Body Fat Analysis

Stat 628 Thursday Group 5

UW-Madison

October 10, 2019

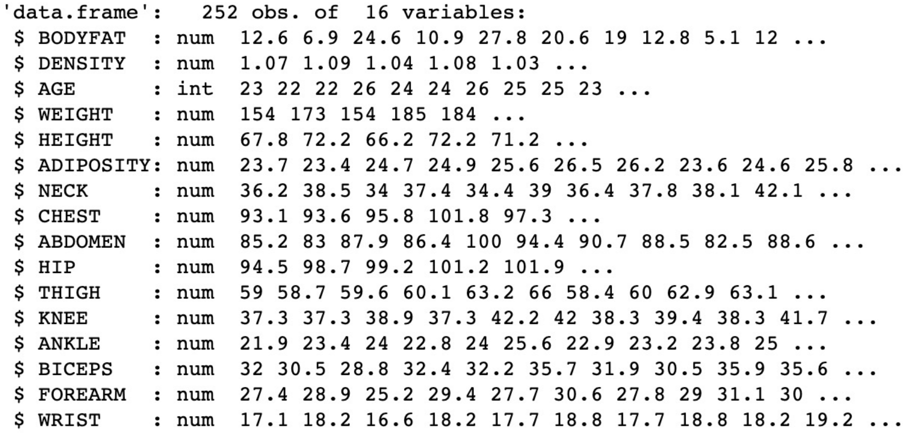
Outline

1. Data Preprocessing
2. Linear Model Construction
   - Multicolinearity
   - Model Selection
   - Diagnostics
3. Bodyfat and BMI
   - Bodyfat and BMI
   - Model Construct
   - Diagnostics
4. Models Evaluation
5. Summary
6. Reference

Introduction

- Body fat percentage (BFP) is a measurement of fitness level.
- BFP plays an important role in various health outcomes.
- It is quite difficult and costly to calculate.
- It is very practical to come up with a simple and precise method to estimate BFP.

Body Fat Data

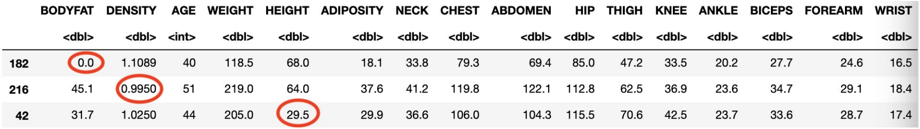


```
'data.frame':   252 obs. of  16 variables:
 $ BODYFAT  : num  12.6 6.9 24.6 10.9 27.8 20.6 19 12.8 5.1 12 ...
 $ DENSITY  : num  1.07 1.09 1.04 1.08 1.03 ...
 $ AGE      : int  23 22 22 26 24 24 26 25 25 23 ...
 $ WEIGHT   : num  154 173 154 185 184 ...
 $ HEIGHT   : num  67.8 72.2 66.2 72.2 71.2 ...
 $ ADIPOSITY: num  23.7 23.4 24.7 24.9 25.6 26.5 26.2 23.6 24.6 25.8 ...
 $ NECK     : num  36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...
 $ CHEST    : num  93.1 93.6 95.8 101.8 97.3 ...
 $ ABDOMEN  : num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
 $ HIP      : num  94.5 98.7 99.2 101.2 101.9 ...
 $ THIGH    : num  59 58.7 59.6 60.1 63.2 66 58.4 60 62.9 63.1 ...
 $ KNEE     : num  37.3 37.3 38.9 37.3 42.2 42 38.3 39.4 38.3 41.7 ...
 $ ANKLE    : num  21.9 23.4 24 22.8 24 25.6 22.9 23.2 23.8 25 ...
 $ BICEPS   : num  32 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 ...
 $ FOREARM  : num  27.4 28.9 25.2 29.4 27.7 30.6 27.8 29 31.1 30 ...
 $ WRIST    : num  17.1 18.2 16.6 18.2 17.7 18.8 17.7 18.8 18.2 19.2 ...
```

Body Fat Data

The data set we used to analyze contains age, weight, height and 10 circumference measurements on 252 men. Each man's percentage of body fat was accurately estimated by Siri's equation $100B = \frac{495}{D} - 450$, where B $=$ proportion of fat tissue, D $=$ Body Density.

Data Cleaning



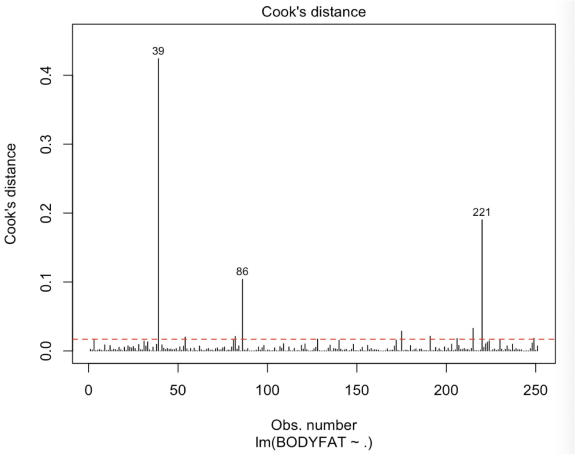
	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
182	0.0	1.1089	40	118.5	68.0	18.1	33.8	79.3	69.4	85.0	47.2	33.5	20.2	27.7	24.6	16.5
216	45.1	0.9950	51	219.0	64.0	37.6	41.2	119.8	122.1	112.8	62.5	36.9	23.6	34.7	29.1	18.4
42	31.7	1.0250	44	205.0	29.5	29.9	36.6	106.0	104.3	115.5	70.6	42.5	23.7	33.6	28.7	17.4

There are three abnormal observations.

Data Cleaning

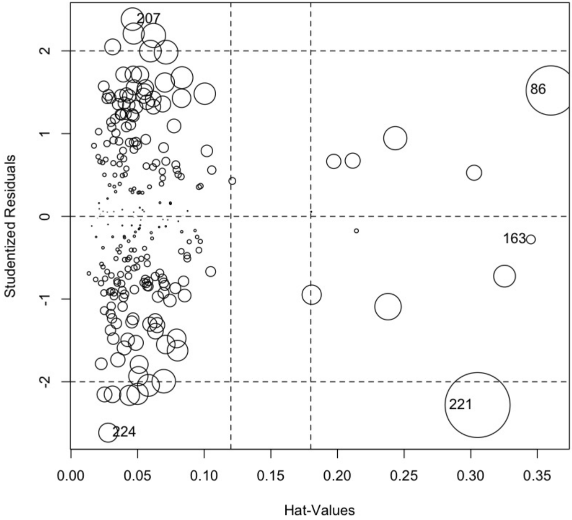
- For observation 182, since there may exist an error, we decide to delete it.
- For observation 216, we won't use density and then we keep it.
- For observation 42, its height is 29.5 inch, but his age is 44 years old and weights 205 pounds. Since all the other features seem quite normal, we consider impute the height according to the BMI calculation, $bmi = weight/height^2$, which is 69.5 inch.

Data Cleaning



From the *Cook's distance*[1] points. The subject of observation 39 weights 363 pounds which is very abnormal and we consider to delete it.

Diagnostics



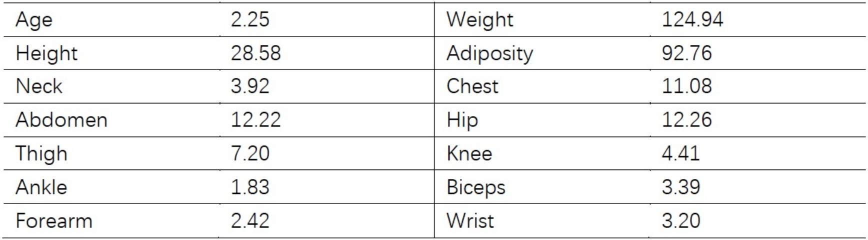
There are no obvious outliers and influential points.

Outline

1. Data Preprocessing
2. Linear Model Construction
   - Multicolinearity
   - Model Selection
   - Diagnostics
3. Bodyfat and BMI
   - Bodyfat and BMI
   - Model Construct
   - Diagnostics
4. Models Evaluation
5. Summary
6. Reference

Multicollinearity

VIF of each variable

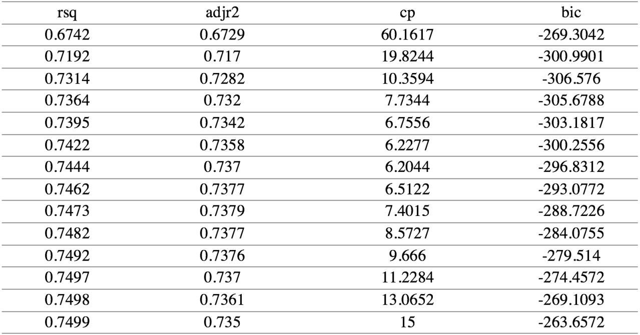


Age	2.25	Weight	124.94
Height	28.58	Adiposity	92.76
Neck	3.92	Chest	11.08
Abdomen	12.22	Hip	12.26
Thigh	7.20	Knee	4.41
Ankle	1.83	Biceps	3.39
Forearm	2.42	Wrist	3.20

Since some of the VIFs are greater than 10, and the mean VIF is 22.18, which is greater than 1.Multicollinearity may have a large impact on the inference.

Model Selection

R^2, $AdjR^2$,$Mallow's\ C_p$ and BIC criterions. (Exhaustive Search)

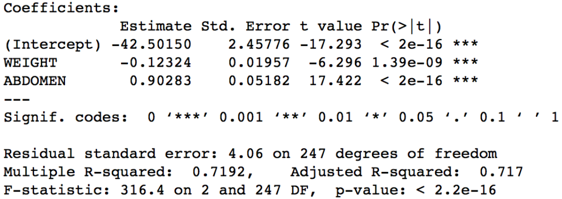


rsq	adjr2	cp	bic
0.6742	0.6729	60.1617	-269.3042
0.7192	0.717	19.8244	-300.9901
0.7314	0.7282	10.3594	-306.576
0.7364	0.732	7.7344	-305.6788
0.7395	0.7342	6.7556	-303.1817
0.7422	0.7358	6.2277	-300.2556
0.7444	0.737	6.2044	-296.8312
0.7462	0.7377	6.5122	-293.0772
0.7473	0.7379	7.4015	-288.7226
0.7482	0.7377	8.5727	-284.0755
0.7492	0.7376	9.666	-279.514
0.7497	0.737	11.2284	-274.4572
0.7498	0.7361	13.0652	-269.1093
0.7499	0.735	15	-263.6572

$$Model : BODYFAT \sim WEIGHT + ABDOMEN$$

Model Selection

Significance test for coefficients



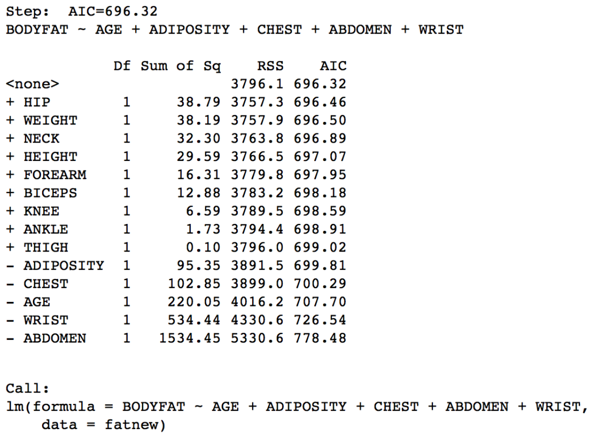
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.50150    2.45776 -17.293  < 2e-16 ***
WEIGHT       -0.12324    0.01957  -6.296 1.39e-09 ***
ABDOMEN       0.90283    0.05182  17.422  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.06 on 247 degrees of freedom
Multiple R-squared:  0.7192,	Adjusted R-squared:  0.717
F-statistic: 316.4 on 2 and 247 DF,  p-value: < 2.2e-16
```

The coefficients and the model are significant and the R^2 and Adj-R^2 are both greater than 0.7. Hence, the model is effective.

Model Selection

BIC criterion (Stepwise Search)

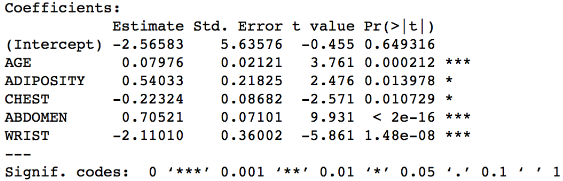
- step(fit2,direction=" both" ,trace=2,k=log(n))



```
Step:  AIC=696.32
BODYFAT ~ AGE + ADIPOSITY + CHEST + ABDOMEN + WRIST

            Df Sum of Sq    RSS    AIC
<none>                   3796.1 696.32
+ HIP        1     38.79 3757.3 696.46
+ WEIGHT     1     38.19 3757.9 696.50
+ NECK       1     32.30 3763.8 696.89
+ HEIGHT     1     29.59 3766.5 697.07
+ FOREARM    1     16.31 3779.8 697.95
+ BICEPS     1     12.88 3783.2 698.18
+ KNEE       1      6.59 3789.5 698.59
+ ANKLE      1      1.73 3794.4 698.91
+ THIGH      1      0.10 3796.0 699.02
- ADIPOSITY  1     95.35 3891.5 699.81
- CHEST      1    102.85 3899.0 700.29
- AGE        1    220.05 4016.2 707.70
- WRIST      1    534.44 4330.6 726.54
- ABDOMEN    1   1534.45 5330.6 778.48


Call:
lm(formula = BODYFAT ~ AGE + ADIPOSITY + CHEST + ABDOMEN + WRIST,
    data = fatnew)
```

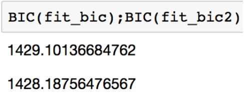
Model Selection

BIC criterion (Stepwise Search)



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.56583    5.63576  -0.455 0.649316
AGE          0.07976    0.02121   3.761 0.000212 ***
ADIPOSITY    0.54033    0.21825   2.476 0.013978 *
CHEST       -0.22324    0.08682  -2.571 0.010729 *
ABDOMEN      0.70521    0.07101   9.931  < 2e-16 ***
WRIST       -2.11010    0.36002  -5.861 1.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The predictors are all significant. Then Compare BIC with the first model of exhaustive search again.



```
BIC(fit_bic);BIC(fit_bic2)
```

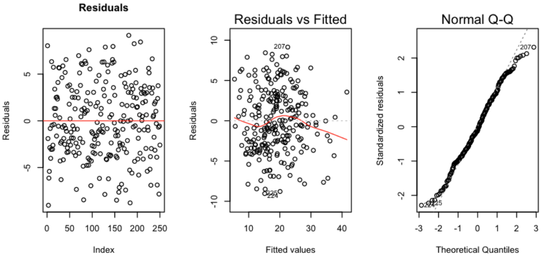
```
1429.10136684762

1428.18756476567
```

The BIC of the exhaustive search and for the stepwise search are quite similar. Considering the simplicity, we choose the model:

$$BODYFAT \sim WEIGHT + ABDOMEN$$

.

Diagnostics



Residuals vs Index plot:satisfy the independence.
Residuals vs Fitted plot:satisfy the homoscedasticity.
Normal $Q - Q$ plot:satisfy the normality.

Outline

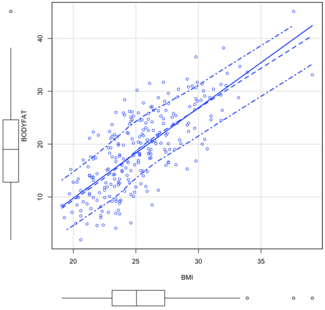
1. Data Preprocessing
2. Linear Model Construction
   - Multicolinearity
   - Model Selection
   - Diagnostics
3. Bodyfat and BMI
   - Bodyfat and BMI
   - Model Construct
   - Diagnostics
4. Models Evaluation
5. Summary
6. Reference

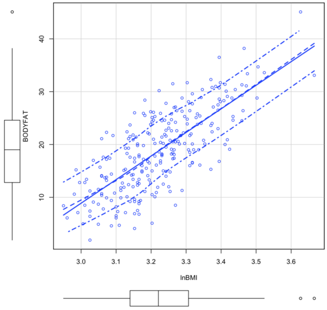
Bodyfat and BMI

- Many materials have talked about relationships between *BODYFAT* and *BMI*, like *TerenceC.Mills*[2] provides a semi-logarithmic relation between bodyfat and BMI.
- *BMI* is much easier to measure than *ABDOMEN*

Thus, we consider to use *lnBMI* to make our model much easily to achieve. Firstly, we explore the relation between *BODYFAT* and *lnBMI*.

Bodyfat and BMI



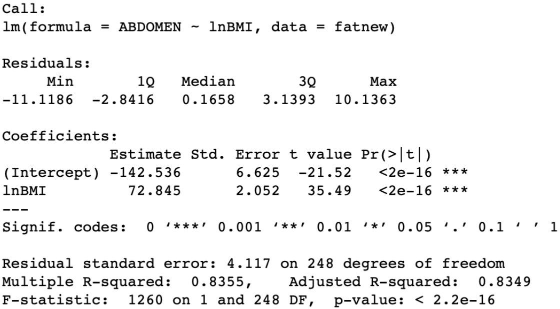


From the scatter plot of *BMI* and *lnBMI*, we can find linear relation between *BODYFAT* and *lnBMI* is much better.

Bodyfat and BMI

Then, we need to test whether *lnBMI* can explain *ABDOMEN* well.

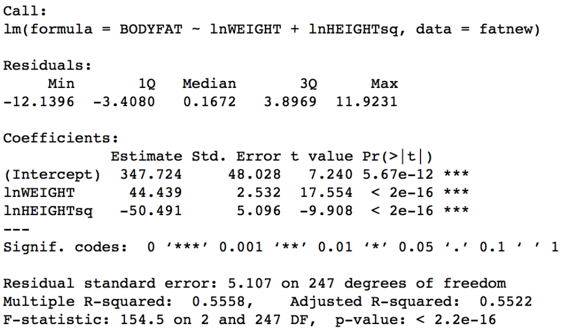
Bodyfat and BMI



```
Call:
lm(formula = ABDOMEN ~ lnBMI, data = fatnew)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1186  -2.8416   0.1658   3.1393  10.1363 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -142.536      6.625  -21.52   <2e-16 ***
lnBMI         72.845      2.052   35.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.117 on 248 degrees of freedom
Multiple R-squared:  0.8355,    Adjusted R-squared:  0.8349 
F-statistic:  1260 on 1 and 248 DF,  p-value: < 2.2e-16
```

We can see R^2 statistics is 0.8355, which means $lnBMI$ can explain most information of $ABDOMEN$. So $lnBMI$ is great! And we know $BMI = \frac{WEIGHT}{HEIGHT^2}$. Thus, for reducing approximation error, we use $WEIGHT$ and $HEIGHT$ to construct a linear model.

Model Construct



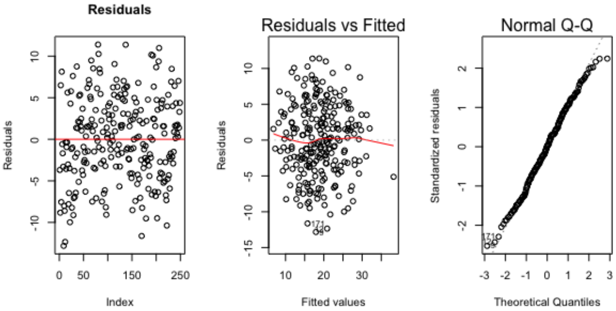
```
Call:
lm(formula = BODYFAT ~ lnWEIGHT + lnHEIGHTsq, data = fatnew)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.1396  -3.4080   0.1672   3.8969  11.9231 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  347.724     48.028   7.240 5.67e-12 ***
lnWEIGHT      44.439      2.532  17.554  < 2e-16 ***
lnHEIGHTsq   -50.491      5.096  -9.908  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.107 on 247 degrees of freedom
Multiple R-squared:  0.5558,	Adjusted R-squared:  0.5522 
F-statistic: 154.5 on 2 and 247 DF,  p-value: < 2.2e-16
```

The R^2 is 0.5558, which is not bad under only one variable. So we get the model,

$$BODYFAT \sim ln(WEIGHT) + ln(HEIGHT^2)$$

Next, we check the model assumption.

Diagnostics



Residuals vs Index plot:satisfy the independence.
Residuals vs Fitted plot:satisfy the homoscedasticity.
Normal $Q - Q$ plot:satisfy the normality.

Outline

1. Data Preprocessing
2. Linear Model Construction
   - Multicolinearity
   - Model Selection
   - Diagnostics
3. Bodyfat and BMI
   - Bodyfat and BMI
   - Model Construct
   - Diagnostics
4. Models Evaluation
5. Summary
6. Reference

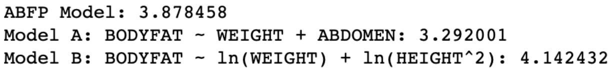
Models Evaluation

Here, we get two models,

$$
\begin{aligned}
Model\ A: &\ BODYFAT \sim WEIGHT + ABDOMEN \\
Model\ B: &\ BODYFAT \sim ln(WEIGHT) + ln(HEIGHT^2)
\end{aligned}
$$

We know *Model B* is easier than *Model A*, and we need to test whether these two models are useful compared to the real model. We use official *US Army bodyfat percentage*[3] calculator to evaluate models under the same data.

Models Evaluation



```
ABFP Model: 3.878458
Model A: BODYFAT ~ WEIGHT + ABDOMEN: 3.292001
Model B: BODYFAT ~ ln(WEIGHT) + ln(HEIGHT^2): 4.142432
```

We use Mean absolute error to compare three models. From the result, we can find that model A and B perform well under acceptable error. Especially, model A perform much better than others.

Outline

1. Data Preprocessing

2. Linear Model Construction
   - Multicolinearity
   - Model Selection
   - Diagnostics

3. Bodyfat and BMI
   - Bodyfat and BMI
   - Model Construct
   - Diagnostics

4. Models Evaluation

5. Summary

6. Reference

Summary

From the analysis above, we decide these three variables: *WEIGHT*, *HEIGHT* and *ABDOMEN*.

Under the consideration of convenience, if someone only can provide *WEIGHT* and *HEIGHT*, we will use,

$$Model\ B:\ \ BODYFAT \sim ln(WEIGHT) + ln(HEIGHT^2)$$

And if we want a more accurate result, adding *ABDOMEN*, we will use,

$$Model\ A:\ \ BODYFAT \sim WEIGHT + ABDOMEN$$

Summary

Rule of thumb:
Model A: BODYFAT$=$ -42.5-0.12 WEIGHT$+$0.90 ABDOMEN$+\epsilon$
Model B: BODYFAT$=$ 347.724$+$44.439 ln(WEIGHT) -50.491 ln(HEIGHT^2) $+\,\epsilon$, *where* $\epsilon \sim N(0, \sigma^2)$
For model A, the mean change in BODYFAT will be 0.12 % per pound change in WEIGHT,will be 0.90 % change per cm change in ABDOMEN, holding each other predictor fixed.
For model B, the mean BODYFAT will be 44.439 % change per exponential pound unit change in WEIGHT, will be 50.491*2 % change per exponential cm unit change in height.

Outline

1. Data Preprocessing
2. Linear Model Construction
   - Multicolinearity
   - Model Selection
   - Diagnostics
3. Bodyfat and BMI
   - Bodyfat and BMI
   - Model Construct
   - Diagnostics
4. Models Evaluation
5. Summary
6. Reference

Reference

[1] Shalabh, IIT Kanpur *Regression Analysis*. 2002

[2] Terence C. Mills *Predicting Body Fat Using Data on the BMI*. 2005

[3] Army Regulation 600–9 *Army Regulation 600–9*. 2013

[4] AceFitness *AceFitness*. 2009