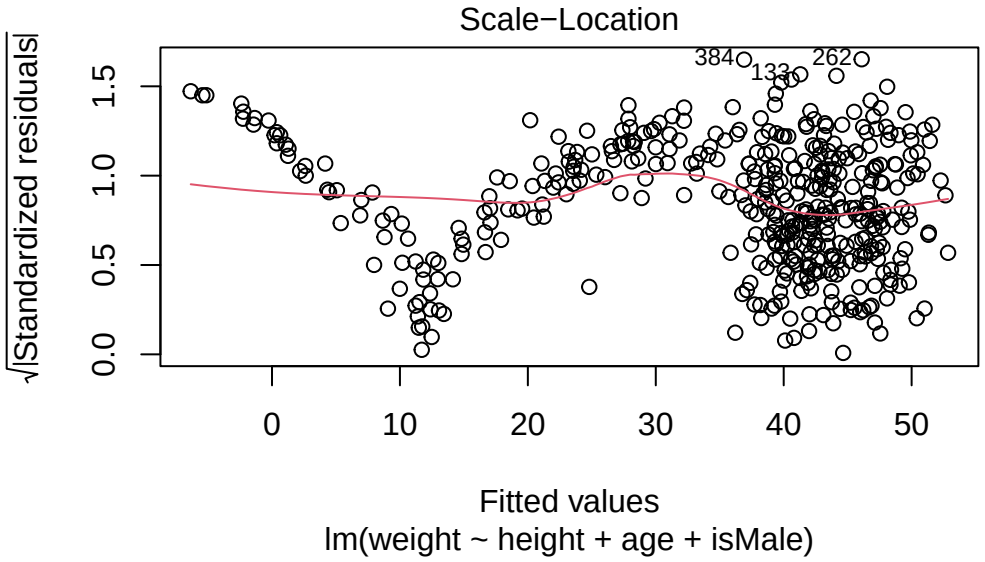
As we can immediately see is that we have managed to increase the R-squared value from 89,6% to 96,5% which is a significant upgrade. The RMSE value has also been massively improved, reducing the average error from 4.849264 to a mere 0.1086593 meaning our model has very little prediction error now.

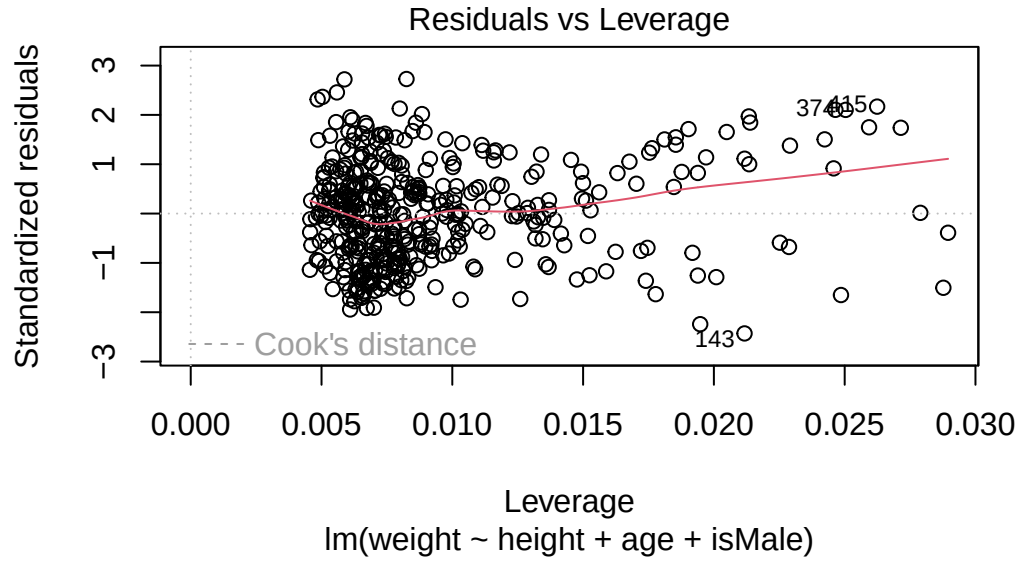
Now let us plot the Residuals vs fitted values and qq plots to comfirm that the second model is in fact better.

```
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(weight ~ height + age + isMale)
384
133
262


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(weight ~ height + age + isMale)
384
262

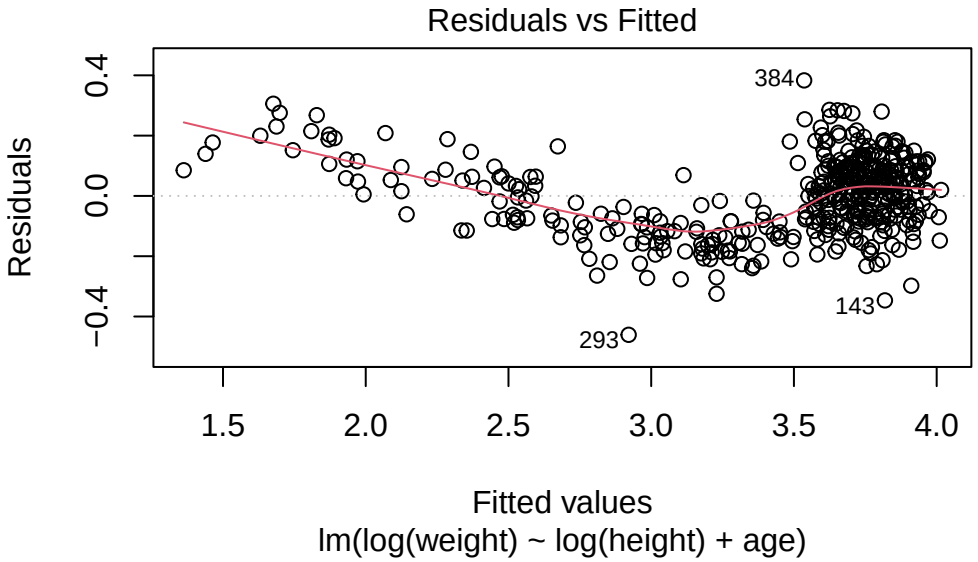


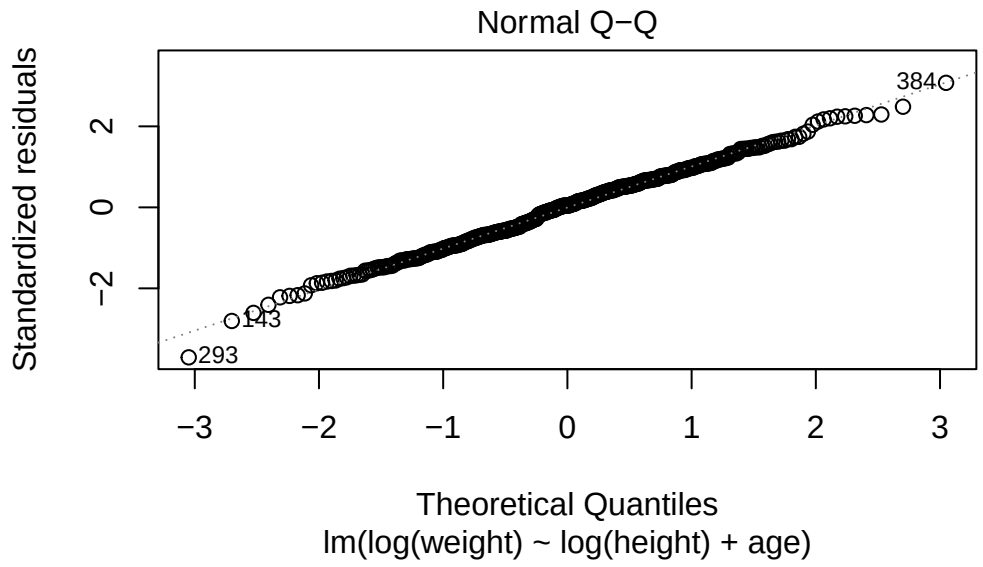


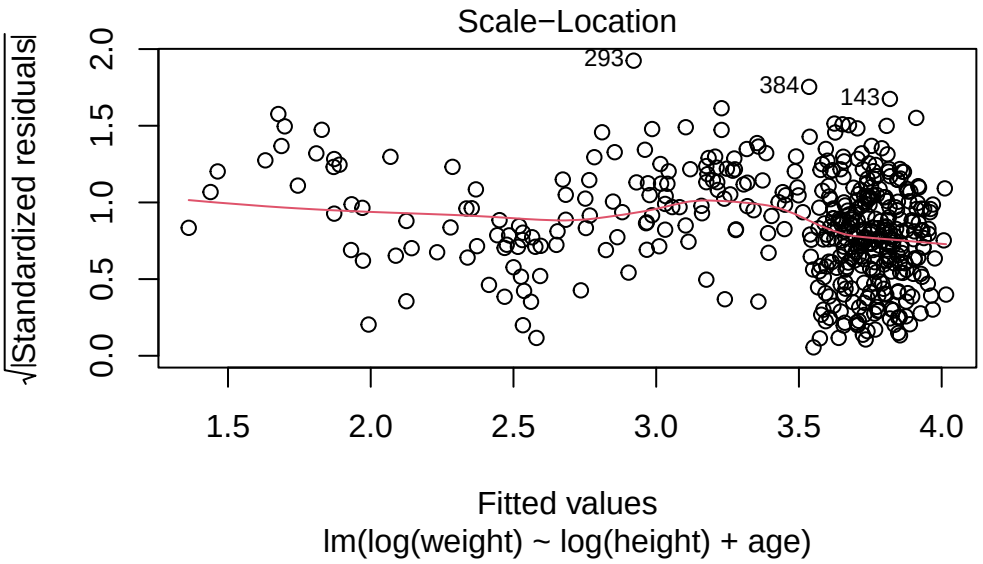
The residuals vs fitted plot is a graphical representation of the residuals (the difference between the observed value and the predicted value) of a regression model as a function of the fitted

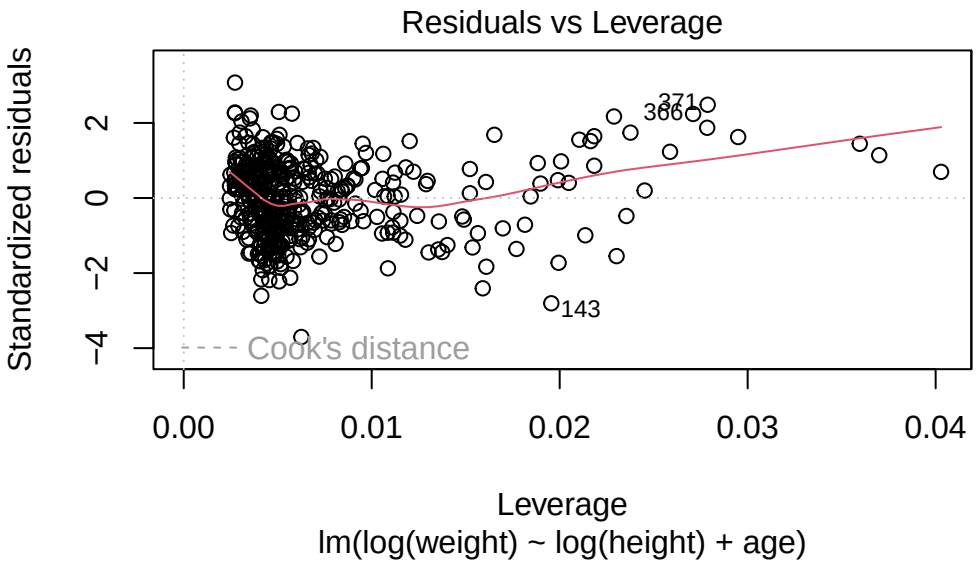
(predicted) values. As such the closer the residuals are from the horizontal line y = 0 the better the fit is the model, has its predictions are closers to the real values.

```
plot(model2)
```




Normal Q−Q
Standardized residuals
384
143
293
−3
−2
−1
0
1
2
3
−2
0
2
Theoretical Quantiles
lm(log(weight) ~ log(height) + age)


Scale−Location
√|Standardized residuals|
293
384
143
0.0
0.5
1.0
1.5
2.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(log(weight) ~ log(height) + age)



As we can see from the residuas vs fitted values we have a significantly better residuals as the distance from the y = 0 is incredibly small and therefore the model has excellent prediction power with very minimal prediction error on all stages. We can also see how much more concentrated around y = 0 at the end of the second model compared to the first one which greatly exemplifies how much more accurate this model is with its reduced heteroscedasticity (non-constant variance of the errors), especially considering that the simplicity has been mantained in terms of model construction.