Report exploring the relationship between blood cholesterol levels and dietary intake using a simulated dataset

Student 202078239

1 Introduction

Prediction modelling in health sciences is used to screen for diseases and other health issues such as obesity, high cholesterol and heart disease that may be prevented with early intervention. In this report I use variable selection techniques to build a model to predict weekly cholesterol levels based on the weekly intake of a chosen selection of dietary components. Ideally, we want to minimise the number of independent variables in the model whilst retaining enough information to predict accurately. There are many reasons to simplify a model including improved model performance, cost, practicality and ease of interpretation and generalisation. Knowledge and research usually help inform variable selection decisions but here we use a simulated dataset so the model will be formulated using statistical techniques alone. The independent variables for consideration include $Patient$, $Week$ and 27 dietary components named D_1 up to D_{27}. These readings are taken over the 7 days preceding the day of the cholesterol, $CHOL$ reading.

2 Method

The Coefficient of Determination, R^2 is a measure of the goodness of fit of a linear model. It is the proportion of the variance in the data that is explained by the regression model and is related to Pearson's correlation coefficient, r, by $R^2 = r^2$. First I will investigate the strength of correlations between the covariates and the $CHOL$ readings. I will then check for any strong correlations between covariates and consider options to avoid any collinearity issues that can multiply or cancel out the effects of the covariates due to holding the same information. Akaike's Information Criterion (AIC) will be used to quantify and compare the fit of models within forward, backward and stepwise model selection, which consider adding and removing variables one at a time and Adjusted R^2 and Mallow's Cp will be used in Leaps and Bounds selection. This uses the "leaps" package in R and approximates the process of all possible subsets by considering non-nested models. Larger values of Adjusted R^2, and smaller values of Cp and AIC indicate better models. Finally, the prediction sum of squares, PRESS statistic will be calculated using the prediction residuals, as a form of cross validation of the model to quantify how well the model should generalise to new data. We want this prediction error to be as small as possible and this is the statistic that I will use to compare between all the models.

3 Results

The data has been standardised so I add a constant value of 3 to all the data, except for $Patient$ and $Week$ to make all values positive for any transformations that I may make. The correlation coefficient between the weekly intake of each dietary component and its corresponding weekly $CHOL$ readings is calculated and summarised in Figure 1.

The dietary components that are most strongly correlated with the $CHOL$ readings are D_{14}, D_{19}, D_5, D_4 and D_{26} and their associated correlation coefficients are -0.868, -0.818, 0.768, 0.736 and 0.725 respectively to 3 d.p.. The linear regression model for $CHOL$ using these 5 covariates, has intercept and gradient parameters:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.07548    0.20751  19.640  < 2e-16 ***
D14         -0.40739    0.01637 -24.888  < 2e-16 ***
D19         -0.40851    0.02592 -15.760  < 2e-16 ***
D5           0.12321    0.03628   3.396 0.000711 ***
D4           0.25687    0.02463  10.428  < 2e-16 ***
D26          0.18737    0.04413   4.246 2.38e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.346 on 990 degrees of freedom
Multiple R-squared:  0.8795,Adjusted R-squared:  0.8789
F-statistic:  1446 on 5 and 990 DF,  p-value: < 2.2e-16
```

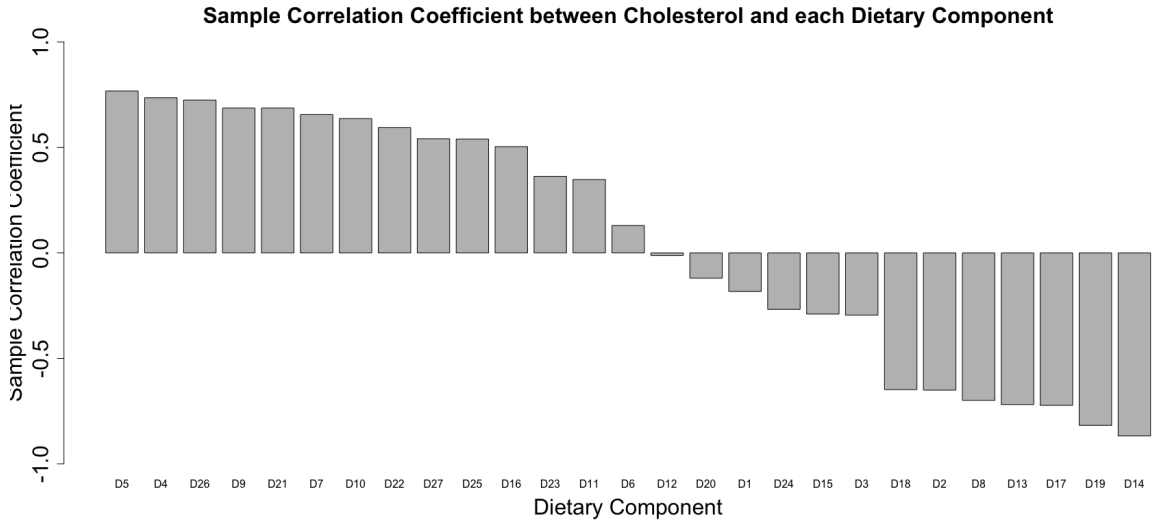


Figure 1: Graph to show which dietary components are most highly correlated to cholesterol, and whether they increase or decrease levels of cholesterol.

These parameters give the equation of the model to be:

$$CHOL = 4.075 - 0.407D_{14} - 0.409D_{19} + 0.123D_5 + 0.257D_4 + 0.187D_{26} \tag{1}$$

In other words, for example, if D_{14} increases in value by 1 unit, $CHOL$ will increase by 0.407 units and if all model covariate values equal 0, then the background $CHOL$ would be 4.075 units. As shown in the model summary output, all the covariates are highly significant, with a p-value < 0.05 and the Adjusted R^2 value of the model is 0.879, indicating that the model is a reasonably good fit. The PRESS statistic for later comparison is 119.8094. D_{14} and D_{19} have the strongest correlations with $CHOL$, and their associated parameter values in the model also provide the greatest effect on $CHOL$. However, D_5 which has the next highest correlation coefficient has the lowest parameter value in the regression equation. This is because correlations describe how closely two variables are linearly related, i.e. how closely the data follow their regression line, but not how large an effect they have on each other, which is described by the gradient of the simple linear regression line, or the parameter value of the covariate in multiple linear regression. As seen in Figure 2, D_5 shows a non-linear relationship. To use this variable successfully in a linear model we should transform it to look more linear to avoid misleading predictions. So, despite D_5 being highly correlated with $CHOL$, a linear model of these variables ignores the polynomial nature of the distribution.

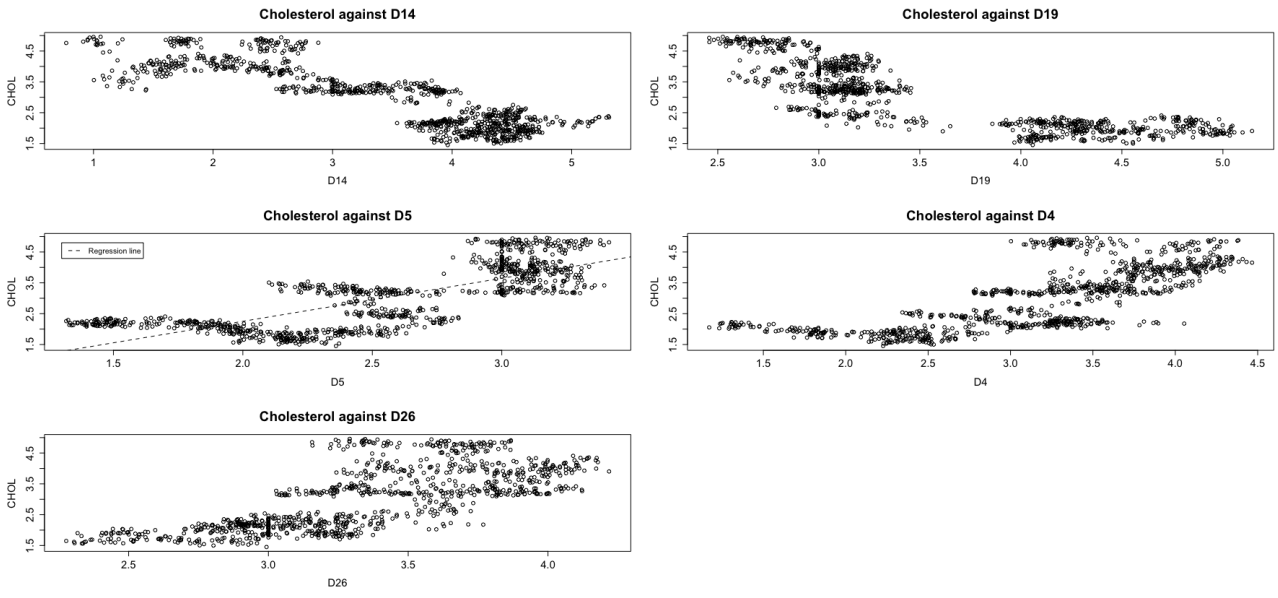


Figure 2: Individual effects on cholesterol of the 5 dietary components with the highest correlation coefficients.

We can test how well this model describes the observed data using ANOVA (Analysis of Variance) to partition the variability into variability explained by the model and variability that is not accounted for. The calculated F-statistic, F^*, shown in Table 1 is 1445.725. This is the ratio of the explained and the unexplained variability, corrected for degrees of freedom. We conduct an hypothesis test, comparing F^* to the critical value of the $F_{5,990(\alpha)}$ distribution, using a significance value of 0.05. This critical value, F is 2.223. Since $F^* > F$, we conclude that our model is explaining variability well and is much better than an intercept only model.

Table 1: ANOVA table to partition variability into explained and unexplained.

Source	Degrees of Freedom, d.o.f.	Sum of Squares, SS	Mean Square, MS (SS/d.o.f.)	F Statistic, F*
Regression	5	865.469	173.094	1445.725
Residual	990	118.531	0.120	
Total	995	984 .000		

Variable Selection

Before beginning variable selection, I construct a boxplot of the $CHOL$ levels for each *Patient*, shown in Appendix A, to check if there are any *Patients* that have significantly different levels from the others. The $CHOL$ levels are similar for all *Patient*s so I won't include *Patient* in the model.

It is important to visualise the data to see if any variables need transformed before using linear regression. Plots of $CHOL$ against each dietary component can be found in Appendix B. Looking at the plots and considering Tukey's Ladder of transformations, it appears that many of the covariates would benefit from being reduced in power, so an alternative would be to consider a transformation of $CHOL$ which may help several of the covariates at once. I also consider using $1/D_{19}$ and polynomial transformations of D_4 and D_5 but these may not be needed if I transform $CHOL$. On inspection of the plots in Appendix B, there appears to be no strong relationship between $CHOL$ and each of D_1, D_6, D_{15} and D_{20}. I therefore remove these from consideration. D_{12} has the lowest correlation coefficient with $CHOL$, however, looking at its scatterplot with $CHOL$, there does appear to be a noticeable relationship, so I will not remove D_{12} at this point. I now check for any covariates that have a correlation coefficient with another covariate of > 0.85 to avoid any collinearity problems. D_{21} and D_9 are identical as they have a correlation coefficient of 1, so I exclude D_{21}. This may be a duplicate set of results for a single dietary component that has been recorded in error. I also remove D_{22}, which is highly correlated with D_{10}, D_{10} having the higher correlation of the two. I then remove D_2, as it is highly correlated with more than one other covariate, and remove D_{16} as it is highly correlated with D_{23} and it looks easier to transform D_{23} into linear form than D_{16}.

The first automatic selection tool I use is the stepwise function using AIC, in the stepwise, forward and backward directions. The stepwise and forward directions give the same model and the backward model differs slightly, having D_3 instead of D_8. The backward model has a slightly lower PRESS value of 28.717 so I proceed with this model:

$$CHOL \sim Week + D_3 + D_4 + D_5 + D_7 + D_9 + D_{11} + D_{12} + D_{13} + D_{14} + D_{17} + D_{18} + D_{19} + D_{25} + D_{26} + D_{27} \quad (2)$$

Following my thought that a transformation of $CHOL$ might be useful, I use the Box-cox function in the MASS package in R to plot a maximum likelihood graph against the parameter of the Box-cox power family. We see which λ value lies between the dashed lines in Figure 3 and decide to transform to $(CHOL)^{1/2}$. This reduces the PRESS statistic to 27.900. I check the model summary to see if there are any non-significant parameter values with a p-value of > 0.05. I remove D_{17} and subsequently D_{13} and D_{18} until all covariates are significant.

Next, I use Leaps and Bounds using Adjusted R^2 to see what model size and variable selection looks promising. From Figure 4 we see that, without transforming $CHOL$, a model with 9 covariates looks like it should hold most of the information.

The 9 covariate model with the highest Adjusted R^2 of 0.970 is found in the leaps output to be:

$$CHOL \sim Week + D_4 + D_9 + D_{11} + D_{14} + D_{19} + D_{25} + D_{26} + D_{27} \quad (3)$$

I repeated using the transformed $CHOL$ and the same covariates were selected with an Adjusted R^2 value of 0.972 and a PRESS value of 28.757. Using Leaps and Bounds I then considered Mallow's Cp criterion. It's best 9 covariate model choice with a Cp of 68.969 is the same as that given by Adjusted R^2. However, the lowest Cp value of 14.377 was using 16 covariates and the best 15 covariate model has only a slightly higher Cp value at 14.427. This 15 covariate model includes the same variables as chosen using AIC.

Next, working with the 9 covariate model, I consider transformations of the covariates. Taking the inverse of D_{19} reduces the PRESS statistic from 28.757 to 25.919. Adding a 3rd order polynomial to D_4 reduces PRESS

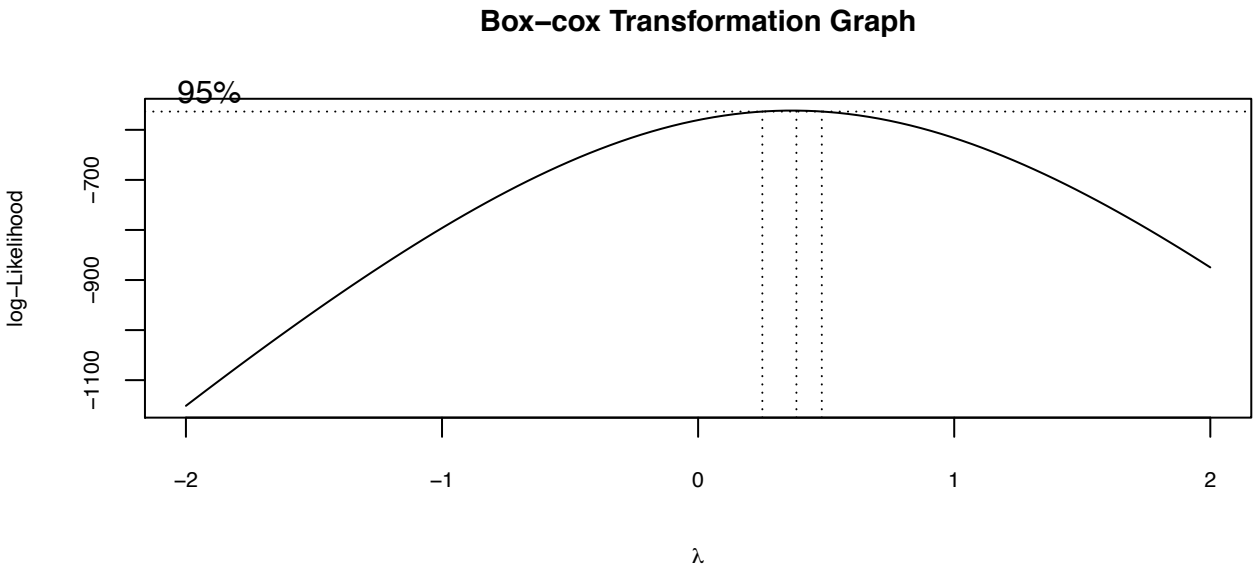


Figure 3: Box-cox transformation graph to indicate which λ power to transform the cholesterol reading to.

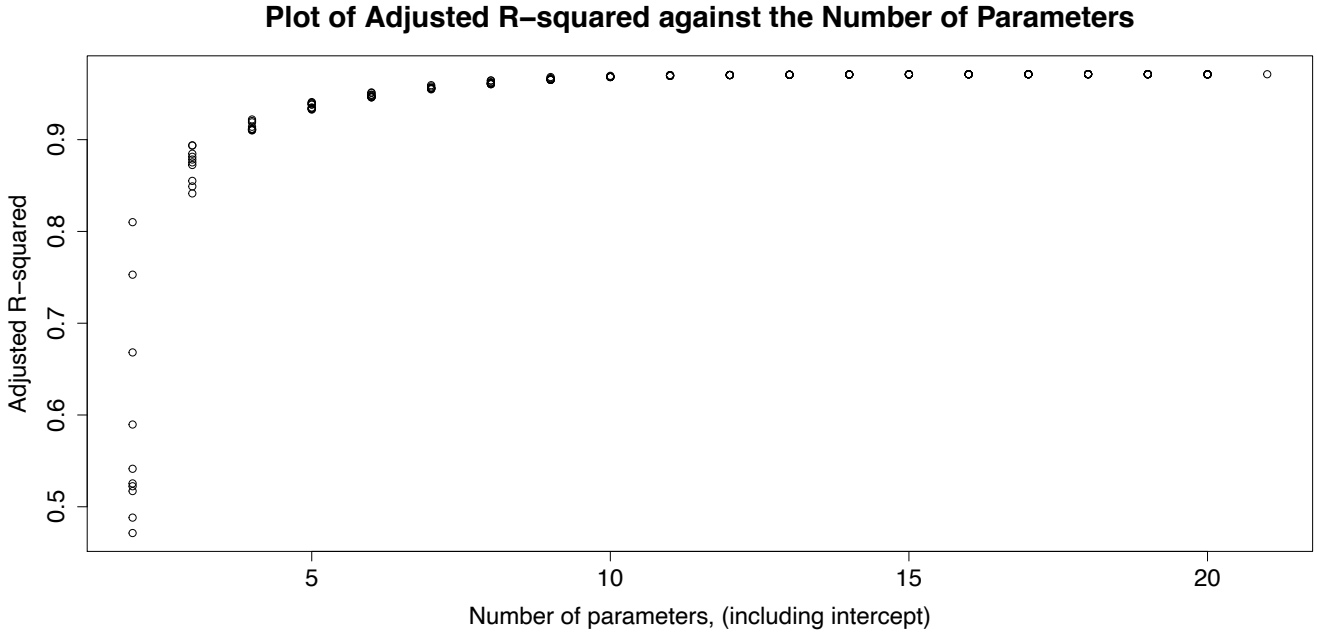


Figure 4: Plot of Adjusted R^2 against the number of parameters in the model, generated using Leaps and Bounds.

further to 24.832. I then use forward selection using hypothesis testing, adding D_7. D_{19} is significant and should be added but I already have the inverse of D_{19} in the model so I replace the inverse with a polynomial for this component, which improves the PRESS statistic to 24.517. I continue with forward selection until there are no more significant variables to add. This builds the model up to 20 covariates with a PRESS value of 22.031. However, on checking the model summary, D_5 is no longer significant so is removed and backward elimination continues until all covariates are significant to 0.05 again. This leaves us with the following 18 covariate model with a PRESS value of 22.948:

$$\begin{aligned}\sqrt{CHOL} \sim Week + D_1 + D_3 + poly(D_4, 3) + D_7 + D_9 + D_{11} + D_{12} + D_{13} + D_{14} \\ + D_{16} + poly(D_{19}, 3) + D_{20} + D_{22} + D_{24} + D_{25} + D_{26} + D_{27}\end{aligned} \tag{4}$$

For ease of viewing, my notation of, for example, $poly(D_4, 3)$ relates to $D_4 + D_4^2 + D_4^3$. I compare this large model with my 9 covariate model, now including the D_{19} polynomial, which has a PRESS value of 24.640:

$$\sqrt{CHOL} \sim Week + poly(D_4, 3) + D_9 + D_{11} + D_{14} + poly(D_{19}, 3) + D_{25} + D_{26} + D_{27} \tag{5}$$

To investigate further, I compare the diagnostic plots for each of these models in Figure 5. For a linear regression model to be valid the residuals must be normally distributed, independent and have constant variance. The residuals versus fitted plot for the 9 covariate model, shows a slight curve suggesting a quadratic term may be missing but the residuals are fairly constant. The QQ-plot shows normally distributed residuals. The spread-location plot emphasises that the residuals are not quite spread equally along the range of predictors. This and the Normal QQ-plot show that the model fits less well for more extreme data, outside 2 standard deviations of the mean. The plots also highlight some outlying data points. The residuals versus leverage plot shows that there are no points with residuals with high Cook's distance scores so these outliers are not influential to the regression line. The diagnostic plots for the 18 covariate model are very similar. The QQ plot shows the residuals following the straight line through to 3 standard deviations this time. However, the variability in the scale-location plot increases slightly with the more complicated model. This suggests to me that the 9 covariate model is better and that the extra variables added in the other model are not improving the model.

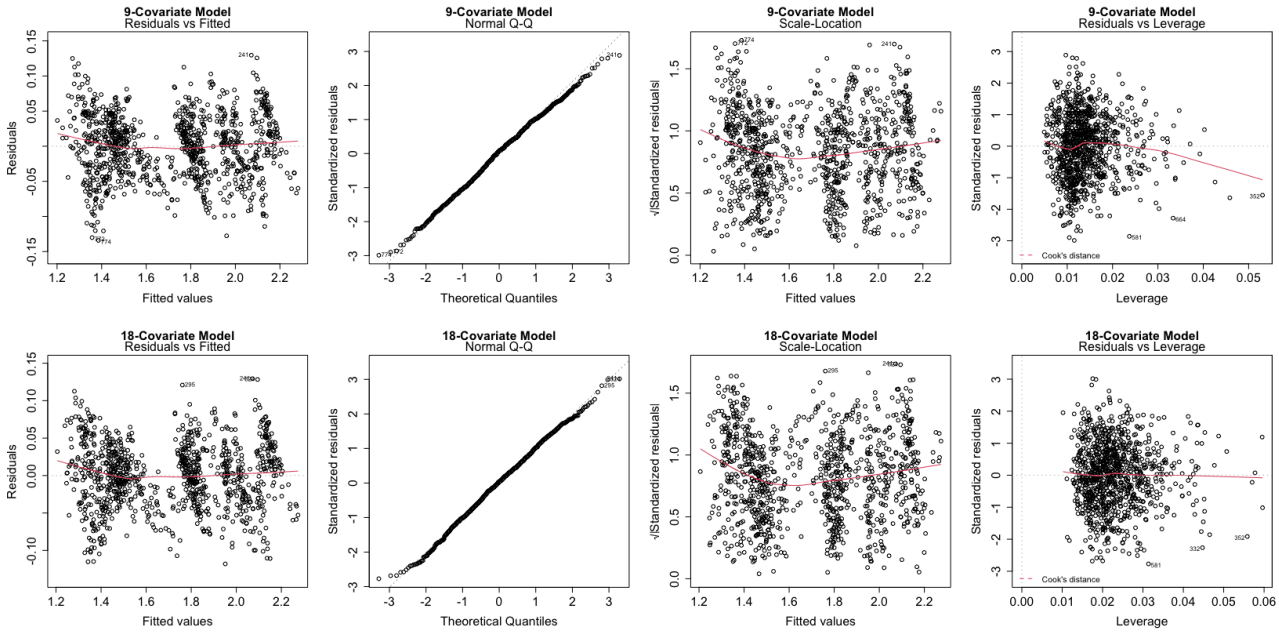


Figure 5: Residual diagnostic plots for the 9-covariate model and the 18-covariate model.

I remove the current transformation of $CHOL$ and check for any better Box-cox transformations. It appears that a cube root transformation is worth considering. However, the diagnostic plots do not improve. So next I check the plots of residuals of the 9 covariate model against all the covariates to see if there are any patterns showing that need to be addressed by including that variable. D_9 shows a definite pattern but D_9 is already in the model so I try transformations on it so see if this helps. Adding increasing orders of polynomial transformation to D_9 reduces the PRESS score. Whilst a 3rd order transformation removes the model residuals more, the residuals become less normally distributed. A 2nd order polynomial transformation of D_9 makes the residuals more normal but the residuals are slightly larger and the PRESS value is not as good as with the 3rd order transformation. Since linear regression models must have normally distributed residuals, I use a 2nd order transformation of D_9 as the increase in residuals is minimal. I check the plots of residuals versus covariates

again with the new 9 covariate model. The residual plots all look fine except D_9 still, so I reluctantly increase D_9 to a 4th order polynomial as this satisfies the independence requirement for the linear regression. This keeps the residuals more normally distributed and minimises the residuals but I fear that it is simply modelling noisy data. The diagnostic plots are not markedly different though. On checking for Box-cox transformations again, taking the square root of $CHOL$ is still the most appropriate transformation for the current model. PRESS with the 4th order polynomial in D_9 is 21.653.

4 Conclusion

I have used the PRESS statistic as a means of comparing the models throughout the investigation as shown in Appendix C, minimising the value as much as feasible. I tried to improve the prediction accuracy by transforming variables into a more linear form and by investigating whether increasing or decreasing the number of variables in the model improved prediction. I decided that a 9 covariate model seemed an optimal size. Using the prediction tool on Myplace highlighted the fact that simply aiming for the lowest PRESS value was insufficient for the model to generalise well to other data. The prediction ability of the PRESS statistic is still restricted by the data on which it is calculated, and therefore a simpler, more generalised model is still likely to be more successful on other datasets. One must also remember that a prediction model is only suitable to be used on data that lies within the same range as used in the model. I was concerned that my simplest model breaks the assumption rule of independence due to the D_9 plot with the model residuals. However, I no longer believe that there is a significant pattern to be seen and therefore conclude my final model, with a PRESS value of 24.640 to be:

$$\begin{aligned}\sqrt{CHOL} = {} & 0.006Week + 1.289D_4 + 0.270D_4^2 - 0.379D_4^3 + 0.069D_9 - 0.258D_{11} + 0.034D_{14} \\ & - 2.540D_{19} + 0.587D_{19}^2 - 0.183D_{19}^3 + 0.061D_{25} + 0.073D_{26} + 0.126D_{27}\end{aligned} \tag{6}$$

The 95% confidence intervals for these coefficients are:

	2.5 %	97.5 %		2.5 %	97.5 %
(Intercept)	0.912423148	1.062138064	D14	0.025722197	0.041296453
Week	0.005856027	0.007077612	poly(D19, 3)1	-2.822575297	-2.257801686
poly(D4, 3)1	1.109625459	1.469262852	poly(D19, 3)2	0.449162805	0.725476248
poly(D4, 3)2	0.119260501	0.421031249	poly(D19, 3)3	-0.328802814	-0.036386234
poly(D4, 3)3	-0.489696926	-0.267653067	D25	0.049885676	0.072384738
D9	0.059324440	0.079293111	D26	0.058157606	0.088760503
D11	-0.274765376	-0.240394036	D27	0.110070297	0.141915300

I decided that adding a 4th order polynomial to D_9 would likely be fitting to noise. The simplicity of a simpler model is likely to be not only more generalisable to further datasets, but also for reasons of practicality, it is cheaper to use and easier for other people to understand.

Variable selection is a complex process where we can build a variety of different models using the same dataset, and following the same methodology. At the start we exclude variables to avoid collinearity issues, but our choice of which to exclude is not always obvious. This step of exclusion itself can have a large effect on what variables end up in your model. Equally, the times at which you trial transformations of variables will also have a significant effect of your final model. No doubt, model selection success improves with experience.

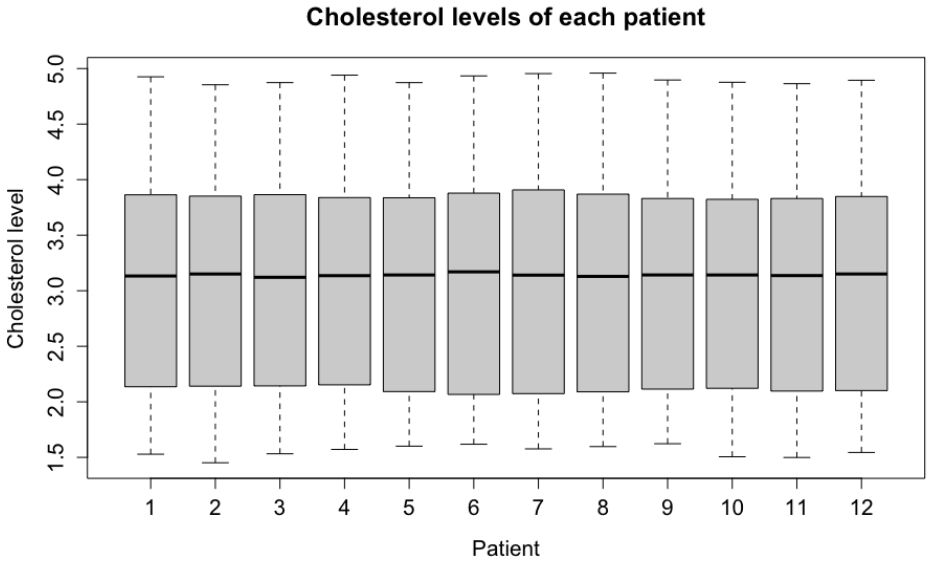
The benefits of forward selection include being less susceptible to *collinearity* issues and that it begins with a smaller model. The benefit of backward selection, however, is that it can assess covariates for their *joint* behaviour. Using an automatic forward or backward selection process either only adds variables until there are no more significant ones to add or removes variables until there are no insignificant variables to remove. Stepwise selection helps with this by adding and removing variables in turn as required, however whether you start this process going forwards or backwards will still bias your model by either focussing more on collinearity issues or variables with joint behaviour. In future investigation I may prefer to start with stepwise selection starting from a backwards direction so my model is more likely to pick up on any joint behaviour there may be, that isn't visible from looking at the data. Knowledge of the subject area would also improve model choice, as you would have prior knowledge of which variables are likely to have an effect on a response variable.

The use that the model is intended for will also affect variables chosen: If we want to know the dietary components that have the largest effect on cholesterol, we would be wise to consider more the strength of the effect (gradient) of the dietary components as well as the strength of correlation between the dietary components and cholesterol. My confidence in the final model lies in the fact that there was agreement between the Mallow's Cp, Adjusted R^2 and AIC statistics when choosing which covariates to include. However, I am not happy with the prediction accuracy of this model and believe there is still visible banding in the data that needs to be addressed.

References

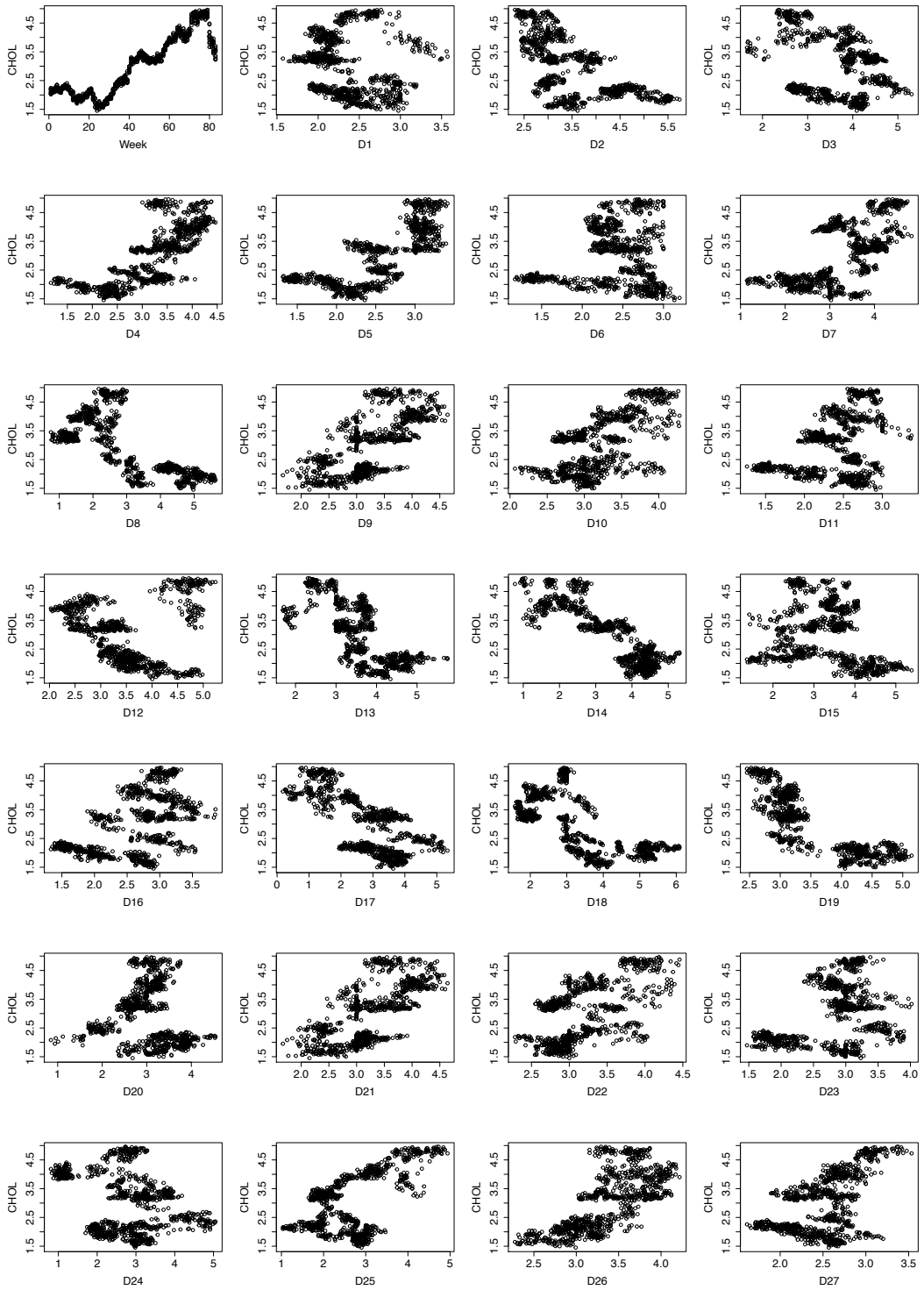
[1] Mohammad Ziaul Islam Chowdhury and Tanvir C Turin. "Variable selection strategies and its importance in clinical prediction modelling". In: *Family Medicine and Community Health* 8.1 (2020).

Appendix A: Graph to show cholesterol levels of each patient



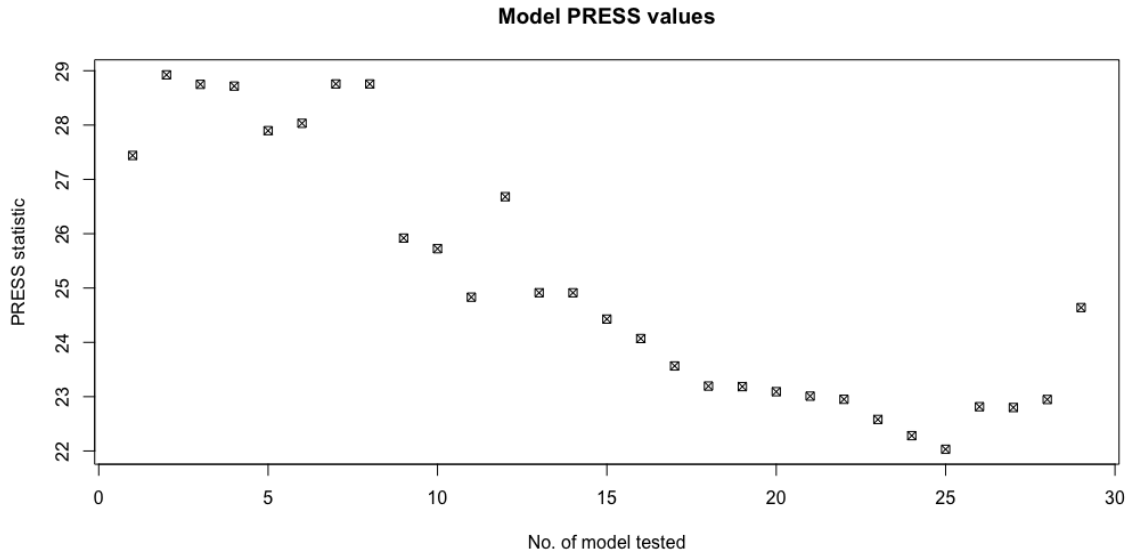
This shows that all the patients have similar cholesterol levels, so we don't need to consider any of them separately.

Appendix B: Plots of cholesterol against each independent variable



Plots of Cholesterol against each individual dietary component to visualise their distributions to see which show an obvious pattern in the data and which data look randomly scattered. It also allows us to see which variables may need transformed for linear regression.

Appendix C: Graph of PRESS over the course of the investigation



This graph shows how the PRESS of the model changed over the course of variable selection. The model with minimum PRESS did not generalise well to the new data using the prediction tool on MyPlace.