Multiple Linear Regression for Male Body Fat

Brandon Walters, Brian Jagoda, Mark Smith

Introduction:

This report will create a multiple linear regression model to predict the percent body fat of a man based on their measurements for certain physical characteristics. The results of our study showed that the most significant predictors for a male's percent body fat are their age, height, waist, and wrist measurements.

Data Description:

For this report, our population of interest consists of all males. The variables we will use as possible predictors for body fat percentage include the age, weight, height, neck, chest, waist, hip, thigh, knee, ankle, bicep, forearm, and wrist measurements of subjects. Table 1 below shows the type and role of each variable.

Table 1. Variable Information

Variable	Type (Units)	Role
Percent Body Fat	Quantitative (Percentage)	Response
Age	Quantitative (Years)	Explanatory
Weight	Quantitative (Pounds)	Explanatory
Height	Quantitative (Inches)	Explanatory
Neck	Quantitative (Centimeters)	Explanatory
Chest	Quantitative (Centimeters)	Explanatory
Waist	Quantitative (Inches)	Explanatory
Hip	Quantitative (Centimeters)	Explanatory
Thigh	Quantitative (Centimeters)	Explanatory
Knee	Quantitative (Centimeters)	Explanatory
Ankle	Quantitative (Centimeters)	Explanatory
Bicep	Quantitative (Centimeters)	Explanatory
Forearm	Quantitative (Centimeters)	Explanatory
Wrist	Quantitative (Centimeters)	Explanatory

The data for this report was collected from a study where 250 men of various ages were randomly selected and had measurements for each variable taken. This was an observational study that randomly chose participants, took observations, and applied no treatments.

The correlation between variables can be seen in Table 2 below. Notice that the response variable of percent body fat possesses a significant correlation with the explanatory variables of weight, chest, waist, hip, and thigh. Thus, these variables are potential good predictors for the response variable.

Additionally, as can be seen in Table 2, there is a strong correlation between the weight variable and the variables of neck, chest, waist, hip, thigh, and knee. Also, the variable of hip possesses a strong correlation with chest, waist, thigh, and knee. Finally, the chest variable is strongly correlated with waist. Therefore, there is potential for multicollinearity among these predictors given the correlation between these variables. This multicollinearity can lead to having wider confidence intervals thus producing less reliable data in relation to the effect of the independent variables in a model.

Table 2. Correlation Matrix for Full Model

	Percent Body Fat	**Age**	**Weight**	**Height**	**Neck**	**Chest**	**Waist**	**Hip**	**Thigh**	**Knee**	**Ankle**	**Bicep**	**Forearm**	**Wrist**
Percent Body Fat	1.000	0.295	0.617	-0.029	0.489	0.701	0.824	0.633	0.549	0.492	0.245	0.482	0.365	0.339
Age	0.295	1.000	-0.016	-0.246	0.119	0.182	0.243	-0.058	-0.216	0.017	-0.110	-0.044	-0.085	0.218
Weight	0.617	-0.016	1.000	0.513	0.810	0.891	0.874	0.933	0.852	0.843	0.581	0.785	0.683	0.725
Height	-0.029	-0.246	0.513	1.000	0.325	0.224	0.187	0.397	0.350	0.513	0.395	0.319	0.322	0.397
Neck	0.489	0.119	0.810	0.325	1.000	0.769	0.728	0.708	0.669	0.648	0.434	0.709	0.661	0.731
Chest	0.701	0.182	0.891	0.224	0.769	1.000	0.910	0.825	0.708	0.698	0.447	0.707	0.599	0.644
Waist	0.824	0.243	0.874	0.187	0.728	0.910	1.000	0.861	0.737	0.710	0.407	0.656	0.530	0.602
Hip	0.633	-0.058	0.933	0.397	0.708	0.825	0.861	1.000	0.881	0.809	0.521	0.722	0.603	0.626
Thigh	0.549	-0.216	0.852	0.350	0.669	0.708	0.737	0.881	1.000	0.777	0.504	0.744	0.604	0.544
Knee	0.492	0.017	0.843	0.513	0.648	0.698	0.710	0.809	0.777	1.000	0.585	0.654	0.579	0.656
Ankle	0.245	-0.110	0.581	0.395	0.434	0.447	0.407	0.521	0.504	0.585	1.000	0.449	0.429	0.545
Bicep	0.482	-0.044	0.785	0.319	0.709	0.707	0.656	0.722	0.744	0.654	0.449	1.000	0.701	0.614
Forearm	0.365	-0.085	0.683	0.322	0.661	0.599	0.530	0.603	0.604	0.579	0.429	0.701	1.000	0.598
Wrist	0.339	0.218	0.725	0.397	0.731	0.644	0.602	0.626	0.544	0.656	0.545	0.614	0.598	1.000

Full Regression Model:

Next, we ran the multiple linear regression for the full model, including all explanatory variables. The analysis of variance (ANOVA) table can be seen below in Table 3 and the coefficients table is shown in Table 4. The multiple linear regression produced an F statistic of $F(13,236) = 54.61$ and a p-value of $p = 2.2 * 10^{-16}$. Since the $2.2 * 10^{-16} < 0.05$, we believe that at least one of the variables is not equal to zero and that the full model is significant. The value for the coefficient of determination is $R^2 = 0.7501$. Thus, 75.01% of the total variance in percent body fat can be explained by the linear model using all variables. As can be seen in Table 2, the least significant predictor in the full model is the knee variable.

Table 3. Full Model ANOVA Table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic	P-value
Model (Regression)	12855.1	13	988.84	54.61	$2.2 * 10^{-16}$
Error (Residual)	4273.7	236	18.1		
Total	17128.8	249			
Coefficient of determination: $R^2 = 0.7501$			Residual standard error: $SE = 4.255$		

Table 4. Full Model Coefficients Table

Term	Estimate	Standard Error	T-Value	p-value
(Intercept)	1.68516	23.37412	0.072	0.942587
Age	0.07189	0.03217	2.234	0.026389
Weight	-0.01762	0.06714	-0.263	0.793153
Height	-0.24675	0.19114	-1.291	0.197989
Neck	-0.38682	0.23486	-1.647	0.100887
Chest	-0.11919	0.10825	-1.101	0.272004
Waist	2.29748	0.23215	9.897	2×10^{-16}
Hip	-0.15878	0.14586	-1.089	0.277446
Thigh	0.17299	0.14683	1.178	0.239926
Knee	-0.04580	0.24560	-0.186	0.852230
Ankle	0.18502	0.21985	0.842	0.400862
Bicep	0.17968	0.17039	1.054	0.292733
Forearm	0.27605	0.20692	1.334	0.183454
Wrist	-1.80162	0.53304	-3.380	0.000848

Reduced Model:

Next, we continuously took away the least significant predictor and then reran the model until all predictors were significant. Table 5 shows each of the steps in the reduction, the variables that remained during that step, and the value of R^2 for the step.

Table 5. Reduction Table

Step	Variable Name (Still in Model)	Variable Removed	Value of R^2
1.	• Age • Weight	Knee	0.7505

	• Height • Neck • Chest • Waist • Hip • Thigh • Ankle • Bicep • Forearm • Wrist		
2.	• Age • Height • Neck • Chest • Waist • Hip • Thigh • Ankle • Bicep • Forearm • Wrist	Weight	0.7504
3.	• Age • Height • Neck • Chest • Waist • Hip • Thigh • Bicep • Forearm • Wrist	Ankle	0.7497
4.	• Age • Height • Neck • Chest • Waist • Hip • Thigh • Forearm • Wrist	Bicep	0.7486
5.	• Age	Chest	0.7469

	• Height • Neck • Hip, • Waist • Thigh • Forearm • Wrist		
6.	• Age • Height • Neck • Waist • Thigh • Forearm • Wrist	Hip	0.7445
7.	• Age • Neck • Waist • Forearm • Wrist • Height	Thigh	0.7433
8.	• Age • Neck • Waist • Wrist • Height	Forearm	0.7404
9.	• Age • Waist • Height • Wrist	Neck	0.7383
	Variables of Final Reduced Model • Age • Waist • Height • Wrist		

The variables that remain significant throughout reduction are Age, Waist, Height, and Wrist. Thus, the other variables were removed, to form the reduced model:

$$Percent\ body\ fat = \beta_0 + \beta_{Age}(Age) + \beta_{Waist}(Waist) + \beta_{Height}(Height) + \beta_{Wrist}(Wrist) + \varepsilon$$

where ε is the residual error and needs to be normally distributed with a mean of zero, standard deviation σ_ε, and it is independent of each of the explanatory variables. The parameters of the reduced model include β_{Age}, which represents the coefficient for the variable of Age, β_{Waist}, which represents the coefficient for the variable of Waist, β_{Height}, which represents the coefficient for the variable of Height, β_{Wrist}, which represents the coefficient for the variable of Wrist, and σ_ε, which represents the standard deviation of the residual error. The conditions to use this reduced model include the data being obtained through simple random sampling, there being a linear relationship between the response variable and these explanatory variables, the explanatory variables not being highly correlated with each other, the residuals must be normally distributed, and the variance of error terms must be similar across the values of the independent variables.

Our hypotheses for determining if the reduced model is significant were $H_O: \beta_{Age} = \beta_{Waist} = \beta_{Height} = \beta_{Wrist} = 0$ versus $H_a: At\ least\ one\ of\ these\ are\ not\ equal\ to\ zero$. As can be seen in Table 6, the reduced model produced a test statistic of $F(4, 245) = 172.772$ with a p-value of $p = 2.2 * 10^{-16}$. Since our p-value of $p = 2.2 * 10^{-16} < 0.05$, we can reject our null hypothesis and conclude that at least one of the predictors in the reduced model is not equal to zero and that our reduced model is significant in predicting the percent body fat for the population of all men.

Table 6. Reduced Model ANOVA Table

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F Statistic**	**P-value**
Model (Regression)	12646.9	4	3161.725	172.772	$2.2 * 10^{-16}$
Error (Residual)	4482.0	245	18.3		
Total	17128.9	249	68.791		
Coefficient of determination: $R^2 = 0.7383$			Residual standard error: $SE = 4.277$		

Since the reduced model produced a coefficient of determination of $R^2 = 0.7383$, we can say that 73.83% of the total variation in percent body fat can be explained by the final reduced model. This is not a significant reduction in explained variation from the full model of 75.01%. To verify this, we ran the Nested F-test and produced a result of $F(9, 236) = 1.278$ with a p-value of p=0.2496. Since $p = 0.2497 > 0.05$, we can conclude that the reduced model did not cause a significant reduction in the explained variation of the full model.

The prediction equation for the reduced model is $\widehat{Percent\ body\ fat} = 2.900 + 0.056(Age) - 0.323(Height) + 1.958(Waist) - 1.911(Wrist)$. The coefficients table, with the 95% confidence intervals of the coefficients, can be seen below in Table 7. Each of the explanatory variables contribute to predicting the response variable, percent body fat. An increase in Age of 10 years will result in an increase in percent body fat by $10(0.056) = 0.5602$ points, holding all other variables constant. An increase in Height of 10 inches will result in a decrease in percent

body fat by 3.23 points holding all other variables constant. An increase in Waist of one inch will result in an increase in percent body fat by 1.958 points, holding all other variables constant. An increase in Wrist of one centimeter will result in a decrease in percent body fat by 1.911 points, holding all other variables constant.

Table 7. Reduced Model Coefficients Table

Term	**Estimate**	**Standard Error**	**T-Value**	**P-Value**	**(CI) 2.5%**	**(CI) 97.5%**
(Intercept)	2.90033	8.08402	0.359	0.7201	-13.022	18.8234
Age	0.05602	0.02382	2.351	0.0195	0.00909	0.10294
Height	-0.32299	0.12155	-2.657	0.0084	-0.5624	-0.0836
Waist	1.95825	0.08539	22.932	$2 * 10^{-16}$	1.790053	2.12645
Wrist	-1.91138	0.40953	-4.667	$5.03 * 10^{-6}$	-2.71802	-1.1047

As can be seen in Table 9, the Waist variable is highly correlated with the response variable. There are no explanatory variables that are highly correlated with each other. The only variable that was highly correlated with the response variable that remained in the reduced model was the Waist variable. We see further evidence of this within Table 8. (Pairwise Scatterplot) as shown below. Where Waist and the response variable show evidence of high correlation. In the full model, the Waist variable was highly correlated with the Weight, Chest, and Hip variables, however each of these were removed during the reduction of the model. All other explanatory variables in the reduced model do not possess strong correlation with another explanatory variable. Thus, there are no highly correlated explanatory variables in the reduced model and there are no potential issues with multicollinearity between variables.

Table 8. Pairwise Scatterplot

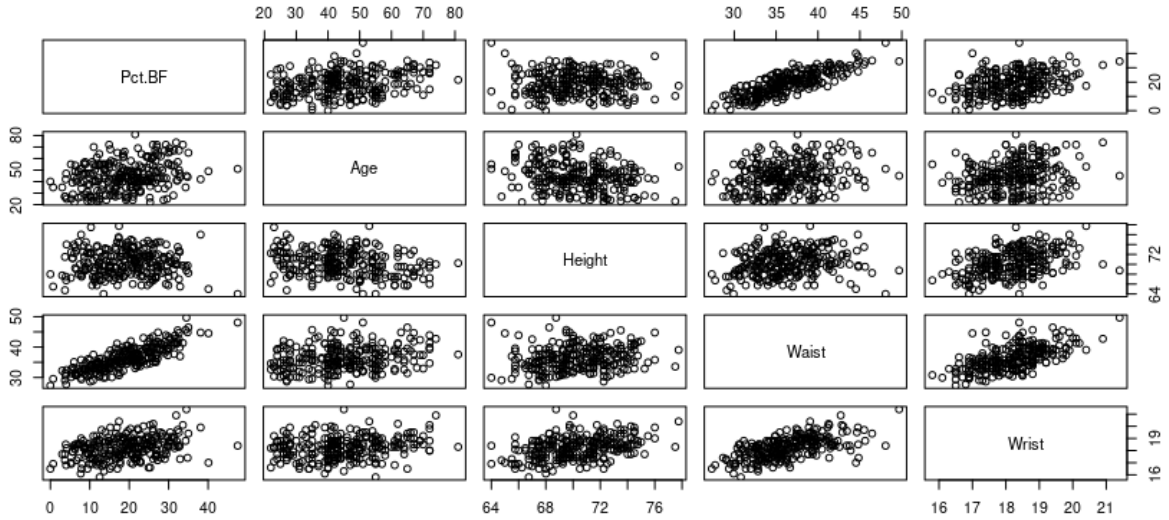


Table 9. Correlation Matrix for Reduced Model

	Percent Body Fat	**Age**	**Height**	**Waist**	**Wrist**
Percent Body Fat	1.000	0.295	-0.029	0.824	0.339
Age	0.295	1.000	-0.246	0.243	0.218
Height	-0.029	-0.246	1.000	0.187	0.397
Waist	0.824	0.243	0.187	1.000	0.602
Wrist	0.339	0.218	0.397	0.602	1.000

Next, we ran the Shapiro-Wilkes Test to test for non-normality for the residuals. The hypotheses for this test are H_O: $The\ residuals\ are\ normal\ for\ the\ population$ and H_a: $The\ residuals\ are\ not\ normal\ for\ the\ population$. This test produced a p-value of $p = 0.02278$, so we reject the null hypothesis. Thus, we have evidence that the residuals are not normal for the population of all men. Also, in Tables 10 and 11 below, we can also see that there are multiple residuals that are greater than two standard deviations from the mean. Therefore, the reduced model does not meet the condition of normality for residuals.

Table 10. Histogram of Standardized Residuals

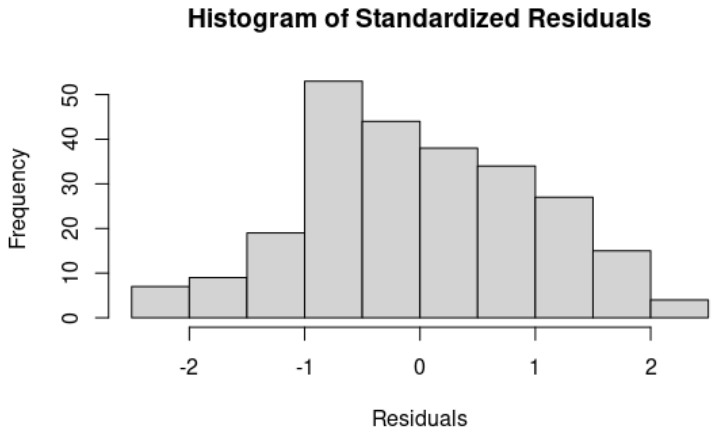
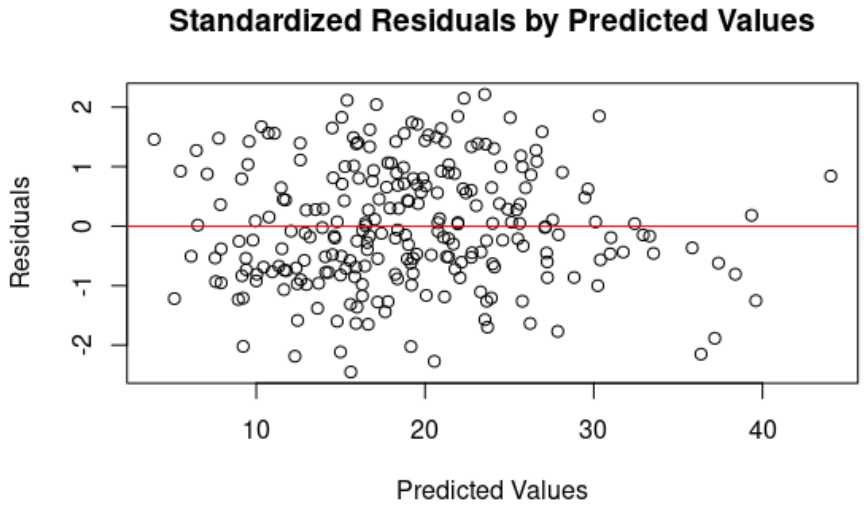


Table 11. Residuals vs Predicted Values Scatterplot



Using the Model:

A certain male subject who lives in close vicinity to Dr. L. graciously volunteered to be measured. He is 6 foot, 3 inches tall (or 75 inches tall), is 58 years old, has a waist measurement of 37.5 inches and a wrist measurement of 16.5 centimeters. The prediction equation for the reduced model produced a result of 23.82% percent body fat for a man with these measurements.

$$\widehat{Percent\ body\ fat} = 2.900 + 0.056(58) - 0.323(75) + 1.958(37.5) - 1.911(16.5) = 23.82$$

A 95% prediction interval for the percent body fat of this man is $15.029 < Percent\ body\ fat < 32.615$. A 95% confidence interval for the average percent body fat of all men with the same measurements as this man is $21.303 < Percent\ body\ fat < 26.341$. The margin of error for the prediction interval of the individual response is 8.793 and the margin of error for the confidence interval of the average response is 2.519. The margin of error for the confidence interval is smaller because the standard error for the mean response in always smaller than the standard error of an individual response.