The Relationship Between Type A Behavior and Public Health

Indrajit Choudhury

Introduction:

Studying human behavior is a fascinating exploration of what makes a difference in the human mind. In this project we conduct an analysis on the effect of certain physical, demographic and health characteristics and particular lifestyle choices on Type A behavior in humans. An individual with a strong Type A personality would display hostility, impatience, difficulty in expressing emotions, competitiveness, perfectionism and an unhealthy and strong dependence on wealth, status or power. For the benefit of the study the Type A trait in each individual is scored, which forms the response variable in the data. A high score being a stronger presence of Type A behavior. The data consists of 462 individuals and 10 variables characterizing them.

The final objective is to obtain a set of factors among all the given factors in the data which best influence Type A behavior in individuals.

Our analysis is divided into five major parts, namely, Model Selection, Model Adequacy Checking, Tests for Interaction, Tests for Influential points and Interpretation of Final Model.

Data Source: https://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.info.txt

The data is taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal. It is a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments.

Definition of Variables:

Response Variable:

1. typea; Y: This is a continuous variable in which a higher value indicates an individual being more likely to be a Type A Personality and a lower value indicating that an individual is more likely to be a Type B Personality (relaxed, patient and friendly).

Covariates:

1. sbp; X_1: Systolic blood pressure (mmHg); continuous variable

2. tobacco; X_2: Total Lifetime usage of tobacco (kg); continuous variable

3. ldl; X_3: low density lipoprotein cholesterol level (mg/Dl); continuous variable

4. adiposity; X_4: index number to measure amount of body fat; continuous variable

5. famhist; X_5: family history of heart disease (Present/Absent); binary variable

6. obesity; X_6: BMI value; continuous variable

7. alcohol; X_7: total lifetime alcohol consumption (liters): continuous variable

8. age; X_8: current age in years; continuous variable

9. chd; X_9: coronary heart disease (1 means patient has it, 0 means they do not); binary variable

Model Selection:

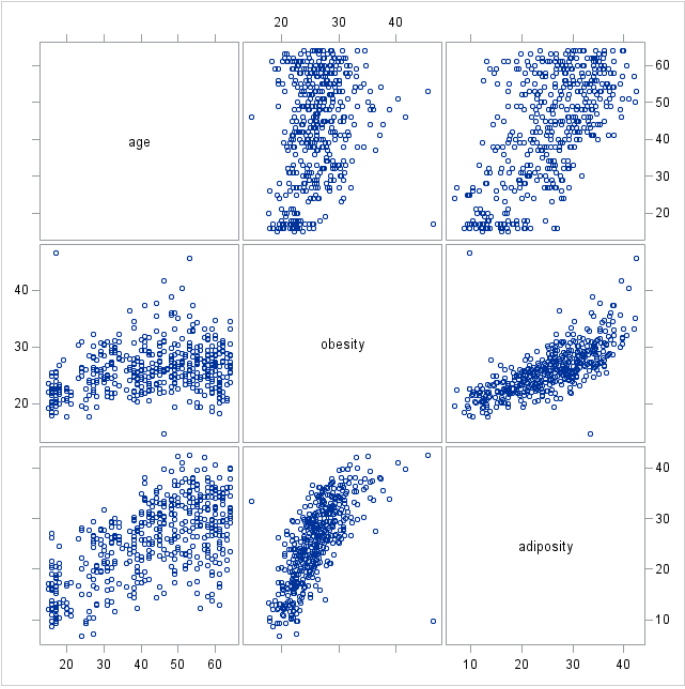
For the purpose of selecting the optimum multiple linear regression model we use the stepwise regression method with a variable entry cutoff of 0.15 significance and an exit cutoff of 0.25. The selection process was run with typea as the response variable and sbp,ldl, adiposity,famhist, obesity, age,chd, tobacco and alcohol as the potential covariates. We verified the model selection by performing Backward Selection and Forward selection methods on the same set of variables. In the end the same set of covariates appeared to be significant.

From the analysis, the chosen model was one with covariates chd, age, obesity and adiposity. To gauge the presence or absence of relationships between the final set of continuous variables we constructed a correlation matrix and in addition analyzed their scatterplots.

	age	obesity	adiposity
age	1	0.29178	0.62595
obesity	0.29178	1	0.71656
adiposity	0.62595	0.71656	1

(Table 1)

The moderate to high positive correlation between adiposity and obesity and between adiposity and age is a reason for concern as it could defy the independence in covariates assumption.



The scatterplots verify the observed high correlation coefficients by displaying a strong linear relationship between adiposity and obesity and between adiposity and age.

(Figure 1)

The following table gives the p-value to test significance of parameter estimate of each variable.

Variable	Pr > |t|
Intercept	<.0001
chd	0.0006
age	0.0371
obesity	0.0033
adiposity	0.0774

The p-value for adiposity in the model is higher than the rest when tested for significance of its regression coefficient. The high p-value along with the high correlation between adiposity and obesity and adiposity and age, lead use to exclude adiposity and run a multiple linear regression of typea on chd, age and obesity.

(Table 2)

This final chosen Main Effects Model is:

$typea = \beta_0 + \beta_6 obesity + \beta_8 age + \beta_9 chd + \varepsilon_i,\ \varepsilon \sim N(0, \sigma^2)$

Where $\hat{\beta}_0$= 50.62770 is an estimate for the intercept.

Estimates of the slope parameters are:-

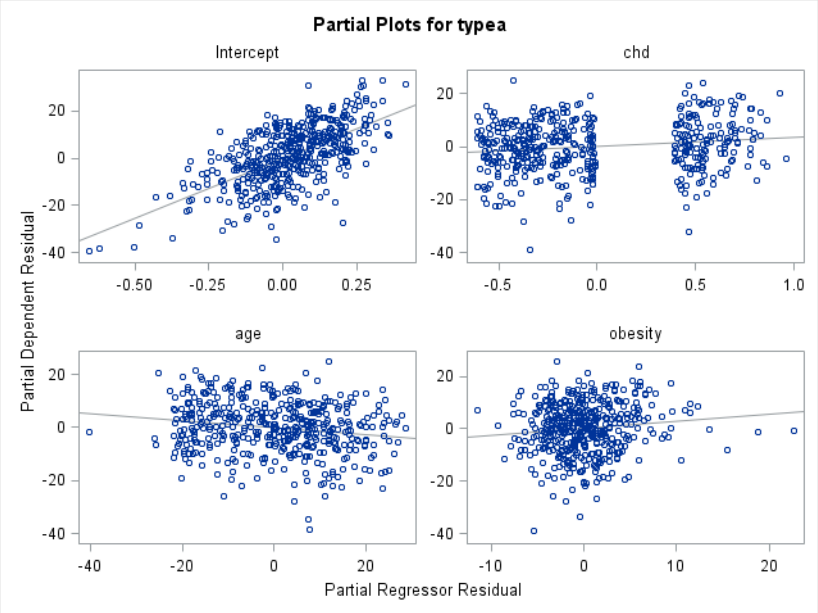
$\hat{\beta}_6$ = 0.26840, which gives the change of mean typea for each unit increase in obesity, adjusted for age and chd.

$\hat{\beta}_8$ =-0.13301, which gives the change of mean typea for each year increase in age, adjusted for obesity and chd.

$\hat{\beta}_9$ =3.41010, which shows that the mean change in typea for a patient with coronary heart disease is 3.41010 times higher than mean change typea for a patient without coronary heart disease, adjusted for obesity and age.

Model Adequacy Checking:

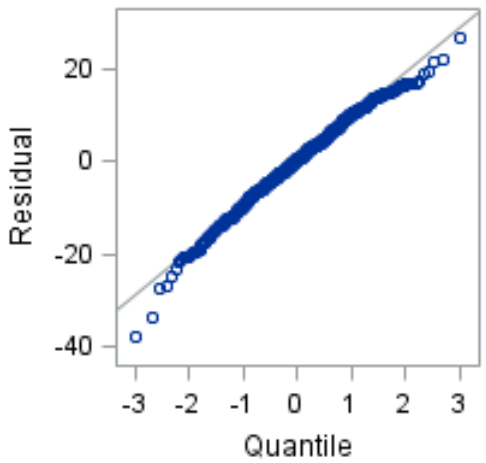
Linearity check:-



(Figure 2)

From the Added Variable Plots above we observe roughly linear relationship between typea and age, adjusting for obesity and chd and between typea and obesity, adjusting for age and chd. Since chd is a categorical variable we cannot comment on the pattern for its Added Variable Plot. This satisfies the assumption of linearity between response variable and covariates in multiple linear regression.

Normality Check :



The Q-Q Plot displays heavy tails to some extent but the data seems to be roughly normal. To further confirm the presence of normality in the data, the Shapiro Wilk's Normality test was conducted.

(Figure 3)

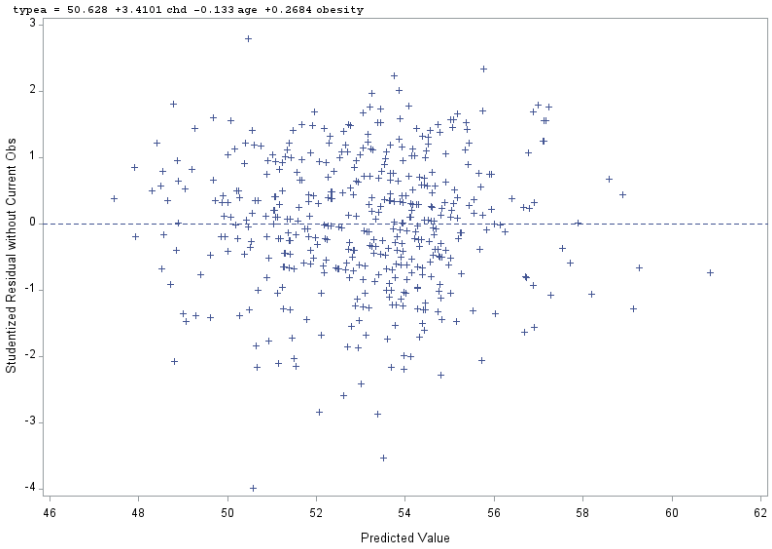
The null hypothesis for the test is, H_0: Data is normally distributed , which is tested against the alternative hypothesis, H_A: The data is not normally distributed

Test	Statistic		p Value	
Shapiro-Wilk	**W**	0.99046	**Pr < W**	0.0044

(Table 3)

The test statistics is 0.99046 which is almost 1 and the p-value= 0.0044 which is greater than 0.001. Hence, we accept H_0at 99% Confidence Level and conclude that the data is normally distributed. This satisfied the normality assumption for the data.

Independence of residuals:-



(Figure 4)

On plotting the studentized residuals from the regression model against the predicted values we observe randomness. This lack of pattern indicates independence between the residuals.

Constant Variance in residuals:-

Figure 4. displays constant variance in residuals throughout. This satisfies the constant error variance assumption in Linear Regression.

Multicollinearity:-

Variable	Variance Inflation Factor
chd	1.1617
age	1.25708
obesity	1.09316

(Table 4)

A high value of Variance Inflation Factor(VIF) indicates presence of multicollinearity. A standard cutoff to determine the presence of multicollinearity is a VIF value of 10. In case of our chosen model, the VIF values of all the variables are observed to be very low. This helps us exclude the possibility of multicollinearity in the model.

Tests for Interaction:

Upon the model selection procedure, we have decided to keep the main effect covariates of our multiple regression model as obesity (Measure of BMI), age (Current Age in Years) and chd (Presence of Coronary Heart Disease). This gives us the Main Effects Model:

$Y_i = \beta_0 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon_i,\ \varepsilon \sim N(0, \sigma^2)$

Next, we attempt to look for any potential interactions among these main effects. Logically, given the nature of the remaining covariates it would make sense to explore possible interaction between the presence of Coronary Heart Disease and Obesity and the presence of Coronary Heart Disease and Age.

We expand the Main Effects into an Interaction Model with these two effects in mind:

$Y_i = \beta_0 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_6 X_9 + \beta_{11} X_8 X_9 + \varepsilon_i,\ \varepsilon \sim N(0, \sigma^2)$

When $X_9 = 0$ (person without Coronary Heart Disease), the model becomes $Y_i = \beta_0 + \beta_6 X_6 + \beta_8 X_8$ and When $X_9 = 1$ (person with Coronary Heart Disease), the model becomes$Y_i = \beta_0 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 + \beta_{10} X_6 + \beta_{11} X_8 = (\beta_0 + \beta_9) + (\beta_6 + \beta_{10}) X_6 + (\beta_8 + \beta_{11}) X_8$.

Given this, the interpretations of the coefficients would be: β_0 is the expected Type A Score for an individual without Coronary Heart Disease whose BMI and Age are both set equal to 0. β_6 is the expected change in Type A Score for a one unit increase in BMI adjusting for age for an individual without Coronary Heart Disease. β_8 is the expected change in Type A Score for a one year increase in age adjusting for BMI for an individual without Coronary Heart Disease. β_9 is the expected change in Type A Score when switching from an individual who does not have Coronary Heart Disease to an individual who does have Coronary Heart Disease, holding BMI and Age constant. β_{10} is the expected change in Type A Score for a one unit increase in BMI adjusting for age for an individual with Coronary Heart Disease compared to an individual without Coronary Heart Disease. β_{11} is the expected change in Type A Score for a one year increase in age adjusting for BMI for an individual with Coronary Heart Disease compared to an individual without Coronary Heart Disease.

To test if these interaction terms are significant or not, we use the Multiple Partial F-Test. The hypotheses are:$H_0: \beta_{10} = \beta_{11} = 0$ versus H_a: at least one of $\beta_{10}\ or\ \beta_{11}$ are non-zero. We use $\alpha = 0.05$. If we consider $X = X^1 + X^2$, then we can re-write our full model as $Y = X^1\beta^1 + X^2\beta^2 + \varepsilon$ where $X^1 = [X_6, X_8, X_9], X^2 = [X_6X_9, X_8X_9]$ and $\beta = \beta^1 + \beta^2, where\ \beta^1 = [\beta_0, \beta_6, \beta_8, \beta_9]\ and\ \beta^2 = [\beta_{10}, \beta_{11}]$. Using this, we re-write the hypotheses as $H_0: \beta^2 = 0_2$ versus $H_a: \beta^2 = 0_2$. The test statistic for this is $F = \frac{SS(X_2|X_1)/q_1}{MSE_L}$, where $SS(X_2|X_1) = SSR(Full) - SSR(Reduced)$, q_1 = number of covariates associated with β^2 = 2 and MSE_L is the MSE for the Full Model. Under the null hypothesis, F follows an $F_{q_1, n-(q_1+1)} = F_{2,459}$ Distribution. Therefore, we would reject H_0 if $F_{.05,2,459} = 3.015$. The information needed to calculate F can be found from the ANOVA outputs from the Full and Reduced Models. Below is the ANOVA table for the Full Model:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2342.52241	468.50448	5.08	0.0002
Error	456	42090	92.30371		
Corrected Total	461	44433			

(Table 5)

Next is the ANOVA table for the Reduced Model:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2039.48523	679.82841	7.34	<.0001
Error	458	42394	92.56229		
Corrected Total	461	44433			

(Table 6)

From this, we calculate the test statistic as $F = \frac{SSR(Full) - SSR(Reduced)/q_1}{MSE_L} = \frac{[2342.52241 - 2039.48523]/2}{93.30371} = 1.642$. Since $F = 1.642 < 3.015 = F_{0.05,2,459}$, the data does not provide significance evidence at $\alpha = 0.05$ to reject H_0. Therefore, we conclude that our interaction terms between the presence of Coronary Heart Disease and Obesity and the presence of Coronary Heart Disease and Age are not significance. As a result, our final model for analysis will be our original main effects model:

$Y_i = \beta_0 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon_i, \varepsilon \sim N(0, \sigma^2)$

Tests for Influential Points:

Now that we have decided upon our final model for analysis, our final step before interpretation of our coefficients is to test for influential points in our data. By indicating the influence option within our PROC REG procedure, we generated a table that listed the raw residuals, jackknife residuals, leverage values, covariance ratio values, Cook's D Values, DFBETAS and DFFITS values for each observation. These will be used to determine potential outliers, leverage points and influential points. Sorting of these values was done in Excel.

First, we will look for outliers. We do this by noting which observations have the largest residual jackknife residual values. We define a large jackknife residual value as an observation with an absolute jackknife residual of greater than or equal to 2.5. The observations that fit this criterion are observations 118, 337, 219, 376, 166 and 334 with jackknife residual values of -3.9825,-3.5396,-2.8759, -2.8394, 2.7873 and -2.5859 respectively. We will consider these 7 points as outliers.

Next, we will look for high leverage points. We do this by noting which observations have the largest leverage values, which are indicated by the Hat Diagonal H for each observation. We consider any observation that has a leverage value of greater than 0.05 as a high leverage point (as this value is significantly larger than the leverage values of most the sample). The observations that fit this criterion are observations 45 and 82 with leverage values of 0.0778 and 0.0539 respectively. We will consider these 2 points as high leverage points.

We note that there aren't any points that are both outliers and high leverage points by the criterion we defined above. We will now attempt to determine whether any points in our data can be considered influential points. We sort the table by values of Covariance Ratio, Cook's D, DFFITS and DFBETAS for each observation and look for any unusually high values. The observations with Covariance Ratios furthest away from 1 are Observations 118, 45 and 337 with Covariance Ratios of 0.8866, 1.0886 and 0.9117 respectively. The observations with the highest values of Cook's D are Observations 118 and 337 with Cook's D Values of 0.028 and 0.021 respectively. The observations with the highest absolute value of DFFITS are Observations 118, 337 and 222 with DFFITS values of -0.3422, -0.2968 and -0.2577 respectively. We already note that even though we are listing the most extreme values for these influence statistics, none of them appear to be particularly large. We find similar patterns when we sort values of DFBETAS for each individual covariate.

Based on this influential statistic analysis, we conclude that since there are no observations that are both outliers and high leverage points and since there are no points with significantly large values for the various influential statistics, that our data does not contain any influential observations. As a result, there is no need to remove any of the points from our data.

Interpretation of Final Model:

To reiterate, the final model for our analysis is:

$Y_i = \beta_0 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon_i, \varepsilon \sim N(0, \sigma^2)$

Where X_6 = Obesity (measured as BMI), X_8 = Age (in years) and X_9 = Coronary Heart Disease (binary variable, where 1 means individual has disease and 0 means that individual does not).

Using PROC REG, this model is fit with the results in the table below:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2039.48523	679.82841	7.34	<.0001
Error	458	42394	92.56229		
Corrected Total	461	44433			

(Table 7)

Root MSE	9.62093	R-Square	0.0459
Dependent Mean	53.10390	Adj R-Sq	0.0397
Coeff Var	18.11718		

(Table 8)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50.62770	2.85886	17.71	<.0001
obesity	1	0.26840	0.11119	2.41	0.0162
age	1	-0.13301	0.03439	-3.87	0.0001
chd	1	3.41010	1.01396	3.36	0.0008

(Table 9)

From here, we find the estimated parameters to be:

$\hat{\sigma}^2 = 92.56229, \hat{\beta}_0 = 50.62770, \hat{\beta}_6 = 0.26840, \hat{\beta}_8 = -0.13301 \; and \; \hat{\beta}_9 = 3.41010.$

The interpretation of these parameters in the context of the problem are as follows:

The expected value of the Type A Score, when BMI and Age are set to 0 for an individual who does not have Coronary Heart Disease is equal to 50.62270 units. For every one unit increase in BMI for an individual who does not have Coronary Heart Disease, adjusting for Age, there is an expected increase in Type A Score of 0.26840 units. For every one year increase in Age for an individual who does not have Coronary Heart Disease, adjusting for BMI, there is an expected decrease in Type A Score of 0.13301 units. When switching from an individual who does not have Coronary Heart Disease to an individual who does have Coronary Heart Disease, adjusting for BMI and Age, there is an expected increase in Type A Score of 3.41010 units.

Additionally, we can test for significance of regression. First, we test the hypotheses $H_0: \beta_1 = \beta_6 = \beta_8 = \beta_9 = 0$ versus H_a: at least one of the coefficients is nonzero using $\alpha = 0.05$. The test statistic is $F = \frac{MSR}{MSE}$, which follows a $F_{3,458}$ distribution under the null hypothesis. We reject the null hypothesis if $F > F_{0.05,3,458} = 2.624$. The value of the test statistic is $F = \frac{679.82841}{92.56229} = 7.34 > 2.624 = F_{0.05,3,458}$. Therefore, the data provides significant evidence at $\alpha = 0.05$ to reject the null hypothesis in favor of the alternative hypothesis. Therefore, we state that at least one of the covariates must be significant.

Next, we test for significance of each individual covariate. This is done in general in the following way:

$H_0: \beta_j = 0 \; vs. H_a: \beta_{j \neq} 0 \; at \; \alpha = 0.05$. The test statistic is $T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$. Under the null hypothesis this follows a t_{458} distribution. Therefore, we would reject the null hypothesis in favor of the alterative hypothesis if $|t| > t_{0.025,458} = 1.96$. We note that in the SAS output, the t-values for each of the estimated coefficients are 17.71, 2.41, -3.87 and 3.36 respectively. Therefore, for the intercept and all three of our covariates, the data provides significant evidence at $\alpha = 0.05$ to reject the null hypothesis in favor of the alternative hypothesis. Therefore, all the regressors and the intercept in this regression model are significant.

Conclusion:

The main objective of this report was to find any potential relationships between Type A Behavior and the covariates: Systolic Blood Pressure, Lifetime Tobacco Usage, individual LDL, individual adiposity, family history of heart disease, BMI, Lifetime Alcohol Usage, Age and Coronary Heart Disease. In order to accomplish this goal, we first performed Stepwise Regression for Proper Model Selection. After this we were left with the covariates Coronary Heart Disease, Age, BMI and Adiposity. We then performed the necessary Model Adequacy Checks in order to ensure that our model would not violate any assumptions. We removed the variable adiposity to ensure independence of variables and also checked for the assumptions of Linearity, Normality, Independence of Residuals and Constant Variance. Then we attempted to see if there were any significant interaction terms relating to Coronary Heart Disease among our remaining three main effects but after using the Overall F-Test, we found that none of these interactions were significant. Finally, we determined that our optimal model was response Type A versus covariates BMI, Age and Coronary Heart Disease. After fitting this model, we conducted the F-Test to determine that least one of these covariates is significant

and we also conducted the individual T-test to come up with the same conclusion for each of the individual covariates. From this, we can say that there is a significant association between Type A Score (which attempts to measure the degree to which an individual exhibits Type A Behavior) and BMI, Age and Coronary Heart Disease where larger values of BMI are associated with larger Type A Scores, smaller values of Age are associated with larger type A Scores and individuals with Coronary Heart Disease are associated with larger type A Scores when compared to individuals without Coronary Heart Disease.

Limitations of the study:

- We were unaware if the exact scoring rules used to formulate the response variable in the data.
- The method of sampling prior to data collection was not specified, hence we are unable to comment on the randomness in sampling.