BODY FAT

Dorothy Miruka

2020-10-20

Contents

1 Introduction

The dataset `Bodyfat4` shows the body fat percentage for 252 individuals. There are 13 variables. First, there is an identifier variable (IDNO) that uniquely identifies each observation. The body fat (BODYFAT) content is the dependent variable. The remaining eleven (11) are the independent variables that could explain the variability of body fat among the different individuals. The dataset has no missing values. The task is to build a regression model that can be useful in explaining and predicting the body fat of individuals. The data analysis proceeds as follows. In the next section, I visualize the data and then present summary statistics. Next, I present the correlation analysis and then the results of the regression.

2 Analysis and Interpretation of Results

2.1 Descriptive Statistics

Table () below shows a summary of the variables. Among the dependent variables, the highest variability(SD-standard deviation) is observed between individuals' abdomen, age, and weight. For the dependent variable body fat, the mean is 18.9, with an unusual observation of zero body fat as the minimum value. Given that there is a base level of body fat necessary for humans to survive, this is a peculiar observation. For robustness, it would be advisable to drop this case.

Variable	Complete	Mean	SD	Min	Q1	Median	Q3	Max
bodyfat	1	18.938492	7.7508557	0.000	12.8000	19.0000	24.6000	45.1000
density	1	1.055574	0.0190314	0.995	1.0414	1.0549	1.0704	1.1089
age	1	44.884921	12.6020397	22.000	35.7500	43.0000	54.0000	81.0000
weight	1	178.924405	29.3891599	118.500	159.0000	176.5000	197.0000	363.1500

Variable	Complete	Mean	SD	Min	Q1	Median	Q3	Max
height	1	70.148809	3.6628558	29.500	68.2500	70.0000	72.2500	77.7500
adiposity	1	25.436905	3.6481108	18.100	23.1000	25.0500	27.3250	48.9000
thigh	1	59.405952	5.2499520	47.200	56.0000	59.0000	62.3500	87.3000
abdomen	1	92.555952	10.7830768	69.400	84.5750	90.9500	99.3250	148.1000
ankle	1	23.102381	1.6948934	19.100	22.0000	22.8000	24.0000	33.9000
hip	1	99.904762	7.1640577	85.000	95.5000	99.3000	103.5250	147.7000
wrist	1	18.229762	0.9335849	15.800	17.6000	18.3000	18.8000	21.4000
knee	1	38.590476	2.4118046	33.000	36.9750	38.5000	39.9250	49.1000

2.2 Correlation Analysis

In this section, I present the pairwise plots for all the variables, excluding the identifier variable (IDNO). The upper part of the figure shows the pairwise correlation between the variables. The main diagonal shows the distribution of the respective variable. For instance, the plot on the intersection of the first row and the first column shows the distribution of the body fat (BODYFAT) variable. The lower half shows the scatter plots for each pair of variables.

Figure 3.1:

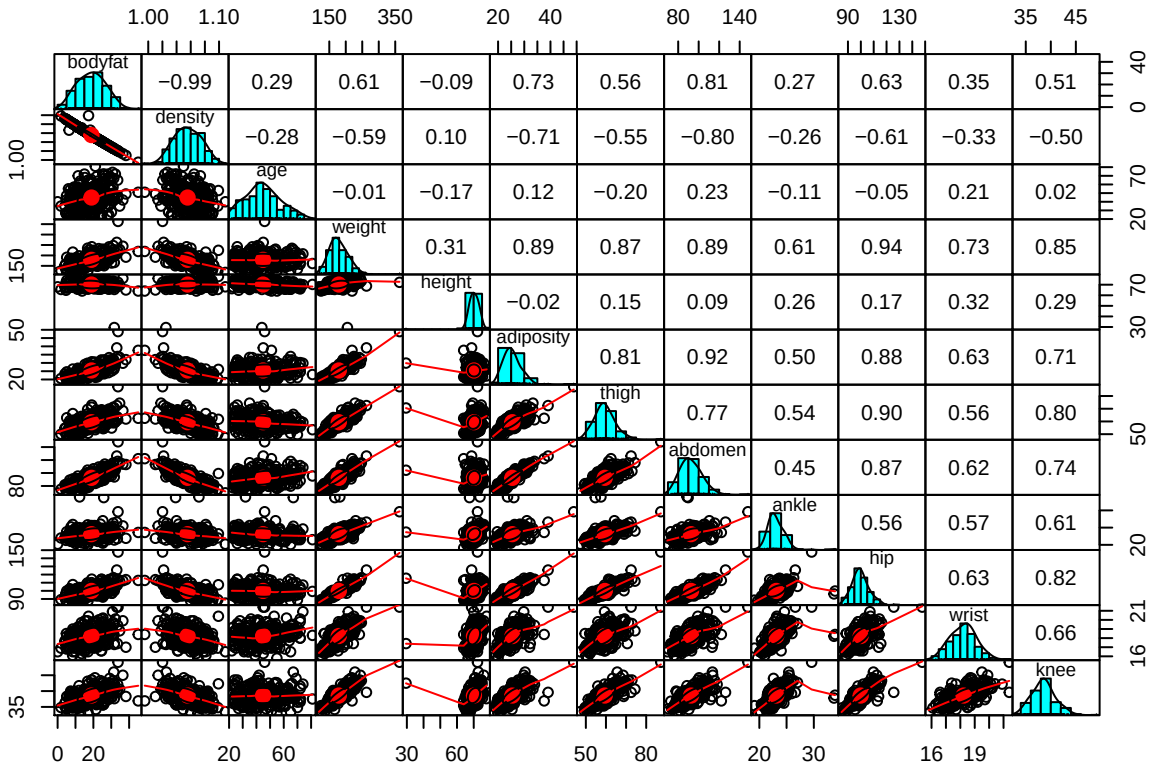
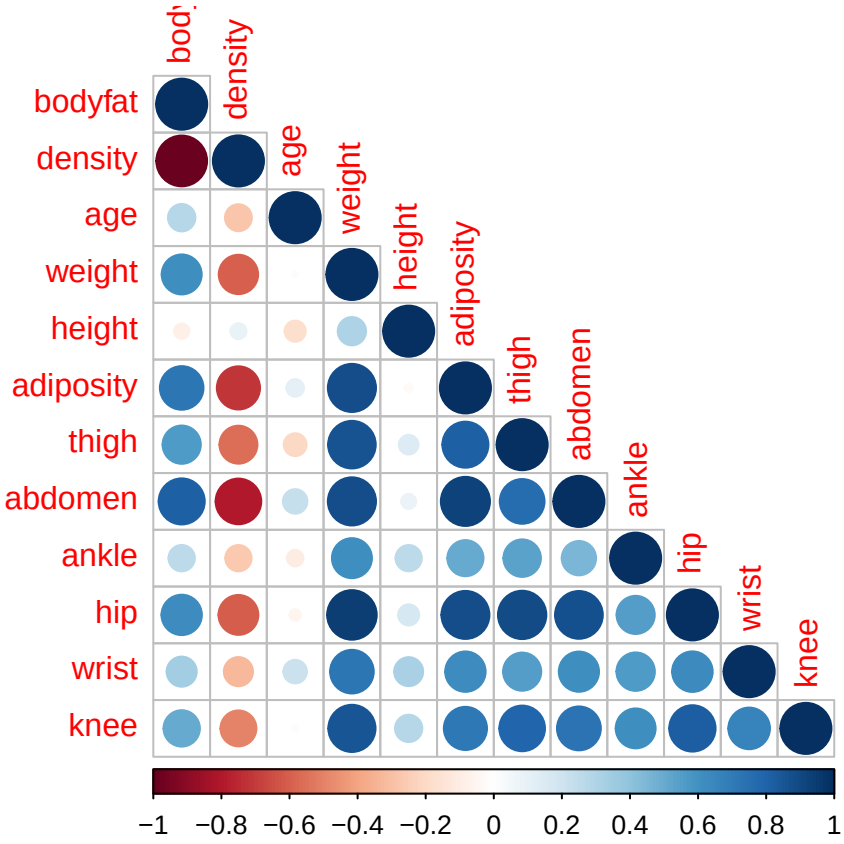


Figure () below is an extension of Figure () above and visualizes the correlation between the variables with the circle sizes showing the size of the correlation while the colour shows the direction (positive or negative).The red colouring corresponds to negative correlation while blue shows positive correlation. White is neutral indicating little correlation in either direction. It is important to note that in interpreting this visualization, I am Ignoring the main diagonal (which is the correlation of a variable with itself).



The correlation shows the high levels of correlation between most of the variables in the dataset. I focus on the dependent variables as a high degree of correlation between them leads to the problem of multicollinearity. Using a cut off of 0.7 that some researchers have used (). The correlation analysis shows 18 pairs of independent variables have a correlation beyond the threshold of 0.7. In the cases where correlation exceeds the threshold, the variables involved are Adiposity, weight, and abdomen (6 times each), knee, thigh and hip (5 times each), density (twice), and wrist (once). when we lower the threshold to 0.6, we get 25 pairs of variables with correlation between the threshold.

Multicollinearity is a problem because the independent variables in a regression model should ideally be independent. If not, the regression coefficient estimates become unstable (that is they vary a lot). The reduced precision of the estimates reduces the predictive power of the model (Frost 2019a). Scholars have suggested several solutions to the multicollinearity problem. The first is to drop the variables that exhibit multicollinearity. The second common technique is to use principal component analysis (PCA) to generate a new set of new variables (called components) that are linear combinations of the original variables but that have low correlation between them. In the next section, I describe and run the principal components analysis.

2.3 Principal Components Analysis (PCA)

Principal components analysis (PCA) is a technique for dimension reduction, increasing interpretability while minimizing the loss of information (Jolliffe and Cadima 2016). PCA is especially useful where there are many variables in a dataset or as in this case where variables are highly correlated. The dimensions of the body fat dataset could be reduced to get rid of the highly correlated dependent variables by creating a new set of fewer variables called principle components that are not correlated. The principal components are linear combinations of the original variables that capture much of the variation in the original dataset. In this section, I construct principal components and discuss the applicability and limitations of the PCA analysis.

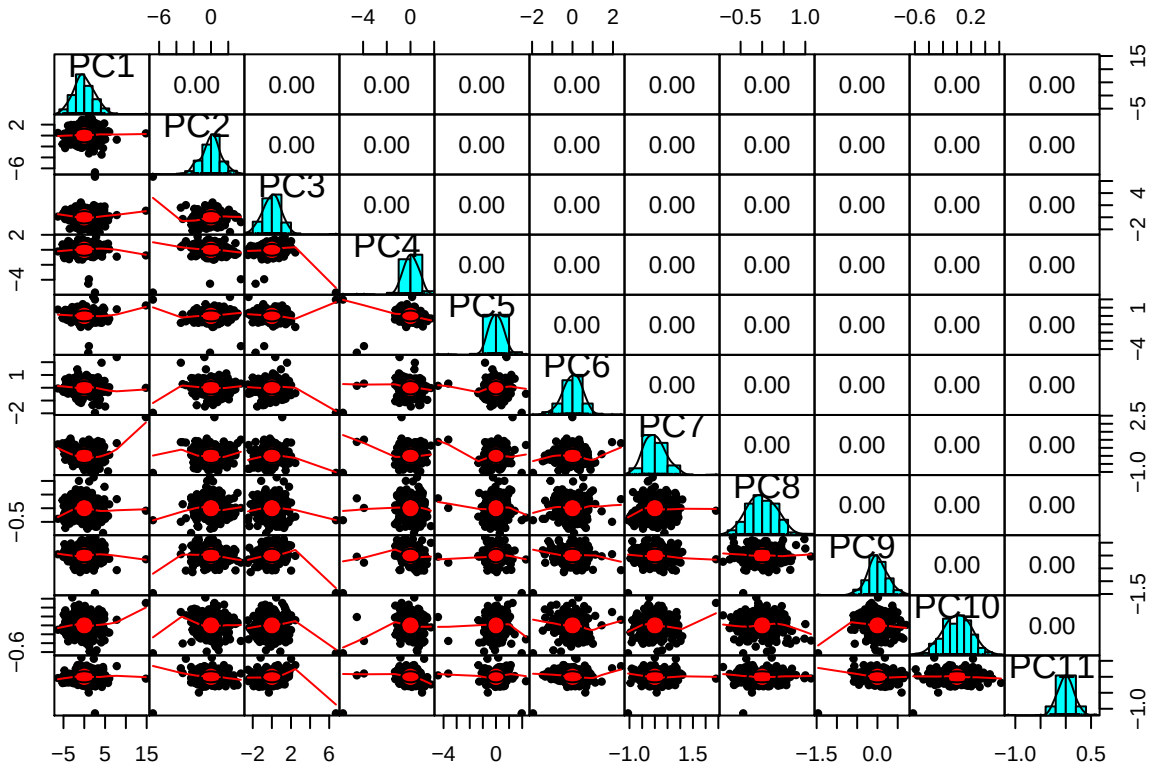
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	
density	-0.2680344	0.4082056	-0.1055719	-0.3038013	0.5643716	-0.1637701	0.5083826	-0.0536521	0.10
age	0.0237235	-0.5610130	-0.6643508	-0.0569772	0.0024084	-0.2972527	0.2584667	0.2769649	-0.01
weight	0.3772631	0.0823905	-0.0003205	0.0886518	0.1049909	0.0514025	0.2051173	-0.1631618	-0.03
height	0.0853037	0.5586899	-0.4611461	0.5938833	-0.2237492	0.0558422	0.1253866	0.0934278	0.10
adiposity	0.3577718	-0.1841652	0.1293031	-0.0331106	0.1083146	0.2605237	0.3115525	-0.1569064	0.56
thigh	0.3461356	0.1012755	0.2906921	0.0387737	0.1639463	-0.1722729	-0.0688549	0.8095963	0.20
abdomen	0.3593055	-0.2180146	-0.0010456	0.1806343	-0.0447768	0.1295240	0.2098289	-0.2458256	0.03
ankle	0.2535765	0.2867944	-0.1080678	-0.6610140	-0.6093952	0.0258560	0.1602675	0.0634257	-0.01
hip	0.3683677	0.0255776	0.1521391	0.0651091	0.1484966	-0.0724567	0.2782068	0.0060735	-0.75
wrist	0.2903686	0.0983681	-0.4420528	-0.2572835	0.4238267	0.4564693	-0.4922601	0.0220862	-0.09
knee	0.3417161	0.1241960	-0.0581284	-0.0360538	0.0820601	-0.7419231	-0.3545537	-0.3741894	0.16

2.3.1 Running PCA

PCA is sensitive to differences in scale between the variables used. Hence, it is recommended to first scale the variables (Zhu, Ge, and Song 2017). In this case, I scale the variables by subtracting the mean and dividing by the standard deviation.

```
## $sdev
##  [1] 2.5769889 1.2466680 1.0195907 0.7856985 0.6512723 0.5231075 0.4405919
##  [8] 0.3389034 0.2627786 0.2057657 0.1728994
```

Note that the PCA allows us to deal with the problem of correlation between variables. Table () below shows that the correlations between the principal components is now zero.



2.3.2 Composition of PCA

The PCA output has two components

- Rotations

Table () shows the rotations where the original variables have reduced to the 11 principal components. As an example, principal component one (PC1) is formed by combining the respective variables using the weights given.

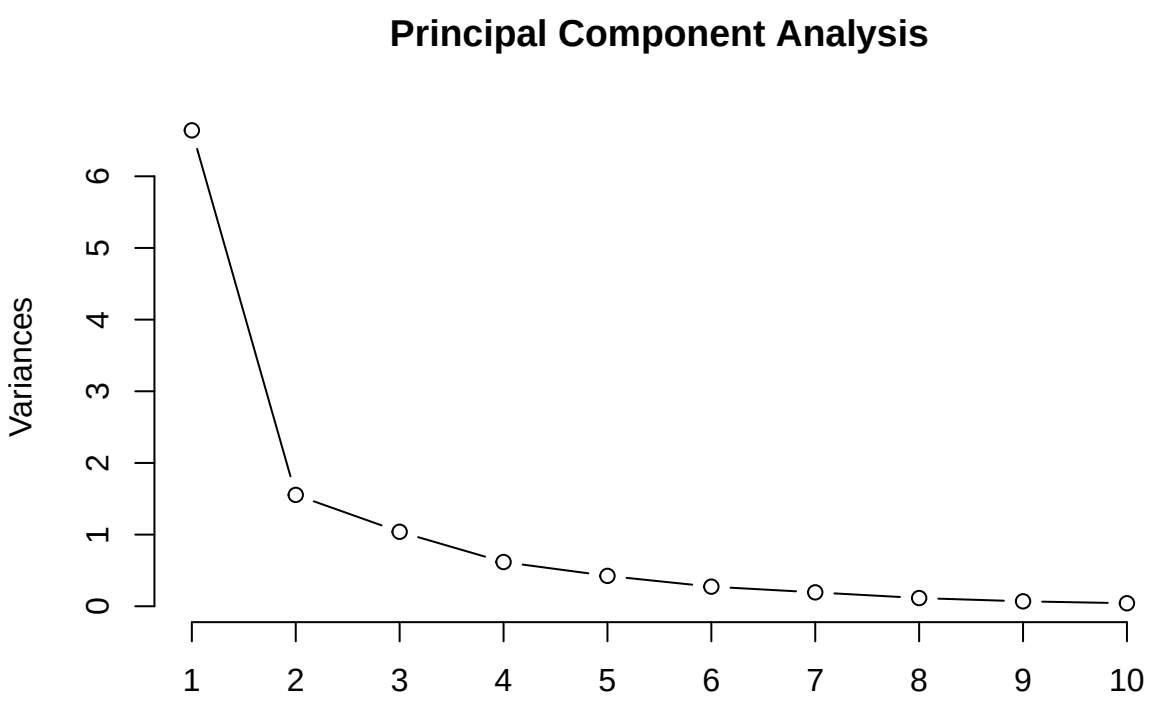
PC1 = -0.27 x density + 0.02 x age + ... + 0.34 x knee

Note that I have rounded the coefficients to 2 decimal places.

- Standard Deviations and importance of components

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5770 1.2467 1.01959 0.78570 0.65127 0.52311 0.44059
## Proportion of Variance 0.6037 0.1413 0.09451 0.05612 0.03856 0.02488 0.01765
## Cumulative Proportion  0.6037 0.7450 0.83951 0.89563 0.93419 0.95907 0.97671
##                           PC8     PC9    PC10    PC11
## Standard deviation     0.33890 0.26278 0.20577 0.17290
## Proportion of Variance 0.01044 0.00628 0.00385 0.00272
## Cumulative Proportion  0.98716 0.99343 0.99728 1.00000
```

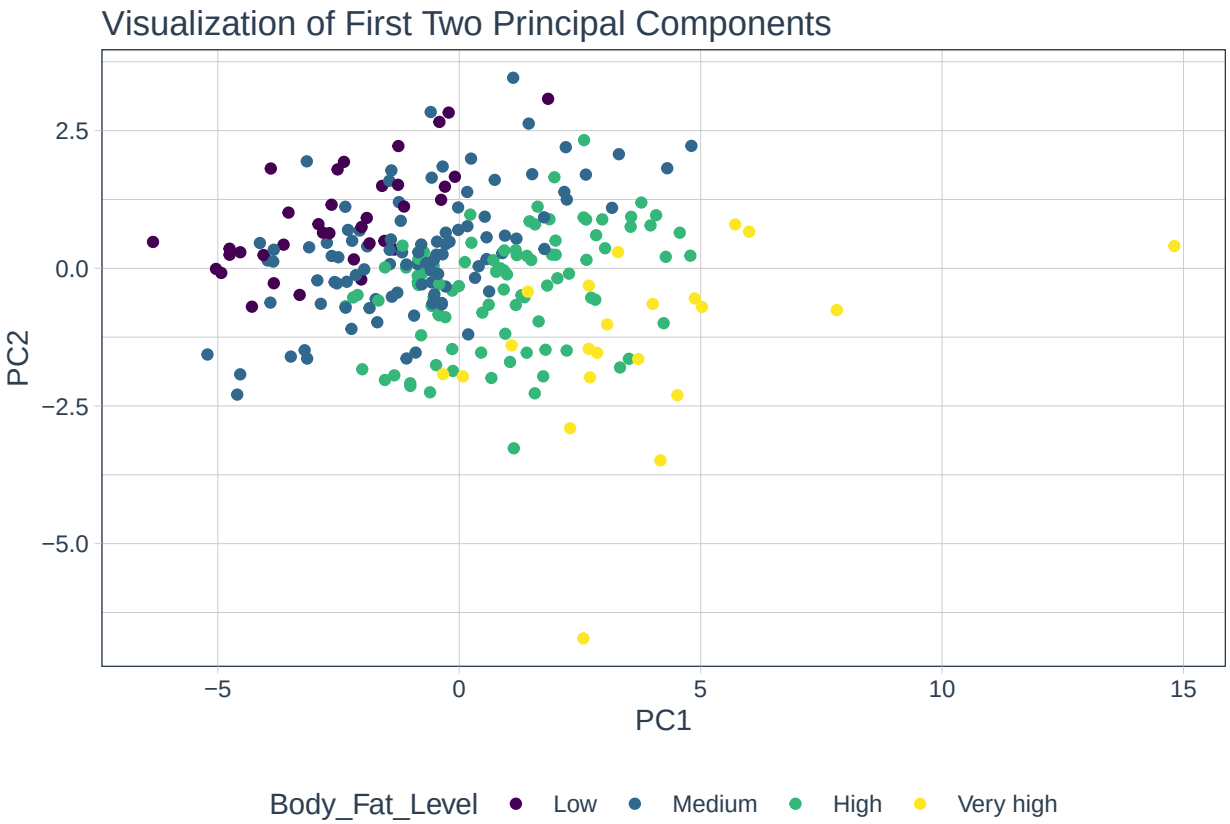
The standard deviation refers to the variability of the data along a principal component. The proportion of variance is the proportion of total variability in the data captured by the first principal component. For instance, principal component 1 accounts for 60.37% (0.6037) of the data. The cumulative proportion sums up the variability captured by the respective principal components. In column 3, for instance, PC1, PC2, and PC3 together account for 89.951% (0.83951) of the variation in the data. Similarly, the first two principal components capture 74.5% of the variation in the data. In the figure () below, I visualize the proportion of variance explained away by the respective principal component (on the x-axis), starting with PC1. The figure illustrates that PC1 and PC2 account for most of the variability in the data. However, the other principal components also do account for a significant but smaller proportion.



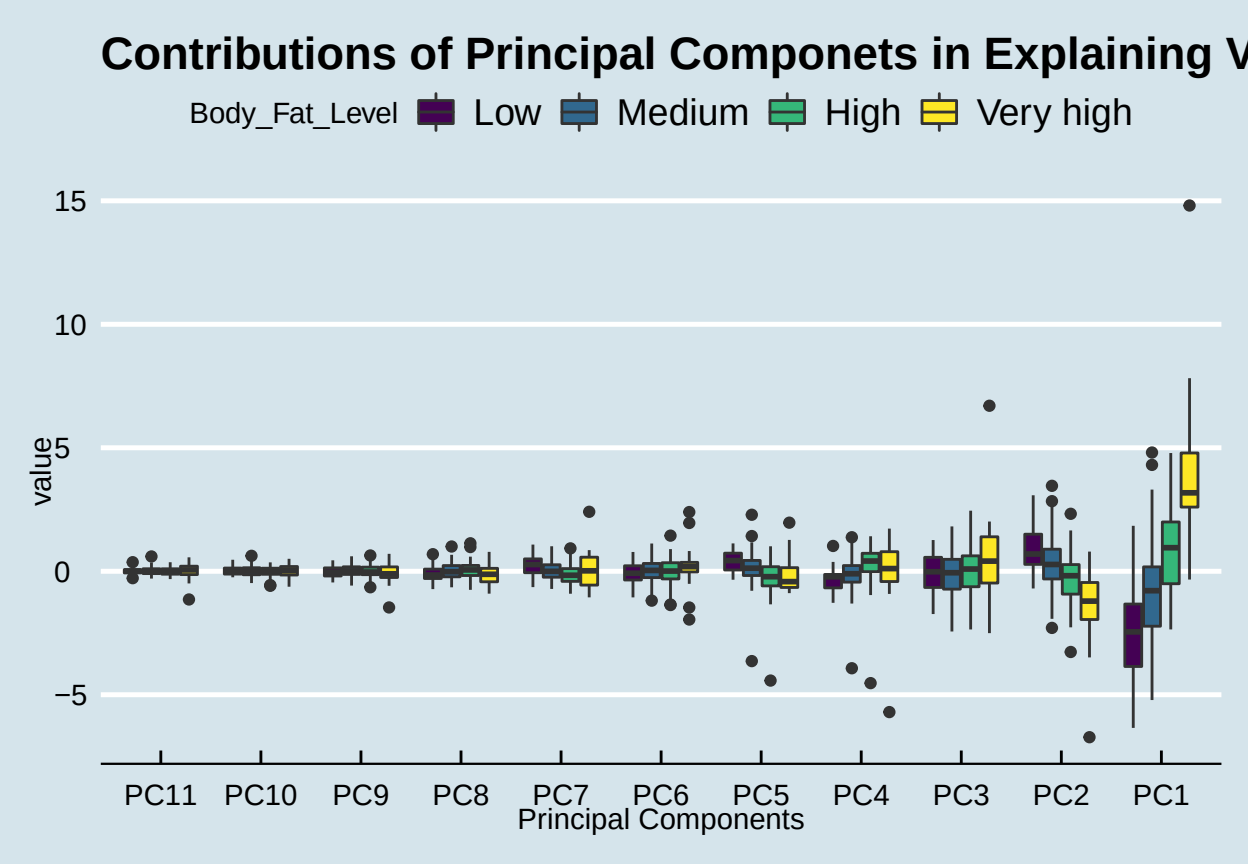
2.3.3 Exploring the Principal Components

Given that the first two principal components account for almost 75% of the variation, I will explore them more. I extract the principal components and attach them to the dataset BodyFat. Plotting PC1 against PC2 shows that PC2 captures people with higher body fat compared to PC1. For ease of comparison, I categorize individuals into four categories according to the level of body fat.

- Low: 0 to 10 inclusive
- Medium: Greater than 10 but less than or equal to 20.
- High Greater than 20 but less than or equal to 30.
- Very high: Greater than 30.



I also visualize each of the principal components faceted by the `Body_Fat_Level` defined above.



lastly, we examine the correlation between PC1, PC2 and PC3 and the variables of BodyFat dataset. Table () shows that PC1 has very high correlation with most of the variables. For instance, PC1 has a correlation greater than 0.7 (absolute value) with 7 of the 11 independent variables. PC2 is not so strong but still has a correlation above 0.6 (absolute) with 2 variables, while PC3 has a correlation greater than 0.6 with one variable. In this case, it appears like PC1 and PC2 are adequate to capture most of the variability in the data.

	PC1	PC2	PC3
density	-0.6907217	0.5088969	-0.1076402
age	0.0611353	-0.6993969	-0.6773658
weight	0.9722029	0.1027136	-0.0003268
height	0.2198266	0.6965008	-0.4701802
adiposity	0.9219741	-0.2295929	0.1318363
thigh	0.8919875	0.1262570	0.2963869
abdomen	0.9259263	-0.2717918	-0.0010661
ankle	0.6534638	0.3575373	-0.1101849
hip	0.9492794	0.0318868	0.1551197
wrist	0.7482767	0.1226324	-0.4507129
knee	0.8805986	0.1548312	-0.0592672

2.4 Regression Analysis

In this section, I start by doing a forward stepwise regression followed by the required backward stepwise regression. Forward stepwise regression starts with no independent variables in the model and iteratively adds predictors that are the most significant. The process stops when the improvement is no longer statistically

significant. On the contrary, the backward stepwise regression starts with all the independent variables in the model and removes the least contributing variables until all variables remaining are significant (Bruce, Bruce, and Gedeck 2020)(James et al. 2013).

2.4.1 Forward Stepwise Regression

I include the forward stepwise selection method for purposes of comparison. Note that the optimal model here includes all variables, although only the density is significant. Overall, the model captures 97.74% of the variation in body fat (see adjusted R squared in the table below).

```
## Start:  AIC=88.26
## bodyfat ~ density + age + weight + height + adiposity + thigh +
##     abdomen + ankle + hip + wrist + knee
```

```
##
## Call:
## lm(formula = bodyfat ~ density + age + weight + height + adiposity +
##     thigh + abdomen + ankle + hip + wrist + knee, data = BodyFat4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7873 -0.3008 -0.0746  0.1748 14.2447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.202e+02  9.315e+00  45.107   <2e-16 ***
## density     -3.814e+02  7.365e+00 -51.782   <2e-16 ***
## age          9.340e-03  8.623e-03   1.083    0.280
## weight       1.324e-02  1.273e-02   1.040    0.299
## height      -1.936e-02  3.009e-02  -0.643    0.521
## adiposity   -4.078e-02  7.374e-02  -0.553    0.581
## thigh       -1.931e-02  3.763e-02  -0.513    0.608
## abdomen      3.499e-02  2.820e-02   1.240    0.216
## ankle       -7.042e-02  5.979e-02  -1.178    0.240
## hip          1.182e-02  3.745e-02   0.315    0.753
## wrist        2.449e-02  1.373e-01   0.178    0.859
## knee        -3.059e-02  6.630e-02  -0.461    0.645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.164 on 240 degrees of freedom
## Multiple R-squared:  0.9784, Adjusted R-squared:  0.9774
## F-statistic: 989.9 on 11 and 240 DF,  p-value: < 2.2e-16
```

2.4.2 Backward Stepwise Regression

As noted, backward step selection begins with all the variables in the model and iteratively removes the least contributive ones. The Akaike Information Citerion (AIC) informs the selection of the best model, with the optimal model being the one with the least AIC. The series of tables below show the procedure (note the AIC value stated at the beginning of each model). In the end, the model consisting of density, age, and abdomen as independent variables has the least AIC (75.39). Note that the model selected in this case is as follows ;

bodyfat = 0.041 - (0.038 * density) + (9.51 * age) + (0.048 * abdomen)

The AIC captures the extent to which the model fits the data without being overly complex. Complexity refers to the number of parameters in the model and fittingly, AIC is the difference between the number of parameters in the model (k) and the maximum value of the likelihood function of the model (L).

AIC = (2 * k) - (2 * ln(L)), where ln stands for natural log.

The central idea of AIC is to balance between fitting the data and the need not to overfit the dataset (Portet 2020).

The model is captured at the bottom of the series of tables. In the model, only age is not significant at 10% significance level. As expected, the model explains 97.79% of the variation in body fat - the adjusted R-Squared - which is a high level of in sample sensitivity. The adjusted R-Squared derives from the R-squared, a goodness-of-fit measure for linear regression models that indicates the percentage of the variance in the dependent variable that the independent variables explain collectively (Frost 2019b). Like the AIC, the adjusted R-squared penalizes the R-squared as the model gets more complex, that is, gets more predictors.

In our body fat case, body fat is negatively related to body fat. Specifically, for every one unit increase in density, body fat reduces by 0.038 units (rounded to two decimal places). Abdomen relates positively with body fat with a one unit rise in abdomen associated with 0.048 units rise in body fat. Although not significant at 10%, age has a positive relationship with body fat, with a one unit increase in age associated with a 9.5 units rise in body fat, a relatively large amount.

Furthermore the model as a whole is highly significant going by the F-test . The F-test of overall significance tests how well the linear regression model better fits the data compared to a model that has no independent variables or in other words where the coefficients of the independent variables are all zero. In out case the F-test is highly significant meaning that the model better explains the model better than a model with no predictors. If our model fits the data just as well as a model with no predictors, then the F_statistic should be close to one. In this case it is 3703.

```
## Start:  AIC=88.26
## bodyfat ~ density + age + weight + height + adiposity + thigh +
##     abdomen + ankle + hip + wrist + knee
##
##             Df Sum of Sq    RSS    AIC
## - wrist      1       0.0  325.2  86.29
## - hip        1       0.1  325.3  86.36
## - knee       1       0.3  325.5  86.48
## - thigh      1       0.4  325.5  86.53
## - adiposity  1       0.4  325.6  86.58
## - height     1       0.6  325.8  86.69
## - weight     1       1.5  326.7  87.39
## - age        1       1.6  326.8  87.49
## - ankle      1       1.9  327.1  87.71
## - abdomen    1       2.1  327.3  87.87
## <none>                    325.2  88.26
## - density    1    3633.1 3958.3 716.04
##
## Step:  AIC=86.29
## bodyfat ~ density + age + weight + height + adiposity + thigh +
##     abdomen + ankle + hip + knee
##
##             Df Sum of Sq    RSS    AIC
## - hip        1       0.1  325.4  84.38
## - knee       1       0.3  325.5  84.50
## - thigh      1       0.4  325.6  84.56
## - adiposity  1       0.4  325.6  84.59
## - height     1       0.5  325.8  84.70
## - weight     1       1.7  327.0  85.63
## - ankle      1       1.8  327.1  85.72
## - abdomen    1       2.1  327.3  85.88
```

```
## - age        1       2.1  327.3  85.90
## <none>                    325.2  86.29
## - density    1    3792.0 4117.3 723.97
##
## Step:  AIC=84.38
## bodyfat ~ density + age + weight + height + adiposity + thigh +
##     abdomen + ankle + knee
##
##             Df Sum of Sq    RSS    AIC
## - knee       1       0.2  325.6  82.57
## - thigh      1       0.3  325.6  82.58
## - adiposity  1       0.4  325.7  82.68
## - height     1       0.6  326.0  82.88
## - ankle      1       1.9  327.2  83.83
## - age        1       2.0  327.3  83.92
## - abdomen    1       2.4  327.7  84.22
## - weight     1       2.5  327.8  84.28
## <none>                    325.4  84.38
## - density    1    3823.2 4148.6 723.87
##
## Step:  AIC=82.57
## bodyfat ~ density + age + weight + height + adiposity + thigh +
##     abdomen + ankle
##
##             Df Sum of Sq    RSS    AIC
## - adiposity  1       0.3  325.9  80.78
## - thigh      1       0.5  326.1  80.98
## - height     1       0.6  326.2  81.04
## - age        1       1.8  327.4  81.93
## - weight     1       2.2  327.8  82.28
## - ankle      1       2.3  327.9  82.35
## - abdomen    1       2.4  328.0  82.40
## <none>                    325.6  82.57
## - density    1    3823.0 4148.6 721.87
##
## Step:  AIC=80.78
## bodyfat ~ density + age + weight + height + thigh + abdomen +
##     ankle
##
##           Df Sum of Sq    RSS    AIC
## - height   1       0.4  326.2  79.05
## - thigh    1       0.6  326.5  79.24
## - age      1       1.8  327.6  80.15
## - weight   1       2.0  327.8  80.29
## - abdomen  1       2.1  328.0  80.41
## - ankle    1       2.4  328.3  80.64
## <none>                   325.9  80.78
## - density  1    3822.9 4148.8 719.89
##
## Step:  AIC=79.05
## bodyfat ~ density + age + weight + thigh + abdomen + ankle
##
##           Df Sum of Sq    RSS    AIC
## - thigh    1       0.4  326.6  77.36
```

```
## - weight   1       1.6  327.8  78.29
## - age      1       1.9  328.1  78.52
## - ankle    1       2.4  328.6  78.91
## <none>                  326.2  79.05
## - abdomen  1       2.7  328.9  79.12
## - density  1    3834.0 4160.2 718.58
##
## Step:  AIC=77.36
## bodyfat ~ density + age + weight + abdomen + ankle
##
##           Df Sum of Sq    RSS    AIC
## - weight   1       1.2  327.8  76.29
## - ankle    1       2.4  329.1  77.24
## - abdomen  1       2.6  329.2  77.33
## <none>                  326.6  77.36
## - age      1       3.4  330.0  77.96
## - density  1    3910.9 4237.5 721.22
##
## Step:  AIC=76.29
## bodyfat ~ density + age + abdomen + ankle
##
##           Df Sum of Sq    RSS    AIC
## - ankle    1       1.4  329.3  75.39
## - age      1       2.3  330.1  76.06
## <none>                  327.8  76.29
## - abdomen  1      24.7  352.5  92.57
## - density  1    4475.0 4802.9 750.78
##
## Step:  AIC=75.39
## bodyfat ~ density + age + abdomen
##
##           Df Sum of Sq    RSS    AIC
## <none>                  329.3  75.39
## - age      1       3.3  332.6  75.92
## - abdomen  1      24.3  353.5  91.32
## - density  1    4601.0 4930.3 755.38
##
## Call:
## lm(formula = bodyfat ~ density + age + abdomen, data = BodyFat4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7157 -0.3081 -0.0784  0.2066 14.4599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.141e+02  7.681e+00  53.910  < 2e-16 ***
## density     -3.790e+02  6.437e+00 -58.869  < 2e-16 ***
## age          9.510e-03  6.008e-03   1.583    0.115
## abdomen      4.797e-02  1.122e-02   4.277 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.152 on 248 degrees of freedom
## Multiple R-squared:  0.9782, Adjusted R-squared:  0.9779
## F-statistic:  3703 on 3 and 248 DF,  p-value: < 2.2e-16
```

2.4.3 Checking the Regression Assumptions

In this section, I work with the regression model selected in section 5 above- which uses density, age and abdomen as independent variables. The regression has five key assumptions:

- *Linear relationship*: here, the assumption is that the relationship between the dependent and the independent variables is linear. a visual inspection of the figure 3.1 (see section 3) shows that with the exception of age, the relationship between body fat and the independent variables is approximately linear. Furthermore in the figure below, the plot titled residuals versus fitted also largely coloborates the linear relationship between dependent and independent variables.
- *Multivariate normality*: The presumption here is that all variables are multivariate normal. I check this assumption with the normal QQ-plot below. In this case, the plotted residuals versus theoretical quantiles show that the assumption is reasonably met as they lie on approximately the straight line.
- *No or little multicollinearity.* There appears to be a violation of this assumption as density and abdomen have a high negative correlation coefficient.

	age	abdomen
density	-0.2776372	-0.7989546

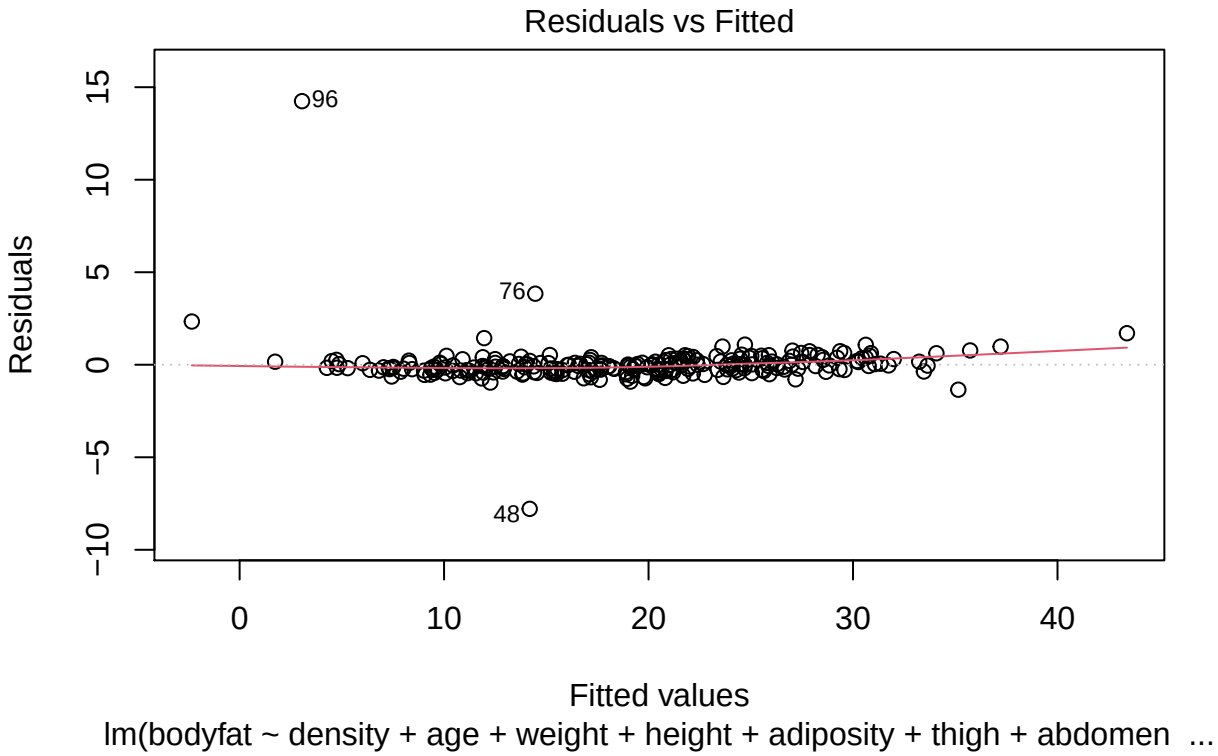
In this case, **I remove the least contributive variable**, abdomen, and rerun the model.

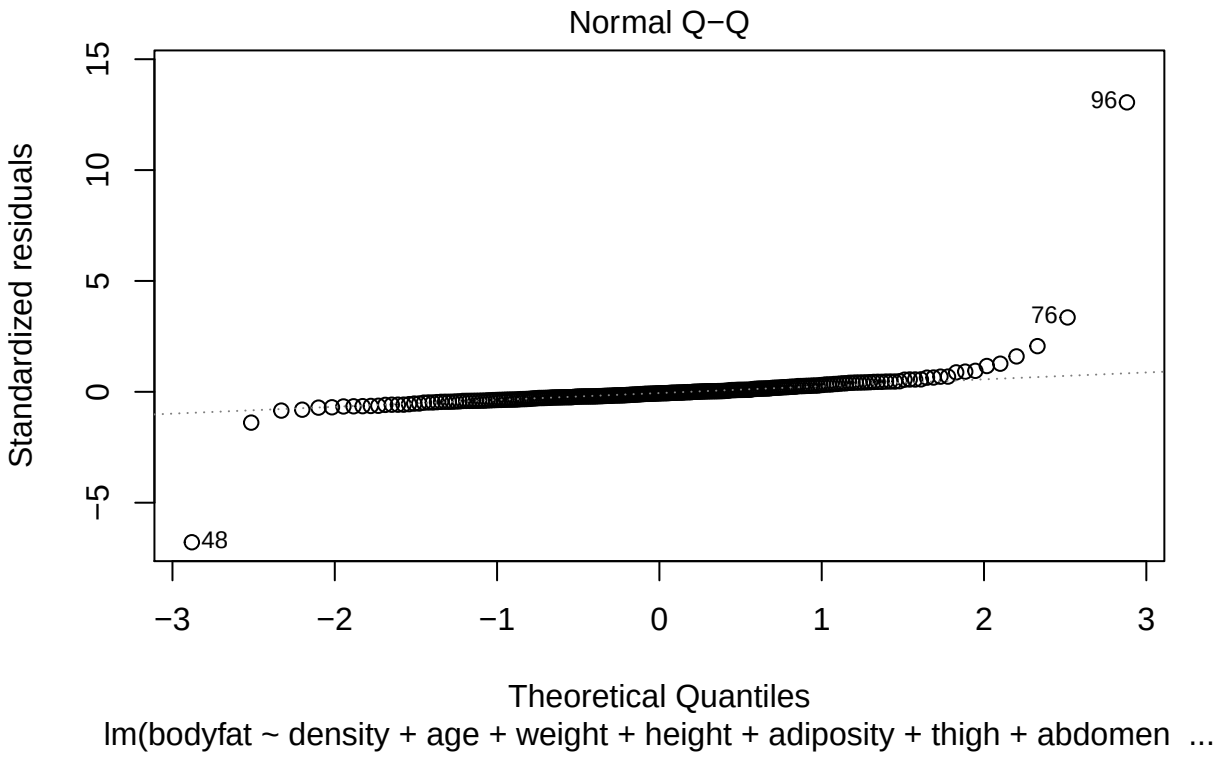
```
##
## Call:
## lm(formula = bodyfat ~ density + age, data = BodyFat4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1033 -0.2150 -0.0945  0.0513 15.7176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.414e+02  4.428e+00  99.663   <2e-16 ***
## density     -4.006e+02  4.114e+00 -97.381   <2e-16 ***
## age          9.892e-03  6.212e-03   1.592    0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 249 degrees of freedom
## Multiple R-squared:  0.9766, Adjusted R-squared:  0.9764
## F-statistic:  5186 on 2 and 249 DF,  p-value: < 2.2e-16
```

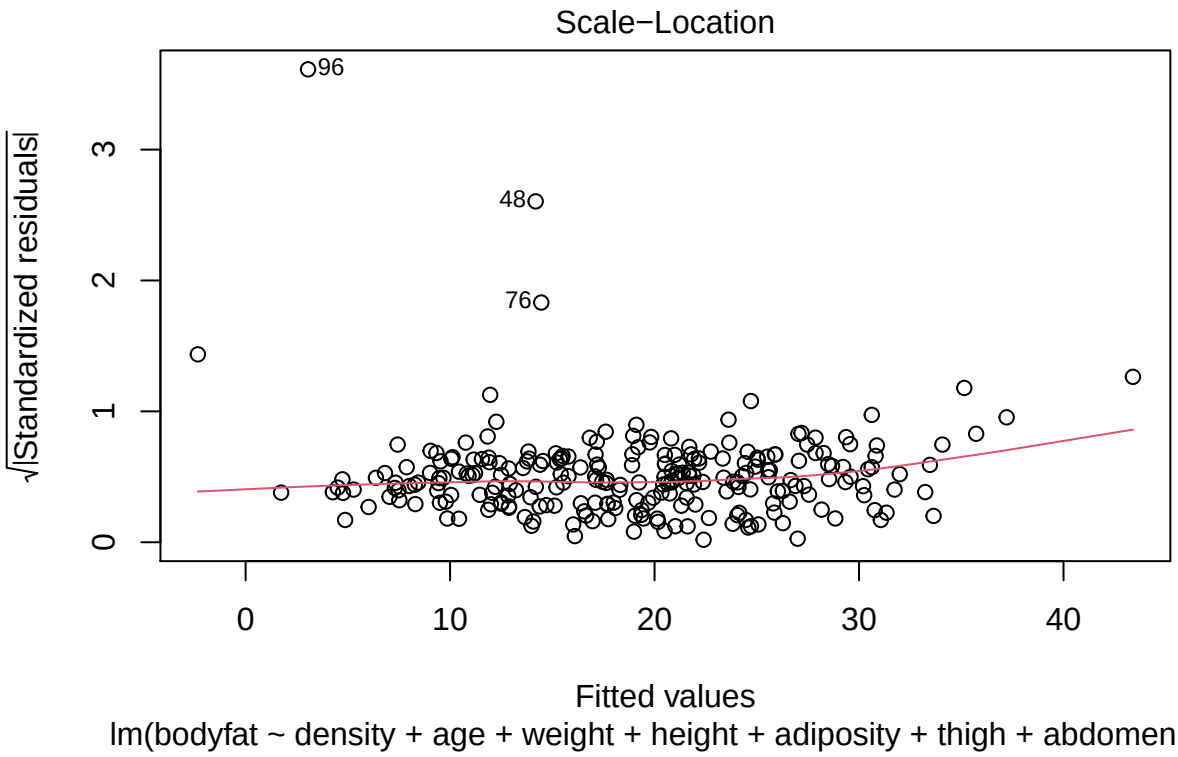
Note that this simplified model consisting of density and age (to the exclusion of abdomen) captures most of the information. The adjusted R-squared is 97.64% compared to 97.79% of the previous model that included abdomen. Age remains insignificant in both cases (10% significant level).

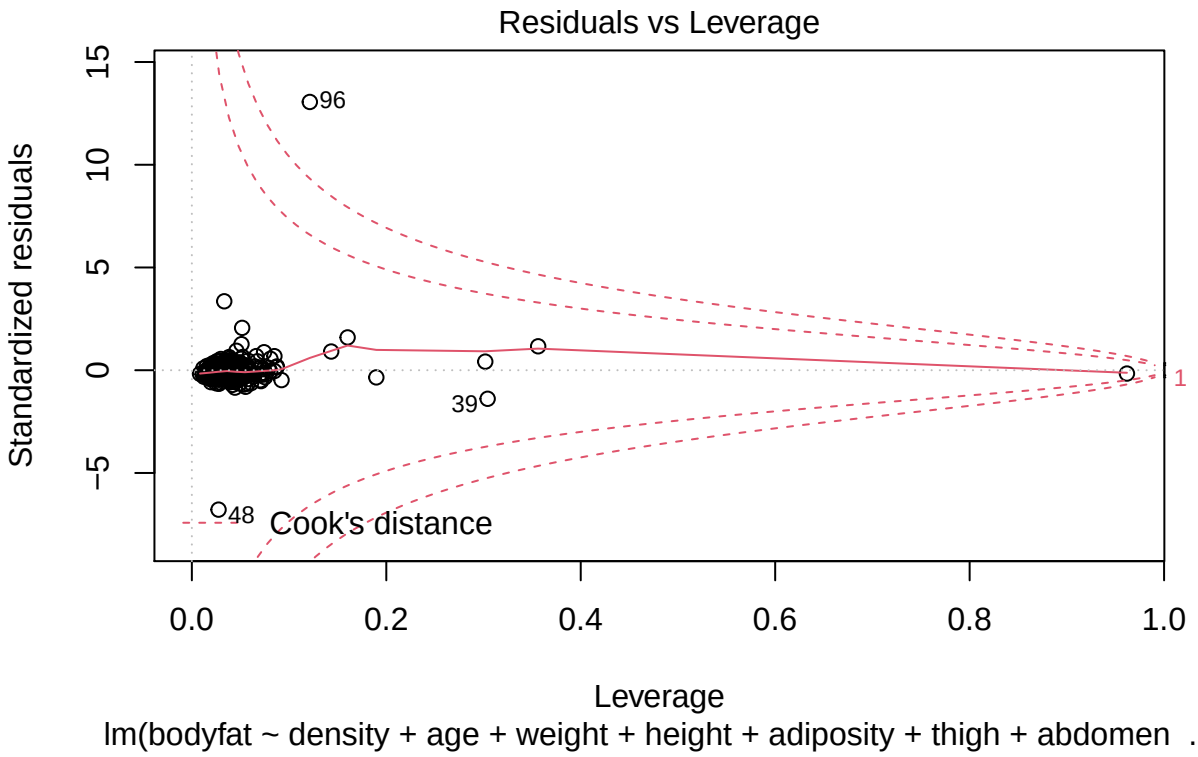
- *No auto-correlation*: The data has not time element so we do not evaluate this assumption.
- *Homoscedasticity*: In the figure below, the graph titled "scale location" provides a useful assessment of how residuals are evenly spread across the predictors. In this case, the roughly horizontal fit line indicates that the residuals are homoscedastic.

Another important issue arising here is that of some influential observations (outliers) the graph titled residuals versus leverage shows. In this case, there is an influential observation that may affect our inferences.




Normal Q−Q
Standardized residuals
15
10
5
0
−5
96
76
48
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bodyfat ~ density + age + weight + height + adiposity + thigh + abdomen ...


Scale−Location
√|Standardized residuals|
96
48
76
0
1
2
3
0
10
20
30
40
Fitted values
lm(bodyfat ~ density + age + weight + height + adiposity + thigh + abdomen ...



2.4.4 Hypothesis Tests for the Parameters

The final model selected is as folows.

bodyfat = a + b1(density) + bn(factor(age_group)) + error

where a is a constant and b1 and bn are respective coefficients. Note that bn coefficients will be for each of the categories representing age less the reference category (age 22 years). I test the foloowing hypothesis.

- For density: NULL: b1 = 0, versus alternative b1 != 0 P-value is close to zero meaning that there is very little probability that the null hypothesis is true. having failed to accept the null hypothesis we go with the alternative hypothesis.
- For age: NULL: bn = 0, versus alternative bn != 0 Likewise, for each age category, the probabilities are quite large (> 10%). Hence in this case we accept the null hypothesis that age is not a significant determinant of body fat.
- The overall model fit: The F-test As noted the F-test evaluates whether our model does better than a model with no predictors. As a ratio, if our model does just as well as that with no predictors, the F-ratio would be close to one.

NULL: b1 = bn = 0 Alternative: b1 != 0 and bn != 0

The F-statistic is 274.6 and the p-value is close to zero, meaning that we cannot accept the null hypothesis and hence go with the alternative hypothesis that our model is better than a model with no predictors.

The associated degree of freedom in our case are 51 and 200.

2.4.5 The Final Model- Making age a Categorical Variable

One of the steps in selecting the model is dropping one of the independent variable, abdomen, that is highly related to another independent variable, density. A final procedure is to convert age to a categorical variable. Technically, age is a continuous variable measured on a ratio scale as we have a true zero for age. Age maybe considered discrete because of the way it is measured. Specifically, age suffers from the "discretization" problem where people do not record their exact age, preferring instead to round up to the nearest year. In the case of body fat, the effect of age on body fat may not be apparent at exact age points but on a range of age groups. We could say, for instance that people between ages 20-30 years have lower body fat than those between 30 - 40 years. In this case, I translate the age to a categorical variable and re-run the analysis. The analysis shows that encoding age as a categorical variable does improve the model to one with an adjusted R-squared of 97.64%. Thus the model with density and age explains 97.64% of the variation in body fat.

I encode the age variable into the following categories and rerun the regression model; 20 - 29 year 30 - 39 years 40 - 49 years 50 - 59 years Over 60 years

I call the new variable age_group.

```
##
## Call:
## lm(formula = bodyfat ~ density + factor(age_group), data = BodyFat4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8433 -0.2212 -0.0867  0.0993 15.5408
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               442.19616    4.46290  99.083   <2e-16 ***
## density                  -401.09029    4.18038 -95.946   <2e-16 ***
## factor(age_group)30-39     -0.19003    0.27718  -0.686    0.494
## factor(age_group)40-49      0.09873    0.23923   0.413    0.680
## factor(age_group)50-59      0.40139    0.26630   1.507    0.133
## factor(age_group)Over 60    0.28321    0.29587   0.957    0.339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.19 on 246 degrees of freedom
## Multiple R-squared:  0.9769, Adjusted R-squared:  0.9764
## F-statistic:  2081 on 5 and 246 DF,  p-value: < 2.2e-16
```

2.4.6 Interpreting the Results

For the final model that consists of density and age (coded as factors) the interpretation of the results is as follows.

Density: Density is a significant determinant of body fat. Age also appears to be an important factor but is not significant. Although abdomen is significant, density alone appears to subsume the effects of abdomen.

Limitations

Disusss limitations here - for instance is the dataset too small? Can we generalize this to the whole world given genetic differences? You are the expert in the area so you should get the limitations in the books.

3 Conclusion

Overall it appears that the best predictors of body fat are density and age. However, only density is significant, with a 1 unit rise in density associating with a 401 unit decrease in body fat.

References

Bruce, Peter, Andrew Bruce, and Peter Gedeck. 2020. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media.

Frost, J. 2019a. "Introduction to Statistics: An Intuitive Guide." *Statistics by Jim Publishing*, 196–204.

———. 2019b. "Regression Analysis: An Intuitive Guide." E-book.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Jolliffe, Ian T, and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065): 20150202.

Portet, Stéphanie. 2020. "A Primer on Model Selection Using the Akaike Information Criterion." *Infectious Disease Modelling* 5: 111–28.

Zhu, J., Z. Ge, and Z. Song. 2017. "Distributed Parallel Pca for Modeling and Monitoring of Large-Scale Plant-Wide Processes with Big Data." *IEEE Transactions on Industrial Informatics* 13 (4): 1877–85.