Regression Project (Part 2)

Ramya Dhatri Vunikili
April 21, 2017

1 Introduction Section

1. Introduction

After the World War II, one of the most vital new arenas of social science research was the impact of massive social change on people's lives. There were studies that suggested the particular types of child rearing and socialization led to the breeding of rigid and punitive attitudes. Also, few works studied what cultural and psychological factors made persons such as Hitler, Luther, and Gandhi as the leaders of giant movements for social change. Since then, tremendous advances have occurred in the ways in which changing lives over time could be studied. To link earlier and later events in an individual's life or show the evolving structure of the life course, we need valid and reliable reports about people's life histories. During this time, longitudinal projects such as the Terman studies of gifted children, the Grant study of Harvard University men, and the Gluecks' accounts of adolescent delinquents began. And they have now run for decades and demonstrated what valuable insights can be obtained from following respondents over time. Between 1928 and 1931, three remarkable studies were started by a diverse group of researchers at the University of California, Berkeley with similar objectives. These were the Oakland Growth Study, the Berkeley Guidance Study and the Berkeley Growth Study.

2. Berkeley Guidance Study

The Guidance Study (GS), was begun by Jean Walker Macfarlane in January 1928. Planned as a 6-year prospective study of a normal sample, the original purposes were to assess
(a) how prevalent and severe were behavior problems of the kind reported for preschool children brought to therapeutic clinics,
(b) biological and environmental factors associated with the presence or absence of such behaviors, and

(c) the influence of intensive discussions with parents about child-rearing practices on children's problem behavior
By the end of the 6 years, more general questions about personality development were intriguing psychologists and psychiatrists, and so the GS continued but with the intent of examining the interactions of psychological, social, and biological factors in personality development.

The 248 original participants were drawn from a socioeconomic survey of every third birth in Berkeley between January 1, 1928 and June 30, 1929. And by the age of 18 years, 150 study members still were being seen. This study developed into an examination of interactions of psychological, social, and biological factors in personality development, with the mothers, children, and spouses of the respondents being included.

The detailed demographic and socioeconomic data on the family at the time of the child's birth were obtained from the parents by an interviewer trained in economics and social work. Pre and perinatal data were obtained by a public health nurse from mothers, physicians, and hospital records. From the infant's third through eighteenth month, a public health nurse also visited the home every 3 months to measure height and weight and make systematic records of progress, including health, diet and behavior. Both parents filled out detailed health histories for themselves and their parents (the infant's grandparents). Because one purpose of the research was to assess the effect of parental "guidance" by professional staff, the 248 infants selected for the GS were assigned, at 21 months of age, to one of two subsamples - a Guidance or a Control group, each with 124 infants. These two subgroups were matched for sex of child, size of family, family income at birth of the study child, occupation of father, neighborhood, and age, education, nativity, and ethnic derivation of the parents. The groups did not differ in condition at birth nor in developmental status and number of behavior problems at 21 months.

When the study members were 21 months old, intensive data collection at the Institute of Human Development (IHD) began. They were then assessed every 6 months from 2 to 4 years, and annually from 5 to 18 years. At each visit their mothers accompanied them and were interviewed. Medical examinations, health histories, and a small battery of anthropometric measurements were obtained at 21 months and annually from 3 through 18 years. From 8 to 18, anthropometric measures, strength measures, body photographs and hand-wrist X-rays for assessing skeletal maturity were taken semi annually. Alternate forms of the California Preschool Schedule, an intelligence test standardized in Berkeley, were administered at 21 months, semi annually through 4 years, and again at 5. The 1916 Stanford-Binet was given at 6 and 7, alternate forms of the 1937 revision at 8, 9, 10, 12 or 13 and 14 years, and the Wechsler Bellevue Intelligence Scale (WBIS), Form I, at 18.

3. SOMATOTYPE

William H. Sheldon, introduced the concept of body types, or somatotypes, in the 1940s and used it in explaining different types of criminal behavior. The gist is that everyone falls, though not altogether neatly, into the three categories. People are born with an inherited body type based on skeletal frame and body composition. Most people are unique combinations of the

three body types: ectomorph, mesomorph, and endomorph. Sheldon's research implied that mesomorphic individuals were more prone to committing violent and aggressive acts.

Ectomorphs are long and lean, with little body fat, and little muscle. They have a hard time gaining weight. Fashion models and basketball players fit this category. Endomorphs, on the other hand, have lots of body fat, lots of muscle, and gain weight easily. Mesomorphs are athletic, solid, and strong.

Sheldon evaluated the degree a body type was present on a 1 to 7 scale where 1 is the minimum and 7 is the maximum. All athletes are made up of the three extreme body types so we are all part endomorph, part mesomorph and part ectomorph. Using a score of one to seven, we can grade our bodies on each of the extreme body types. e.g. two, three, six means: two (low endomorphy); three (low ectomorphy); and six (high mesomorphy). In this way, we can compare our body type with that of the athletes.

A new method of somatotyping can be done using the Body Mass Index (BMI). This is proven to be a better method than the traditional somatotyping and also it takes height into consideration while evaluation.

4. Research Questions

1) If Somatotype is considered to be high in people with high mesomorphy, does it have a strong positive relationship with strength parameters (ST2, ST9 and ST18)?
2) As the physical structure of humans evolves over time, should Somatotype be more accurately determined by parameters at older ages rather than at age 2?
3) As height is not a prominent parameter in determining Somatotype, do the HT2, HT9 and HT18 have a weak relationship with Somatotype?
4) Do children of both the gender exhibit same trend of Somatotype at all the three ages?

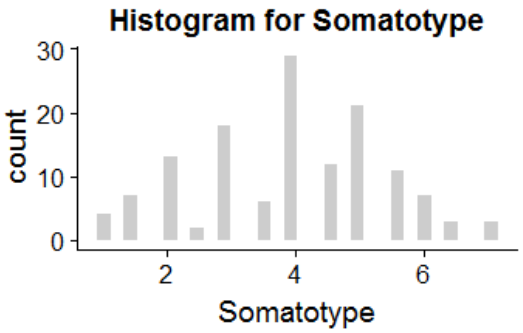
2 DESCRIPTIVE SECTION

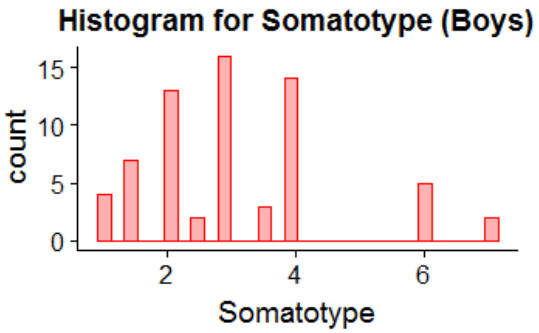
Table 2.1: Descriptive Statistics Of The Continuous Regressors

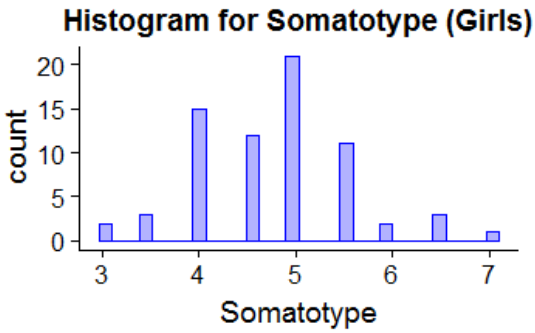
Statistic	Mean	St. Dev.	Median	Min	Max
WT2	13.21	1.61	13.20	10.10	18.60
HT2	87.80	3.36	87.70	80.90	98.20
WT9	31.63	5.97	30.90	19.90	66.80
HT9	135.49	5.50	135.70	121.40	152.50
LG9	27.68	2.46	27.30	21.80	40.40
ST9	64.57	15.45	64	22	121
WT18	64.87	10.67	65.10	42.90	110.20
HT18	172.58	8.84	172.50	153.60	195.10
LG18	35.84	2.57	35.75	30.00	44.10
ST18	167.13	49.72	150.5	77	260
Soma	3.96	1.44	4.00	1.00	7.00

Table 2.2: Descriptive Statistics Of The Categorical Regressors

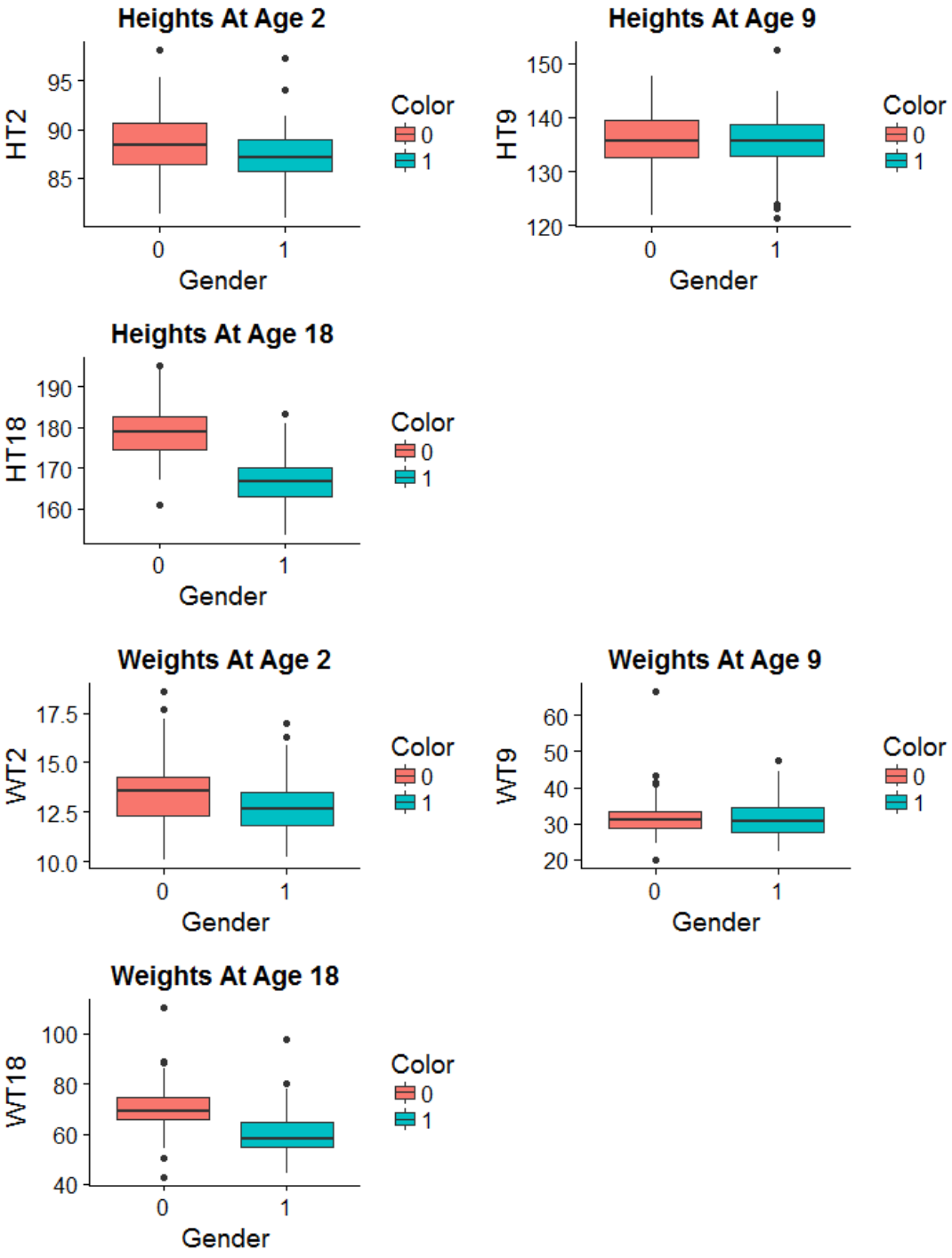
Statistic	No. Of Individuals	Sample Proportion
Boys	66	66/136 = 0.485
Girls	70	70/136 = 0.515



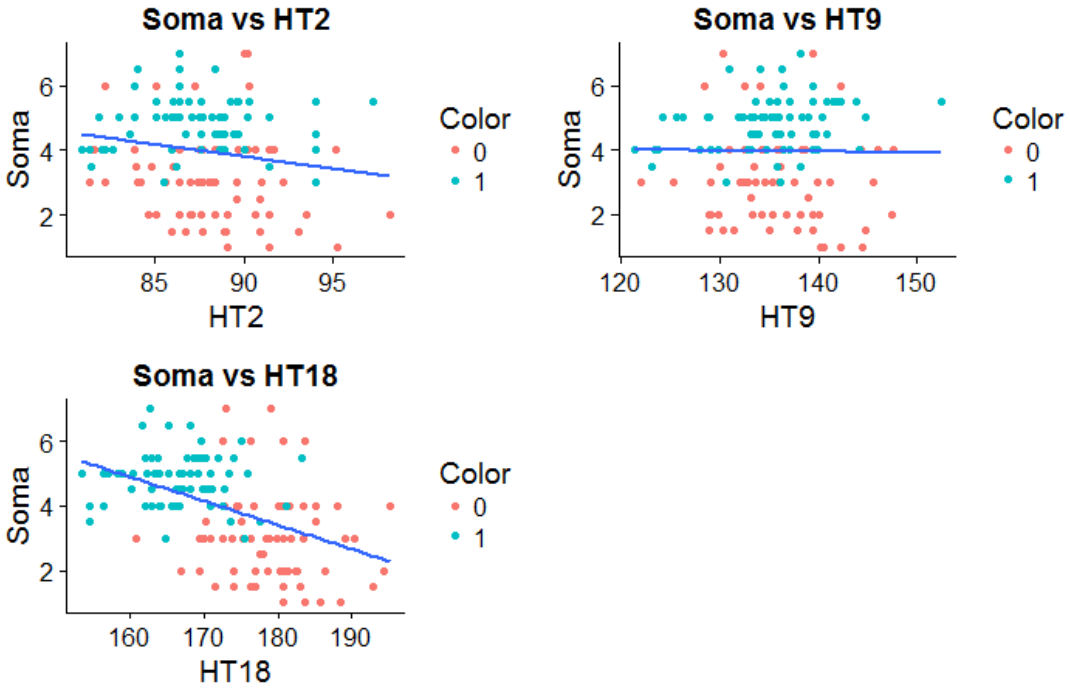




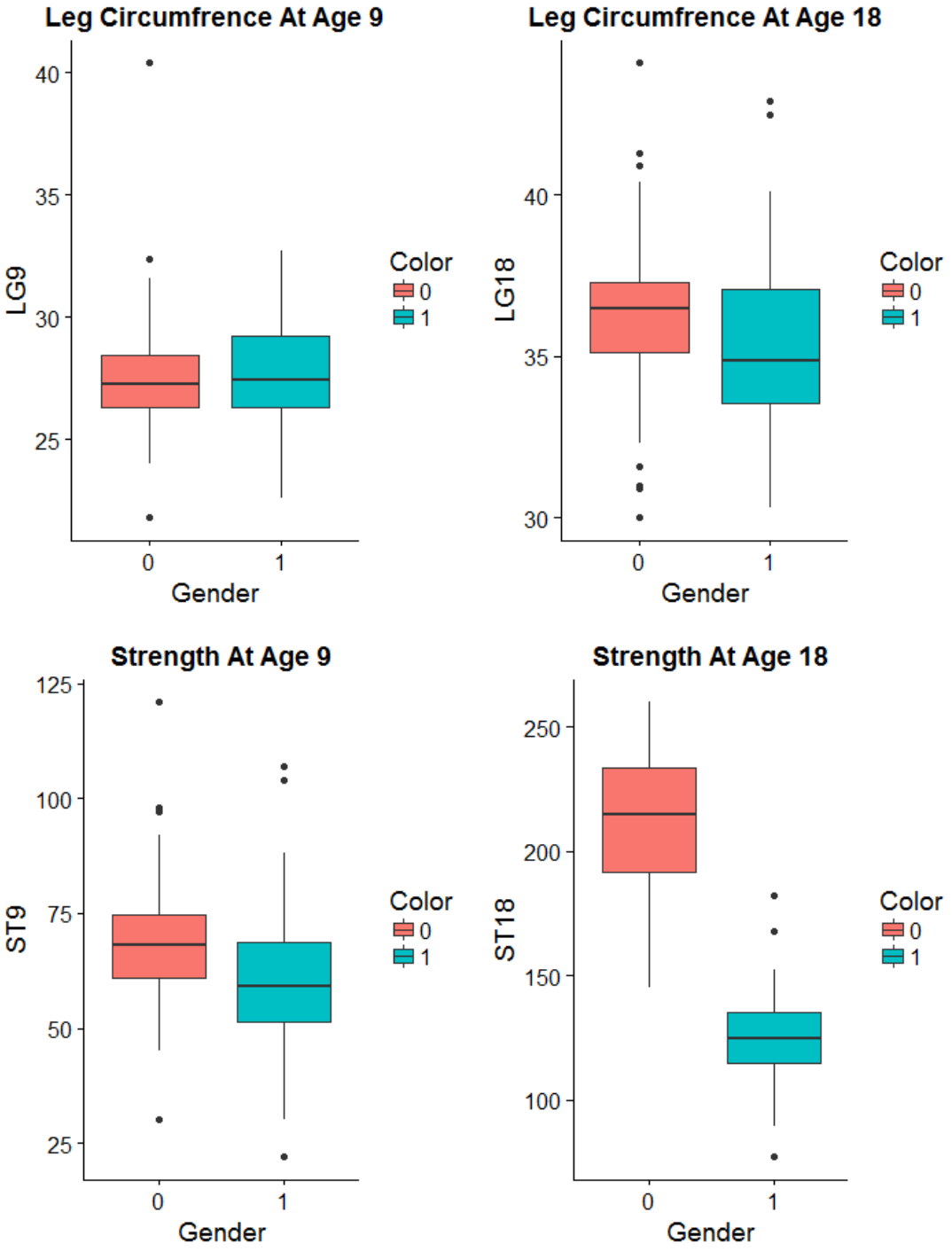
The Soma for boys varies from 1 to 7 with a median of 3. According to Sheldon's theory, this implies that boys are less prone to criminal activities as most of them have a Somatotype < 4. Whereas, girls have a median Soma of 5 with a few them having a Soma of even 7. Thus, sex could be an important predictor in analyzing the aggressive behavior of people. However, the fact that girls outnumber boys (70 vs 66) has to be taken into consideration while generalizing the point that few girls exhibit higher levels of Somatotype when compared to boys. Also, the number of boys and girls having a Soma >= 6 is fairly equal (7 vs 6).



Though there is no much difference between the mean height and weight for boys and girls at ages 2 and 9, there is a considerable difference between their heights and weights at age 18. The third quartile height and weight of girls at age 18 is lesser than that of the first quartile of boys.



The regression of Soma on height at different ages is shown in the above plots. There is a very weak negative relationship between the Soma and height at age 2 and 18. On the other hand, the slope of the regression line of Soma on HT9 is almost 0. It is interesting to note that there is no change in Somatotype even with the change in height in children at the age of 9. Also, one more point to be noted is that most of the girls and boys lie on either side of the regression line. This supports our earlier result of separate intercepts for boys and girls. On the whole, height can be regarded as NOT such a prominent factor in determining Soma.



Looking at the box plots of leg circumference and strength of boys and girls at different ages, there seems to be a predominant difference between ST18 of boys and girls.

3 Explanatory Section

In order to choose the best model, I've chosen to perform forward selection and backward selection to obtain AIC, BIC and C_p statistics. They have been tabulated below:

Table 3.1: Model Selection By Forward Elimination

Model	Selection Procedure	Statistic
M1	AIC	AIC=-87.36
M2	BIC	BIC=-57.7
M3	C_p	C_p=-88.61

Forward AIC Model (M1):

$$Soma \sim as.factor(Sex) + LIN + HT18 + ST18 + DW9 + HT2 + ST18 : Sex + LIN : Sex + DW9 : Sex + HT2 : Sex$$

Forward BIC Model (M2):

$$Soma \sim as.factor(Sex) + LIN + HT18 + ST18 + DW9 + ST18 : Sex + LIN : Sex$$

Forward C_p Model (M3):

$$Soma \sim as.factor(Sex) + LIN + HT18 + ST18 + ST18 : Sex$$

Table 3.2: Model Selection By Backward Elimination

Model	Selection Procedure	Statistic
M4	AIC	AIC=-88.69
M5	BIC	BIC=-58.15
M6	C_p	C_p=9.83

Backward AIC Model (M4):

$$Soma \sim WT2 + WT9 + HT9 + ST9 + WT18 + HT18 + LG18 + ST18 + WT2 : Sex + HT9 : Sex + ST9 : Sex + HT18 : Sex + LG18 : Sex$$

Backward BIC Model (M5):

$$Soma \sim WT2 + WT9 + HT9 + ST9 + WT18 + LG18 + ST18 + WT2 : Sex + ST9 : Sex + LG18 : Sex$$

Backward C_p Model (M6):

$$Soma \sim WT2 + WT9 + HT9 + ST9 + WT18 + HT18 + LG18 + ST18 + WT2 : Sex + HT9 : Sex + ST9 : Sex + HT18 : Sex + LG18 : Sex$$

2. Best Subset Selection

By looking at the statistics for both the elimination procedures, it can be observed that backward elimination provides better values than forward elimination. This implies that models with weights at different ages are better than those with AVE, LIN, QUAD and the change in weights. Hence, models M4, M5 and M6 are of our interest. Also, M4 and M6 given by AIC and C_p are the same models. The C_p value of M5 is 9.83 where the number of predictors in the model is 8. Hence, $C_p = 1 + p \approx 9$.

I've chosen to go with model M5 (BIC) as it is more parsimonious than the models M4/M6. Also, in support of one of the earlier results, the models M4, M5 and M6 have WT9 as a predictor but not LG9 as these are highly correlated with a correlation coefficient of 92%.

Reasons To Choose M5 As The Best Model:
1) It is parsimonious than the models M4/M6.
2) The Somatotype in model M5 is determined less by the height terms which is in accordance with Sheldon's study.
3) P Values:
When a regression of Soma is run on the full model the t values of WT18, WT9, HT9:Sex, LG18 and ST18:Sex are the highest. However, the p values of ST18:Sex and HT9:Sex are over 0.21 and model M5 has correctly dropped them.
Comparing the p values for models M5 and M4, it can be noted that most of the predictors in M5 have a p-value <0.05 except LG18 and ST9. Whereas the p values of many predictors exceed 0.05 with one of them even being 0.8. This suggests that M5 is a better model over M4.

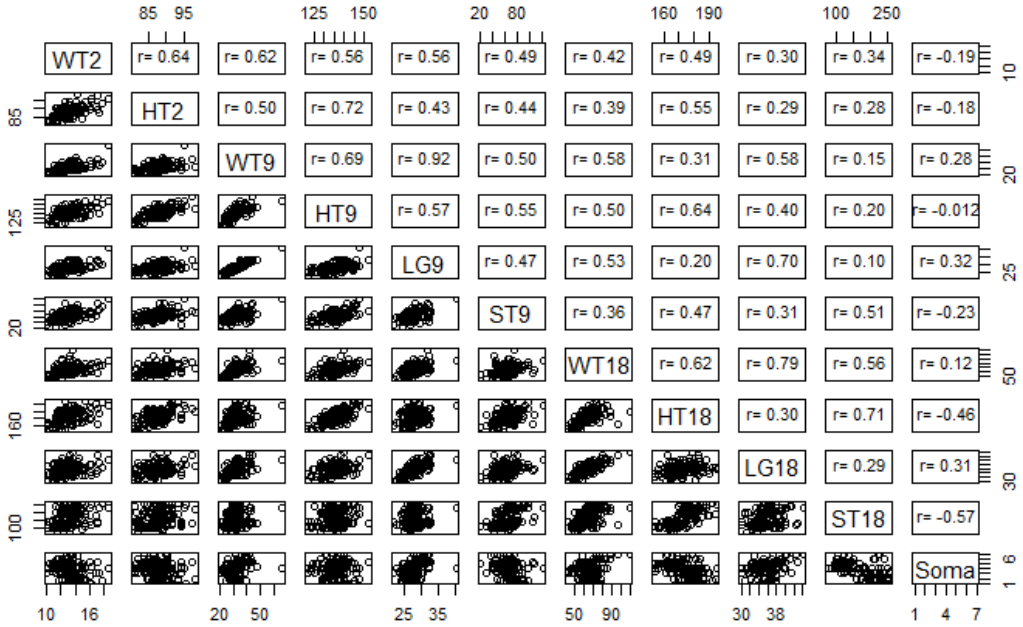
Table 3.3: Statistics of Model M5 (BIC) - Best Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.9252	2.3294	5.12	0.0000
WT2	-0.3128	0.0642	-4.87	0.0000
WT9	0.0856	0.0186	4.61	0.0000
HT9	-0.0877	0.0171	-5.12	0.0000
ST9	-0.0154	0.0082	-1.87	0.0636
WT18	0.0804	0.0137	5.88	0.0000
LG18	0.0806	0.0498	1.62	0.1079
ST18	-0.0136	0.0033	-4.10	0.0001
WT2:Sex	0.2525	0.0847	2.98	0.0034
ST9:Sex	0.0297	0.0095	3.13	0.0022
LG18:Sex	-0.1149	0.0294	-3.91	0.0001

Table 3.4: Statistic of Model M4 (AIC)

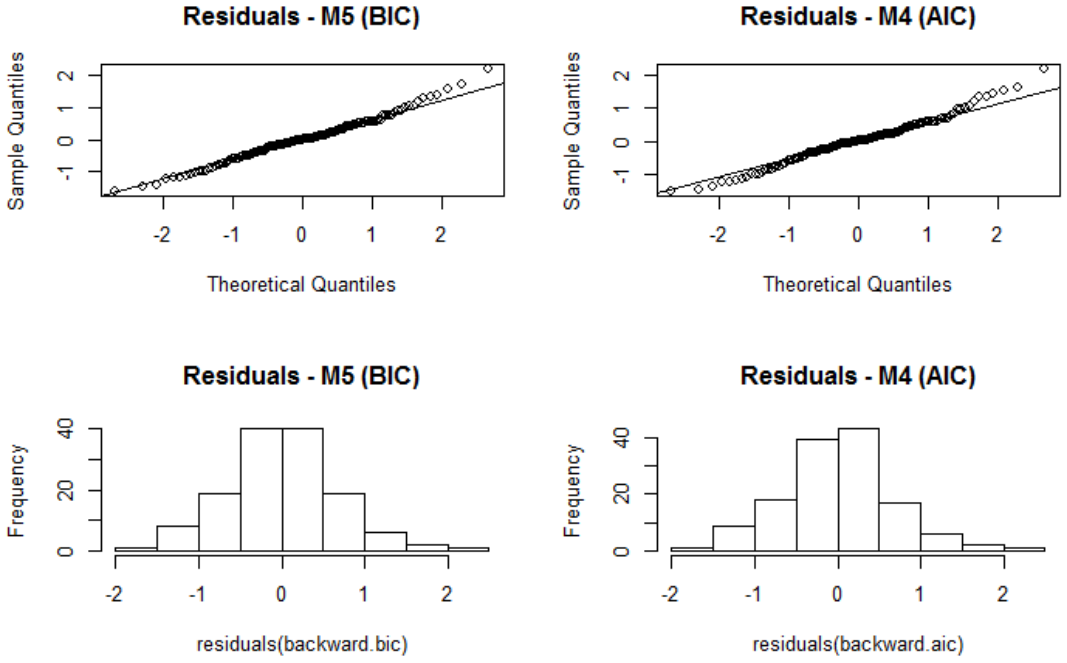
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.0644	2.4300	4.96	0.0000
WT2	-0.2882	0.0664	-4.34	0.0000
WT9	0.0781	0.0202	3.87	0.0002
HT9	-0.1038	0.0386	-2.69	0.0081
ST9	-0.0135	0.0082	-1.64	0.1041
WT18	0.0812	0.0143	5.67	0.0000
HT18	0.0067	0.0277	0.24	0.8108
LG18	0.0970	0.0574	1.69	0.0934
ST18	-0.0136	0.0033	-4.12	0.0001
WT2:Sex	0.2021	0.0951	2.13	0.0356
HT9:Sex	0.0810	0.0490	1.65	0.1006
ST9:Sex	0.0229	0.0100	2.28	0.0243
HT18:Sex	-0.0503	0.0338	-1.49	0.1397
LG18:Sex	-0.1541	0.0532	-2.89	0.0045

The p-value of predictor LG18 in M5 is 0.11, far greater than 0.05, can be attributed to the fact that there is a high correlation of LG18 with WT18. This can be observed from the scatter plot matrix below. Also, ST9 is also having a noticeable correlation with HT9. Hence, the p-values of LG18 and ST9 are greater than 0.05 and those of WT18 and HT9 are 0.

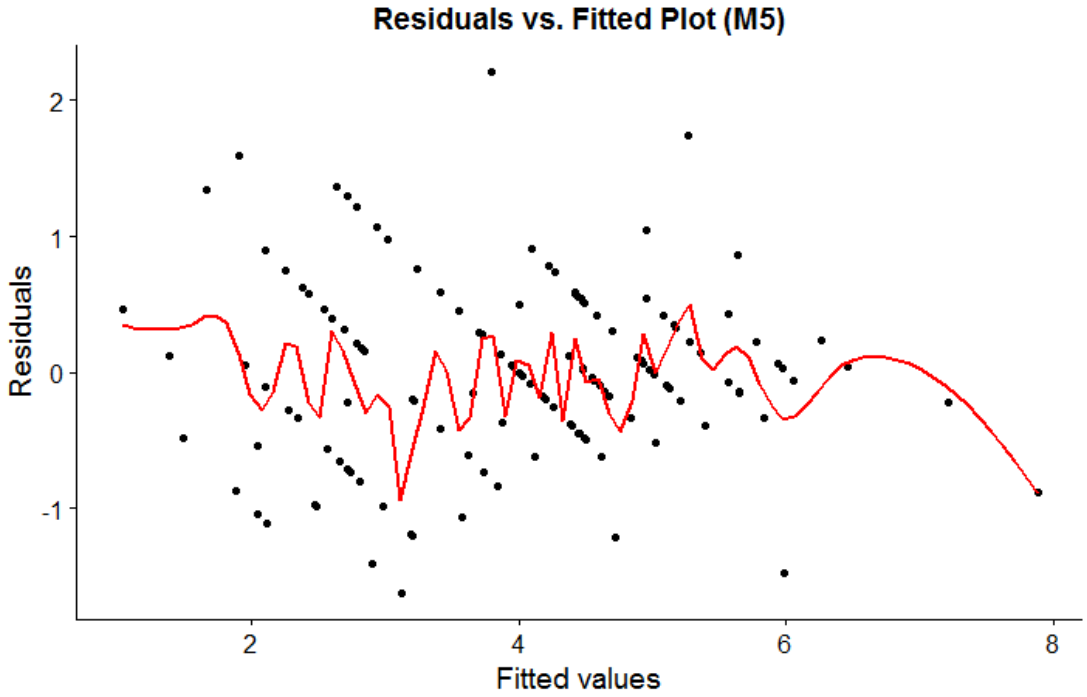


4) Normality:

The lower part of the residuals in M5 tend to be more normal than those in M4. The histogram of residuals of model M5 appears to be almost a perfect one.



5) Homoscedasticity:
The residual plot for M5 shows that residuals are lower for higher fitted values and higher for lower fitted values. The model exhibits collinearity as discussed above because of correlation between LG18 and WT18.



3. Conclusion

1. Research Questions

1) If Somatotype is considered to be high in people with high mesomorphy, does it have a strong positive relationship with strength parameters (ST2, ST9 and ST18)?
No, Somatotype shows a weak correlation between ST18 and ST9. This leads us to think about the relation between performance and Soma rather than physique itself.

2) As the physical structure of humans evolves over time, should Somatotype be more accurately determined by parameters at older ages rather than at age 2?
Yes, Somatotype is predominantly determined by the weight, height and strength parameters at ages 9 and 18. The best model selected has only the weight parameter considered at age 2.

3) As height is not a prominent parameter in determining Somatotype, do the HT2, HT9 and HT18 have a weak relationship with Somatotype?
Yes. The Somatotype is less dependent on the height predictor when compared to the weight predictor. Thus, best model is supported by Sheldon's study.

4) Do children of both gender exhibit same trend of Somatotype at all the three ages?
Somatotype does not directly depend on the gender as previously shown using histograms. Instead, interactions of weight, strength and leg circumference with gender are noteworthy.

2. Limitations

1) Though the Berkeley Guidance Study has started with 248 infants the sample data provided for the study consists of only 136 observations. There is a possibility that the remaining infants have grown up to exhibit characteristics far different from the existing ones. In which case, the outliers for this model might make more sense and could have plausibly led to the selection of altogether a different model.

2) There is not enough data provided to study the behavioral traits of the children. Though the relation of Soma with different physical traits could be studied, it could not be extended to validate Sheldon's study that the children with high Soma indeed have extreme behavioral attributes.

3. Conclusion

The best model has been obtained by the BIC criteria which is parsimonious than the AIC. The model selection could also be done without the high leverage point. But this could be an influential observation representing the behavior of remaining infants from the initially collected 248 records. Somatotype is highly determined by the weight parameters rather than the height parameters which is in accordance with Sheldon's study. However, Somatotype alone cannot be an efficient predictor for judging the extreme behavior of children. Recent

studies have shown that body mass index (BMI) along with the mental traits like IQ would lead to a better conclusion on extreme behavior in children.