Predicting Blood Pressure

Arden Miller

Executive Summary

A predictive model for systolic blood pressure was constructed. The fitted model is:

$$\text{blood pressure} = 132.33 - 2.753\ \text{age} + 0.0355\ \text{age}^2 + 0.2243\ \text{weight}$$

This model has some value for predicting the average blood pressure for men of a given age (between 28 and 68) and a given weight (between 121 and 247 pounds). It would not be very useful for predicting the blood pressure of an individuals since too much of the variability in blood pressures is not explained by this model.

The Heart Study Data

The data used to create the predictive model for systolic blood pressure was taken from the Los Angeles Heart Study (supervised by J. M. Chapman). The data consists of measurements of the following variables taken for 60 men:

	Variable	Mean	Range
sbp:	systolic blood pressure in mm Hg	121.5	90 – 190
age:	age in years	44	28 – 68
chl:	cholesterol in mg per dl	317.5	240 – 520
ht:	height in inches	69	62 – 74
wt:	weight in pounds	167	121 – 247

Figure 1 contains pairwise scatter plots of the variables. These plot indicate only weak relationships between systolic blood pressure (sbp) and the remaining variables. There are two men that have usually high values for sbp. The plot for weight and height indicates a weak positive relationship between these variables.

The Predictive Model

The best predictive model that I can identify using this data is:

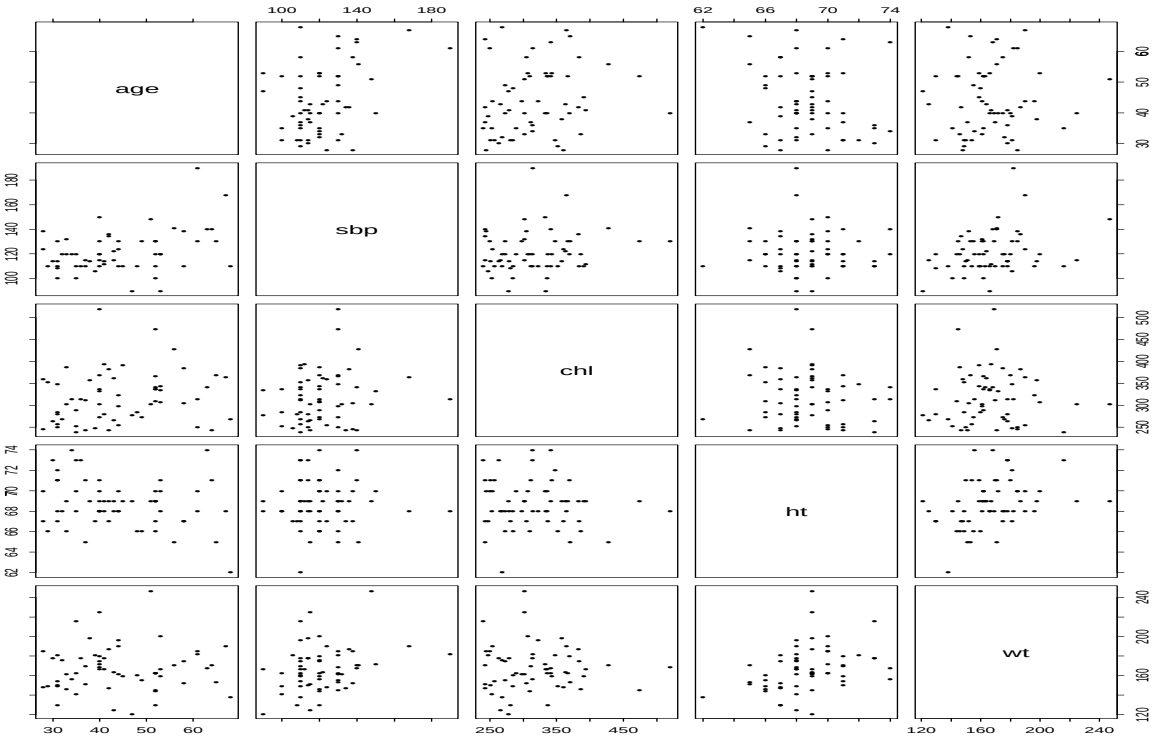

Figure 1: Pairwise Scatter Plots of Heart Study Variables

$$\text{blood pressure} = 132.33 - 2.753\ \text{age} + 0.0355\ \text{age}^2 + 0.2243\ \text{weight}$$

This model does not use either cholesterol or height as predictors since they do not increase the predictive power enough to warrant their inclusions. The model is only valid for the range of conditions encompassed by the data. This means that the model should only be used to predict the blood pressure for men aged 28 to 68 who weigh between 121 and 247 pounds.

This model indicates that for men of a fixed age, blood pressure increases in a linear manner with weight. For each increase in weight of 1 pound, blood pressure increases by an average of 0.22 units. The relationship between age and blood pressure is more complicated. Figure 2 contains a plot of systolic blood versus age given that weight is fixed at 167 pounds (the mean weight for this data). This plot suggests that blood pressure is lowest for men in their late thirties and early forties. After this point blood pressure tends to increase with age.

This model explains only about 30 percent of the variability in the systolic blood pressure measurements for the data. This isn't surprising as measurements taken on people are often highly variable and therefore hard to predict. The model will not be very useful for predicting the systolic blood pressure of individuals. In order to get an interval estimate that captures the true blood pressure of an individual at least 95% of the time we need to take the estimate from our model ± 33 units. This is quite large considering that the range of systolic blood pressure values in the data is only 100 units. The model is more useful for predicting the average blood pressure for men of a specified age and weight. To get an interval estimate that captures this average blood pressure at least 95% of the time we need to take the estimate from our model ± 11 units.

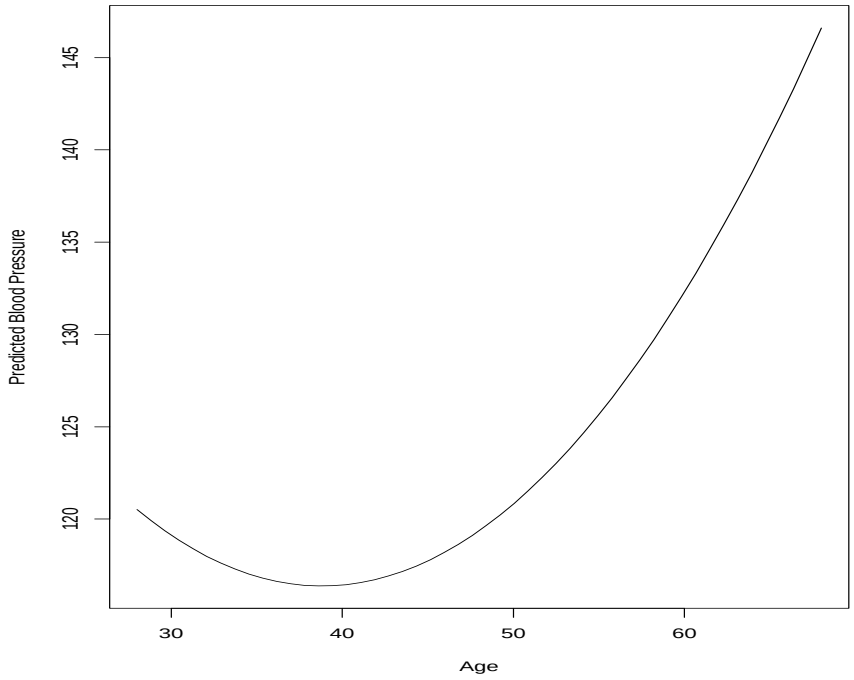

Figure 2: Relationship between Predicted Blood Pressure and Age

There are 2 observations in the data that are having a large impact on the fitted model. One of these corresponds to an individual with very high blood pressure and the other to the oldest subject who happened to have very low blood pressure given his age. If these observations are dropped then the fitted model becomes:

$$\text{blood pressure} = 155.13 - 3.393\ \text{age} + 0.0422\ \text{age}^2 + 0.1717\ \text{weight}$$

Predictions made using this model can vary by up to 5 units from the previous model over the ranges of age and weight in the data. This is not a large difference considering the amount of uncertainty in our predictions (see previous paragraph). Thus I would recommend using the initial model.

Statistical Appendix

First I tried fitting a basic regression model that simply contained linear terms for each of the explanatory variables. A partial plot indicated that it might be beneficial to add a quadratic model for age (see Figure 3). I tried adding the quadratic term for age and found that it was significant. However, neither ht nor chl was significant so I tried dropping these regressors one at a time (ht first and then chl). Thus I ended up with a model that contained a linear term for wt and a linear and quadratic term for age.

To evaluate the precision of predictions made using this model, I created 95% CI's for $E(Y)$ and 95% predictions intervals for individual observations at each point in the data set. I found the margin of error for the CI's varied from ± 5 to ± 11 except for 1 point which was ± 15. The margin of error for the PI's varied from ± 30 to ± 33. Given that the entire range of sbp for the

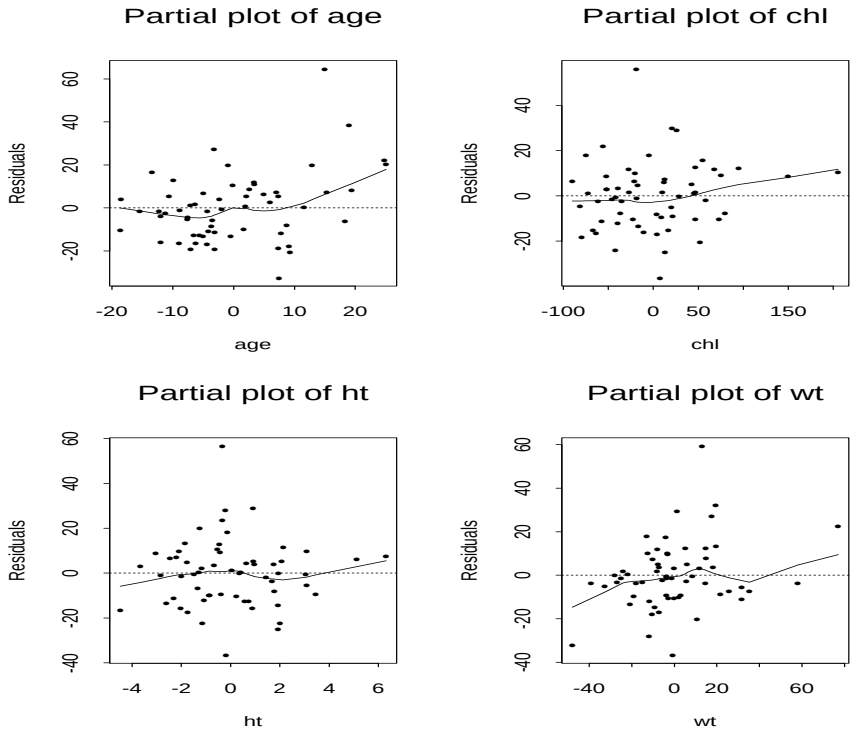

Figure 3: Partial Plots

data is only 100 units, this model is not very useful in predicting blood pressure for individuals but has some use in estimating $E(Y)$.

The residual plots for this model are given in Figure 4. These plots do not indicate problems with linearity, non-constant variance or normality. However, it appears there is evidence that observation 5 is an outlier and that observation 60 may be influential. The results from influence.measures indicate that these 2 points both have very large values of Cook's Distance indicating a large impact on the fitted regression coefficients. Clearly observations 5 and 60 have significantly more influence on the fitted model than any of the other observations – some the other observations were "starred" but these all had much smaller values of Cook's Distance.

	Intercept.	poly.age..2.1	poly.age..2.2	wt	dffits	cov.ratio	cooks.d	hats	inf
5	-0.304	0.871	0.435	0.384	1.188	0.375	0.271	0.073	*
60	-0.435	-0.819	-0.976	0.390	-1.406	0.943	0.454	0.247	*

The DFBETAs indicate that the coefficients for age and age squared will be affected most by dropping these points. Observation 5 also has an extreme value for Covariance Ratio indicating it may have a large impact on the size of confidence intervals produced using the model. I tried refitting the model without these 2 observations. The fitted coefficients changed enough that I decided to report this model as well (see report). I compared the fitted values (for all points in the original dataset) for this model to those for my original model and found that the biggest difference was about 5 units but all the rest were less that 3 units. Given that the margins of error for 95% CI vary from ± 5 to ± 15 the models are not that different from each other. Therefore I recommended using the original model.

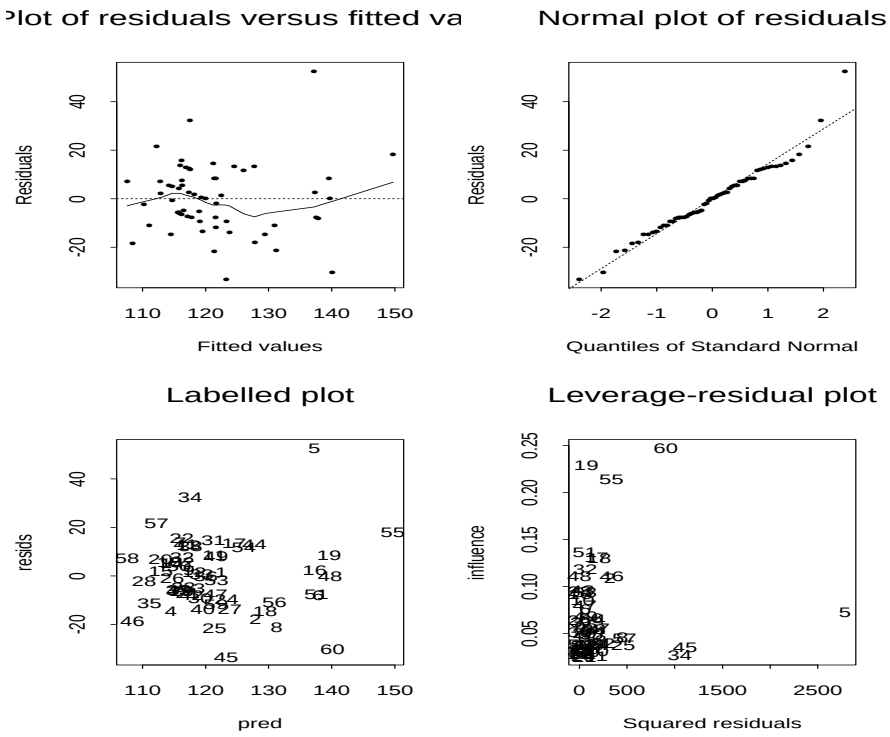

Figure 4: Residual Plots

475.330 Assignment 2: Marking Guide

This assignment asks the students to build a predictive model for systolic blood pressure using four explanatory variables. The assignment is worth a total of 20 marks.

The assignment should consist of two parts: a report that is understandable by a non-statistics major and a statistical appendix that explains their analysis.

The Report

The report should contain the following:

1. An executive summary. This should be a short paragraph that summarises their report. In this case it should indicate that they have created a predictive model and give a brief assessment of the model.
2. A brief description of the data. They should at least have a table that contains the means and ranges of the variables. A pairs plot is helpful but not absolutely necessary for this data.
3. Their fitted model. It should be clear how the model can be used to obtain estimates for blood pressure. If they have transformed the response they should indicate how to get estimates in the original units. There are number of possible models they may select:
   (a) Response is sbp; regressors are age, age^2 and wt. I suspect that this is the model most will use.
   (b) Response is sbp; regressors are age, age^2, ht, ht^2, and wt. The P-value for ht^2 in this model is about 0.06 – if ht^2 is dropped ht is not significant.
   (c) Response is log(sbp) or sbp$^{(-1)}$; regressors are age, age^2 and wt. Some people will have transformed the response to improve the residual plots. Although, I don't think this is necessary it is not wrong (unless they use a transformation that makes the residual plots worse).
4. They should explain what their fitted model indicates about the relationship between blood pressure and the explanatory variables.
   - If they use a polynomial model for age then they should use a graph to illustrate how sbp changes with age.
   - If they have transformed the response, then they should use graphs to illustrate the relationships between blood pressure (not transformed) and the explanatory variables.
5. They should evaluate (in non-technical language) the precision of predictions made using their model.
6. If they identify any points that have a large impact on the model, they should explain the consequences of deleting those observations.

Statistical Appendix

The statistical appendix should contain:

1. A justification of the model they selected. This should consist of a brief description of how they identified their model – it should not consist of computer output for all the models they tried.
2. A discussion of diagnostic plots and influence measures for their model. If they identify high influence points they need to discuss the impact of dropping these on the fitted model.
3. An explanation of how they evaluated the predictive ability of the model. They should have looked at the width of confidence intervals and prediction intervals created using their model.

Allotment of marks

- Give 5 marks for presentation. The writing should be clear and concise. The report and the statistical appendix should include all the parts listed above. In the report, look for explanations that would be understandable to non-statistics students. Graphs should have informative captions. The statistical appendix should give a explanation of their analysis.
- Give 4 marks for identifying a suitable model, presenting the model clearly in the report, and explaining how they selected their model in the statistical appendix.
- Give 3 marks for an explanation of their model. They should explain how each of the explanatory variables affects blood pressure (they will need to use a graph(s) if they have transformed either the response or any of the regressors). They also need to comment on the range of values for the explanatory variables over which the model is valid.
- Give 4 marks for a sensible evaluation of the precision of predictions. To do this properly they will need to look at the size of confidence intervals and prediction intervals produced by their model in their statistical appendix. They should summarise their findings in non-technical terms in their report.
- Give 4 marks for a discussion of diagnostic plots and influential points. They should discuss the impact of deleting influential points – they should not simply delete points because they are "starred" by influence.measures.

Note:

- Include short comments indicating why a student has lost marks.