Predicting systolic blood pressure using a linear regression

Jamin Kiukkonen

1 Introduction

The MONICA (Multinational Monitoring of trends and determinants in Cardiovascular disease) Project was established in the early 1980s around the globe to monitor trends in cardiovascular diseases, and to relate these to risk factor changes in the population over a ten year period. It aimed to explain the diverse trends in cardiovascular disease mortality. Altogether 21 countries participated in the project. The total population in this study was ten million men and women, who were 25-64 years old. This paper considers only the data collected in Finland. The data consists of relevant state of health information regarding cardiovascular diseases.

The motivation behind this paper was to investigate the association between a person's background factors and cardiovascular diseases using the linear regression algorithm. Furthermore, I also seek to answer the question that, if there is an association, what are these background factors that promotes the prevalence of the cardiovascular diseases.

2 Description

Figure 1 shows a boxplot of systolic blood pressure plotted against the four age groups used in this study. based on this plot, there seems to be a somewhat clear trend; older people tend to have a higher systolic blood pressure. Figure 2 presents a graphical gender-wise comparison of the systolic blood pressure between men and women, all age groups together, using a swarm plot. A swarm plot is a type of scatter plot where a categorical variable can be plotted against a continuous variable. In a swarm plot, each category is represented by a swarm of data points, and the data points have been adjusted so that they don't overlap with each other. A swarm plot is particularly useful when the sample size is not too large, since it displays all of the data points of each category as its own distribution. It is worth noting that on figure 2 the "center of the swarm" (i.e. the center of the distribution representing men) is shifted more towards right, indicating that, on average, men could have a higher systolic blood pressure. This will be discussed in chapter 3 on statistical testing. The mean systolic blood pressure is 138.1 mmHg for men, and 132.12 mmHg for women.

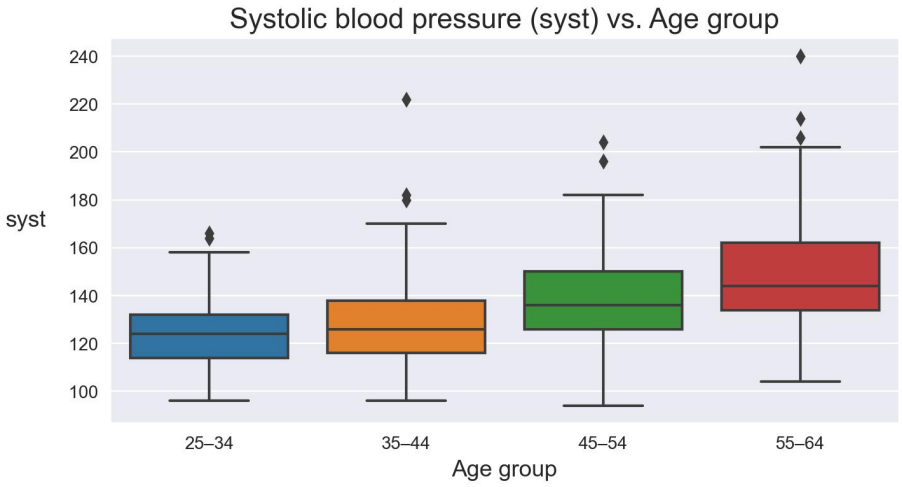

Figure 1: Systolic blood pressure plotted against age group in a boxplot. The boxplot shows that older people tend to have a higher systolic blood pressure. The diamond-shaped black dots represent outliers.

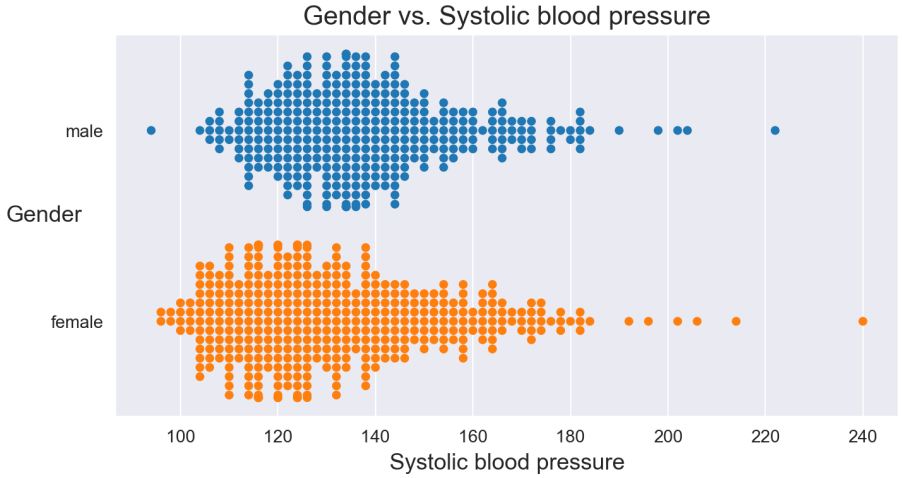

Figure 2: A gender-wise comparison of systolic blood pressure using a swarm plot. The swarm plot does a great job illustrating the slight difference in the mean systolic blood pressures between men (138.1 mmHg) and women (132.12 mmHg).

A more interesting comparison turned out to be a gender-wise comparison of the systolic blood pressure across the age groups. Figure 3 shows that there is a clear trend between the systolic blood pressure and age; as age increases, the systolic blood pressure tends to increase as well. Figure 3 also reveals that the increment of the systolic blood pressure is more even among ageing men than among ageing women. It is also worth noting that women actually have higher mean blood pressure in the oldest age group (55-64). However, again, the reader should note that in this oldest age group, there are two women that have a really high systolic blood pressure, which might affect the overall mean of this age group. Therefore, median, which is barely affected by outliers, might be a better approach to compare the age groups, at least in the case of the oldest age group. The presence of outliers in the women's groups can also be seen in the standard deviations. All of the age groups together, the standard deviation of the systolic blood pressure is 21.84 for women, and 19.11 for men. Women also have higher standard deviation of the systolic blood pressure in every age group except one. The interquartile range of the systolic blood pressure for the whole data set is 26.0. Means, medians, modes, standard deviations, and interquartile ranges of the systolic blood pressure grouped by age group and gender are reported in table 1. Table 1 shows that the mean, median and mode of the systolic blood pressure is higher for men in every age category except the oldest, in which the three previously mentioned measure of location parameters are higher for women than for men. On the other hand, women have higher standard deviation of the systolic blood pressure in every age group except the age group of 35-44.

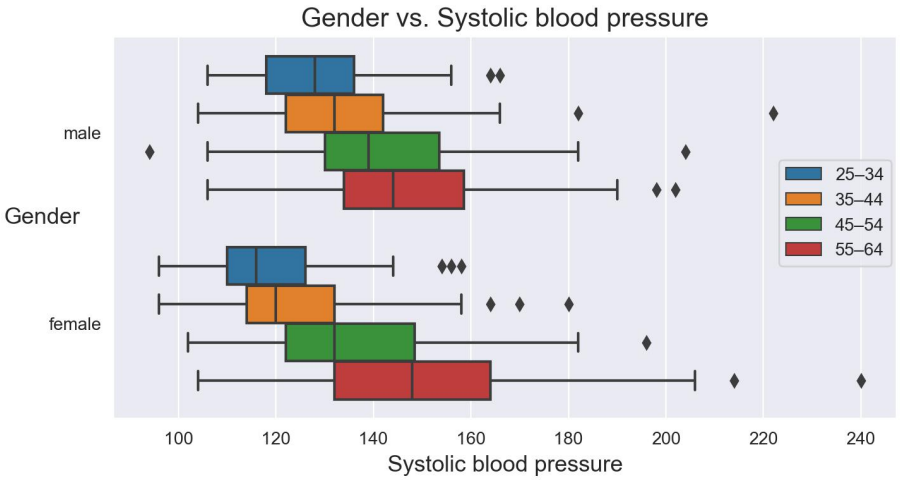

Figure 3: A gender-wise comparison of the systolic blood pressure across different age groups. The graph revealed that the increment of the systolic blood pressure is more even among ageing men than among ageing women. Men have higher systolic blood pressure in every age group except the oldest (55-64).

Table 1: Means, medians, modes, standard deviations, and interquartile ranges of the systolic blood pressure, grouped by age group and gender.

Age group	Gender	Mean	Median	Mode	Std	IQR
25-34	male	128.5	128.0	128.0	12.2	18.0
	female	119.1	116.0	116.0	12.8	16.0
35-44	male	134.0	132.0	122.0	17.3	20.0
	female	123.5	120.0	120.0	15.5	18.0
45-54	male	142.1	139.0	136.0	19.7	23.5
	female	136.4	132.0	122.0	20.0	26.5
55-64	male	146.8	144.0	134.0	20.6	24.5
	female	149.1	148.0	138.0	23.8	32.0

3 Statistical testing

Statistical testing was done to evaluate the relationships between the variables used in the linear regression model. According to the Kolmogorov-Smirnov test, the response variable (systolic blood pressure) is not normally distributed (statistic = 1.0, $p < 0.0005$). The Kolmogorov-Smirnov test was done for the whole sample population together, and separately for men and women. All three tests yielded the same test statistic and p value. Therefore, the non-parametric version of the t-test, the Mann-Whitney U test, was used to test whether the two distributions associated with the systolic blood pressure of men and women have the same location. The Mann-Whitney U test revealed that the differences in the systolic blood pressure between men and women are statistically significant (statistic = 55289.0, $p < 0.0005$). In other words, a randomly chosen male from the data set is more likely to have a higher systolic blood pressure than a randomly chosen female.

Since the response variable is not normally distributed, the correlation between the response variable and the explanatory variables was measured with Spearman's rho, which is a non-parametric measure of the strength and direction of the association between the two variables in question. Spearman's rho is based on the ranks of the data rather than the actual values, making it a useful tool when dealing with non-normally distributed data or when there may be outliers in the data. The correlation coefficients between the explanatory variables were also measured using Spearman's rho. Table 2 presents a correlation matrix of the Spearman's rank correlation coefficients. All of the correlations that were significant, turned out to be highly significant, and are denoted with three asterisk symbols, as per practice.

Table 2: Spearman's rank correlation coefficients. Highly significant correlations ($p < 0.001$) are denoted with three asterisk symbols.

	Blood pressure	Age group	Sex	BMI
syst	1.000	.459***	-.179***	.422***
agegr	.459***	1.000	-.006	.420***
sex	-.179***	-.006	1.000	-.131***

4 Regression analysis

Regression analysis was conducted to evaluate whether person's background factors are associated with cardiovascular diseases. Linear relationship, no multicollinearity, independency, homoscedasticity, and normality, were assumed for the residuals. As for the linear regression model, the response variable was the systolic blood pressure, and the explanatory variables, after a careful selection, ended up being age group, gender, and BMI. BMI was centered to interpret the intercept term. The explanatory variables were chosen based on factors contributing to the model's performance, while trying to avoid increasing the complexity of the model too much. These factors included the values of R-squared, residual standard error, AIC (Akaike information criterion), BIC (Bayesian information criterion), VIF (variance inflation factor), and p-values associated with the F-test statistic acquired using ANOVA (analysis of variance). Equation (1) is the final model equation used in this paper. The model's regression coefficients and the corresponding 95 % confidence intervals and p are reported in table 3.

$$\begin{aligned} syst_i = \beta_0 + \beta_1\, agegr[35-44]_i \,+\, \beta_2\, agegr[45-54]_i \,+\, \beta_3\, agegr[55-64]_i \\ +\, \beta_4\, sex[female]_i \,+\, \beta_5\, bmiC_i \,+\, \epsilon_i \end{aligned} \quad (1)$$

Table 3: The regression coefficients and the corresponding confidence intervals of the linear regression model.

	Coefficients	Lower CI	Upper CI	p-value
Intercept	128.89	125.89	131.90	< 0.000
agegr[35-44]	3.53	-0.12	7.18	0.0582
agegr[45-54]	11.53	7.76	15.29	< 0.000
agegr[55-64]	19.13	15.22	23.03	< 0.000
sex[female]	-5.04	-7.60	-2.48	0.00012
bmiC	1.17	0.87	1.48	< 0.000

For the regression model in question, the centered BMI variable (bmiC) allows one to interpret the intercept term. The intercept term in a regression model represents the expected value of the response variable when all predictor variables are zero-centered. By centering the BMI variable, it redefines the zero point to the mean value of the BMI variable. Without the centering of the BMI term, one could not meaningfully interpret the intercept term, since BMI can't be zero. As for the interpretation of the intercept term; a male (the reference category) who's age is 25-34 (the reference category) and BMI is 26 (26 is the mean value of the BMI variable), is expected to have a systolic blood pressure of 128.89 mmHg. The interpretation for a continuous predictor variable and a categorical predictor variable is somewhat different. The regression coefficients for the different age groups are denoted by β_1, β_2, *and* β_3. The interpretation for

all of these regression coefficients is the same. Let X_2 be a categorical predictor variable and β_2 it's regression coefficient, and let Y be the response variable. Now, β_2 is interpreted as the difference in Y for each one-unit difference in X_2, when all of the other predictor variables are held constant. However, since X_2 is a categorical variable coded as 0 or 1, a one unit difference represents switching from one category to another. This means that β_2 is then the average difference in Y between the category for which $X_2 = 0$ (the reference group) and the category for which $X_2 = 1$ (the comparison group). In simpler terms, the regression coefficients β_1, β_2, *and* β_3 represent the increase (increase because β_1, β_2, *and* β_3 are all positive) in the expected systolic blood pressure for the corresponding age group compared to the reference group, which is the youngest age group (25-34). Note that since we also have another categorical predictor variable in the model, when we are holding it constant, it means that we are fixing its value in the equation to zero, which is it's reference group. So for example, the interpretation for β_2 is such that a randomly chosen male from the data set belonging to the age group of 45-54, is expected to have a systolic blood pressure of $128.89 + 11.53 = 140.42$. If we wanted to find out the same statistic for a randomly chosen female belonging to the same age group, we would have to add the β_4 coefficient to this equation, giving us $128.89 + 11.53 + (-5.04) = 135.38$. The β coefficient of a continuous predictor variable can be interpreted such that it represents the difference in the response variable for each one-unit difference in the continuous predictor variable in question, while holding the other predictor variables constant. β_5 can be interpreted such that when the value of bmiC increases by one unit, the systolic blood pressure increases by 1.17 units. It is worth noting that every β coefficient except β_1 was highly significant, and even β_1 was just on the edge of being significant with a p-value of 0.0582.

If the model assumptions hold, then the observed residuals $e_i = Y_i - \widehat{Y_i}$ (where Y_i is the ith observed value of the response variable, and $\widehat{Y_i}$ is the predicted value generated from the model) should behave in a similar fashion. In other words, the residuals should be approximately normally distributed with a constant variance for all the different X values. Figure 5 displays a scatter plot of the residuals against the fitted (predicted) values. From this it can be seen that the residuals seem to be scattered approximately evenly across their mean. Hence, it can be concluded that the residuals are homoskedastic. Also, since there are no clear patterns of any sort, it is safe to conclude that the residuals are independent. Figure 6 presents a normal quantile-quantile plot (probability plot), which is being used to test for the normality assumption. Should the residuals be approximately normally distributed, then the Q-Q plot of those residuals will result in approximately straight line. The model's performance was measured with R-squared and the residual standard error (RSE). An R-squared value of 0.282 was obtained, which states that 28.2 % of the variance in the response variable can be explained by the explanatory variables. RSE was 17.7, which indicates that the true value lies typically around 17.7 units (mmHg) away from the regression line. In other words, the typical difference between a prediction made by the model and an observed response is 17.7 units.

Figure 3 displays two scatter plots side by side of the same graph. In the left-hand graph the data points have been colored by gender, and on the right-hand graph the data points have been colored by age group. Based on this scatter plot, it is not clear whether higher BMI increases systolic blood pressure.

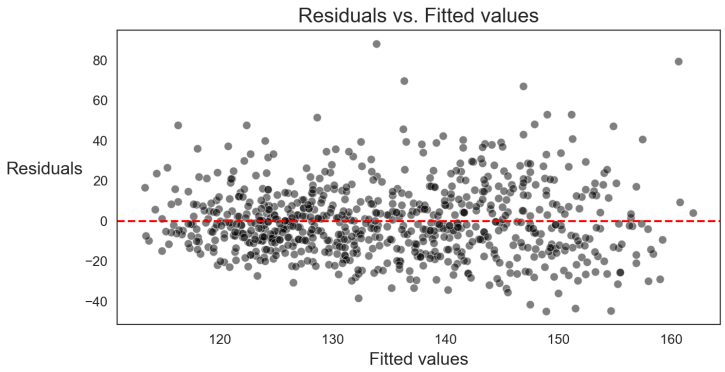

Figure 4: Residuals against fitted (predicted) values. A dotted red line has been added to emphasize that the residuals are approximately evenly scattered around their mean. The evenly distributed residuals imply that they are homoskedastic. Since the residuals are not following any clear pattern, it can be concluded that they are independent. The transparency of the data points has been adjusted to a better visualization of the distribution.

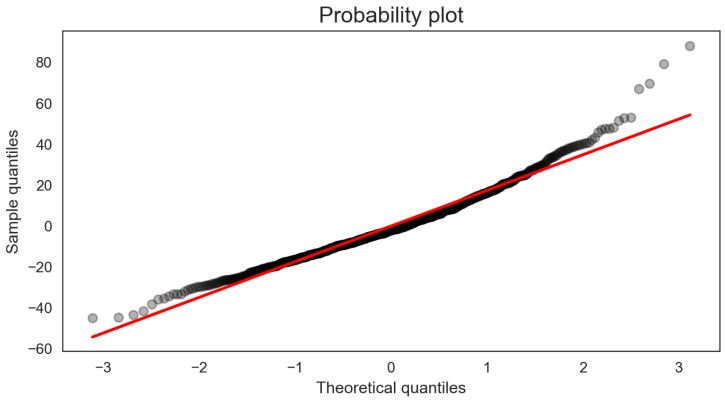

Figure 5: A Q-Q plot of the sample quantiles against the theoretical quantiles from a normal distribution. Since the data points in the Q-Q plot lie approximately on the identity line $y = x$, this indicates that the two distributions being compared are similar. Hence, we can conclude that the residuals are approximately normally distributed.

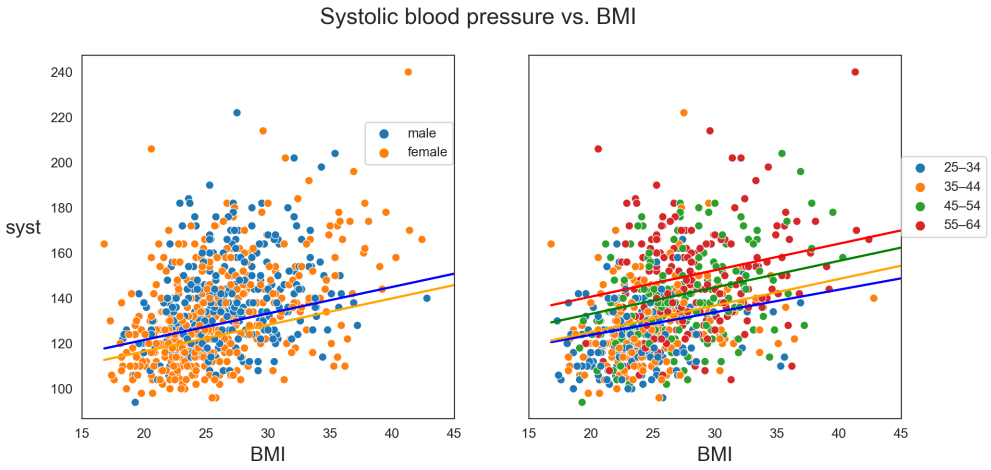

Figure 6: Two scatter plots of the systolic blood pressure vs. BMI side by side with a shared y-axis. In the left-hand graph the data points have been colored by gender, and in the right-hand graph the data points have been colored by age group. The best-fit lines have been added for both graphs, and they have been colored to match the legends. So for example, in the left-hand graph the blue regression line represents the best-fit line for males.

The regression model was used to predict the systolic blood pressure of an arbitrary person. Let's say that we wanted to predict the systolic blood pressure of a female belonging to the age group of 25-34 and has a BMI of 25. Now, according to our model, the expected systolic blood pressure of this person would be 153 mmHg, with a 95 % prediction interval of (117 mmHg, 189 mmHg). As expected, when moving up on age group (while holding sex and BMI the same), the model will predict a higher systolic blood pressure. The same goes for BMI, although it seems that it doesn't raise the value of the systolic blood pressure as much as age. Also, the model predicts a higher systolic blood pressure for males than for females in all of the age groups with varying BMI's. It is worth noting that in the data set an observation was made that in the oldest age group the mean value of the systolic blood pressure was higher for women than for men. Still, the model predicts otherwise.

5 Conclusion

In conclusion, the linear regression model with the given explanatory variables did not predict the systolic blood pressure well. This can be seen in the values of the performance metrics (R-squared, RSE) as well as in Figure 6. Also, the model wrongly predicted a higher systolic blood pressure for males than for females in the last age group, even tho this was not the case in the actual data set. Age group, BMI and gender were found to be the best predictors of the systolic blood pressure. Interaction between any of the variables used in the model did not increase the R-squared, AIC or BIC values, so the author of this paper does not know what adjustments he could have made to the model to make it perform better. In future papers, a more profound knowledge of statistical testing is desirable for a more scientific approach.