Explaining Somatype

Benjamin Jakubowski

April 16, 2016

1 Introduction

In this report, we build the best explanatory model for somatotype using data from the Berkley Guidance Study. As noted in the assignment introduction,

> *The Berkeley Guidance Study, under the direction of Jean Macfarlane, started with a sample of infants who were born in Berkeley, California in 1928-1929. Most of the children were Caucasian and Protestant, and two-thirds came from middle-class families. The basic cohort includes 136 of these children who participated in the study through the 1930s and up to the end of World War II. Annual data collection ended in 1946.*

Our objective in this analysis is to use the features collected during this study to predict participants' somatypes. Features available include

- **Sex**: 0 = males, 1 = females
- **WT2**: Age 2 weight (kg)
- **HT2**: Age 2 height (cm)
- **WT9**: Age 9 weight (kg)
- **HT9**: Age 9 height (cm)
- **LG9**: Age 9 leg circumference (cm)
- **ST9**: Age 9 strength (kg)
- **WT18**: Age 18 weight (kg)
- **HT18**: Age 18 height (cm)
- **LG18**: Age 18 leg circumference (cm)
- **ST18**: Age 18 strength (kg)
- **Soma**: Somatotype, a 1 to 7 scale of body type.

2 Exploratory Analysis

2.1 Tabular summaries

First, we present tabular summaries of the features in this dataset. We first present tabular summaries for continous features for the entire dataset (Table 1), then we present summaries for continuous features disaggregated by gender (Tables 2 and 3). Finally, we present summaries for our discrete features (Tables 4 and 5):

Table 1: Summary of continuous features- both genders

	WT2	HT2	WT9	HT9	LG9	ST9	WT18	HT18	LG18	ST18
Sample mean	13.21	87.80	31.63	135.49	27.68	64.57	64.87	172.58	35.84	167.13
Sample SD	1.61	3.36	5.97	5.50	2.46	15.45	10.67	8.84	2.57	49.72
Sample median	13.20	87.70	30.90	135.70	27.30	64.00	65.10	172.50	35.75	150.50
Sample min	10.10	80.90	19.90	121.40	21.80	22.00	42.90	153.60	30.00	77.00
Sample max	18.60	98.20	66.80	152.50	40.40	121.00	110.20	195.10	44.10	260.00

Table 2: Summary of continuous features- boys

	WT2	HT2	WT9	HT9	LG9	ST9	WT18	HT18	LG18	ST18
Sample mean	13.63	88.37	31.63	135.89	27.50	68.92	70.27	178.98	36.29	212.09
Sample SD	1.63	3.32	6.16	5.38	2.48	14.67	9.98	6.52	2.49	28.59
Sample median	13.60	88.35	31.00	135.60	27.25	68.00	69.35	178.90	36.50	214.50
Sample min	10.10	81.30	19.90	122.00	21.80	30.00	42.90	160.90	30.00	145.00
Sample max	18.60	98.20	66.80	147.50	40.40	121.00	110.20	195.10	44.10	260.00

Table 3: Summary of continuous features- girls

	WT2	HT2	WT9	HT9	LG9	ST9	WT18	HT18	LG18	ST18
Sample mean	12.82	87.25	31.62	135.12	27.84	60.46	59.78	166.54	35.42	124.74
Sample SD	1.49	3.33	5.82	5.61	2.45	15.13	8.66	6.07	2.58	17.61
Sample median	12.70	87.10	30.65	135.70	27.45	59.00	58.30	166.75	34.85	124.50
Sample min	10.20	80.90	22.00	121.40	22.60	22.00	44.10	153.60	30.30	77.00
Sample max	17.00	97.30	47.40	152.50	32.70	107.00	97.70	183.20	42.90	182.00

Table 4: Summary of soma distribution- both genders

Soma	Count- girls	Sample proportion- girls	Count- boys	Sample proportion- boys
1.0	0	0.00	4	0.06
1.5	0	0.00	7	0.11
2.0	0	0.00	13	0.20
2.5	0	0.00	2	0.03
3.0	2	0.03	16	0.24
3.5	3	0.04	3	0.05
4.0	15	0.21	14	0.21
4.5	12	0.17	0	0.00
5.0	21	0.30	0	0.00
5.5	11	0.16	0	0.00
6.0	2	0.03	5	0.08
6.5	3	0.04	0	0.00
7.0	1	0.01	2	0.03

Table 5: Summary of sex distribution

	Count	Sample proportion
0	66	0.49
1	70	0.51

2.2 Figures for exploratory analysis

Next, we present figures to support further exploration of the data. We present two figures:

- **Figure 1**: This figure summarizes the conditional distributions of each numeric feature (conditioning on sex) using boxplots.
- **Figure 2**: This figure summarizes the pairwise distributions of somatype and numeric feature (coloring by sex) using scatter plots.

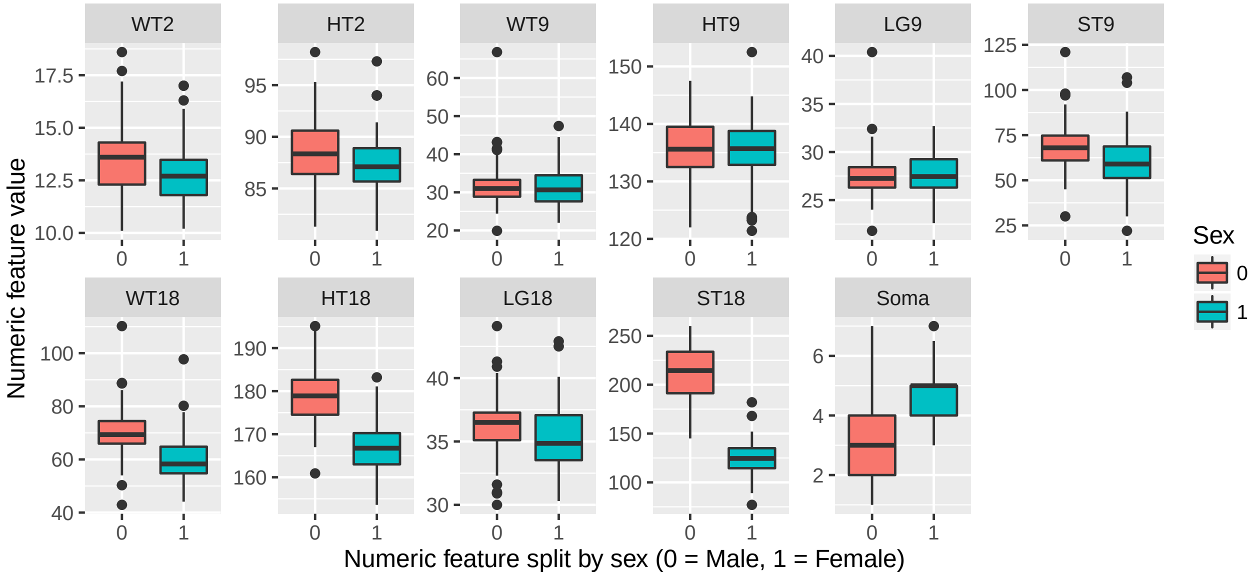

Figure 1: Comparing numeric feature distributions by sex (0 = Male, 1 = Female)

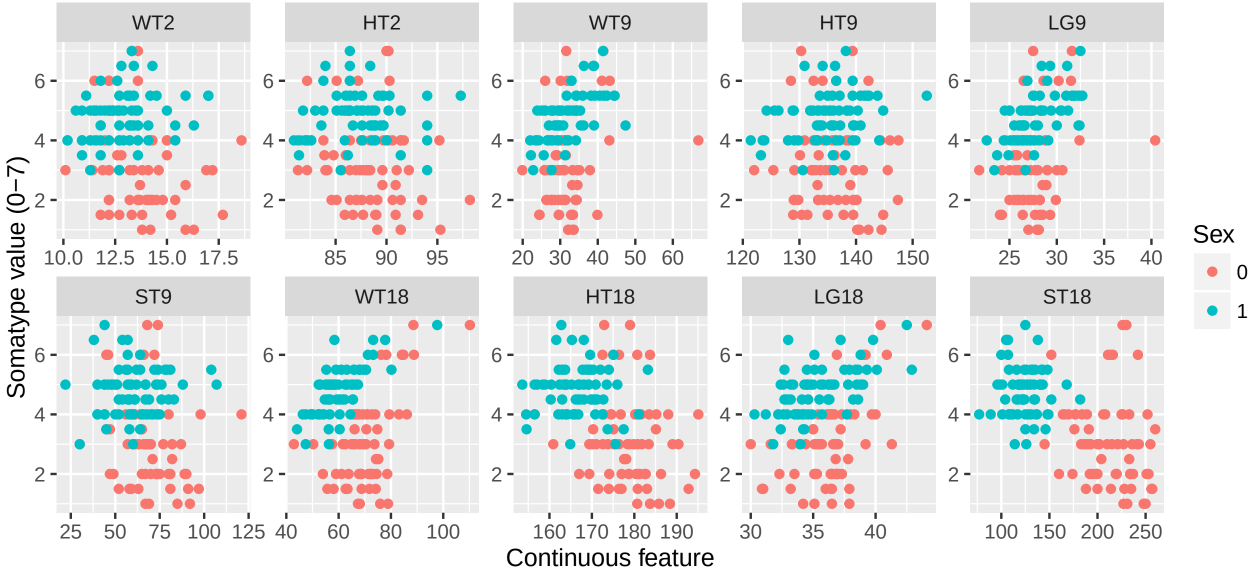

Figure 2: Somatype value vs numeric feature value by sex (0 = Male, 1 = Female)

We can interpret these plots as follows:

- **Figure 1**: Looking at this figure, it is apparent a number of features differ significantly based on sex. Most significantly, it appears HT18 (height at age 18), ST18 (strength at age 18), and Soma (somatype) all have substantially different distributions conditioning on sex.
- **Figure 2**: In this figure we can see the pairwise relationships between each predictor and the target variable (Soma), split by sex. It is apparent from these plots that these pairwise distributions are signifcantly different between boys and girls.

Overall, based on these figures, it is apparent the conditional distributions $F_{Soma,Feature|Sex}$ differ dramatically for boys and girls. Thus, when modeling we are justified in including all the $Sex : Feature$ interaction terms.

3 Modeling and Inference

Next, we proceed to build an explantory model. Based on the assignment instructions (and our exploratory analysis), we constrain candidate models (i.e. our hypothesis space) to *include additive relationships among any of the variables, as well as any two-way interaction term with gender.* To chose our model from this hypothesis space, we will use the following procedure:

1. We will use lasso (ℓ_1 regularized linear regression), since ℓ_1 regularization promotes sparse coefficient vectors and thus achieves implicit feature selection. Note this model building methodology is preferred to stepwise methods, which are unstable (i.e. greedy) algorithms[1].
2. Recall the general lasso objective is $\min_{\beta\in\mathbb{R}^p}\left[\frac{1}{N}||y - X\beta||_2^2 + \lambda||\beta||_1\right]$. Thus, we need to optimize the regularization hyperparameter λ. To do so, we will use 10-fold cross-validation.
3. For our final model, we will select the model corresponding to the highest regularization (largest λ) such that the mean squared error is within one standard error of the objective minimum. This is shown by the right-most dotted line in Figure 3.
4. We will use the glmnet package [2] to fit the model. This package fits generalized linear models via penalized maximum likelihood.

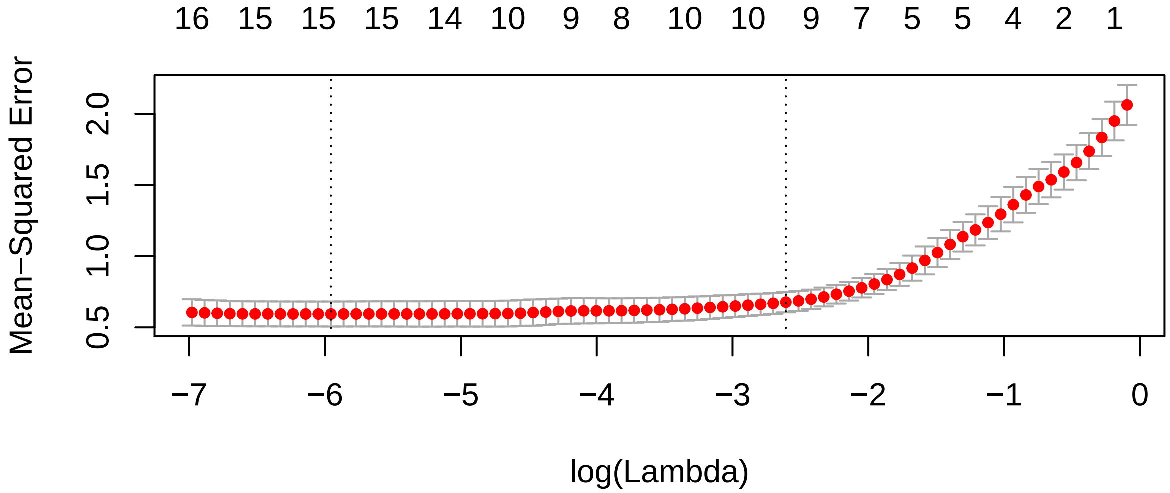

Figure 3: Using cross validation to optimize regularization

The coefficients from the regularized regression model are shown in Table 6.

[1] For justification of this choice, see http://andrewgelman.com/2014/06/02/hate-stepwise-regression/

[2] See https://web.stanford.edu/˜hastie/glmnet/glmnet_alpha.html

Table 6: Non-zero coefficients of final lasso model

(Intercept)	WT2	HT2	WT9	WT18	HT18	LG18	ST18	Sex:WT9	Sex:ST9
9.812	-0.074	-0.027	0.019	0.072	-0.041	0.026	-0.012	0.017	0.003

Next, note we can't retrieve standard errors, t-statistics, or p-values for individual coefficients from our lasso model. To get standard errors (or otherwise estimate uncertainty in our coefficient estimates), we will bootstrap. Specifically, we will

- Generate 500 bootstrap samples
- For each sample, we'll fit the lasso model using our optimal $\lambda_{1SE} = 0.0738$, and determine the model coefficients for the features selected above.
- Finally, we'll determine mean values for these coefficients, as well as the 2.5% and 97.5% quantile of the bootstraped distributions (presented in figure 4 and table 7).

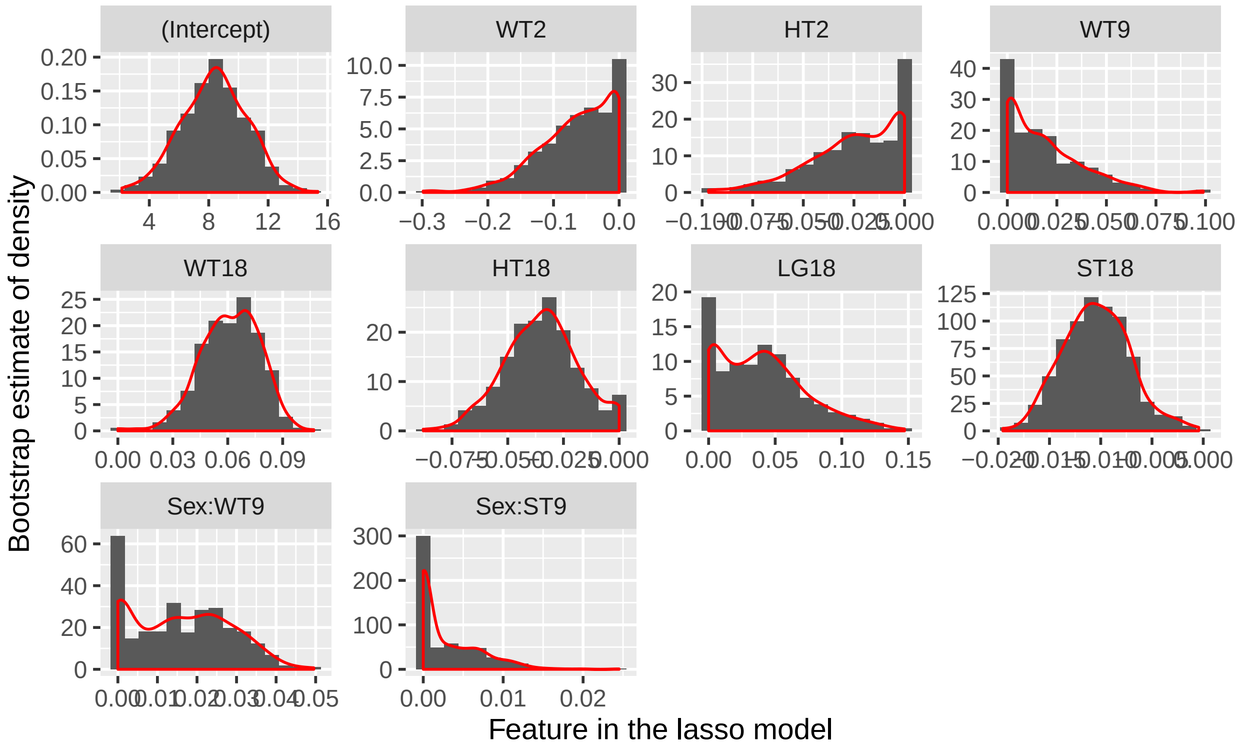

Figure 4: Bootstrap sample estimates of l1 regularized coefficients

Note from these tables it appears a number of features are 'non-significant', in the sense of having 0 weight in a significant proportion of the bootstrapped models. However, we retain these features based on the criterion

Table 7: Bootstrap estimate of the lasso coefficents distribution

	(Intercept)	WT2	HT2	WT9	WT18	HT18	LG18	ST18	Sex:WT9	Sex:ST9
2.5% Quant.	4.009	-0.193	-0.075	0.000	0.028	-0.067	0.000	-0.016	0.000	0.000
97.5% Quant.	12.470	0.000	0.000	0.064	0.088	0.000	0.116	-0.003	0.038	0.012
Mean	8.364	-0.064	-0.024	0.018	0.061	-0.035	0.040	-0.010	0.016	0.003

that this feature set was found to be optimal (i.e. the most regularized model within 1SE of the minimum) through cross validation.

Next, we present residual diagnostics:

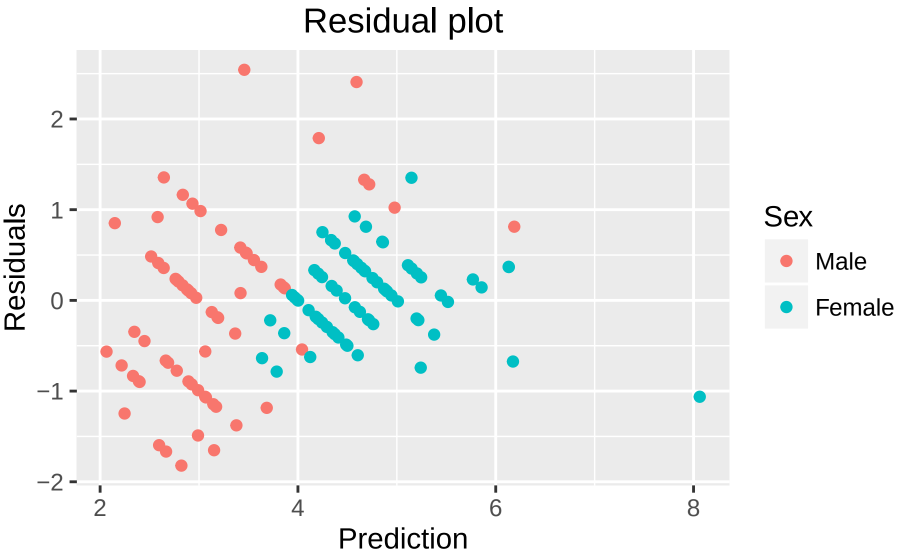

Figure 5: Residual plot for predictive model

This residual plot shows the residuals are fairly evenly distributed around 0 (regardless of predicted value). We don't see constant variance, but again that is not unexpected or unexplained (since it is apparent the variance in residuals differs between boys and girls).

4 Conclusion

In this report, we built a model to predict somatype given the following features:

- **Sex**: 0 = males, 1 = females
- **WT2**: Age 2 weight (kg)
- **HT2**: Age 2 height (cm)
- **WT9**: Age 9 weight (kg)
- **HT9**: Age 9 height (cm)
- **LG9**: Age 9 leg circumference (cm)
- **ST9**: Age 9 strength (kg)
- **WT18**: Age 18 weight (kg)
- **HT18**: Age 18 height (cm)
- **LG18**: Age 18 leg circumference (cm)
- **ST18**: Age 18 strength (kg)
- **Soma**: Somatotype, a 1 to 7 scale of body type.

Recall in our exploratory analysis, we

- Noted the significant differences in the distributions of our predictors conditioned on sex (see Figure 1)
- Noted that the pairwise distribution of our target and predictors were also significantly different between boys and girls (see Figure 2)

Thus, we chose to fit a model using the avialable features plus all interactions between sex and the continuous predictors. To fit an optimal model over this relatively large feature set (20 features on only 136 observations), we needed to use some form of feature selection. Instead of using a stepwise feature selection procedure, we chose to use lasso (which is a more robust method for implicit feature selection).

Using cross validation to select λ, we ultimately fit a model with 10 features (including the intercept). Again, the coefficients are given in table 6.

Unfortunately, the relatively large number of features included in this model make interpretation somewhat difficult. However, this is our best explanatory model for the factors that may be associated with somatotype (in the sense of chosing the greatest regularization without significantly sacrificing predictive performance on cross-validation hold-out sets).