Regression Analysis Student Project, Fall 2011

Maria Elvira David Paderes

mdpaderes@gmail.com

Weight, Age and Blood Fat Content

NOTE: The data used in this student project was taken from the following URL:
http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html[1]

In particular, the dataset is contained in the file entitled <x09.txt>.

The dataset used for this student project contains 25 data points, each of which represents the following information of a hypothetical individual:

1. Weight, in kilograms;
2. Age, in years; and
3. Blood fat content.

The dataset is shown in the appendix at the end of this paper.

This student project attempts to determine if blood fat content (the response variable) is correlated to the weight and the age of a person (the explanatory variables).

[1] This is a webpage that contains linear regression datasets posted by John Burkardt, a research associate in the Department of Scientific Computing of the Florida State University.

Hypothesis

Blood fat content is more correlated to the weight than to the age of a person.

Explanatory Variable: Weight

I ran MS Excel's Analysis ToolPak for regression analysis using weight as the explanatory variable and blood fat content as the response variable. The following are the results:

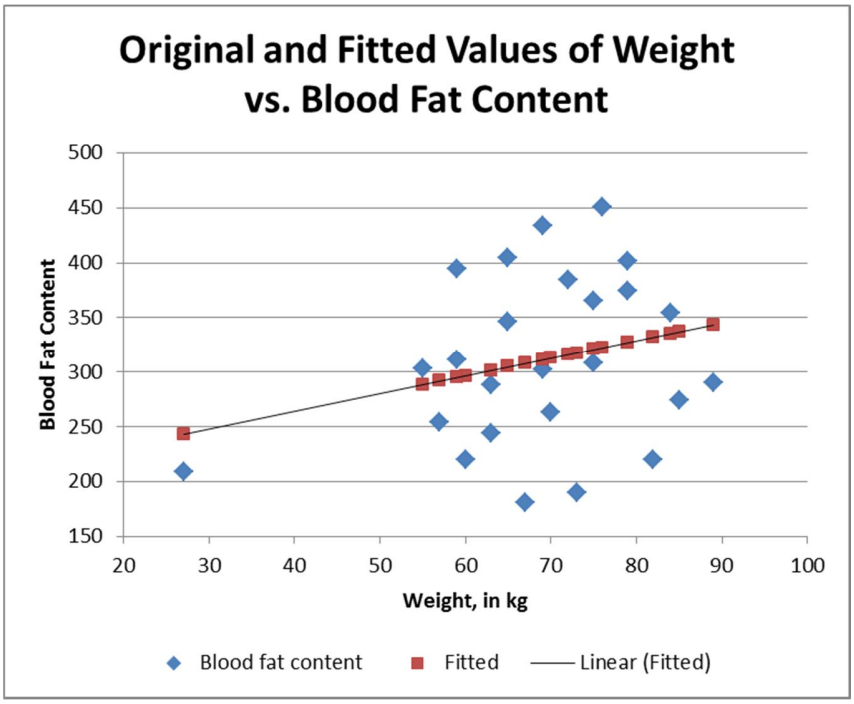
Regression Statistics								
Multiple R	0.265							
R Square	0.070							
Adjusted R Square	0.030							
Standard Error	76.654							
Observations	25							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	10,231.726	10,231.726	1.741	0.200			
Residual	23	135,145.314	5,875.883					
Total	24	145,377.040						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	199.298	85.818	2.322	0.029	21.770	376.825	21.770	376.825
Weight (kg)	1.622	1.229	1.320	0.200	-0.921	4.166	-0.921	4.166

The following may be noted from the information above:

- The R^2 is low, which is only around 0.070.
- The critical value for the two-sided *t*-distribution with 23 degrees of freedom[2] and a 95% confidence level[3] is approximately 2.069. However, the *t* statistic for the null hypothesis $H_0: \beta_1 = 0$ is 1.320[4]. Thus, we cannot reject the null hypothesis.
- The p-value for β_1 is roughly 20%, which is much greater than the significance level of 5%. This further supports the above statement that we cannot reject the null hypothesis.

The following is a scatterplot of the original and fitted values of blood fat content. There is no obvious relationship between weight and blood fat content:

[2] The degrees of freedom is equal to *n-2,* where *n* is equal to the number of data points used. In this case, *n* is equal to 25.
[3] The confidence level used in all regression analysis runs in the MS Excel Analysis ToolPak for this student project is 95%.
[4] β_1 is the β coefficient for weight.



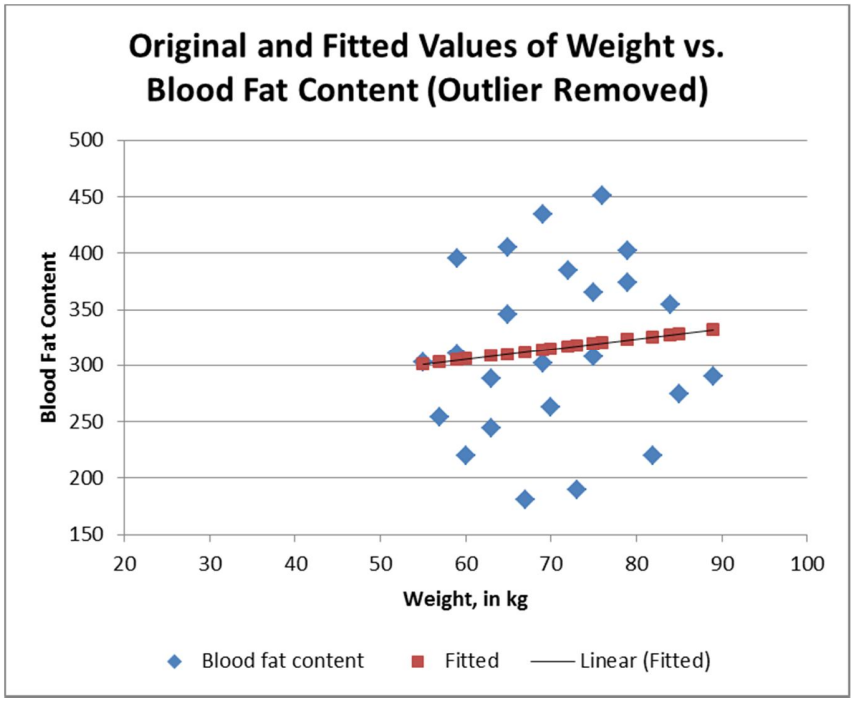
However, notice the outlier data point of 27 kgs. This is corrected by removing the said outlier and running the Analysis ToolPak again. The following are the results of the second run:

Outlier (index 11) removed								
Regression Statistics								
Multiple R	0.113							
R Square	0.013							
Adjusted R Square	-0.032							
Standard Error	77.717							
Observations	24							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	1,719.905	1,719.905	0.285	0.599			
Residual	22	132,879.054	6,039.957					
Total	23	134,598.958						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	250.893	121.100	2.072	0.050	-0.254	502.040	-0.254	502.040
Weight (kg)	0.910	1.705	0.534	0.599	-2.626	4.446	-2.626	4.446

The following may be noted from the information above:

- The R^2 declined from 0.070 to 0.013.
- The critical value for the two-sided *t*-distribution with 22 degrees of freedom and a 95% confidence level is approximately 2.074. However, the *t* statistic for the null hypothesis $H_0: \beta_1 = 0$ is 0.534. Again, from this, we cannot reject the null hypothesis.
- The p-value for β_1 is roughly 60%, which, again, is much greater than the significance level of 5%. This further supports the above statement that we cannot reject the null hypothesis.

The following is a scatterplot of the results of the second run:



As expected, there is still no obvious relationship between weight and blood fat content.

Explanatory Variable: Age

This time, I used age as the explanatory variable. The following are the results of this run:

Regression Statistics								
Multiple R	0.837							
R Square	0.701							
Adjusted R Square	0.688							
Standard Error	43.461							
Observations	25							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	101,932.666	101,932.666	53.964	0.000			
Residual	23	43,444.374	1,888.886					
Total	24	145,377.040						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	102.575	29.638	3.461	0.002	41.265	163.885	41.265	163.885
Age (yrs)	5.321	0.724	7.346	0.000	3.822	6.819	3.822	6.819

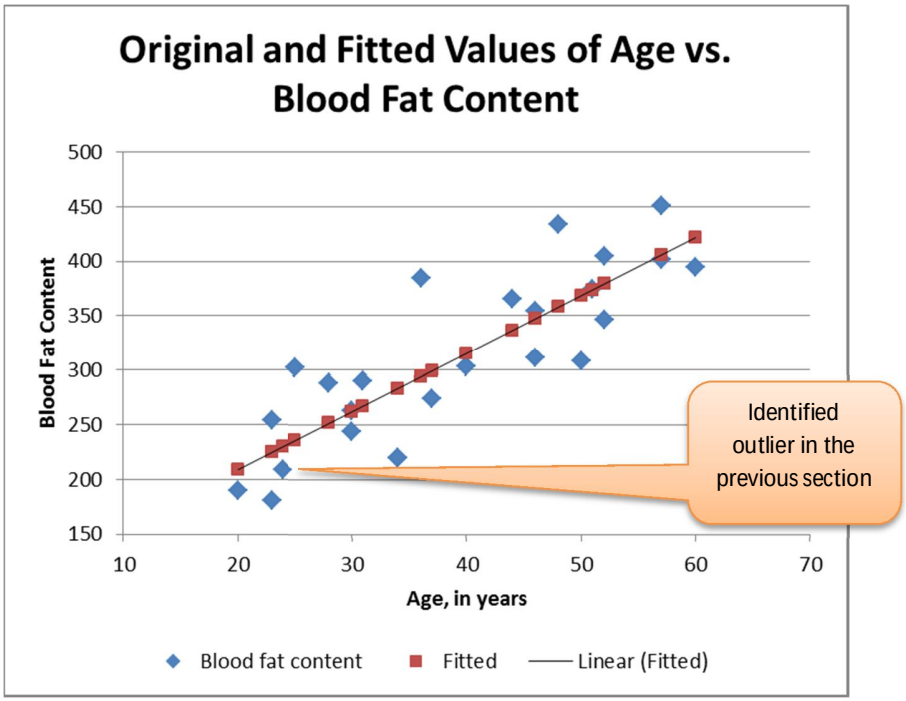
The following may be noted from the information above:

- The R^2 is much higher this time, which is around 0.701.
- As mentioned earlier, the critical value for the two-sided *t*-distribution with 23 degrees of freedom and a 95% confidence level is approximately 2.069. The *t* statistic for the null

hypothesis H_0: $\beta_2 = 0$ is 7.346[5], which far exceeds the critical value. Thus, we can reject the null hypothesis.

- Finally, the p-value for β_2 is roughly 0%, which is obviously less than the significance level of 5%. This further supports the above statement that we can reject the null hypothesis.

The following is a scatterplot of the original and fitted values of blood fat content using age as the explanatory variable:



It is interesting to note that there is a discernible positive correlation between age and blood fat content, even when the outlier identified in the previous section is included.

[5] β_2 is the β coefficient for age.

Explanatory Variables: Weight and Age

Finally, in this section, both weight and age are used as explanatory variables. The following are the results:

Regression Statistics								
Multiple R	0.840							
R Square	0.706							
Adjusted R Square	0.679							
Standard Error	44.111							
Observations	25							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	102,570.815	51,285.407	26.358	0.000			
Residual	22	42,806.225	1,945.738					
Total	24	145,377.040						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.983	52.430	1.487	0.151	-30.750	186.715	-30.750	186.715
Weight (kg)	0.417	0.729	0.573	0.573	-1.094	1.929	-1.094	1.929
Age (yrs)	5.217	0.757	6.889	0.000	3.646	6.787	3.646	6.787

The following may be noted from the information above:

- The R^2 is equal to 0.706. Compare this to the R^2 of 0.701 when only age is the explanatory variable. I interpret this as: Does using weight as an additional explanatory variable contribute to fitting values for blood fat content? My answer is: Not much.
- The β coefficient for age, which is 5.217, is much greater than the β coefficient for weight, which is 0.417.
- The story is the same for the *t* statistics: Compared to the critical value of 2.069, the *t* statistic for weight, 0.573, is too low for the null hypothesis H_0: $\beta_1 = 0$ to be rejected, while the *t* statistic for age, 6.889, is high enough that we can reject the null hypothesis H_0: $\beta_1 = 0$.

Conclusion

Blood fat content is more correlated to the age than to the weight of a person, thus disproving the initial hypothesis set out at the start of this student project.

Appendix

The following is the data set used for this student project:

Index	Weight (in kgs)	Age (in years)	Blood fat content
1	84	46	354
2	73	20	190
3	65	52	405
4	70	30	263
5	76	57	451
6	69	25	302
7	63	28	288
8	72	36	385
9	79	57	402
10	75	44	365
11	27	24	209
12	89	31	390
13	65	52	346
14	57	23	254
15	59	60	395
16	69	48	434
17	60	34	220
18	79	51	374
19	75	50	308
20	82	34	220
21	59	46	311
22	67	23	181
23	85	37	274
24	55	40	303
25	63	30	244