Frequentist Statistical Inference Project 2

Jaejin Kim

Spring 2019

1 Introduction

Somatotype is a concept developed by American psychologist William Herbert Sheldon, which refers to the underlying physique that is not changed by overeating or training. According to his theory, there are three general body types, which he refers to as endomorph, mesomorph and ectomorph. An endomorphic individual typically has stout figure, bigger and thicker frame with wide hips and shoulders, and are easier to gain weight, while an ectomorphic individual is long, slender and thin body frame with narrow shoulders and hips, and is easier to lose weight. An mesomorph person is in the middle, with his/her medium structure and height, and may find it easy to gain muscle and to lose and gain weight.

If the theory of somatotype is true and people can be categorized as ectomorph, mesomorph or ectomorph, this could be potentially useful in many different ways. For example, by accurately predicting one's somatotype at an early age, sports industry could scout and train prospects that have significant significant advantage over other people. Especially in professional sports where innate physical gifts of athletes play a critical role, being able to recruit prospects before other teams can offer advantage over competitors. Endomorphs would be

more likely to succeed in sports of pure strength such as weightlifting or power-lifting, but would be not as successful in gymnastics or marathon. Ectomorphs on the contrary, would be the opposite, having advantages in endurance sports. Mesomorphs can be expected to succeed in any sports.

Naturally, there have been numerous attempts to record the course of growth and study the relationship between variables. The Berkely Guidance study is one of the pioneering works on the subject. The Berkely Gudiance Study was a longitudinal monitoring of 248 infants who were born in Berkeley, California in 1928-1929. The study followed the participants for at least 18 years and recorded periodically their body measures such as height and weight. The participants were mainly caucasian and protestant, and two-thirds were from middle-class families.

However, the theory of somatotype has also been subject to criticism as being subjective and erroneous. Sheldon assessed participants' body type only through visual assessment which left many questioning objectivity of his study. His idea of there being an inherent physique type that is not affected by external factors was also often refuted.

In this project, I will try to study the relationship between somatotype and body frame, to see if somatotype can actually be explained with actual measures of body frame such as weight and height. For this purpose, I will use a sample from the Berkely Guidance Study, which has 136 of the children who participated in the Berekely Guidance Study up until the end of the World War II. The variables in the dataset are:

- Soma - somatotype at age 18 (1 = most slender,..., 7 = least slender)
- HT2, HT9, HT18 - height at age 2, 9, 18 (cm)
- WT2, WT9, WT18 - weight at age 2, 9, 18 (kg)
- ST9, ST18 - strength at age 9, 18
- LG9, LG18 - leg circumference at age 9, 18

2 Descriptive Statistics and Exploratory Data Analysis

Our dataset comprises of 136 datapoints, of which 66 are male and 70 are female. Below is the descriptive statistics of all of the continuous variables.

2.1 Descriptive Statistics of Continuous Variables

Table 1: **Summary Statistics of Continuous Variables**

	All Children (n = 136)					Male (n = 66)					Female (n = 70)				
	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max	Mean	SD	Meadian	Min	Max
WT2	13.21	1.61	13.20	10.1	18.6	13.63	1.63	13.60	10.1	18.6	12.82	1.49	12.70	10.2	17.0
WT9	31.63	5.97	30.90	19.9	66.8	31.63	6.16	31.00	19.9	66.8	31.62	5.82	30.65	22.0	47.4
WT18	64.87	10.67	65.10	42.9	110.2	70.27	9.98	69.35	42.9	110.2	59.78	8.66	58.30	44.1	97.7
HT2	87.80	3.36	87.70	80.9	98.2	88.37	3.32	88.35	81.3	98.2	87.25	3.33	87.10	80.9	97.3
HT9	135.49	5.50	135.70	121.4	152.5	135.89	5.38	135.60	122.0	147.5	135.12	5.61	135.70	121.4	152.5
HT18	172.58	8.84	172.50	153.6	195.1	178.98	6.52	178.90	160.9	195.1	166.54	6.07	166.75	153.6	183.2
LG9	27.68	2.46	27.30	21.8	40.4	27.50	2.48	27.25	21.8	40.4	27.84	2.45	27.45	22.6	32.7
LG18	35.84	2.57	35.75	30.0	44.1	36.29	2.49	36.50	30.0	44.1	35.42	2.58	34.85	30.3	42.9
ST9	64.57	15.45	64.00	22.0	121.0	68.92	14.67	68.00	30.0	121.0	60.46	15.13	59.00	22.0	107.0
ST18	167.13	49.72	150.50	77.0	260.0	212.09	28.59	214.50	145.0	260.0	124.74	17.61	124.50	77.0	182.0

We can see that there are noticeable difference across the variables between male and female, especially for the variables recorded at the age of 18. We will investigate further in the exploratory data analysis section below.

Next is the summary statistics of somatotype.

2.2 Descriptive Statistics of Categorical Variable

Table 2: **Summary Statistics of Categorical Variable**

	All Children (n = 136)		Male (n = 66)		Female (n = 77)	
Somatotype	n	proportion	n	proportion	n	proportion
1	4	0.029	4	0.061	0	0
1.5	7	0.051	7	0.106	0	0
2	13	0.096	13	0.197	0	0
2.5	2	0.015	2	0.03	0	0
3	18	0.132	16	0.242	2	0.029
3.5	6	0.044	3	0.045	3	0.043
4	29	0.213	14	0.212	15	0.214
4.5	12	0.088	0	0	12	0.171
5	21	0.154	0	0	21	0.3
5.5	11	0.081	0	0	11	0.157
6	7	0.051	5	0.076	2	0.029
6.5	3	0.022	0	0	3	0.043
7	3	0.022	2	0.03	1	0.014

Again, we see that males are mostly assigned lower values, while females are generally assigned higher values on somatotype. Let's look at the histogram of somatotype color coded by sex.

2.3 Exploratory Data Analysis

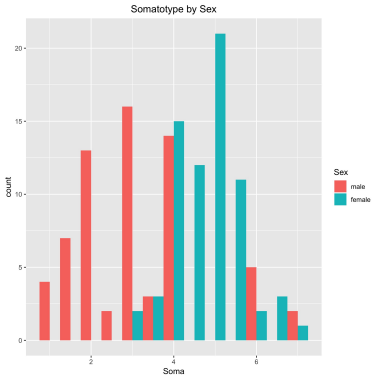

Figure 1: Histogram of Somatotype by Sex

It seems clear that there is different distributions of somatotype for male and female. We will need to introduce sex variable in the model, either by itself or as interaction terms.

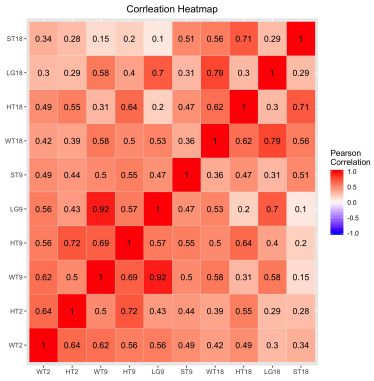

Figure 2: Correlation Heatmap

The heatmap shows that there are high degrees of correlation between variables. Leg circumference and strength in particular are highly correlated with other variables. LG9 has the correlation coefficeint of 0.92 with WT9, and LG18 has 0.79 with WT18. ST18 has the pearson correlation of 0.71 with HT18. Because our goal is to build a explanatory model, introducing high degree of multicollinearity might make it hard to interpret the result. Also, our research question focuses on the relationship between body frame and somatotype, therefore strength is not an appropriate variable to include in our model. Therefore, among the continuous variables, I choose to use only weight and height variables.

Now that we decided on using weight and height variables, we proceed to look into the variables plotted against sex to get an idea of which of the variables might be causing the difference.

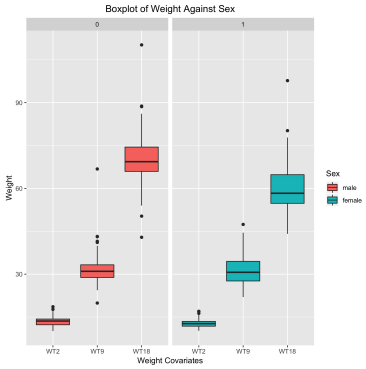

Figure 3: Boxplot of Weight against Sex

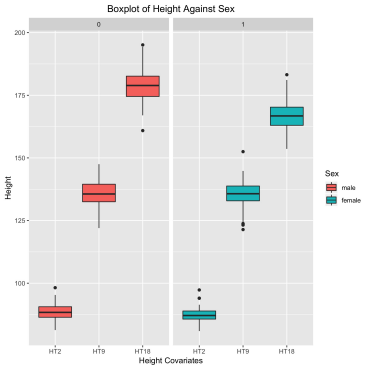

Figure 4: Boxplot of Height against Sex

From the boxplots, we can see that for both variables there is no significant difference between sex at age 2 and 9, but there is noticeable difference at age 18. This conforms to our common knowledge, as the physical difference between male and female become pronounced only after puberty. We would need to introduce interaction terms in our model between the continuous variables at age 18 and sex.

Finally, we will examine the scatterplots between height and weight variables at different time points.

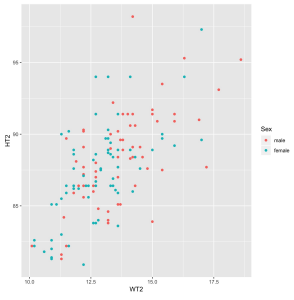

Figure 5: Scatterplot Between WT2 and HT2

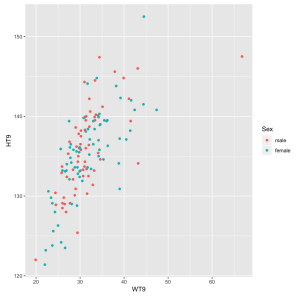

Figure 6: Scatterplot Between WT9 and HT9

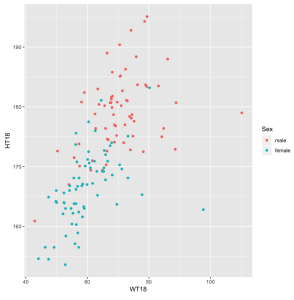

Figure 7: Scatterplot Between WT18 and Ht19

Again, we find similar trend that the difference between male and female are pronounced only at age 18. However, on Figure 8, it seems that there is an outlier, a boy that recorded close to 70kg at age 9. This datapoint sets apart from the others significantly, as other children recorded at most around 45kg at the same age. This may be a measurement error or mistake in recording the data. However, with limited information on the entire data collection process, there is not enough evidence to believe that this is illegitimate data point either. Therefore, we will keep the dataset as is.

3 Model Selection

From our explanatory data analysis from the above section, we have decided that the model would have all 3 weight variables and 3 height variables, of which WT18 and HT18 would be interacting with sex variable. Our model with full set of covariates will look as follows.

$$Somatotype \sim Sex + WT2 + WT9 + WT18 + WT18 \cdot Sex + HT2 + HT9 + HT18 + HT18 \cdot Sex$$

With this model as the base model, I used backward stepwise model selection. For model selection, I used both AIC and BIC then compared the adjusted R-squared of the resulting models from stepwise selection. The resulting models from stepwise model selection using AIC and BIC were as follows.

$$AIC : Somatotype \sim HT2 + HT18 + Sex + WT2 + WT9 + WT18$$

$$BIC : Somatotype \sim HT2 + HT18 + Sex + WT2 + WT9 + WT18 + HT18 \cdot Sex + WT18 \cdot Sex$$

The adjusted R-squared of the AIC selected model was 0.6797 and the BIC selected model was 0.6823. Although it is a slight improvement, we have more variance explaining power with the model selected using BIC. In addition, it also fits more to our exploratory data analysis, because it has interaction term between HT18 and sex, and WT18 and sex. Thus, we use the model selected through stepwise model selection

using BIC as our final model.

3.1 Interpreting the Final Model

Table 3: **Summary Output of the Final Model**

	Estimate	SE	t-value	P-value
Intercept	15.26677	3.17648	4.806	4.30e-06 (***)
HT2	-0.05765	0.03129	-1.843	0.067735 (.)
HT18	-0.07322	0.01840	-3.979	0.000116 (***)
as.factor(Sex)1	-3.88481	3.99918	-0.971	0.333210
WT2	-0.15564	0.06723	-2.315	0.022229 (*)
WT9	0.05695	0.02325	2.450	0.015673 (*)
WT18	0.09041	0.01273	7.102	8.01e-11 (***)
HT18:as.factor(Sex)1	0.03963	0.02542	1.559	0.121538
as.factor(Sex)1:WT18	-0.01991	0.01745	-1.141	0.256099

Significance codes: 0 (***), 0.001 (**), 0.01 (*), 0.05 (.)
Multiple R-squared: 0.07013, Adjusted R-squared: 0.6797
F-statistic: 48.75 on 6 and 129 DF, p-value: ¡ 2.2e-16

The intercept of the model for male is 15.26677 and is statistically significant at $\alpha = 0.05$ significance level. The intercept for female is $15.26677 - 3.88481 = 11.38196$. However, we cannot interpret the intercepts as it would require us to other covariates to 0 and it is impossible to have 0kg of weight or 0cm of height. HT2, WT2 and WT9 all has statistically significant coefficients at $\alpha = 0.0$. Coefficient for HT18 for male is -0.07322 and for female it is $-0.07322 + 0.03963 = -0.03359$. This suggests that if we compare two groups of boys at the age of 18 that are equal for all other covariates, but one group is 1cm taller, the taller group will have -0.07322 lower somatotype value meaning that they are closer to ectomorph than to endomorph. In the same setting, for girls, the taller group will have -0.03359 lower somatotype value. For WT18, the coefficient for male is 0.09041 and for female it is $0.09041 - 0.01991 = 0.0705$. Again, under the same conditions, the group of boys who are 1kg heavier will have 0.09041 higher somatotype value suggesting that they are closer to endomorph than to ectomorph, and for girls with 1kg heavier weight at 18 will have 0.0705 higher somatotype value compared to 1kg lighter girls under equal condition. The F-test of the model is statistically significant at $\alpha = 0.0$ significance level, and the adjusted R-squared suggests that this model explains 67.97% of the variance.

Generally, weight variables have positive coefficients and height variabels have negative coefficients. This suggests that the taller and lighter an individual is, the lower the value of somatotype the individual is likely to have, and shorter and heavier an individual is, the higher the value of somatotype of the individual. Also, for girls the coefficient are loaded so that they have smaller absolute value of coefficient for weight and coefficient for height, suggesting that somatotype of girls tend to be affected less by weight and height than that of boys.

3.2 Residual Analysis

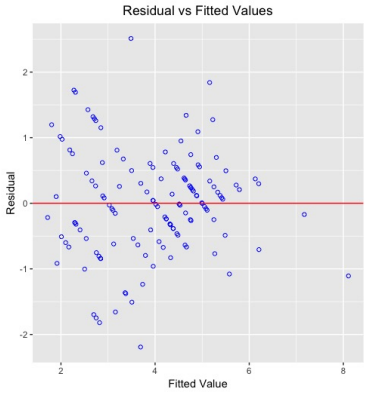

Figure 8: Residual vs. Fitted

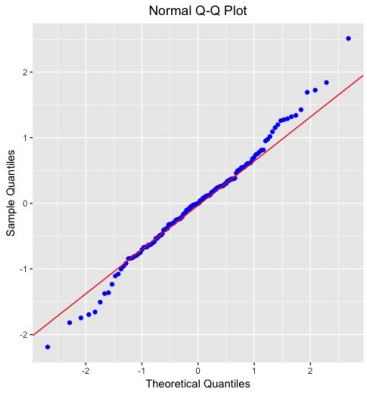

Figure 9: Normal QQ Plot

The residual vs. fitted plot of the left checks if the residuals are homoscedastic. It seems that the residuals are generally homoscedastic in the lower end, but somewhat heteroscedastic in the middle and more so in the higher end. The normal q-q plot on the right checks if the residuals are normally distributed. It seems that the residuals are generally in line with the normal distribution, but on the extreme ends they do not align perfectly.

4 Conclusion

This paper aimed to study the relationship between somatotype and body measures. Through exploratory data analysis, we were able to narrow down the scope on which variables we might want to include in the model, and by using stepwise model selection using BIC, we were able to generate the final model that best explains somatotype using the provided variables. Our model was able to explain close to 68% of the variability in somatotype, explained relationship between somatotype, height and weight, and explained the difference between the somatotype for girls and boys. We found out that the taller an individual is, given all other variables are equal, one is likely to have smaller somatotype value, and the heavier an individual is, given all other variables fixed, one is likely to have bigger somatotype value. We also discovered that for girls, the magnitude of the expected difference in somatotype when there is one unit difference in height and weight is smaller.

However, the normality and homoscedasticity assumptions on the residuals were not perfectly met, whic may be due to small sample size. For future study, it would be preferrable to collect a bigger dataset with more participants, and different body measure variables that are not too highly correlated with weight or height such as wrist and ankle circumference to improve on this model.