CHAPTER 3.

SIMPLE REGRESSION MODELS FOR REE–COMPARISONS

3.1 Introduction

Statistical modeling is one of the most interesting areas of statistics. The applications where statistical modeling is used are numerous and are related to various fields, including econometrics, engineering, and pharmacology. Regression analysis is used to investigate and model the relationship between a response variable and one or more predictors.

In this chapter we will try to derive linear models that predict REE of Greek elite athletes. It is the first time that a real sample of Greek athletes is used for creation of such predictive equations. Also it is a great opportunity to compare the derived equations with established similar equations of the past (see 2.4) and come up with interesting conclusions.

3.2 REE VS weight - Regression

As we can see from the correlation table 2.6.1, body weight is highly correlated with REE. Weight is the most important explanatory variable for REE. Equations 2, 3, 5 and 6 of chapter 2.4 use only weight to predict REE. The scatter plots of REE versus WEIGHT (Figure 3.2.1), for both sexes, indicate obvious linear relationship between them.

Figure 3.2.1 Scatter plots of REE VS Weight by sex
(Regression lines and 95% confidence intervals)

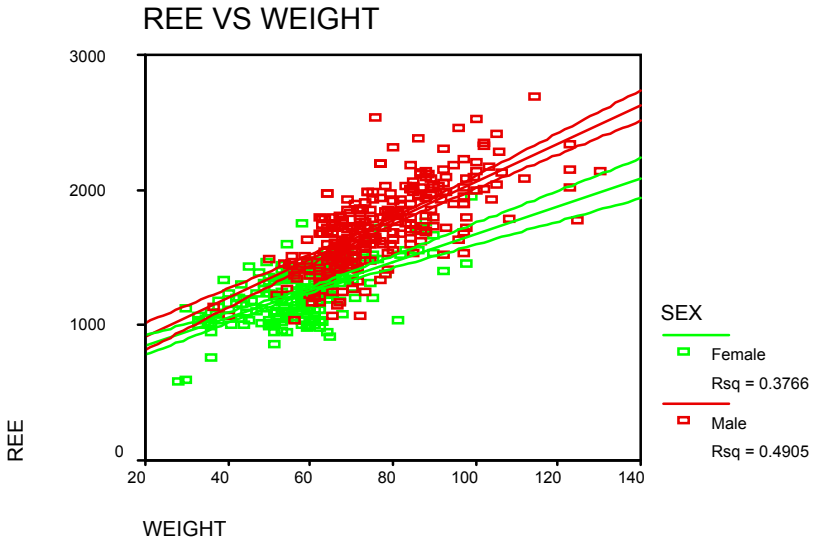


Table 3.2.2 Simple predictive equations-1 for REE

Males	Females
REE = 628.84 + 14.27·Weight	REE = 643.53 + 10.31·Weight
$R^2_{adj} = 0.489$	$R^2_{adj} = 0.374$
95% C.I. of Constant (491.83, 765.84)	95% C.I. of Constant (536.78, 750.28)
95% C.I. of Weight (12.52, 16.01)	95% C.I. of Weight (8.52, 12.11)

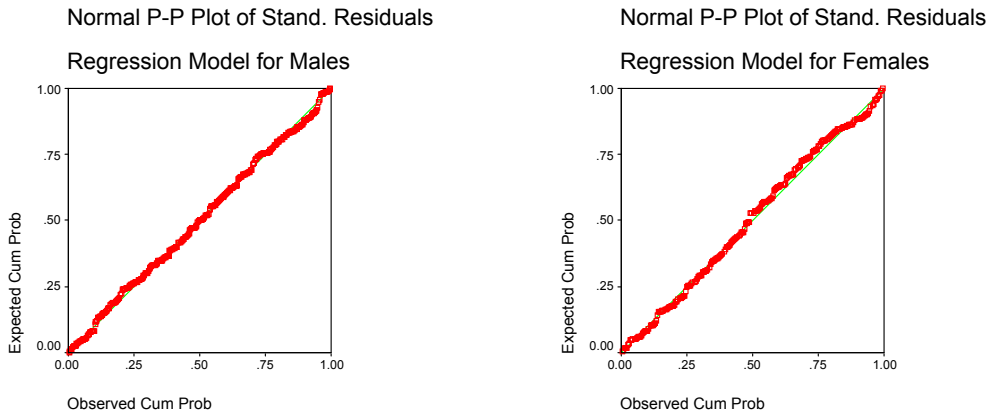
By comparing the two derived equations we can conclude that the 95% confidence intervals of weight-coefficients are distinct since (8.52, 12.11) ∩ (12.52, 16.01) = ∅.

The energy needs per gram of body mass are significantly higher for males than for females, fact that can be explained by the differential of body composition between the two sexes (as has been presented in the table 1.5.1).

3.3 Assumptions of linear model

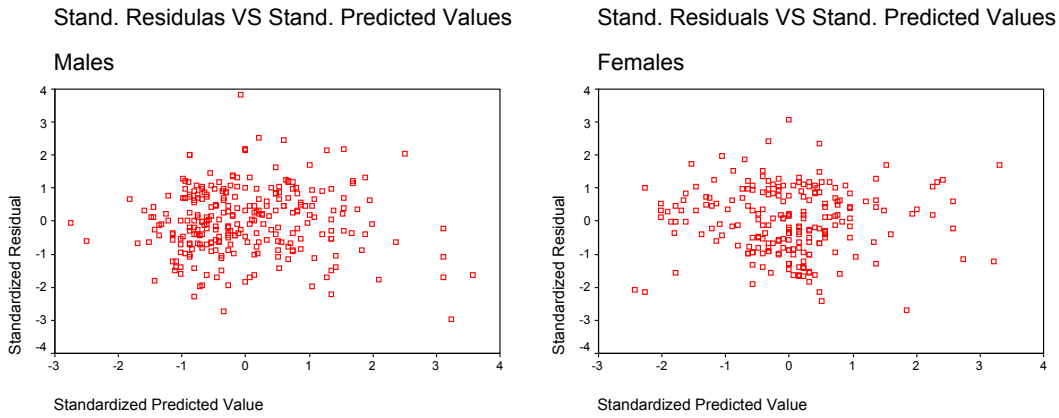
Diagnostics plots and tests are used in linear models in order to check the basic assumptions of the linear regression. The following plots and tests indicate satisfaction of the assumptions for both males and females.

Figure 3.3.1 Normal P-P plots of standardized residuals by sex



The Kolmogorov Smirnov test for normality of errors gives a P-value for Males equal to 0.898 and for Females equal to 0.828. So we do not have evidence to reject the null hypothesis that says that errors follow a Normal distribution. The mean error of both sexes is zero.

Figure 3.3.2 Scatter plots of standardized Residuals by sex



.

In both scatter plots, residuals seem to indicate homoscedasticity, linearity and independency since they look like being between two parallel lines (Draper and Smith, 1981). Also from these Scatter plots one outlier for each model is identified with standardized residual greater than 3.

3.4 REE VS weight, height and age - Regression

We will use weight, height and age, as explanatory variables for REE since are the most easily measured. Harris and Benedict equations (see equation 1 in section 2.4) also use the same explanatory variables. The equations that are derived are the following:

Table 3.4.1 Simple predictive equations-2 for REE

Males	Females
REE = -140.56 + 12.01Weight + 6.09Height - 7.13Age	REE = 729.29 + 11.17Weight - 0.28Height - 4.77Age
$R^2_{adj} = 0.531$	$R^2_{adj} = 0.375$
95% C.I. of Constant (-608.15, 327.03)	95% C.I. of Constant (213.93, 1244.65)
95% C.I. of Weight (9.78, 14.25)	95% C.I. of Weight (8.23, 14.11)
95% C.I. of Height (3.12, 9.06)	95% C.I. of Height (-4.03, 3.46)
95% C.I. of Age (-12.53, -1.72)	95% C.I. of Age (-10.81, 1.26)

We can see that coefficients of height and age for females are not significantly different from zero. Consequently, we should remove one or both of these variables from the model.

Something that should be investigated is the problem of multicollinerity, which refers to high correlation among the explanatory variables that does not allow one to examine the individual effect of each explanatory variable. In the presence of multicollinearity the estimates of the unknown parameters are not stable. This means that small changes in the data may lead to large changes in the parameter estimates. In addition, the regression coefficients may have very large standard errors, while at the same time the R^2 is high. Finally, the coefficients may have a wrong sign.

A very simple method for detecting multicollinearity is based on the calculation of correlation coefficient between any pair of the explanatory variables. If this coefficient takes a value close to ±1 then this is an indication of the presence of multicollinearity. Furthermore, we can compare these correlation coefficients with the R^2 coefficient of the model. If any of the coefficients is greater of R^2 there may be multicollinearity and we should be cautious or just exclude one of the two correlated explanatory variables (Jarrett, 1987).

The condition number is a statistical index that most Statistical tools provide in order to detect multicollinearity. A condition number greater than 15 indicates a possible problem and a number greater than 30 suggests a serious problem with multicollinearity.

For both sexes the problem of serious multicollinearity is suggested since:

Males' Condition Number = 57 > 30

Female' Condition Number = 73 > 30

Previous bibliography has not referred to the multicollinearity problems in equations despite the strong correlation that always exists between weight and height.

An other possible solution is to apply principal components analysis to weight, height and age in order to produce uncorrelated components (PCs) and avoid regression on correlated explanatory variables. Other than dealing with multicollinearity, what does principal components regression have going for it? At the very best, the PC's are so readily interpretable that they become the new variables in the prediction model. At the very worst, they are not

interpretable at all but one can still relate the responses to the original predictors, by transforming back the model (Jackson, 1985).

Table 3.4.2 Principal Components of weight, height and age for Males

Variable	PC1	PC2	PC3
AGE	0,060	0,188	0,980
HEIGHT	0,522	-0,843	0,130
WEIGHT	0,851	0,503	-0,149

The first principal component explains the 78.6% of the total males' variance.

$$PC1 = 0.060\ AGE + 0.522\ HEIGHT + 0.851\ WEIGHT$$

Table 3.4.3 Principal Components of weight, height and age for Females

Variable	PC1	PC2	PC3
AGE	0,126	0,200	0,972
HEIGHT	0,558	-0,824	0,097
WEIGHT	0,820	0,530	-0,215

The first principal component explains the 84.7% of the total females' variance.

$$PC1 = 0.126\ AGE + 0.558\ HEIGHT + 0.820\ WEIGHT$$

The regression equations, using only the first principal components, are the following:

$$\text{Males:} \quad REE = -359 + 13\ PC1_{males} \qquad (R^2_{adj} = 52\ \%)$$

$$\text{Females:} \quad REE = 43 + 8.4\ PC1_{females} \qquad (R^2_{adj} = 35\ \%)$$

In this case principal components do not help in any way since they are not interpretable at all and cannot be used as a replacement of the originals variables.

3.5 Regression models using fat free body mass information

Fat free body mass is, also, very important explanatory variable for REE. Equation 4 in section 2.4 uses only fat free body weight and age for creating an equation that fits for both sexes.

Figure 3.5.1 Scatter plots of REE VS Fat-Free Weight
(Regression lines and 95% confidence intervals)

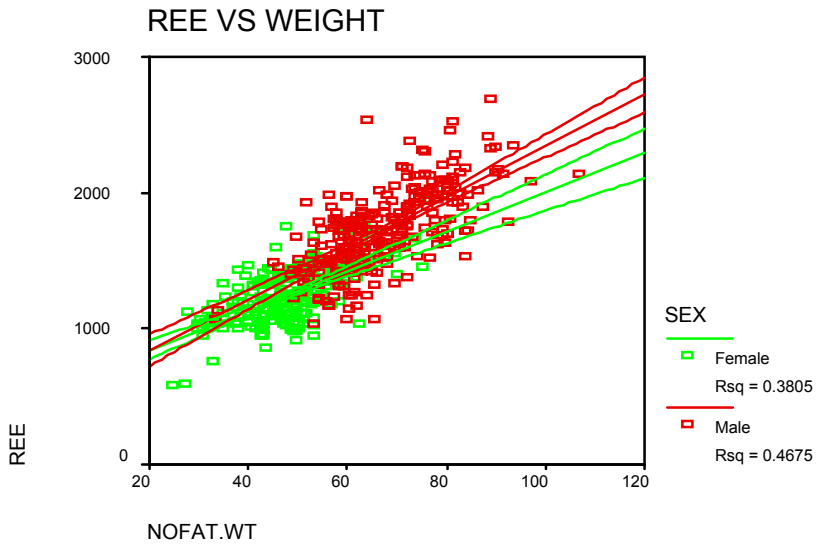


Table 3.5.2 Simple predictive equations-3 for REE

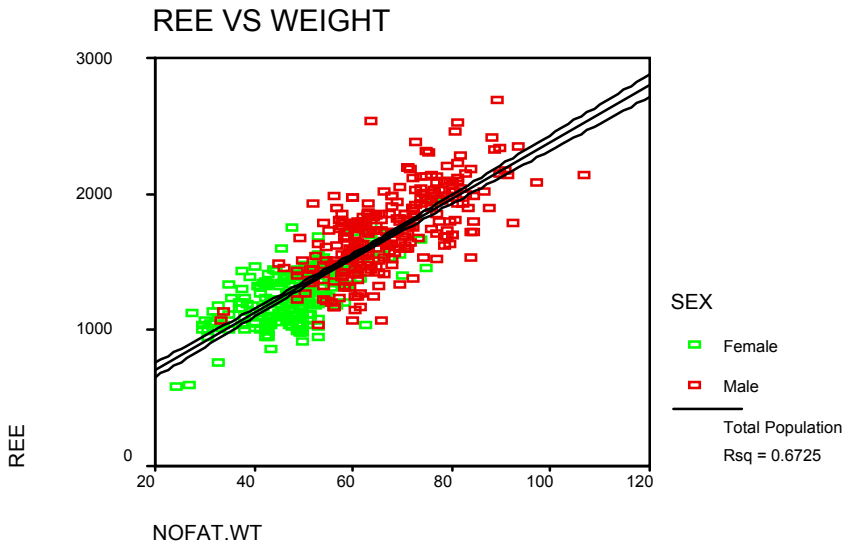
Males	Females
REE = 462 + 19·Fat Free Weight	REE = 552 + 14·Fat Free Weight
$R^2_{adj} = 0.466$	$R^2_{adj} = 0.378$
95% C.I. of Constant (298, 627)	95% C.I. of Constant (431, 674)
95% C.I. of Fat Free Weight (16, 21)	95% C.I. of Fat Free Weight (12, 17)

In 4.2 and 4.3 we will see that using a dummy variable of sex, interacted with Fat Free Weight, we are driven to the conclusion that the two coefficients of Fat Free Weight are significantly different between sexes.

Although we should try deriving a common equation using Fat Free Weight and age in order to compare it with Ravussin and Bogardus equation (see section 2.4).

Figure 3.5.3 Scatter plot of REE VS Fat-Free Weight for both sexes

(Regression lines and 95% confidence intervals)



Using age we manage an even better R^2_{adj} without facing the problem of multicollinearity.

The new equation is the:

$$REE = 394.67 + 22{,}45\ NOFAT.WT - 9.7\ AGE,\ R^2_{adj} = 0.684$$

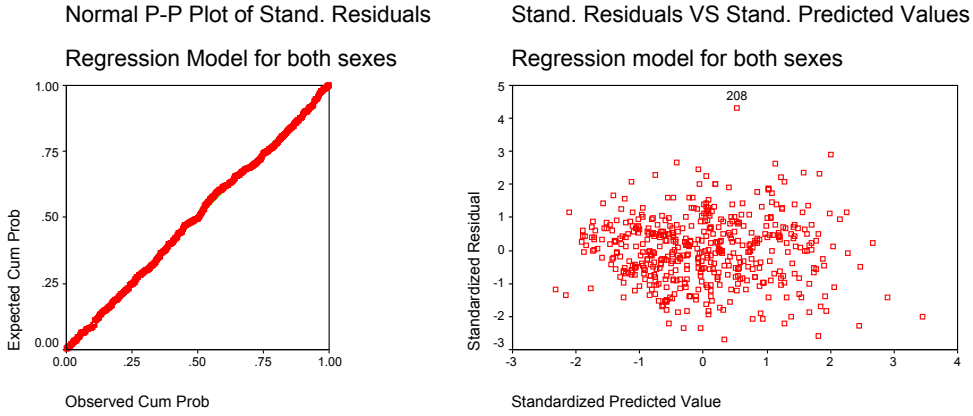
95% C.I. for Constant: (304.21, 485.12)

95% C.I. for Nofat.wt: (21, 24)

95% C.I. for Age: (-14, -5.4)

Diagnostics plots (Figure 3.4.4) indicate that the assumptions of the linear regression are satisfied and the model is statistically acceptable.

Figure 3.5.4 Diagnostic plots



Also the Kolmogorov-Smirnov test for normality of errors gives a P-value equal to 0.942. So we cannot reject the null hypothesis that says that errors

follow a Normal distribution. The scatter plot indicates one outlier (observation 208) with standardized residual greater than 3.

3.6 Comparisons between different equations

In this section we will compare the derived equations with the equations that have been presented in section 2.4. In order to do this we will check if the coefficients of these equations are included in the 95% confidence intervals for the coefficients of our equations. If a coefficient does not belong to the respectively 95% C.I, we conclude that the equations are statistically different. Tables, 3.6.1 and 3.6.2 help in comparing relative equations.

Table 3.6.1 Males' equations

Males	Variables - Coefficients				
Name	Constant	Weight	Height	Age	Fat-Free
Equation 1	629 ±137	14 ±2			
FAO	679 ✓	15.3 ✓			
Schofield	688 ✓	15.1 ✓			
Henry and Rees	672 ✓	13.4 ✓			
Piers and Shetty	849.6	10.6			
Equation 2	-140±467	12 ±2	6 ±3	-7 ±5	
Harris-Benedict	66.5 ✓	13.8 ✓	5 ✓	-6.8 ✓	
Equation 3	395 ±90			-10 ±4	22.5 ±1.5
Rav. & Bogardus	441 ✓			-2.4	21.9 ✓

✓ Included in the relative 95% Confidence Interval of Coefficients.

Table 3.6.2 Females' equations

Females	Variables - Coefficients				
Name	Constant	Weight	Height	Age	Fat-Free
Equation 1	644 ±106	10 ±2			
FAO	496	14.7			
Schofield	688 ✓	15.1			
Henry and Rees	614.8 ✓	11.5 ✓			
Piers and Shetty	595.1 ✓	10.9 ✓			

Equation 2	729 ±516	11 ±3	-0.3 ±4	-5 ±6	
Harris-Benedict	665.1 ✓	9.6 ✓	1.8 ✓	-4.7 ✓	
Equation 3	395 ±90			-10 ±4	22.5 ±1.5
Rav. & Bogardus	441 ✓			-2.4	21.9 ✓

✓ Included in the relative 95% Confidence Interval of Coefficients.

Harris – Benedict equations use weight, height and age as explanatory variables, so they will be compared with the derived equations – 2. Harris – Benedict equations' coefficients are included in the respectively 95% C.I. with some of them almost equal, so we can conclude that Harris – Benedict equations are not statistically different from the derived equations 2. We have to mention the problem of multicollinearity and that the coefficients of Height and Age, of the females' equation, are not significantly different from zero for our derived equations.

FAO equations use only weight as explanatory variable and they can be compared with the derived equations – 1. The coefficients of FAO equation for males are included in the respectively 95% C.I. Contrary to males, females' FAO equation is statistically different from the derived equation since coefficients are not included in the respectively 95% C.I.

Schofield et al' equations show the same comparing results with FAO equations. Again females' equations are statistically different.

Both Henry and Rees equations are not statistically different from the derived equations – 1. The males' equations are almost the same and so Henry and Rees equations can be selected as the most representative for our males' sample.

Piers and Shetty equation for males is statistically different from the derived equations - 1. Contrary to males, females' equation is almost equal with the derived females' equation - 1. Piers and Shetty equation for females can be selected as the most representative equation for our females' sample.

Ravussin and Bogardous equation use fat free body weight and age as explanatory variables and will be compared with the equation - 3. The comparison shows statistically different coefficients for age and almost the same coefficients for fat free body weight.

For ordinary multiple regression models, the R^2 index is a good measure of the model's predictive ability, especially when applying the model to other datasets (Frank and Harrell, 2001). In order to choose the most representative established equation for the total data sample we will calculate $R^2 = 1 - \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2}$ for each of them as shown in the following table.

Table 3.6.3 R^2 on total data sample for foreign equations

Equations	R^2 on total data sample (Males and Females)
Harris – Benedict	0.49
FAO	0.58
Schofield et al	0.58
Ravussin and Bogardous	0.47
Henry and Rees	0.69
Piers and Shetty	0.67

Henry and Rees equations seem to be the most representative of the above equations with $R^2 = 0.69$ on the total sample of Greek athletes.

3.7. Conclusions

As we can see from the previous section some of the equations, presented in 2.4, do not statistically differ from those derived in this analysis.

The information mostly used is weight that seems to be the most important explanatory variable for REE. Some of these equations seem to be rather representative for our sample and especially Henry and Rees equation for males and Piers and Shetty equation for females. On the total data sample Henry and Rees equations are those with the highest R^2.

The Harris – Benedict equations use also height and age but not any information about body composition (fat proportion) is used.

Ravussin and Bogardous equation use fat free body weight and age in one equation that represent both males and females. No fat body weight information is included in the equation. It has to be investigated if fat free body mass can be considered as unisex in terms of energy expenditure.

To summarize Henry and Rees are the best of the equations existing in the literature with all model-coefficients inside the relative 95% confidence intervals of our derive equations-1 (see tables 3.6.1 and 3.6.2).