SF2930 Regression Analysis VT23
Project 1 Report (Scenario I - BFM women)

Xindi Liu(xindi@kth.se)

March 22, 2023

1 Full model

1.1 Residual analysis

First I fitted a model using `DEXfat` as the response and all other columns as regressors. I plotted the normal probability of residuals and the residuals vs fitted values in Figure 1. I plotted the residuals vs all regressor variables in Figure 2.

The left figure in Figure 1 shows a positive skew or light-tailed distribution. The assumption about the normal probability of residuals does not hold.

The right figure is an outward - opening funnel, which indicates the variance is not constant.

`anthro3b`, `anthro4`: curve. `anthro3a`: double bow.

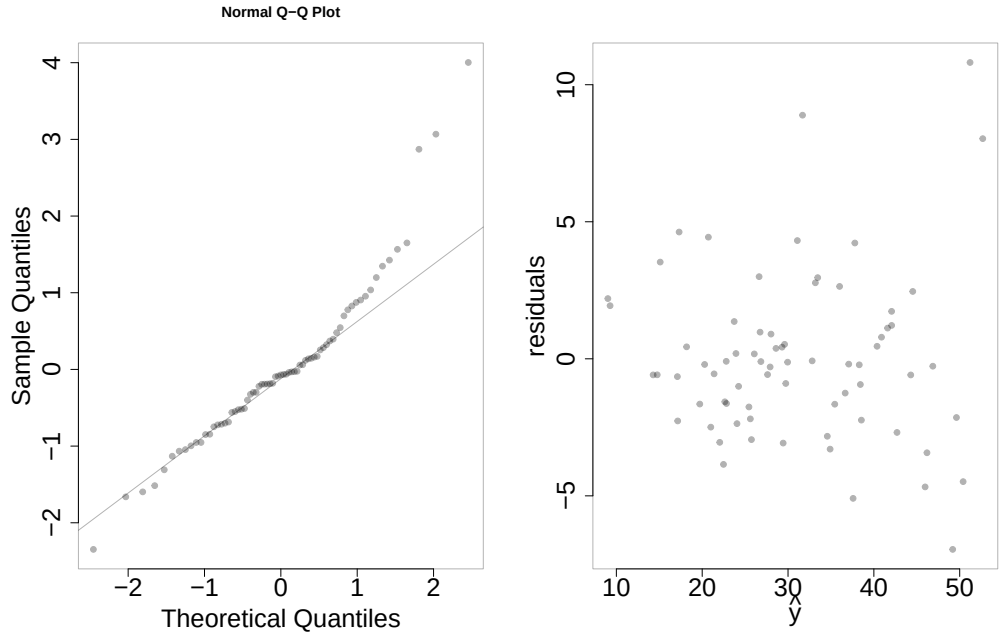

Figure 1: Normal probability of residuals(left) and Residuals vs. fitted values(right)

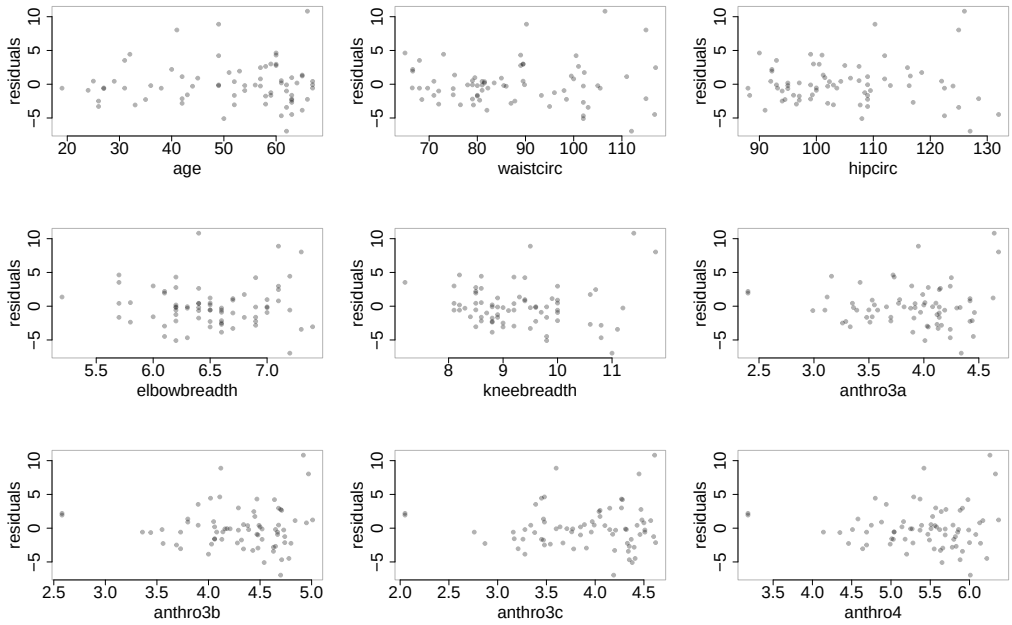

Figure 2: The residuals vs. regressor variables

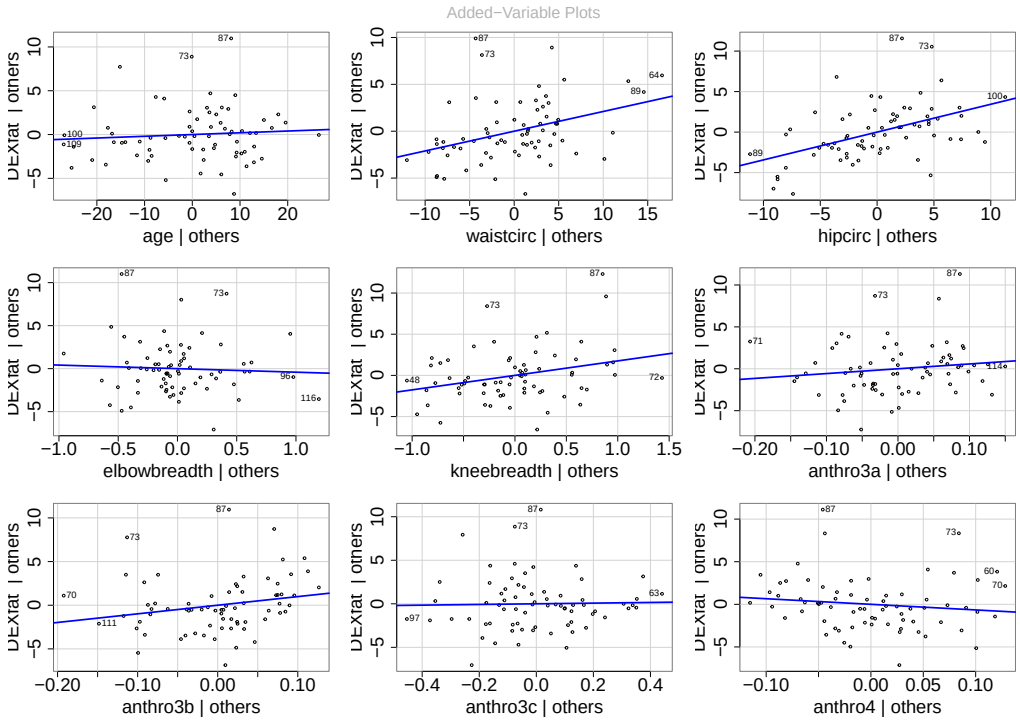

Figure 3: Added variable plots

1.2 Outliers

In the qqplot, datapoint 73 is at the right upper corner outliner while other points lie on the line. So I want to remove it. I tested Cook's distance, DFBETAS and DFFITS. Its Cook's distance is less than the cutoff value but is the largest, DFBETAS is larger than the cut off for 6 variables, DFFITS is larger than the cut off. Therefore it is reasonable to remove 73.

Because there is no other obvious outliner from the plot I don't want to remove any other points.

1.3 Transformation

I tried to transform the response with the logarithm function. The plot of Residuals vs. fitted values and added values seem to be a horizontal band. But the qqplot is still heavy tailed. I applied the logarithm function to the response one more time. Then the qqplot looks much better. The plots after a transformation with the logarithm function two times on the original response are shown in Figure 4 and 5.

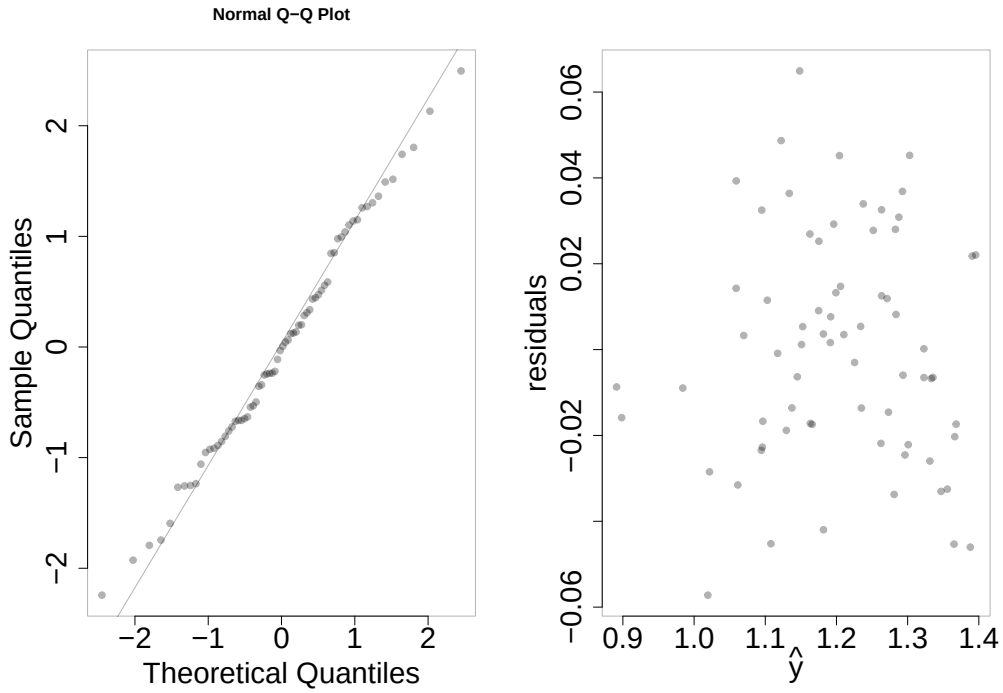

Figure 4: Normal probability of residuals(left) and Residuals vs. fitted values(right) after transformation

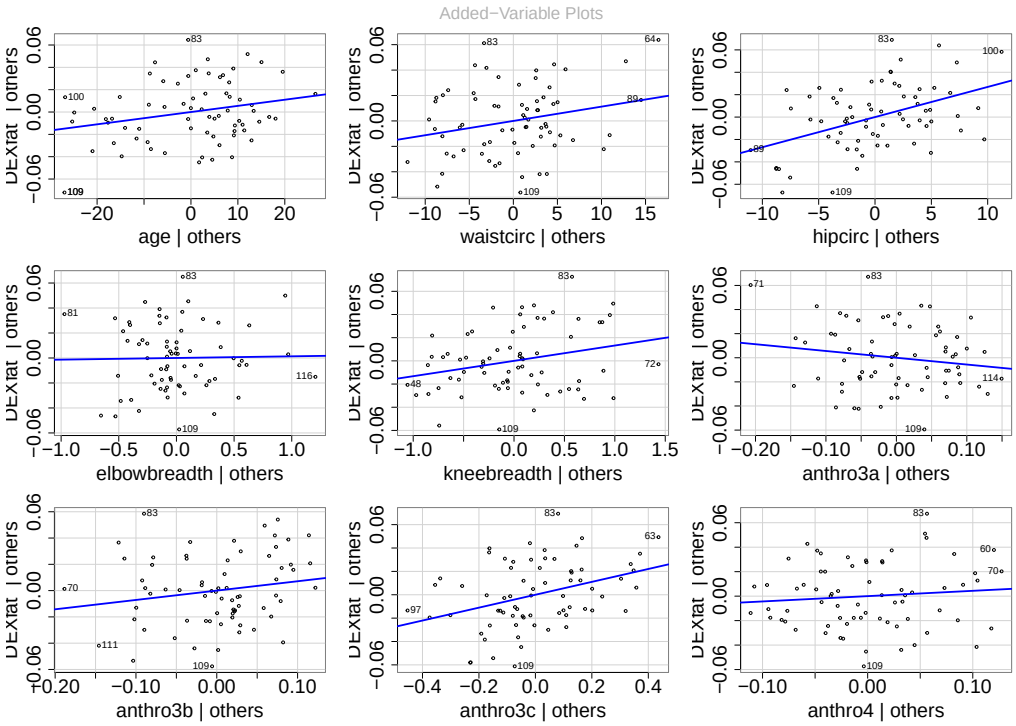

Figure 5: Added variable plots after transformation

2 Variable selection

I used `regsubsets` to get the best models for each number of variables. Then I evaluated their MSE and adjusted R^2 using cross validation. I used 5 folds. I searched on the internet and people said that usually 5 or 10 folds are used. There are only 70 observations, and with 10 folds there will be only 7 observations in the test set. I think it is too few with 7 observations in the test set, so I chose 5 folds. The results are shown in Figure 6

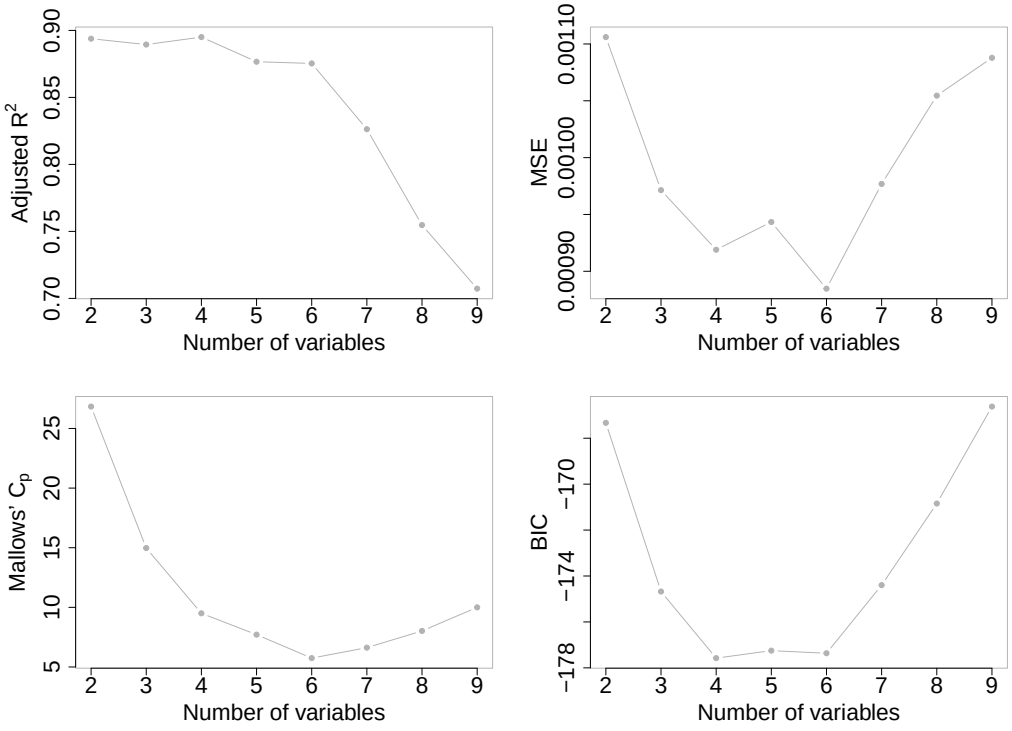

Figure 6: Model evaluation criterias

The model with 6 variables has the lowest MSE, C_p and BIC. Therefore it is the best model. The variables are: `age waistcirc hipcirc kneebreadth anthro3b anthro3c`.

3 Multicollinearity

The variance inflation factor of the variables is shown in the table:

age	waistcirc	hipcirc	kneebreadth	anthro3b	anthro3c
1.145151	5.224639	4.956645	2.493234	8.166823	8.724883

There is a clear indication of multicollinearity on `anthro3b` and `anthro3c`. To remove multicollinearity I should probably remove `anthro3b` or `anthro3c`. I tested both and compared their goodness of fit.

Removed variable	R^2	Adjusted R^2	Residual standard error
anthro3b	0.9361	0.9311	0.03083
anthro3c	0.939	0.9343	0.03011

Remove `anthro3c` results in a better fit of the data. The variance inflation factor after removing `anthro3c` is shown in the table:

age	waistcirc	hipcirc	kneebreadth	anthro3b
1.135767	5.014221	4.954905	2.493082	2.203005

Now the multicollinearity is ok.

4 Bootstraping

I used bootstrapping residual to compute the confidence intervals of the coefficients. The confidence intervals are shown in the table below together with the coefficients of the model.

	2.5%	$\hat{\beta}$	97.5%
(Intercept)	0.0117391581	0.1020043272	0.203897038
age	-0.0001337880	0.0004585406	0.001089084
waistcirc	0.0003627107	0.0014732964	0.002610722
hipcirc	0.0011561786	0.0026426513	0.003965053
kneebreadth	-0.0006484746	0.0116488428	0.023297300
anthro3b	0.1096100516	0.1312228369	0.152792362

5 Correction: remove age or kneebreadth or both

Two of the variables in my final model have confidence intervals containing 0: `age` and `kneebreadth`. What happens if I try removing one or both of these variables from my model?

Here are some comparisons.

	2.5 %	97.5 %
(Intercept)	0.0219701804	0.203270170
waistcirc	0.0003436744	0.002643627
hipcirc	0.0011504619	0.004040392
kneebreadth	-0.0009381160	0.024379368
anthro3b	0.1147521167	0.155406831

Table 1: Confidence intervals if age is removed

	2.5 %	97.5 %
(Intercept)	0.0568010029	0.2355708958
age	-0.0001891465	0.0009786744
waistcirc	0.0006737714	0.0029595776
hipcirc	0.0017912810	0.0045234869
anthro3b	0.1094011307	0.1522259179

Table 2: Confidence intervals if kneebreadth is removed

	2.5 %	97.5 %
(Intercept)	0.0661459880	0.235372019
waistcirc	0.0007878507	0.002974406
hipcirc	0.0016449510	0.004321724
anthro3b	0.1110964321	0.155535335

Table 3: Confidence intervals if both age and kneebreadth are removed

Observation: If only remove `age` or `kneebreadth` the other variable still has confidence an interval that contains 0.