# Project I - SF2930 Regression Analysis

Shadman Ahmed Oguzhan Ugur

20 April, 2018

## Scenario I: – Body fat assessment, BFM men

### Introduction and Project Goals

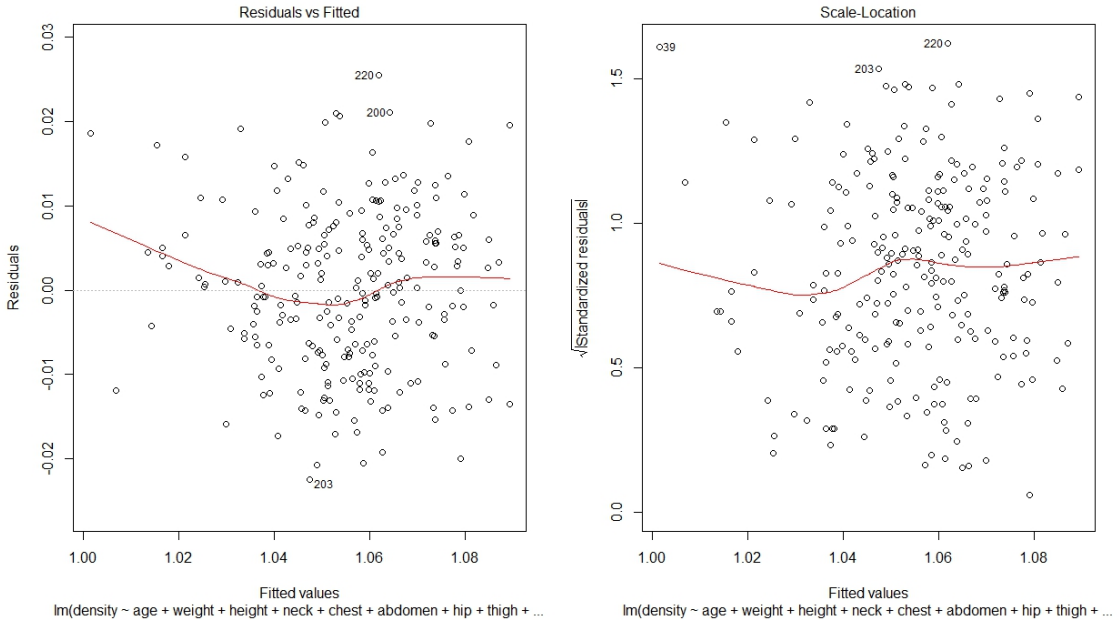
People suffering from obesity are commonly referred to people with a high degree of body fat. When a person consumes more calories than the body requires, it instead accumulates the extra food resource as fat for the purpose of storing the extra energy. Today obesity has been reported to be a major health problem, especially in developed countries, with cheap calorie- and fat rich food. And together with less exercise are one of the reasons behind this epidemic problem. Consequentiality, this can cause heart disease and diabetes.

One of the most popular and common measurements used to check if a person is suffering of obesity, is by using body mass index (BMI), by dividing the weight by the height squared, where people with a value above 25 is considered obese. It is simple to apply and does not require extensive instruments. But not totally reliable and can underestimate the fatness of different types of bodies. For example, it does not take account for the amount of muscle mass. The purpose of this project we will instead use body fat mass (BFM) which is a more effective way to measure fatness. It takes account of lot more attributes and thus more adaptive regarding estimating different types of human bodies.

### Analyses and Model Development

We will create a regression model to predict BFM from data containing hydrostatic weighing along with measurements of the abdomen, wrist circumference, age, weight and many more. A total of 14 attributes from 248 men. Then analyze the model with help of different statistical methods for optimization and validation purposes.

## Residual Analysis

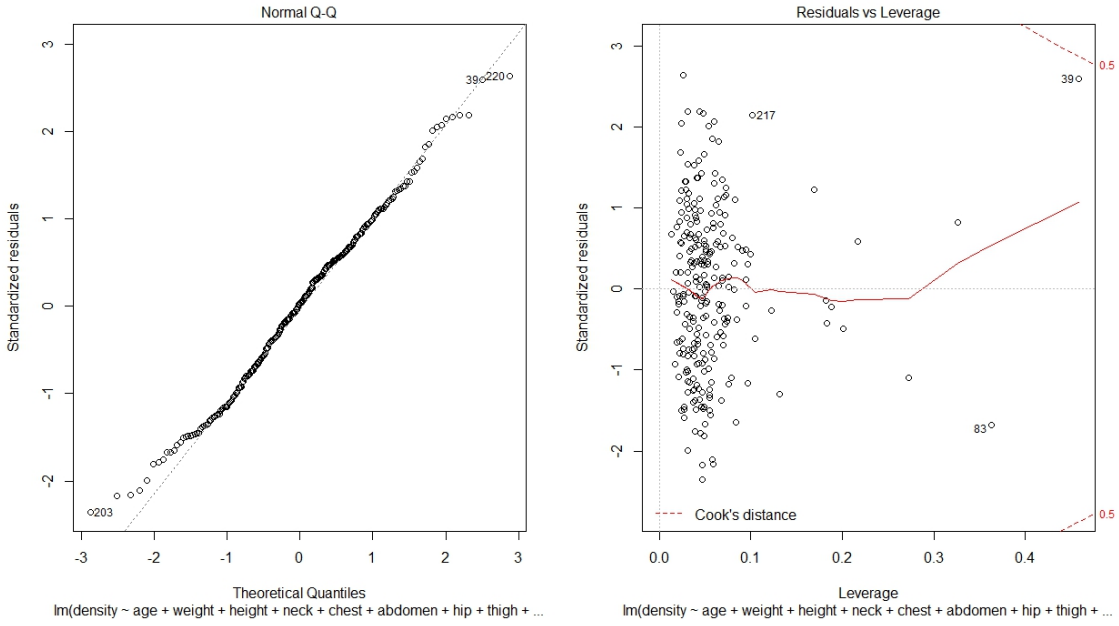



(a) Residuals vs Fitted (b) Scale Location

Figure 1: Residuals plots of the regression model

When building a model you have to check some aspects to confirm that the model shows characteristics of linear regression. One assumption is that the error should have constant variance and have zero mean. By looking at plot 1a, as the fitted values increase, the variance of the error is approximately constant, and the mean is close to zero: mean(error) = -4.05903e-19. The second plot 1b show if the residuals are equally spread along the ranges of predictors. We can here check the assumption of equal variance or in other words the homoscedasticity of the data which the plot confirms. Particularly because the residuals are approximately equally randomly spread on the horizontal line.

## Outliers, Leverage and Influential Observations




(a) Normal QQ plot

(b) Residual vs Leverage

Figure 2: Residuals plots of the regression model

Plot 2a show that most of the residuals are normally distributed. Which is another assumption of a good regression model. But it appears that the data points 39, 203 and 220 are critical points. The next plot 2b helps us to find observations that are influential, meaning points that could affect the determination of a regression model negatively. By looking at the plot 2b, one can barely see the cook's distance line. Observations positioned on and above the line are influential, with a cook's distance value above 1. One of the critical observations from the QQ plot, point 39 is the closest one to the cook's distance with a value of 0.4058267, but it is still inside the boarder thus doesn't has to be excluded.

Another method to detect influential data points is called covariance ratio, which measures the effect on the estimated covariance matrix. Where data points with a covariance value close to 1 tells us that these are not influential. The average of the covariance values from the data points and the predictive function is 1.064994, which is a good indication that the data points overall are not influential. The data points 31 and 39 have the highest covariance values, 1.5136142 and 1.3028652 but still not significance.

**Transformation**

One way to solve the problem of outliers or to correct other model inadequacies is by transforming the variables. Some common methods are logarithmic, squirt root, reciprocal transformations.

Table 1: Adjusted $R^2$ values of different transformations

| | No transform | Log | Sqrt | Reciprocal |
|---|---|---|---|---|
| Adjusted $R^2$ | 0.7310 | 0.7329 | 0.7319 | 0.7346 |

Earlier we showed that our model was in accordance with the assumptions of linear regression. A transformation was not necessary because in table 1 the adjusted R-squares have similiar values. The $R^2$ values tells how well the data points fit the regression model. Thus, a transformation is not used for further analysis of this model.

**Multicollinearity**

Analyzing a correlation matrix of the variables can provide information if the data contain multicollinearity. To check if our assumed independent variables are in fact correlated with each other. Eigensystem analysis provides a condition number where a number below 100 tells that there are no serious problem of multicollinearity, a number above 1000 and we have a severe problem. We got a condition number of $\kappa = 440.5118$ which indicates we have moderate multicollinearity but not severe, thus our data is acceptable for further studies. Next, you can see in particular, which variables that affect the multicollinearity. The method we used is called variance inflation factor where a value above 5 or 10 indicates that the specific regression coefficient is poorly estimated because of multicollinearity.

Table 2: VIF values

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Age | 2.256 | Chest | 10.165 | Knee | 4.744 | Wrist | 3.354 |
| Weight | 43.944 | Abdomen | 12.881 | Ankle | 1.952 | | |
| Height | 2.865 | Hip | 14.546 | Biceps | 3.683 | | |
| Neck | 4.391 | Thigh | 7.815 | Forearm | 2.172 | | |

In table 2 we can clearly see that we have 4 variables: weight, chest, abdomen, hip that are influenced by the multicollinearity. But because our condition number is less than 1000 we did not proceed to use any methods to deal with multicollinearity such as elimination of variables, ridge regression etc.

### Variable selection

In order to build an optimal model, identification of the best subsets among many variables is necessary and important, in other words variable selection is important. The main idea behind variable selection is to choose variables that makes the best fits for the model. In this section we will focus on forward/backward elimination and we will use adjusted $R^2$ and AIC(Akaike information criteria) as the evaluation criteria.

### Forward selection and backward elimination

The forward selection and backward elimination method is two methods of stepwise regression. The idea behind stepwise regression is to build a regression model by either adding or removing possible predictor variables based on some evaluation criteria such as adjusted $R^2$, F-test, AIC, BIC etc.

The forward selection method begins with no variable in the model and then for each variable a measure of contribution to the model is calculated. The measure of contribution are usually any of the evaluation criteria written on the previous paragraph. If the variable improves the fit based on the evaluation criteria then it is added to the model. The variables are added sequentially to the model. The Forward selection method terminates when the remaining variables have no contribution to the improvement of the model.

The backward elimination method on the other hand is working on the opposite direction of forward selection. It includes all the predictors at the beginning and eliminates the variables that has a negative contribution to the model. The variables are eliminated sequentially until all variables remaining in the model has satisfying contribution based on the elimination criteria.

### Evaluation with Akaike information criteria

The AIC(Akaike information criteria) looks into the AIC value and draws conclusion based on this value. If the value increases while adding or eliminating regressors then the regressors will be left outside the model, because it is claimed to have negative influence on the model based on the AIC. The results for both forward selection and backward elimination with AIC is depicted below on table 3.

Table 3: AIC values of the variables

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Forward: | abdomen | weight | wrist | forearm | neck | biceps | age | thigh | hip | AIC = -1582.902 |
| Backward: | abdomen | weight | wrist | forearm | neck | | age | thigh | hip | AIC = -1583.804 |

This result shows us that we should have 9 regressors for forward selection and 8 for backward elimination. So 5 and 6 regressors respectively has been left outside of the model.

The backward method can be seen as better since it has less AIC value then the forward method. The only regressors that differs is the biceps variable and is a borderline case for the backward method.

**Evaluation with adjusted R-Square**

The adjusted R-square evaluation was computed by looking at the regressors that had a p-value less than 0.1. The regressors that had a p-value less than 0.1 was tested by using forward selection and backward elimination by considering the adjusted R-sqaured value. If the value increases while eliminating or adding the regressors then that means that the predictor has a good contribution and should be included into the model.

Table 4: Adjusted R-Squared values of the variables

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Forward: | age | weight | neck | abdomen | thigh | | forearm | wrist | biceps | $adjR^2 = 0.733$ |
| Backward: | age | weight | neck | abdomen | thigh | hip | forearm | wrist | | $adjR^2 = 0.735$ |

The table above shows the regressors that are included in our model with highest R-square, it is also possible to see that these variables are the correct ones by looking at figure 3 which depicts the best subset with highest R-square for the predictors. The black boxes at the top shows the variables with the highest adjusted R-square.

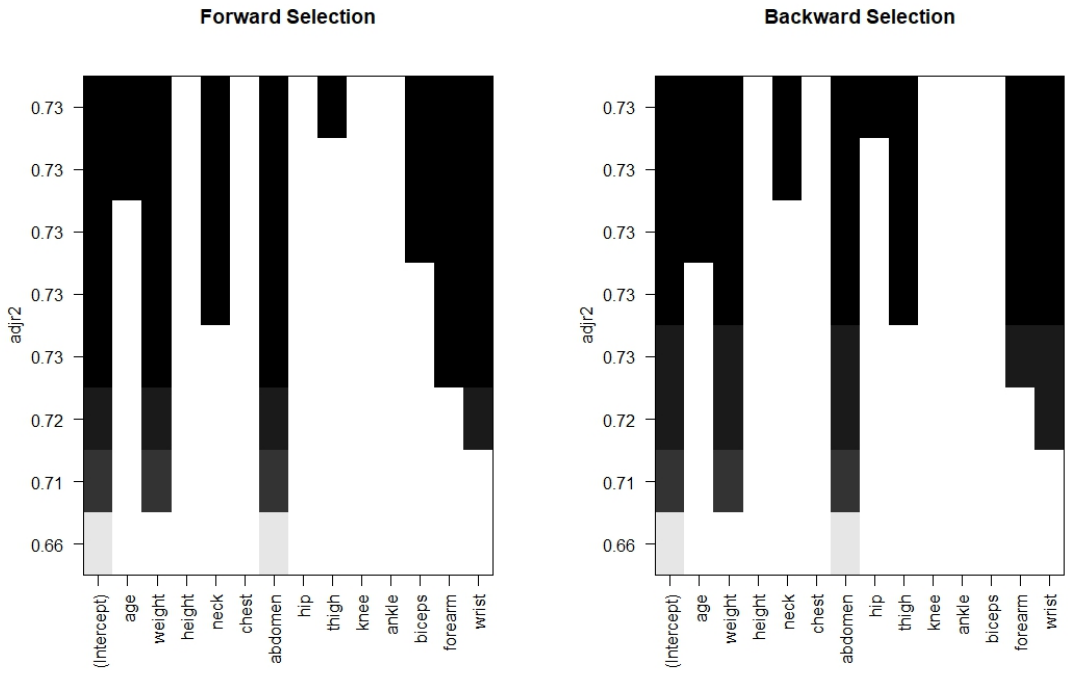



Figure 3: Subset plots with adjusted R square

### Cross validation

Cross validation is an important method in statistics since it measures the predictive performance of a statistical model. The cross validation method basically test how well the results of a statistical model will generalize to an independent set of data, in other words how well it will perform in practice.

The usual way of performing cross validation is by partitioning the dataset into training and validation set. The training set is used to estimate the variables and the validation set is used to test the estimated model. The predictive accuracy of a model can be measured by considering the mean squared error of the estimated model and the validation set. If the the training error is low and the validation error is too high then the estimated model is overfitted and has failed to build a good generalized predictive model. The error of the validation set will usually be higher than the error of the training set since the training set was used to estimate the variables, but if the validation error is too high then that indicates bad predictability. So, the variables has to be estimated in such a way that there has to be a balance between the test and the training error.

However, an improved cross validation method is the so called k-fold cross validation, where the data is divided into k subsets and the cross validation procedure is repeated k times. Each fold is used at least once as a validation set while the other folds are put together to form the training set. The average error over all trials is then computed and this makes it less important of how you split the data, since every data point gets into the test set once. The number of folds k can be chosen differently but large number of folds k gives a less biased estimation and in theory the best value of k is equal to the number of data points N. The N fold cross validation is also called Leave-One-Out Cross-Validation(LOOCV). This gives an unbiased estimate of the accuracy but it is too computational expensive and is therefore not that common in practice. The most common number of folds are k = 5,10,20 since it approximately gives the same estimation accuracy as LOOCV.

We chose to perform the k-fold cross validation only on the models created with the backward elimination method because it had a better AIC and adjusted R-square score. The average MSE error for the k-fold cross validation is written at table 5 for AIC and adjusted R-square models.

Table 5: Sum of MSE

| | AIC model | adj R^2 |
|---|---|---|
| 5-fold | 0.0000978 | 0.0000978 |
| 10-fold | 0.0000984 | 0.0000984 |
| 20-fold | 0.0000991 | 0.0000991 |

As it is possible to see at table 3 the errors are identical and no conclusion can be drawn from the cross validation. The two models are equally good. The reason to this is that the backward elimination method for AIC and adjusted R-square have the same predictors.

**Bootstrap**

Bootstrap is a re-sampling method and is mostly used to provide a measure of estimation accuracy for a parameter estimate. The bootstrap is generally used to estimate the Standard deviation/error or confidence intervals around regression coefficients. Bootstrapping uses random sampling with replacement, this means that observation from the data set will be selected randomly in order to produce a bootstrap data set. The boot strap dataset can contain the same observation more than once since the observed sample is replaced after it is selected.

The bootstrapping residual procedure begins by fitting a linear regression model $y = \mathbf{X}\beta+\epsilon$ and obtaining n residuals where n is the number of samples. Then a random sample of size n will be arranged in a bootstrap residual vector $e^*$ where the random samples are chosen

with replacement. The bootstrapped residuals is then attached to the predictors $\hat{y}$ to form a bootstrapped vector $y^*$.

$$y^* = \hat{y} + e^*$$

This vector is bootstrapped responses which are regressed on the original regressors by the regression procedure used to fit the linear model. This will produce the bootstrap estimate of the vector of regression coefficients.

We tested with different numbers of samples and saw that the standard deviation of the bootstrap parameters stabilized for approximately 900 samples. Looking at table 6, we can see that the standard deviation of the bootstrap parameters is similar to the normal theory based standard deviation of our estimated regression coefficients.

Table 6: Standard deviations of estimated parameters and bootstrap parameters

| | Age | Weight | Neck | Abdomen | Hip | Thigh | Forearm | Wrist |
|---|---|---|---|---|---|---|---|---|
| S($\hat{\beta}$) | 7.078e-5 | 9.14e-5 | 5.128e-4 | 1.650e-4 | 3.198e-4 | 2.953e-4 | 4.232e-4 | 1.168e-3 |
| S($\hat{\beta}^*$) | 7.170e-5 | 8.933e-5 | 5.176e-4 | 1.640e-4 | 3.190e-4 | 2.875e-4 | 4.186e-4 | 1.098e-3 |

## Conclusion

After creating and analysing our regression model to predict BFM, the best fit got the variables: Age, Weight, Neck, Abdomen, Hip, Thigh, Forearm, Wrist. The values of the coefficients are shown in the table below.

Table 7: Coefficients of the regression model

| | Age | Weight | Neck | Abdomen | Hip | Thigh | Forearm | Wrist |
|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}$ | -1.351e-4 | 1.893e-4 | 1.0942e-3 | -2.150e-3 | 5.608e-4 | -7.324e-4 | -1.1778e-3 | 3.3212e-3 |

With a standard error 0.009716, adjusted R-squared 0,7345 and p-value of 2.2e-16, we can conclude that the produces regression model of the BFM is a good fit of the dataset.