Individual project report

Predicting the somatotype of children in age 18

Name: Wong Choi Yan Heki Student ID : 20040850 Data set: 20

Summary

The project is to analyze the data set of 136 children born in 1928-29 in Berkeley in order to identify their somatotype, which is the body types based on their body shape, at age 18, so that they can then match with the most suitable sport or activities according to their somatotype. The objective is developing a new multiple regression model for predicting the somatotype of children, at age 18. Gender, weight gain from age 2 to 9, weight gain from 9 to 18, and height in age 18 are found to be the significant predictors for the somatotype at age 18. Results come out that children who increase their weight gain from age 2 to age 9 and from age 9 to age 18, may have less chance of being thin, and being normal, at age 18. In addition, children who is taller at age 18 and who is a boy, have higher chance of being thin or being normal. Apart from these, height of children at age18 may have same effect on predicting children being thin and children being normal, but other variables may have different effect on predicting.

Background information

The data set contains 12 variables on 136 children born in 1928-29 in Berkeley. The outcomes variables is **soma**, somatotype, a scale from 1, very thin, to 7, obese, of body type. The variables includes **sex**(0=male, 1=female), a binary categorical variables, and nine continuous variables: age 2 weight(**wt2**), age 2 height(**ht2**), age9 weight(**wt9**), age9 height(**ht9**), age9 leg circumference(**lg9**), age9 strength(**st9**), age18 weight(**wt18**), age18 height(**ht18**), age18 leg circumference(**lg18**), age18 strength(**st18**).

Variable	N	Mean	Std Dev	Minimum	Maximum
Sex	116	0.4310345	0.4973694	0	1.0000000
WT2	116	13.2517241	1.6591503	10.1000000	18.6000000
HT2	116	87.8301724	3.3575669	81.3000000	98.2000000
WT9	116	31.5034483	6.0249904	19.9000000	66.8000000
HT9	116	135.4750000	5.5962546	121.4000000	152.5000000
LG9	116	27.5913793	2.4505747	21.8000000	40.4000000
ST9	116	65.5517241	15.3330248	22.0000000	121.0000000
WT18	116	65.4810345	10.6598141	42.9000000	110.2000000
HT18	116	173.5353448	8.8899577	154.6000000	195.1000000
LG18	116	35.8913793	2.5701148	30.0000000	44.1000000
ST18	116	174.2068966	50.2454008	77.0000000	260.0000000

Table1: Weighted means of all of the independent variables

Soma	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4	3.45	4	3.45
1.5	7	6.03	11	9.48
2	13	11.21	24	20.69
2.5	2	1.72	26	22.41
3	17	14.66	43	37.07
3.5	5	4.31	48	41.38
4	25	21.55	73	62.93
4.5	10	8.62	83	71.55
5	16	13.79	99	85.34
5.5	6	5.17	105	90.52
6	7	6.03	112	96.55
6.5	2	1.72	114	98.28
7	2	1.72	116	100.00

Table2: Weighted frequencies for the dependent variables

Objective

The objective is developing a new multiple regression model for predicting the somatotype of children, at age 18. We will do regression analysis in order to determine which predictors are significant to create model to identify the somatotype of children. To this end, a new multiple regression model will be proposed based on the analysis results.

Methodoloy and Results

Since there are 13 possible outcomes variables, which are somatotype, a scale from 1, 1.5, 2, 2.5, 3, … , 7 (from very thin to obese), we will use either proportional odds model or multinomial logit model. It is checked that the proportional odds assumption is not satisfied (P-value of Score test <0.0001, reject the null), which means the slope coefficients depend on the level of the response variables, multinomial logistic regression is thus chosen for modeling, designating last level as reference level.

Step1 :Regroup the responses

As some of the responses have only few observations, we need to regroup the responses in order to get a more precise model. By considering the means of independent variables, we will group the response if the corresponding independent variables have similar mean. The 13 responses are thus regrouped into 3 main responses, which is somatotype, a scale from 1(thin), 2(normal), 3(obese).

soma1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	43	37.07	43	37.07
2	40	34.48	83	71.55
3	33	28.45	116	100.00

Table3: Weighted frequencies for the dependent variables after grouping

In place of the preceding variables, we also consider the following variables:

WTD9 = $\frac{\mathrm{WT9}-\mathrm{WT2}}{9-2}$ = average weight gain from age 2 to 9

WTD18 = $\frac{\mathrm{WT18}-\mathrm{WT9}}{18-9}$ = average weight gain from age 9 to 18

Step2: Model selection

	Model	AIC	BIC
1	**Intercept, wtd18**	253.991	265.005
2	**Intercept, sex, wtd18**	170.682	187.203
3	**Intercept, sex, wtd18, wtd9**	153.911	175.940
4	**Intercept, sex, wtd18, wtd9, ht18**	138.633	166.169
5	**Intercept, sex, wtd18, wtd9, ht18, ht9**	138.349	171.392
6	**Intercept, sex, wtd18, wtd9, ht18, ht9, ht2**	141.525	180.075
7	**Intercept, sex, wtd18, wtd9, ht18, ht9, ht2, wt2**	144.977	189.035

Table4: AIC and BIC of models

We input the variables from the most significant one, to the least significant one, to compare the corresponding AIC and BIC. All the interaction terms are found to be insignificant. From table 4, the 4^{th}have the smallest BIC and 5^{th}model have the smallest AIC. The model with the smallest Akaike Information Criterion (AIC) represents the best approximation to the true model, and BIC is closely related to AIC. For the adjusted R^2 value, which derived to reflect the extent of how the independent variables contribute to the variation in the dependent variable. 4^{th} model have adjusted R^2 =0.4925 and 5^{th} model have adjusted R^2=0.50156. As 4^{th}model only have slightly lower adjusted R^2 and slightly higher AIC than 5^{th}model, we use the sequential methods to decide the finalize model.

We can get the follow result from the SAS,

Forward selection: **Intercept, sex, wtd18, wtd9 and ht18**

Backward elimination: **Intercept, sex, wtd18, wtd9 and ht18**

Stepwise selection: **Intercept, sex, wtd18, wtd9 and ht18**

The all results are consistency. Thus, variables (**sex, wtd18, wtd9 and ht18**) have been entered into the model.

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Sex	2	1	49.6705	<.0001
2	wtd18	2	2	31.5029	<.0001
3	wtd9	2	3	22.5297	<.0001
4	HT18	2	4	15.8160	0.0004

Table5: Forward selection

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	WT2	2	6	0.5362	0.7648
2	HT2	2	5	0.8062	0.6682
3	HT9	2	4	3.9582	0.1382

Table6: Backward elimination

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Sex		2	1	49.6705		<.0001
2	wtd18		2	2	31.5029		<.0001
3	wtd9		2	3	22.5297		<.0001
4	HT18		2	4	15.8160		0.0004

Table7: Stepwise selection

Step3: Measure the multicollinearity of the independent variables

The variance inflation factor (VIF) measures the multicollinearity of the independent variables in a multiple linear regression model. The higher the VIF, the lower the precision in the estimate of the parameter. In the table below, we can see that the VIF of all the independent variables is low (all <5), so the multicollinearity of the variables is low, and thus the parameter estimate in the model is precise.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.78894	1.30205	3.68	0.0004	0
Sex	1	1.10215	0.14018	7.86	<.0001	2.59132
wtd9	1	0.41834	0.06504	6.43	<.0001	1.22947
wtd18	1	0.46368	0.05935	7.81	<.0001	1.72503
HT18	1	-0.03568	0.00781	-4.57	<.0001	2.57132

Table8: VIF of model

Step4: Goodness of fit test

The p-value of Deviance and Pearson Goodness-of-fit statistic >0.05, so we do not reject the null hypothesis and thus the model fitted well at significant level 0.05. In addition, we reject the null hypothesis in Likelihood Ratio test, which means the model provides a good fit to the dependent and independent variables.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	118.6335	222	0.5344	1.0000
Pearson	138.4897	222	0.6238	1.0000

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	134.8564	8	<.0001
Score	87.5737	8	<.0001
Wald	37.2507	8	<.0001

Table9: Deviance and Pearson Goodness-of-fit statistics *Table10: Likelihood Ratio Statistic*

Step5: Fit the model

Hence, the fitted model is:

$$\begin{cases} \log\left(\frac{P(Soma=1)}{P(Soma=3)}\right) = -32.2101 + 5.7389(sex=0) - 4.5391wtd9 - 5.6588wtd18 + 0.3784ht18 \\ log\left(\frac{P(Soma=2)}{P(Soma=3)}\right) = -26.7830 + 2.9962(sex=0) - 3.0614wtd9 - 4.1313wtd18 + 0.3017ht18 \end{cases}$$

Analysis of Maximum Likelihood Estimates							
Parameter		soma1	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1	-32.2101	15.1596	4.5145	0.0336
Intercept		2	1	-26.7830	12.6291	4.4975	0.0339
Sex	0	1	1	5.7389	1.2198	22.1352	<.0001
Sex	0	2	1	2.9962	0.9684	9.5731	0.0020
wtd9		1	1	-4.5391	1.0135	20.0580	<.0001
wtd9		2	1	-3.0614	0.8165	14.0586	0.0002
wtd18		1	1	-5.6588	1.2275	21.2540	<.0001
wtd18		2	1	-4.1313	1.0466	15.5807	<.0001
HT18		1	1	0.3784	0.1091	12.0238	0.0005
HT18		2	1	0.3017	0.0931	10.5075	0.0012

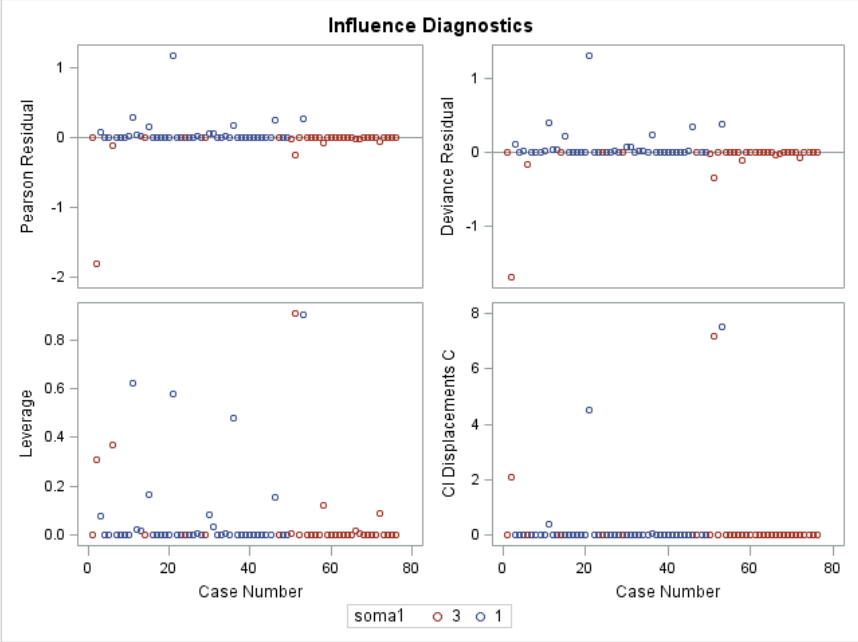
Table11: Parameter Estimates of model

Step 6: Outliers or influential points

Outliers and influential points in the model can be caused by measurement errors or may be the result of inherent variability of the data. Thus, we should find out those data, which are considerably exceptional or inconsistent with the rest of the data after we fitted the model. As multinomial logistic regression in SAS does not compute any diagnostic statistics, we use the Logistic Regression procedure to calculate and examine diagnostic measures. We run two binary logistic regressions (somatotype level1 to level 3, somatotype level2 to level3).

For somatotype level 1 and level 3:

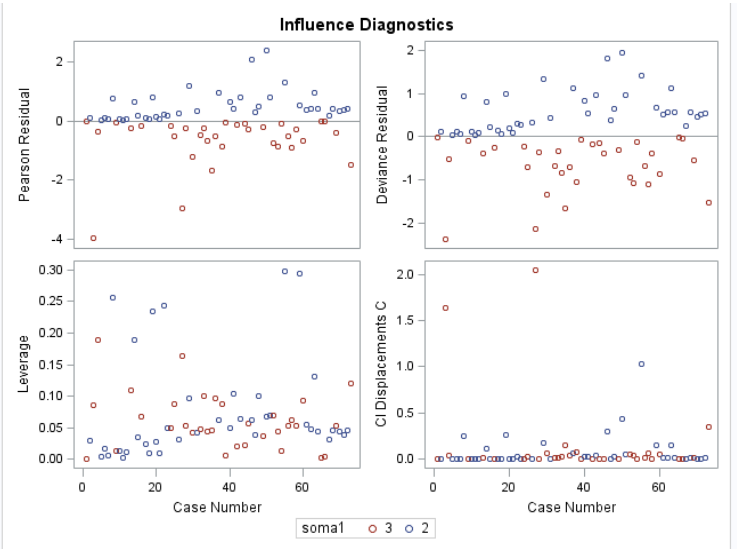
Pearson residual is used as an indicator to determine outlier. We use ± 2 as the cutoff value. From the graph 1, we can see that the highest Pearson Residual is $2^{nd}(1.79977)$and $21^{st}(1.16933)$data, but their value is not big. So they are not regarded as an outlier. For influential point, confidence interval displacement diagnostics measures the influence of individual observations on the regression estimates, and we use 1 as the cutoff value. From the graph, we find $2^{nd}(2.06476)$, $21^{st}(4.51793)$, $51^{st}(7.19684)$ and $53^{rd}(7.49434)$ are greater than 1. However, after deleting the individual observation, we cannot find significant change in the regression estimates, so these four datum are not influential point.



Graph 1: Influence Diagnostics for somatotype 1 and 3

<u>For somatotype level 2 and level 3:</u>

From the graph 2, we can see that $46^{th}(2.06438), 50^{th}(2.37769),\ 27^{th}(2.95559), 3^{rd}(3.97122)$ data have large Pearson residual and so these four datum are outlier. For influential point, we find $3^{rd}(1.63094),\ 27^{th}(2.04406)$ and $55^{th}(1.02264)$also have confidence interval displacement C greater than 1. After deleting the $3^{rd}\ and\ 27^{th}$data, we find significant change in the regression estimates, but we cannot get the significant change of parameter estimates after deleting 55^{th}data. So, only $3^{rd}\ and\ 27^{th}$data are influential point. We thus delete $46^{th}, 50^{th},\ 27^{th} and\ 3^{rd}$observations and fit the model again.



Graph 2: Influence Diagnostics for somatotype 2 and 3

***Step 7:* After remove the outlier**

Repeat Step2 to step6 again. By sequential method, it output the following results:

Forward selection: **Intercept, sex, wtd18, wtd9, ht9 and ht18**

Backward elimination: **Intercept, sex, wtd18, wtd9 and ht18**

Stepwise selection: **Intercept, sex, wtd18, wtd9 and ht18**

As both backward and stepwise selection agree with the same model, thus variables (**sex, wtd18, wtd9 and ht18)** is entered into the model. In addition, from table 15, the p-value of Deviance and Pearson Goodness-of-fit statistic >0.05, so we do not reject the null hypothesis and thus the model fitted well at significant level 0.05. In addition, we reject the null hypothesis in Likelihood Ratio test, which means the model provides a good fit to the dependent and independent variables. From table 17, the overall effects of sex, wtd18, wtd9 and ht18 are listed in this tables. Since P-value of Wald test <0.05, the null is rejected and all of the variables are significant.

For this model, the adjusted R^2 = 0.5313, AIC =124.696 and BIC=151.881. The adjusted R^2 is larger and the AIC and BIC is smaller, so the new model is better.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Sex	2	1	51.5859	<.0001
2	wtd18	2	2	29.8751	<.0001
3	wtd9	2	3	23.4424	<.0001
4	HT18	2	4	18.2251	0.0001
5	HT9	2	5	6.4097	0.0406

Table12: Forward selection

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	WT2	2	6	0.4225	0.8096
2	HT2	2	5	0.2719	0.8729
3	HT9	2	4	5.5964	0.0609

Table13: Backward elimination

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Sex		2	1	51.5859		<.0001
2	wtd18		2	2	29.8751		<.0001
3	wtd9		2	3	23.4424		<.0001
4	HT18		2	4	18.2251		0.0001
5	HT9		2	5	6.4097		0.0406
6		HT9	2	4		5.5964	0.0609

Table14: Stepwise selection

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	104.6965	214	0.4892	1.0000
Pearson	109.0028	214	0.5094	1.0000

Table15: Goodness-of fit statistic

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	140.0194	8	<.0001
Score	88.3795	8	<.0001
Wald	36.7634	8	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Sex	2	21.8972	<.0001
wtd9	2	20.0197	<.0001
wtd18	2	20.3125	<.0001
HT18	2	13.2702	0.0013

Table16: Likelihood Ratio Statistic *Table17: Type 3 Analysis of Effects*

From the SAS result, the new fitted model is

$$\begin{cases} \log\left(\frac{P(S=1)}{P(S=3)}\right) = -45.0757 + 6.3782(sex=0) - 5.3476wtd9 - 6.1126wtd18 + 0.4804ht18 \\ \log\left(\frac{P(S=2)}{P(S=3)}\right) = -38.0598 + 3.6944(sex=0) - 3.8062wtd9 - 4.6906wtd18 + 0.3954ht18 \end{cases}$$

Analysis of Maximum Likelihood Estimates							
Parameter		soma1	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1	-45.0757	17.8508	6.3763	0.0116
Intercept		2	1	-38.0598	15.5478	5.9923	0.0144
Sex	0	1	1	6.3782	1.4418	19.5710	<.0001
Sex	0	2	1	3.6944	1.2308	9.0106	0.0027
wtd9		1	1	-5.3476	1.1954	20.0124	<.0001
wtd9		2	1	-3.8062	1.0149	14.0650	0.0002
wtd18		1	1	-6.1126	1.3563	20.3118	<.0001
wtd18		2	1	-4.6906	1.1930	15.4572	<.0001
HT18		1	1	0.4804	0.1322	13.2152	0.0003
HT18		2	1	0.3954	0.1183	11.1719	0.0008

Table18: Parameter Estimates of model

Step 7: Odds

1. A one unit increase in **wtd9** multiplies the odds of being in somatotype 1 vs somatotype 3 by 0.005($e^{-5.3476}$). Because 100(0.005-1)% = -99.5%, the odds are expected to decrease by about 99.5%. Also, a one unit increase in **wtd9** also multiplies the odds of being in somatotype 2 vs somatotype 3 by 0.022($e^{-3.8062}$). Because 100(0.022-1)% = -97.8%, the odds are expected to decrease by about 97.8%.

In other words, children who increase gain weight from age 2 to age 9, may have less chance of being thin, in somatotype1 relative to in somatotype3 (chance decrease by 99.5%) and being normal, in somatotype2 relative to in somatotype 3(chance decrease by 97.8%), at age18.

2. Similar to the above result, a one unit increase in **wtd18** multiplies the odds of being in somatotype 1 vs somatotype 3 by 0.002($e^{-6.1126}$). Because 100(0.002-1)%=-99.8%, the odds are expected to decrease by about 99.8%. Also, a one unit increase in **wtd18** also multiplies the odds of being in somatotype 2 vs somatotype 3 by 0.009($e^{-4.6906}$). Because 100(0.009-1)%=-99.1%, the odds are expected to decrease by about 99.1%.

In other words, children who increase gain weight from age 9 to age 18, may have less chance of being thin, in somatotype1 relative to somatotype 3(chance

decrease by 99.8%) and being normal, in somatotype2 relative to somatotype 3(chance decrease by 99.1%), at age18.

3. A one unit increase in **ht18** multiplies the odds of being in somatotype 1 vs somatotype 3 by 1.617($e^{0.4804}$). As 100(1.617-1)%= +61.7%, the odds are expected to increase by about 61.7%. A one unit increase in **ht18** also multiplies the odds of being in somatotype 2 vs somatotype 3 by 1.485($e^{0.3954}$). Because 100(1.485-1)%= +48.5%, the odds are expected to increase by about 48.5%.

In other words, children who is taller in age 18, may have higher chance of being thin, in somatotype1 relative to somatotype 3(chance increase by 61.7%) and being normal, in somatotype2 relative to somatotype 3(chance increase by 48.5%), at age18.

4. The relative log odds of being in somatotype 1 vs. in somatotype 3 will increase by 6.3782 if he is a boy instead of a girl. Similarly, the relative log odds of being in somatotype 2 vs. in somatotype 3 will increase by 3.6944 if he is a boy instead of a girl.

In the other words, boy will appear to be thinner in age 18 than girl. And the chance for boy to be thin(in somatotype 1 relative to somatotype 3) is higher than to be normal (in somatotype 2 relative to somatotype 3) at age18.

<u>Step 8: Test the equality of parameter estimate for each variables in two models</u>

To test whether effect of the four variables for predicting somatotype 1 to 3 and for predicting somatotype 2 to 3 is equal or not, we do the following hypotheses testing. The null hypothesis is that the parameter estimate of variables for predicting somatotype 1 to 3 is equal to that for predicting somatotype 2 to 3. The results come out that the effect of **ht18** for predicting somatotype 1 to 3 is not different from that of predicting somatotype 2 to 3 (P-value >0.05). But for the other variables **sex, wtd9 and wtd18**, the effect for predicting is different (P-value <0.05).

In other words, the height of children at age 18, may have same effect on predicting whether he or she is thin (in somatotype1 relative to somatotype 3) or normal (in somatotype2 relative to somatotype 3). However, gender of children, weight gain from age2 to 9, and weight gain from age 9 to 18, may have different effect on predicting whether he or she is thin or normal.

Linear Hypotheses Testing Results			
Label	Wald Chi-Square	DF	Pr > Ch Sq
sex0_1_vs_sex0_2	13.0048	1	0.0003
WTD9_1_vs_WTD9_2	5.9350	1	0.0148
WTD18_1_vs_WTD18_2	4.5032	1	0.0338
HT18_1_vs_HT18_2	1.9890	1	0.1584

Table19: Hypotheses testing of equality of parameter estimate in both model

Prediction

For new observation, we need to identify whether he/she is in somatotype 1, 2, or 3. For each observation, we put the value of sex, wtd9, wtd18, and ht18 into both of the model, in order to get the probability of he/she is in somatotype1, 2 and 3. We thus compare three of the probability and use the highest probability for our prediction result. To measure the accuracy of our model, we use the prediction results of 20 new observations to compare with their original results. It comes out that 14 out of 20 datum get the same result as the original. Our model thus have 70% accuracy.

Conclusion

Significant predictors for the somatotype predicting model will be gender, weight gain from age 2 to 9, weight gain from 9 to 18, and height in age 18. Results appear that children who increase their weight gain from age 2 to age 9 and from age 9 to age 18, may have less chance of being thin, and being normal, at age 18. In addition, children who is taller at age 18 and who is a boy, have higher chance of being thin or being normal. Moreover, height of children at age18 may have same effect on predicting whether he or she is thin or normal, but other variables may have different effect on predicting whether or she is thin or normal. These model can help to predict the somatotype of children at age 18, so that they can match with the most suitable sports and activities according to their somatotype.

Limitation on the interpretation and application

All the data is from children born in 1928-29. As the data were taken long time ago, the model may not suitable for measure the somatotype of children anymore. Moreover, there are only 136 data in the data file and only 112 data left for creating the model after deduce the outliers and prediction data, so more data should be collected in order to get a more accuracy model.