Identified the Best Predictors Of Systolic Blood Pressure Through Investigating The Impact Of Different Genes And Environmental Factors

Maraym Gholami

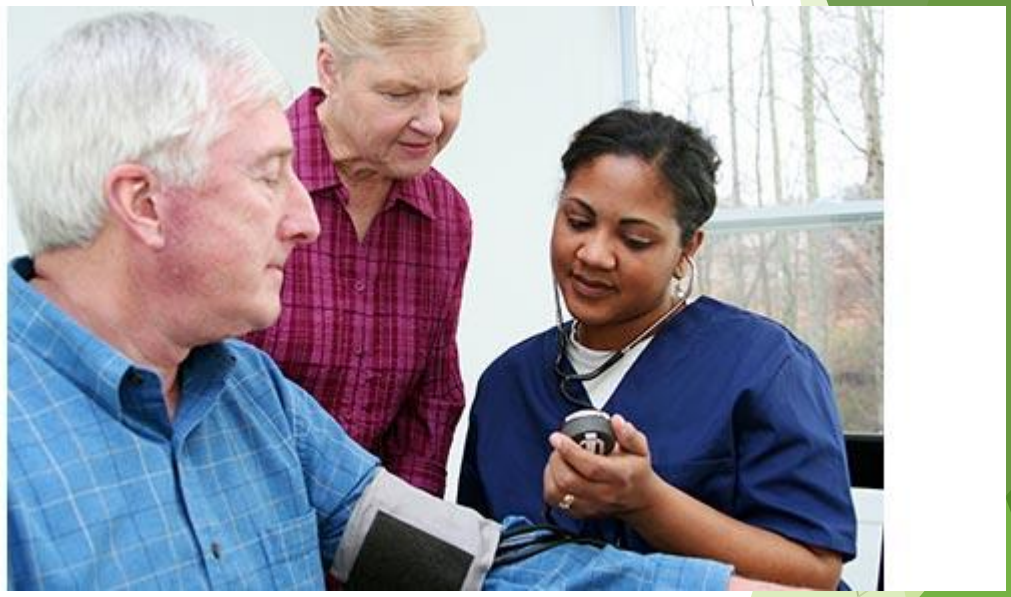
August, 2015

Introduction

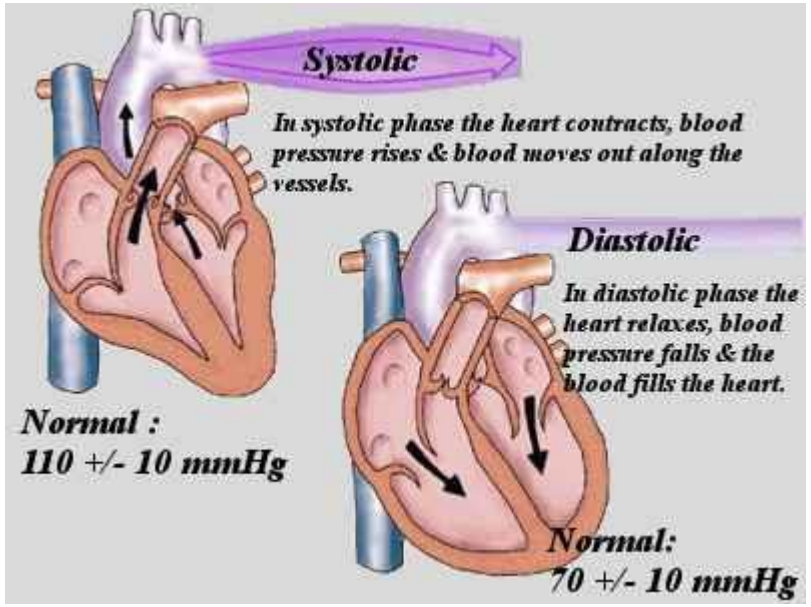
- At GlaxoSmithKline (GSK), a study was conducted for genetic and genomic research to enable the medical community to accurately prescribe the right medicine for the right patient.
- Of the 500 subjects, 250 had blood pressure less than and 250 had high blood pressure. The 501 variables consist of one response variable (systolic blood pressure) and 500 predictors (17 clinical covariates and 483 genetic markers).
- Objective of the present study is to find best predictors for systolic blood pressure.
- We use multiple regression, regularization and logistic regression to analyze the data.

Background

- One billion people suffer with high blood pressure
- 7.1 million deaths each year
- 55-year-old with normal blood pressure (BP) has a 90% lifetime prospect of developing hypertension

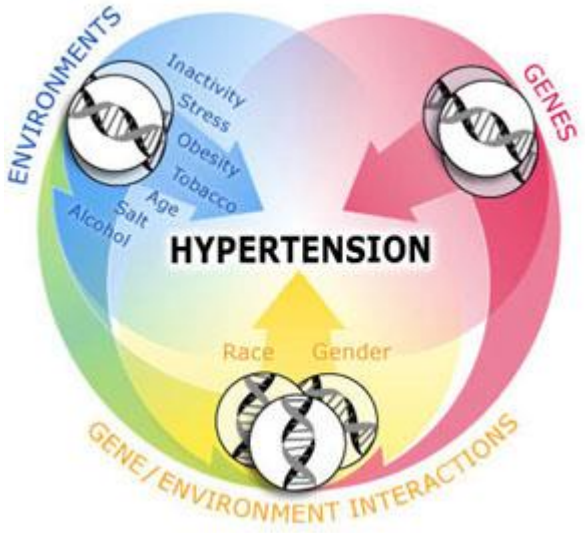


Blood Pressure



High blood pressure	140/90 mmHg or higher
Prehypertension	between 120-139 mmHg and/or 80-89 mmHg
Normal blood pressure	less than 120/80 mmHg

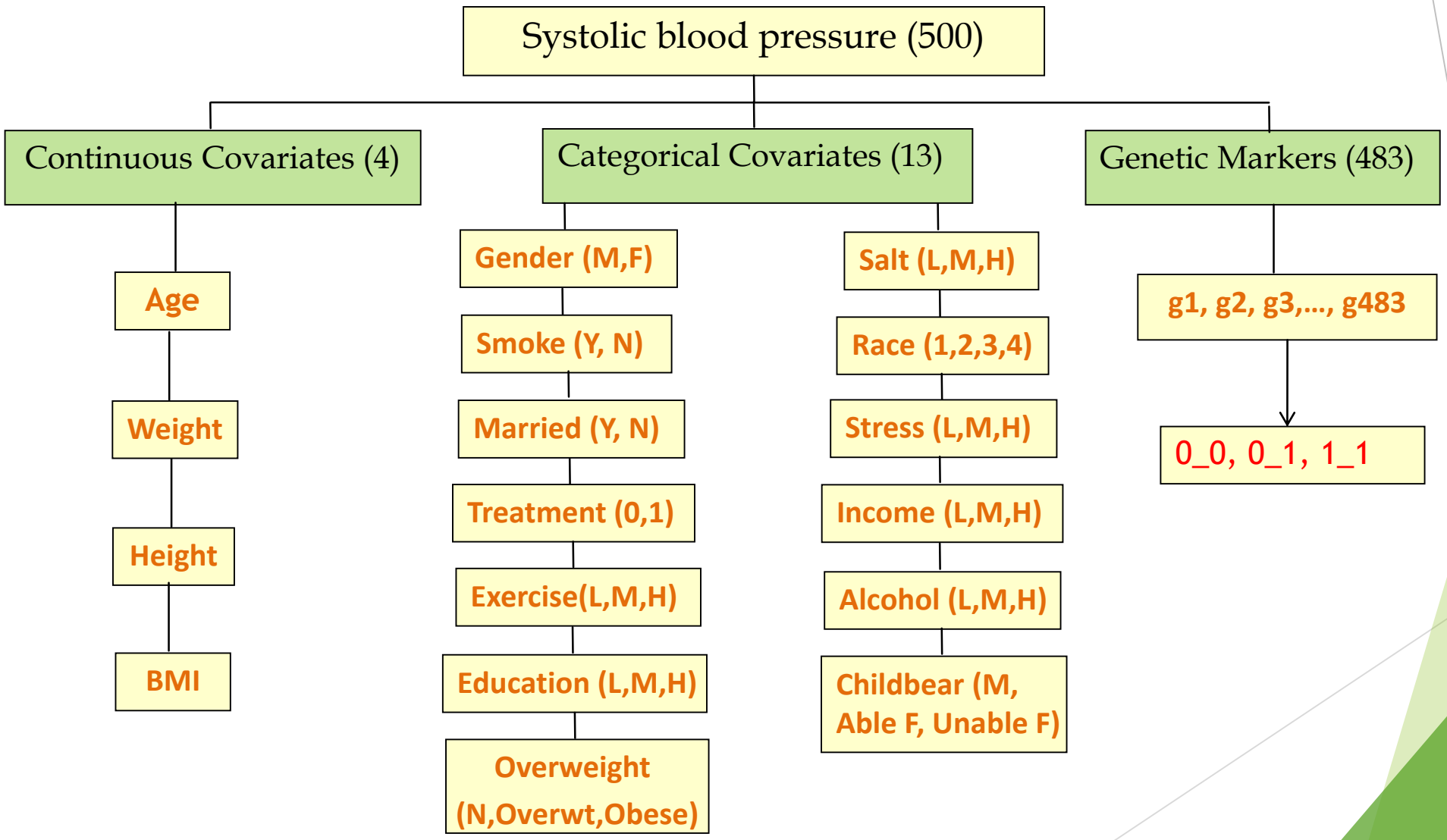
Causes and effects of high blood pressure



Hypertension and risk factors can lead to

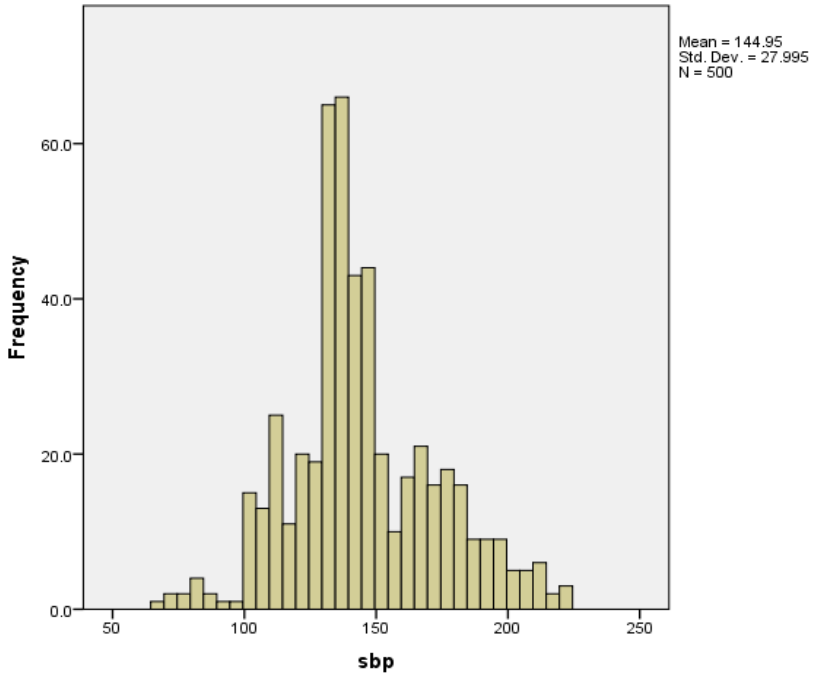
- heart attack
- stroke
- kidney damage
- vision loss
- memory loss
- fluid in the lungs

Variables used in data



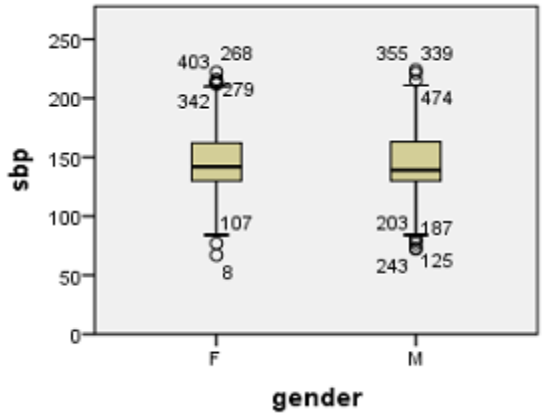
Response Variable (Blood Pressure)

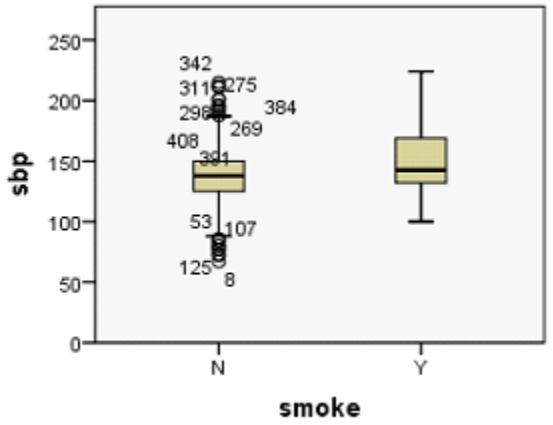
- Of the 500 subjects, 11 are with hypotenstion (SBP less than 90)
- 72 are with normal blood pressure (SBP less than 120)
- 167 are with prehypertension (SBP between 120 and 140).
- 50 percent have high blood pressure (SBP greater than 140) out which 13% can develop life threatening complications (as their SPB is higher than 180).

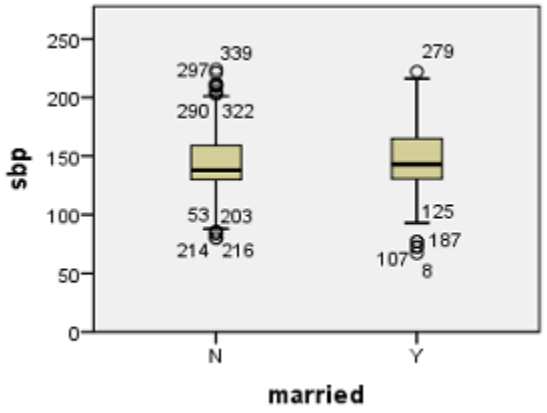


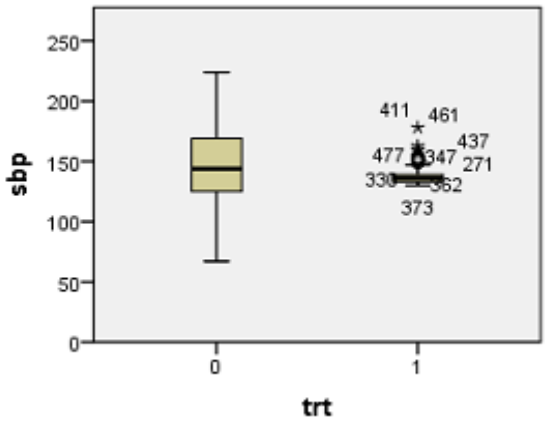
Descriptive Analysis

Box plots for systolic blood pressure by gender, married, smoke and treatment





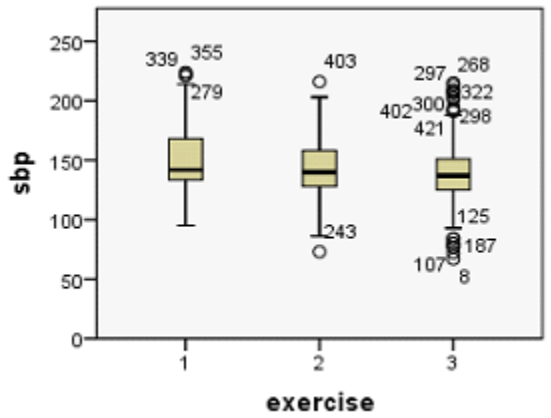


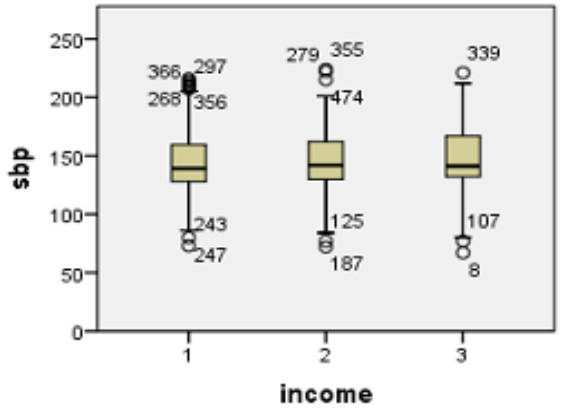


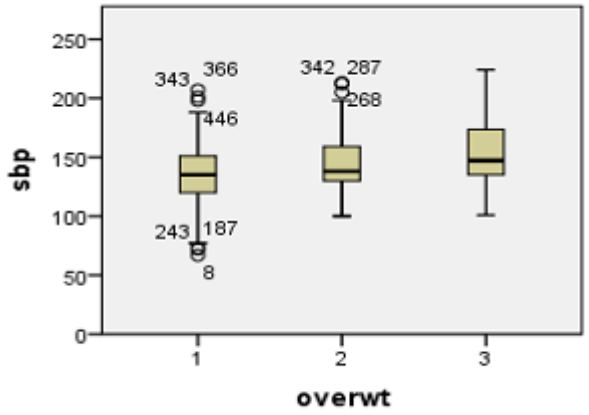
Gender			married		Smoke		Treatment	
	Male	Female	Yes	No	Yes	No	Yes	No
Count	236	264	239	261	266	234	101	399
Mean	145.02	144.89	146.72	143.33	150.03	139.18	137.96	146.72

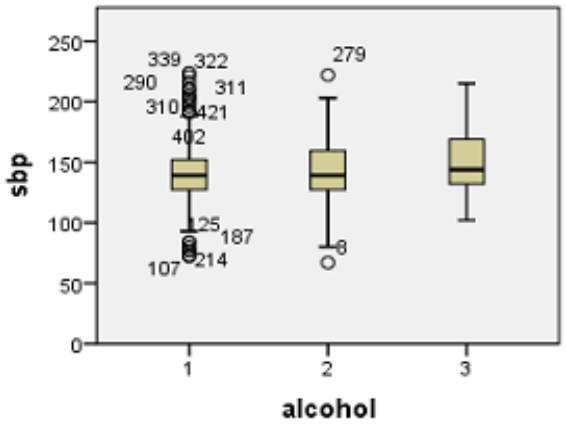
	t value	Pr(>|t|)
Smoke	4.4	0.000
Treatment	2.89	0.000

Box plots for systolic blood pressure by exercise, income, overweight and alcohal



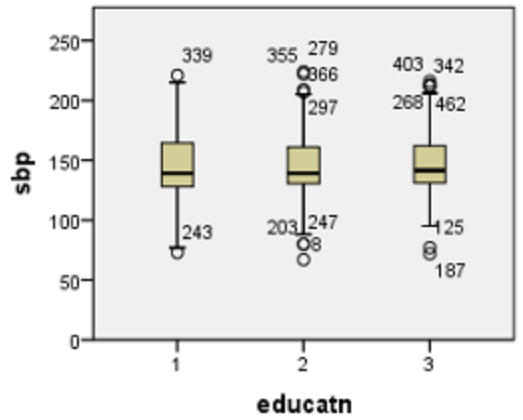


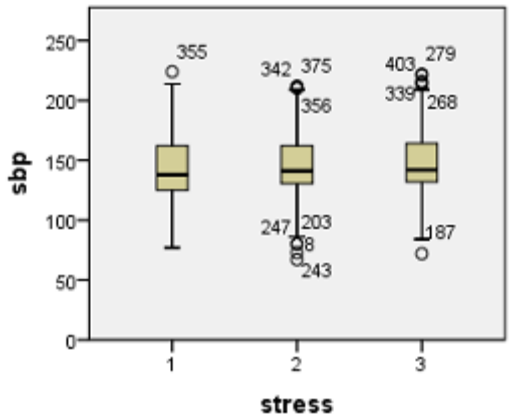


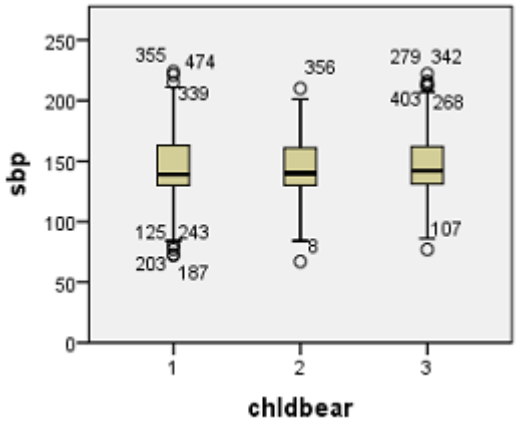


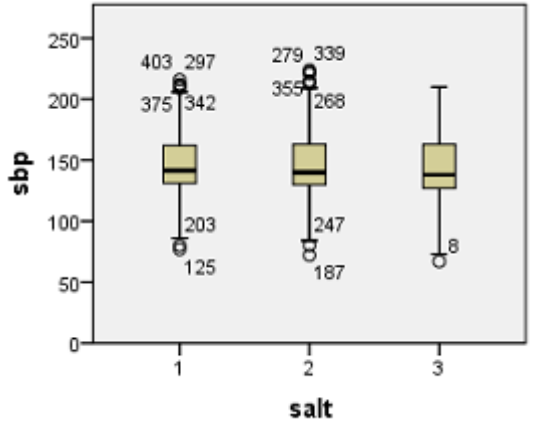
	F-value	Pr(>|F|)
Exercise	5.784	0.003
Over weight	19.017	0.000
Alcohol	5.281	0.005

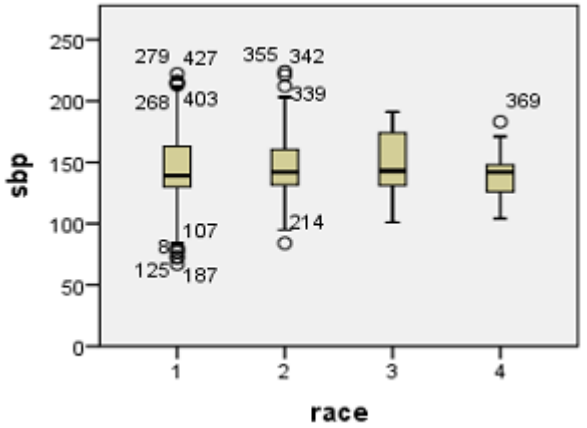
Box plots for systolic blood pressure by education stress salt race and child bearing











Statistical Analysis

Our aim of this study is to identify the best clinical and genetic markers predictors for systolic blood pressure

Approaches

- Clinical covariates
  1. Multiple Regression
  2. Variable selection based on:
     - Forward Method
     - Backward Method
     - Stepwise Method
     - Subset Selection
       - Cp and R-Square Adjusted
       - Adjusted
- Clinical and Genetic Markers covariates
  - All patients
  - High blood pressure Group / Low Blood pressure Group

  Regularization Methods:
  1. LASSO (with and without penalty factor)
  2. Elastic Net (with and without penalty factor)
- Defining New Categorical response Variables with respect to high and low blood pressure

  Regularization Methods with respect to binomial response(logistic regression):
  1. LASSO (with and without penalty factor)
  2. Elastic Net (with and without penalty factor)

- R version 3.2.1/GLMNET/CAR/MASS

All Clinical Covariates Into The Model

- Fitting Multiple Regression Model.
- Four effective covariates on systolic blood pressure are :
  - Smoke
  - Exercise
  - Alcohol
  - Treatment
- R^2 – adjusted equal to 0.2 indicates that only about 20% of the variation in systolic blood pressure can be explained by the relationship to Clinical predictors.
- Complete table are attached to Appendix A

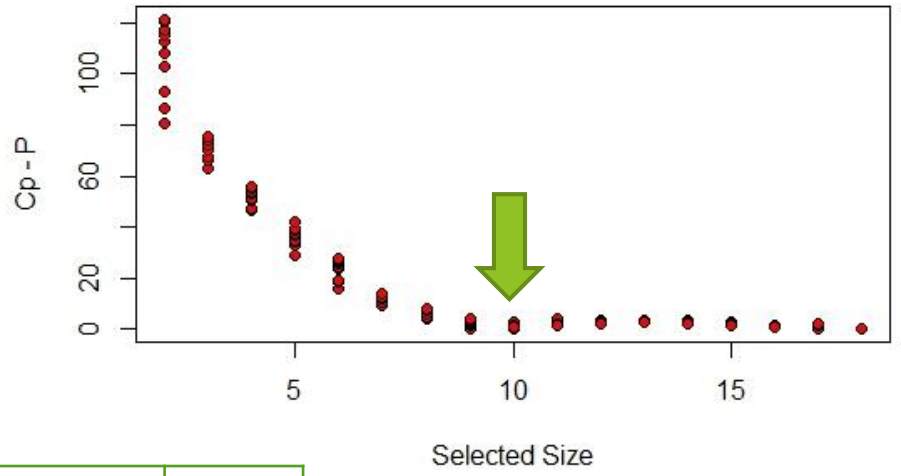
	Pr(>|t|)	Significance
(Intercept)	0.7606	
gender	0.2838	
married	0.174	
smoke	3.47E-06	***
exercise	7.04E-05	***
age	0.1598	
weight	0.6183	
height	0.2406	
overweight	0.0744	.
race	0.4114	
alcohol	1.63E-05	***
Treatment	1.90E-06	***
BMI	0.1965	
stress	0.0933	.
salt	0.8134	
chldbear	0.2289	
income	0.0824	.
educate	0.9935	

Variable Selection

1. Variable selection based on:
   - Forward Method ⟶ All Covariates except BMI
     - Gender/Married /Smoke/ Age/ Weight/Height/Overweight/Race/ Alcohol/ Treatment / Stress/Salt/Childbearing/ Income/ Education.

Similar

- Backward Method:
  - Married /Smoke/ Exercise/ Age/ Height/ Alcohol/ Treatment /BMI/ Stress/ Income
- Stepwise Method
  - Married /Smoke/ Exercise/ Age/ Height/ Alcohol/ Treatment /BMI/ Stress/ Income
- Subset Selection based on Cp:

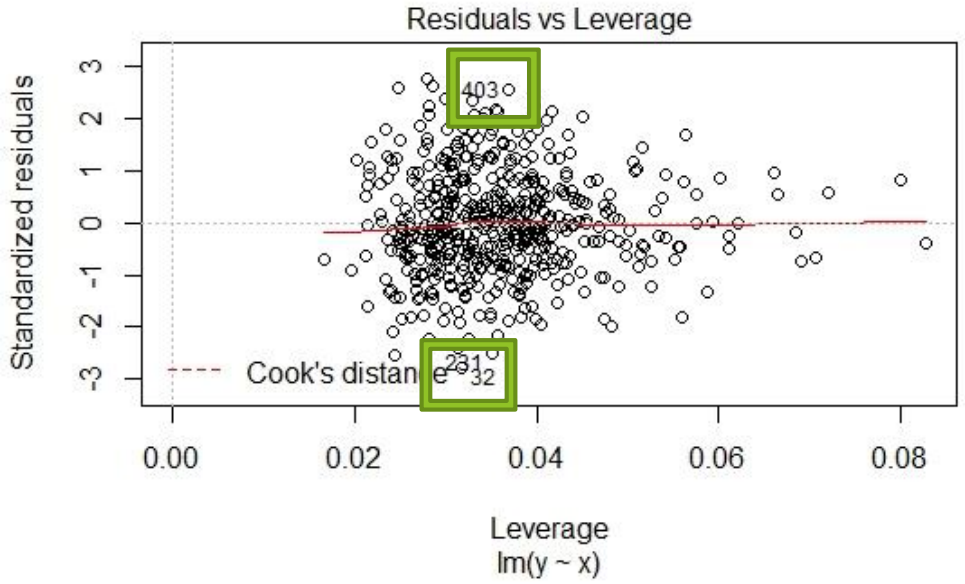


- Subset Selection based on Adjusted R-Square :

Gender	Married	Smoke	Exercise	Age	Weight	Height	Overweight	
FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	**Stress**
Salt	**Childbearing**	**Income**	**Education**	**Race**	**Alcohol**	**Treatment**	**BMI**	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	

Influence Data

- 13 Influence data based on
  - DFBETAS
  - DFFITS
  - Covariance ratios
  - Cook's distances
  - Diagonal elements of the hat matrix
- All those methods are available in “CAR” package in R.



Influence sample	dffit	cov.r	cook.d	hat
8	-0.41	0.83	0.01	0.2
32	-0.51	0.8	0.01	0.3
204	-0.12	1.13	0	0.3
231	-0.48	0.85	0.01	0.4
243	-0.43	0.86	0.01	0.3
339	0.38	0.89	0.01	0.3
355	0.47	0.8	0.01	0.3
356	0.44	0.87	0.01	0.3
366	0.41	0.83	0.01	0.2
375	0.45	0.83	0.01	0.3
403	0.5	0.84	0.01	0.4
474	0.42	0.86	0.01	0.3
485	-0.05	1.11	0	0.7

Final model

- Selecting covariates based on previous variables selection methods.
  - Married /Smoke/ Exercise/ Age/ Height/ Alcohol/ Treatment /BMI/ Stress/ Income
  - Final Model After eliminating influence data
- R-squared adj equals to .22 which is increased by 0.02.
  - There might be some other effective factors
- The MSE is 543.93
- AIC of the model is 3089.52
- There are some Interactions effect, They were negligible because of small effect on the Model.

Interesting results:

- Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	58.87889	17.7727	3.313	0.000994	***
Married/Y	3.84265	2.16786	1.773	0.076948	.
Smoke/Y	10.40449	2.18799	4.755	2.64E-06	***
Exercise/2	-9.52305	2.70537	-3.52	0.000473	***
Exercise/3	-10.12955	2.53552	-3.995	7.50E-05	***
age	0.21189	0.08147	2.601	0.009592	**
height	0.55435	0.21693	2.556	0.010916	*
Alcohol/2	3.34206	2.66318	1.255	0.21013	
Alcohol/3	13.9706	2.6886	5.196	3.03E-07	***
Treatment/1	-13.99667	2.70463	-5.175	3.38E-07	***
BMI	1.148	0.16038	7.158	3.15E-12	***
Stress/2	2.9581	2.68051	1.104	0.270347	
Stress/3	4.30706	2.67563	1.61	0.108124	
Income/2	1.35101	2.59781	0.52	0.603267	
Income/3	5.77526	2.663	2.169	0.030604	*

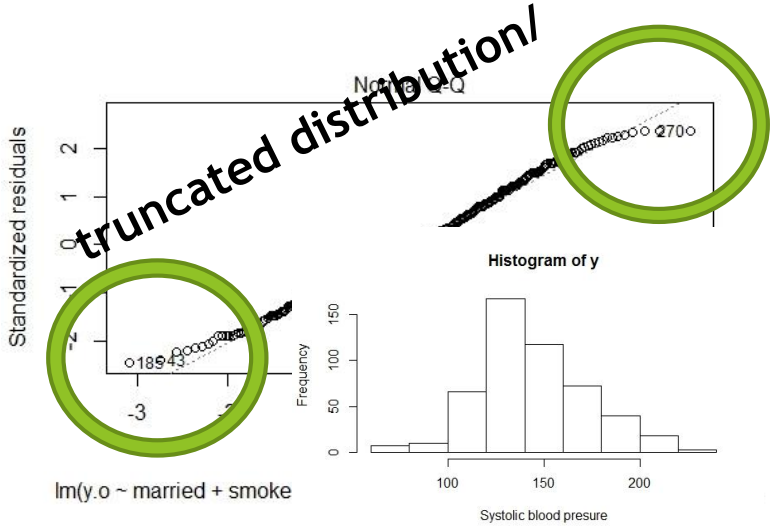
ntinues
riables

Categorical Variable

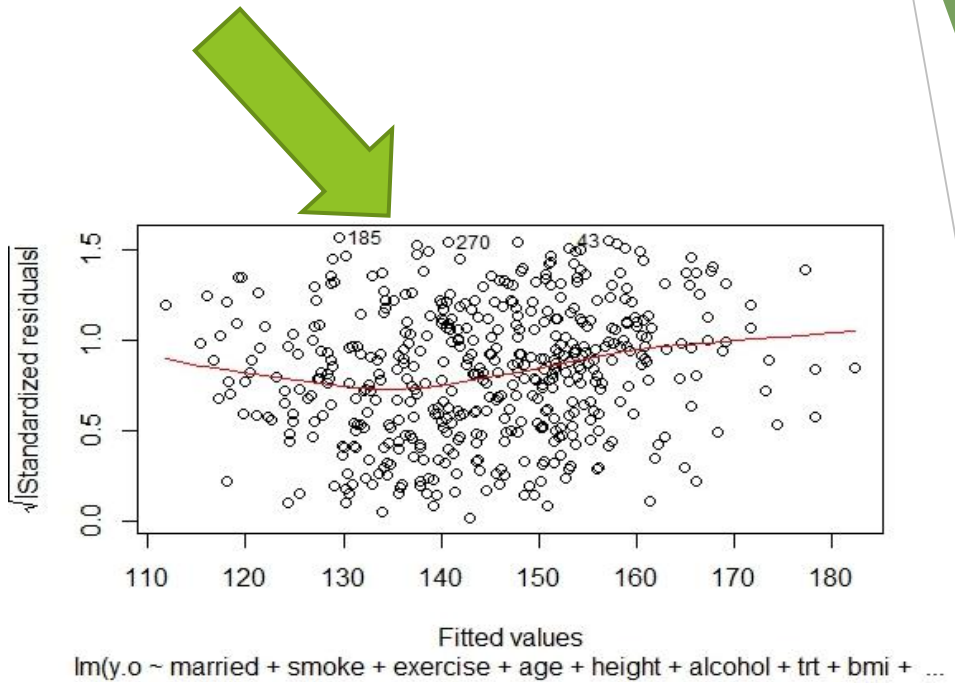
If age or height differed by one unit, and other covariates did not differ, systolic blood pressure will differ by o.2 or 0.5 units, on average, respectively.

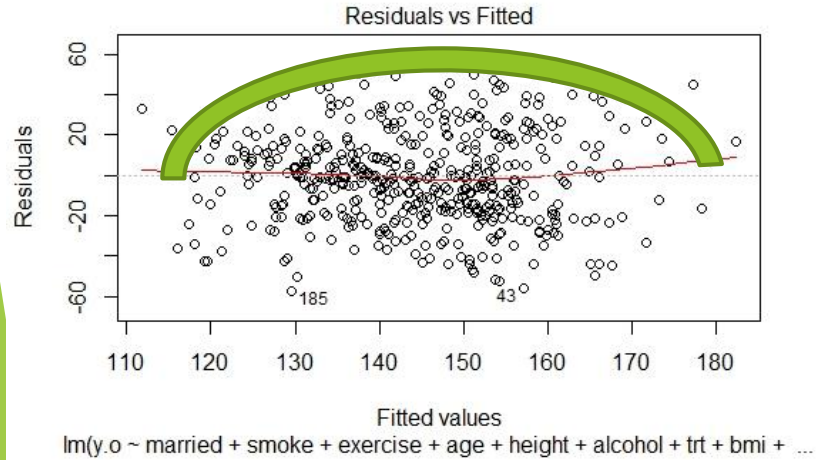
In Smoke group, we would expect that smoker have 10 units higher systolic blood pressure than Non smoker, on average, keeping all other covariates same.

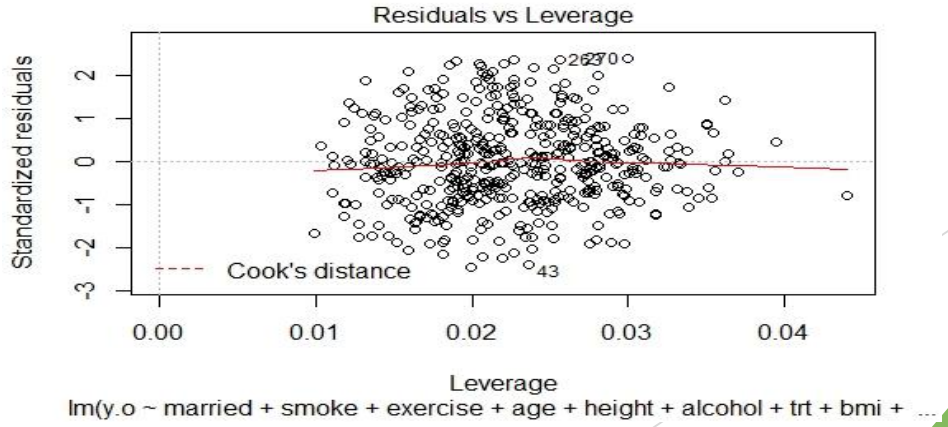
Goodness of fit:



Using Transformation method have not been helpful







Variable Selection based on Regularization Method:

- The LASSO (Least Absolute Shrinkage and Selection Operator):

  LASSO is a regression method that involves penalizing the absolute size of the regression coefficients. This is convenient when we want some automatic variable selection specially with large number of variable, or when dealing with highly correlated predictors.

- Limitation In LASSO Method:

  the LASSO selects at most n variables. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others.

- The Elastic Net:

  This method overcomes the limitations of the LASSO.

Clinical and Genetic Markers covariates

LASSO(without Penalty Factor):

MSE=381.43

R-Square adj = 0.56

Elastic Net (without Penalty Factor):

MSE=385.96

R-Square adj =0.56

Selected Covariates are the same for both Methods:

8 out of 17

- Married/ Smoke /Exercise /Weight/ Overweight/ Alcohol/ Treatment/BMI Stress

58 out of 483

- g7 g9 g10 g46 g48 **g50** g59 g63 g86 g92
- g108 g120 g122 g135 g137 g150 g160 g168 g169 g175 g179 g182 g187 g191
- **g200** g204 g222 g231 g232 g271 g279 g288 g289 g292 g295 **g298**
- g309 g330 g337 g348 g356 g362 g364 g366 g371 g377 g391
- g411 g412 g422 g425 g438 g443 g447 **g453** g465 g469 g480

Genetic Markers	Related Coefficient
G50	13.4
G200	10.65
G298	2.5
G453	2.27

High/low Blood pressure groups

- Based on provided information on website we have divided all patient into two groups.
- Elastic Net methods were applied (n<p)

Significant Covariates

All Patients

High blood pressure group

Smoke / overweight /alcohol
g50 g200 g377
g432 g458 g459

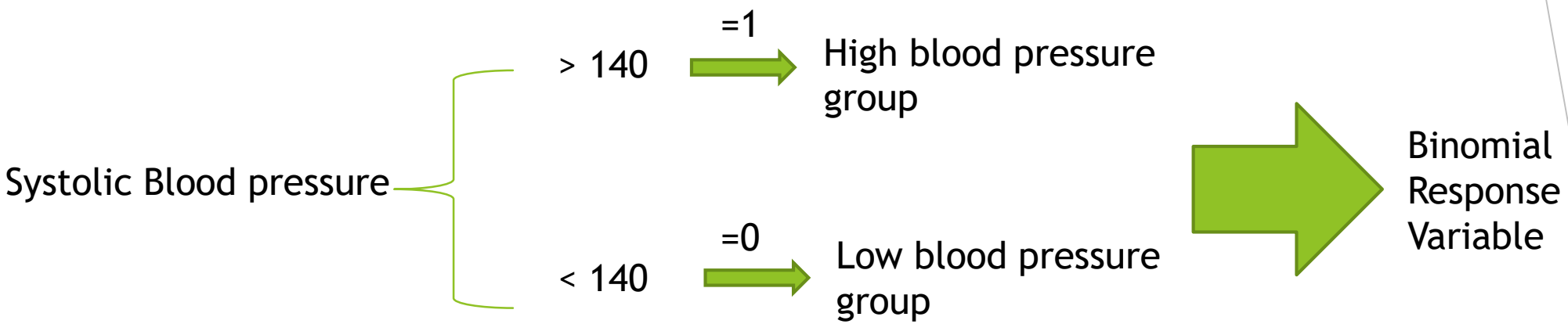
MSE= 217.36

Low blood pressure group

Age /weight / treatment /BMI
g5 g7 g9 g47 g95 g111 g112 g132 g169 g177 g191
g200 g205 g207 g214 g216 g242 g253 g285
g327 g357 g359 g377 g385 g393
g427 g434 g464 g474 g478 g480

MSE= 189.27

Define New Categorical response Variables



Regularization Methods with respect to binomial distribution of y

- Regularization Methods were employed with respect to binomial distribution of y, Selected variables are:
  - Married/ exercise/ overweight/ Treatment/ BMI/ stress.
  - g10 g36 g49 g50 g65 g75 g86 g98
  - g120 g122 g137 g150 g168 g187 g191
  - g200 g204 g231 g279 g298
  - g309 g330g385 g391
  - g412 g425 g447 g450 g453 g460 g469 g475
- In this model, The MSE equals to 0.63
- Based on this model, we can calculate the probability of having High or Low blood pressure given selected covariates.

Conclusion and Recommendations

- According to the results of this study the main predictive factors for systolic Blood pressure were introduced based on regularization method.
- We have found that two genetics markers: “g50” and “g200” were very important predictors for predicting systolic blood pressure(because they have appeared in most models)
- We recommend further study to investigate the association between systolic blood pressure and clinical and genetic markers variables based on principal component analysis and clustering methods to bring out strong patterns in predictors.

References

- ***[1]*** *David W. Hosmer, Jr., Stanley Lemeshow; Rodney X. Sturdivant, "Applied Logistic Regression, 3rd Edition" , John Wiley & Sons, New Jersey, 2013.*
- ***[2]*** *Zou, Hui, Hastie, Trevor, "Regularization and Variable Selection via the Elastic Net". Journal of the Royal Statistical Society, Series B: 301-320, 2005.*
- ***[3]*** *Friedman, Jerome; Trevor Hastie; Rob Tibshirani."Regularization Paths for Generalized Linear Models via Coordinate Descent". Journal of Statistical Software: 1-22, 2010.*
- ***[4]*** *http://cran.r-project.org/web/packages/glmnet/index.html*
- ***[5]*** *http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html*
- ***[6]****Rodríguez, G. Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/wws509/notes/ , 2007.*
- ***[7]*** ***Hastie****, Trevor,* ***Tibshirani****, Robert,* ***Friedman****, Jerome , "The Elements of Statistical Learning, Data Mining, Inference, and Prediction", Second Edition , 2009.*

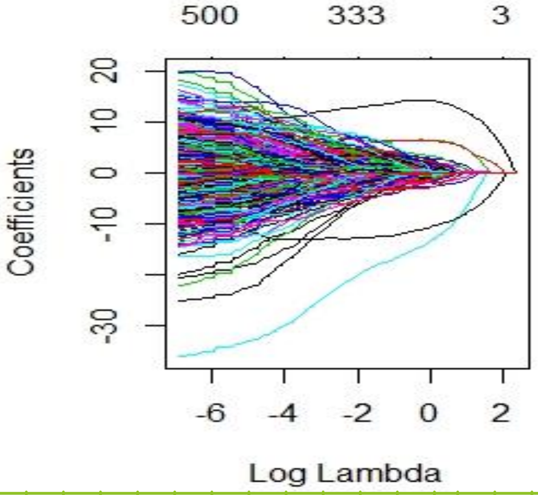
Appendix A

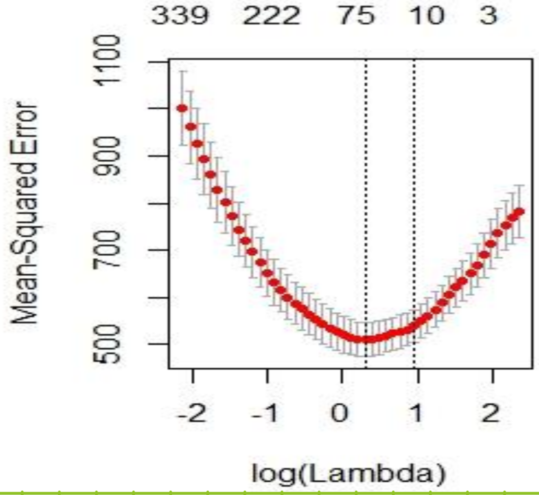
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	36.47519	48.53904	0.751	0.4527	
genderM	-2.85038	3.04392	-0.936	0.3495	
marriedY	2.77287	2.32273	1.194	0.2332	
smokeY	10.74273	2.34442	4.582	5.89E-06	***
as.factor(x$exercise)2	-11.47064	2.9038	-3.95	9.00E-05	***
as.factor(x$exercise)3	-10.72057	2.7085	-3.958	8.71E-05	***
age	0.1095	0.08686	1.261	0.2081	
weight	-0.13328	0.14571	-0.915	0.3608	
height	1.17103	0.7474	1.567	0.1178	
as.factor(x$overwt)2	8.98298	4.336	2.072	0.0388	*
as.factor(x$overwt)3	11.92226	5.92162	2.013	0.0446	*
as.factor(x$race)2	0.3626	2.92517	0.124	0.9014	
as.factor(x$race)3	0.55622	5.34617	0.104	0.9172	
as.factor(x$race)4	-6.72193	5.83917	-1.151	0.2502	
as.factor(x$alcohol)2	1.43863	2.85796	0.503	0.6149	
as.factor(x$alcohol)3	13.1264	2.89709	4.531	7.45E-06	***
as.factor(x$trt)1	-15.26782	2.92514	-5.22	2.69E-07	***
bmi	1.39409	0.83965	1.66	0.0975	.
as.factor(x$stress)2	2.23425	2.88105	0.775	0.4384	
as.factor(x$stress)3	5.04992	2.86234	1.764	0.0783	.
as.factor(x$salt)2	1.52332	2.88632	0.528	0.5979	
as.factor(x$salt)3	0.98388	2.81447	0.35	0.7268	
as.factor(x$chldbear)2	-4.5792	3.20196	-1.43	0.1533	
as.factor(x$chldbear)3	NA	NA	NA	NA	
as.factor(x$income)2	0.97177	2.77538	0.35	0.7264	
as.factor(x$income)3	4.3264	2.84459	1.521	0.1289	
as.factor(x$educatn)2	0.45128	2.82983	0.159	0.8734	
as.factor(x$educatn)3	-0.57115	2.81851	-0.203	0.8395	

Appendix B:

- Min MSE obtained by

Lambda = 2.14

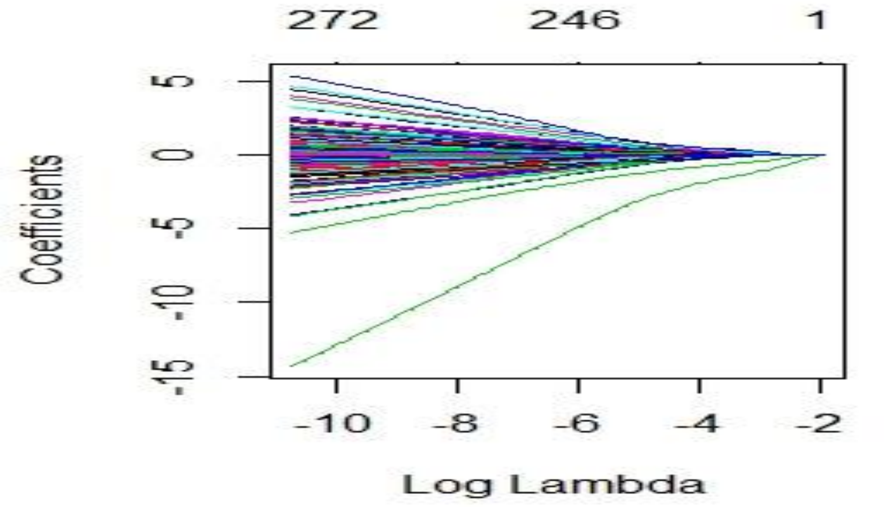


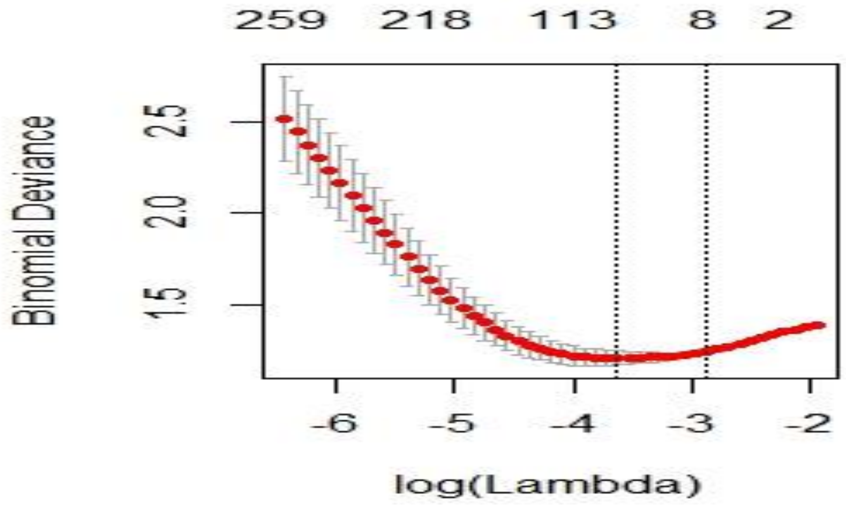


	LASSO	Elastic Net
(Intercept)	98.2521 71	97.5963 81
married	1.2130	1.0814
smoke	6.2416 4	6.0703 5
exercise	2.5617 18	2.4920 82
weight	0.0369	0.0187
overwt	6.1011 2	5.3408 7
alcohol	2.5821	2.4919 6
trt	12.176 74	11.569 08
bmi	0.1273 45	0.1830 7
stress	0.3636 75	0.3960 5
g7	1.8300 39	1.8045 1
g9	0.4702 9	0.4808
g10	0.9438 4	0.9497
g46	0.4585 9	0.4320 3
g48	0.6182 49	0.5524 04
g50	1.3519 8	1.3946
g59	0.7205 8	0.6897 7
g63	0.1393	0.1871
g86	0.5842 8	0.5268
g92	0.0933 04	0.1054 31
g108	0.1967 4	0.1655 3
g120	0.7729	0.7616
g122	0.3943 35	0.3950 14
g135	0.6273	0.9380
g137	1.0469	1.0228
g150	0.8330	0.788 86
g160	0.3675	0.3585
g168	0.3582	0.6738
g169	0.8438	0.7567
g175	0.0185 9	0.0384 8
g179	0.9290 18	0.9161 98
g182	0.2413	0.2057
g187	0.8704	0.9524 4
g191	0.0765 1	0.1435
g200	10.169 92	9.8227 82
g204	0.1448	0.1968
g222	0.9736	0.4781
g231	1.5688	1.5630
g232	0.6335	0.2454
g271	0.0724 8	0.1006 4
g279	1.0015 17	1.0249 92
g288	0.9390	0.9363
g289	0.4871	0.9583
g292	0.3502	0.4123
g295	0.9406	0.9350
g298	2.7889	2.6705
g309	1.7393	1.6383
g330	0.1164 3	0.1572 1
g337	0.9701	0.9689
g348	0.5544 5	-0.532 15
g356	0.6921 3	0.6742
g362	0.0550 3	0.0550 6
g364	0.6061 38	0.6029 8
g366	0.0731	0.0479
g371	0.2017 4	0.2198 9
g377	1.2624	1.2813
g391	1.7921 8	1.7365
g411	1.0940 3	1.0203
g412	0.1456 16	0.1210 38
g422	1.6958 5	1.1493
g425	1.7420 5	1.7040
g438	0.2723 9	0.2830 5
g443	1.9279 1	1.8368
g447	1.1071 9	1.0947
g453	2.2744	2.1281 3
g465	1.8263 5	1.7315 6
g469	0.5102	0.0746
g480	0.1611	0.2214

Appendix C:

Min MSE obtained by
Lambda 0.03516715





(Intercept)	-0.59094
married	0.255988
exercise	-0.08011
overwt	0.036192
trt	-1.41869
bmi	0.028479
stress	0.080703
g10	0.066553
g36	-0.01445
g49	0.003576
g50	0.195049
g65	-0.0936
g75	-0.00088
g86	-0.1118
g98	-0.05993
g120	-0.05367
g122	-0.03582
g137	0.063439
g150	0.016843
g168	0.015018
g187	0.092409
g191	0.00922
g200	-0.57558
g204	0.022184
g231	0.056584
g279	-0.01874
g298	0.07418
g309	0.055583
g330	0.013998
g385	0.199842
g391	-0.07908
g412	-0.00051
g425	0.048352
g447	-0.01751
g450	-0.01224
g453	-0.12179
g460	-0.03033
g469	0.347407
g475	0.018117