Summary MSG500 Fall 2009

Regression, Model Selection, and Classification

1 Introduction

In this class we have focused on regression models, and model selection. This spans quite a large set of tools, and constitute the backbone of applied statistics. In the following summary I will review the most important methods we discussed in class, and I will demonstrate how to use them and how to the interpret the outcome using a heart-disease data set from South Africa (see e.g. Rousseauw et al, 1983, South African Medical Journal, or the book The Elements of Statistical Learning, by Hastie, Friedman and Tibshirani). I will also follow up with a discussion of classification trees and logistic regression.
This is a summary only and should not be used as a replacement of the text or lectures, which covers the methods in more detail. The companion computer code was posted on the class home page earlier in the semester, and you may want to revisit those codes after reading the summary. Similar code is also posted for the pollution data from last lab.
The summary is not an example of a lab report. In a lab report you should keep methods and results separate. Here, however, I am introducing and reviewing methods with the help of the data set, and as you can see there is no separate methods section.

2 Data set

The South African hear disease data set used in this summary is a subset from a larger data set. The data set describes a retrospective sample of males in a high-risk heart-disease region of the Western Cape in South Africa. Each high-risk patient has been monitored and the following patient attributes were obtained: systolic blood pressure (sbp), cumulative tobacco (tobacco), low density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behavior (typea), obesity, alcohol, and age. A total of 462 samples are included in this data set. Adiposity is a measure of % bodyfat, whereas obesity measures weight-to-height rations (body-mass-index, bmi). Type-A behaviour pattern is characterised by an excessive competitive drive, impatience and anger/hostility.

We will try to predict ldl from the other attributes. ldl is the "bad" cholesterol, and we expect high ldl levels to be strongly associated with adiposity and obesity. However, we aim to build a predictive model for ldl using the most informative subset of variables.

Before preceding with the analysis, I split the data set into a training (312 observations) and test set (150 observations).

3 Graphical Exploration

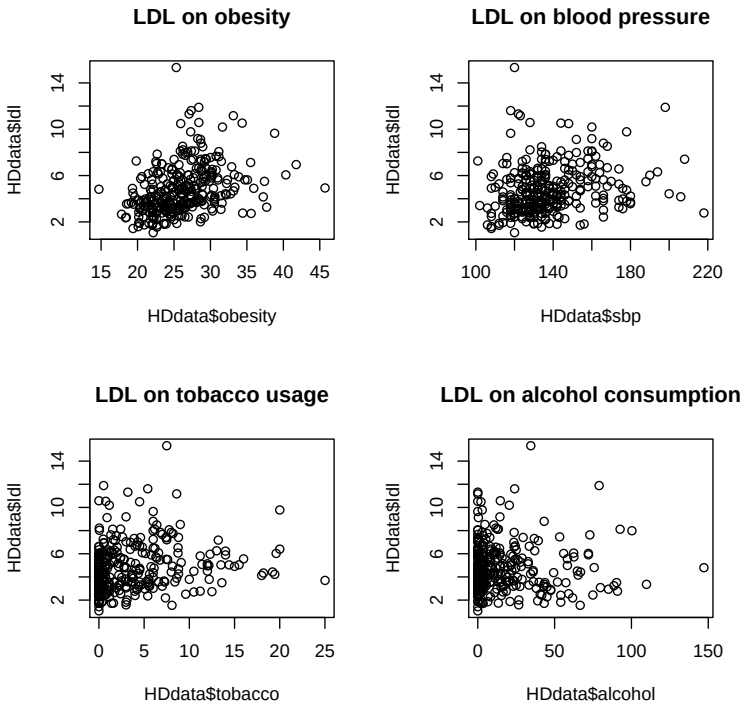


Figure 1: Scatter plots of ldl on obesity, blood pressure, tobacco usage and alcohol consumption. The figures indicate that ldl variance is non-constant across the rage of these attributes, and we may have to transform the data via e.g. log.

We will investigate the relationship between ldl levels and attributes such as smoking status and obesity. We begin with a graphical exploration of the data set.

Scatter plots can be used to detect relationships between numerical variables. You should look for linear vs. non-linear relationships, and use the scatter plots to discover potential modeling challenges such as skewed distributions, or non-constant variance.

In figure 1 we see a clear indication of a non-constant variance of ldl across the range of several attributes. ldl variability increases with obesity levels, blood pressure, tobacco and alcohol consumption. This "fan shape" of the data can usually be suppressed by a log or square-root transform of the response variable. We can also observe in figure 1 that ldl is positively associated with obesity and blood pressure. The relationship between ldl and the tobacco and alcohol consumption is less clear. A possible source for this unclear dependency structure is that the consumption variable isn't truly numerical. The "0" consumption is a special case of no consumption, and this behavior profile should probably be separated from the overall consumption. We achieve this by creating two new variables: indicator variables that are 0 if the reported consumption is 0, and 1 if the reported consumption is non-zero.

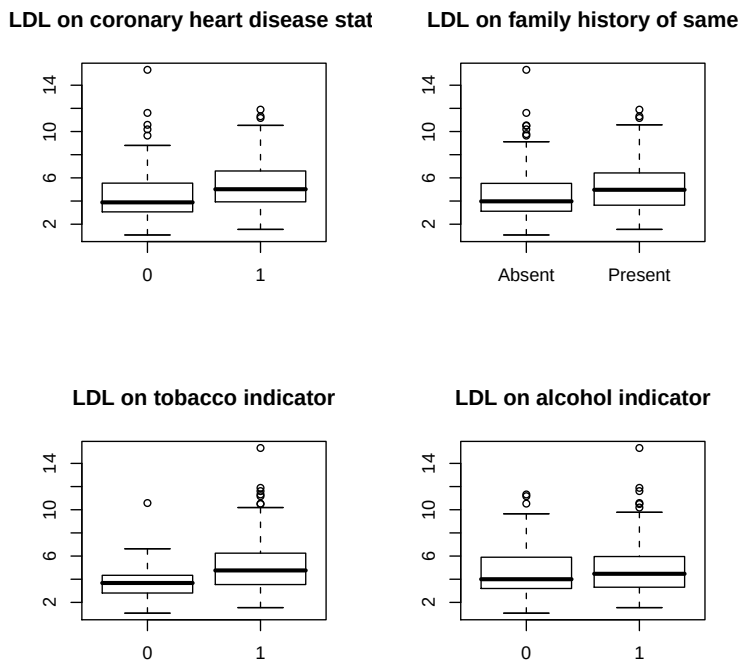


Figure 2: Indicators for heart disease status and family history of heart disease, as well as indicators for tobacco and alcohol usage, are associated with increased ldl levels. The alcohol usage indicator exhibits the weakest association with ldl.

In figure 2 (lower panel) we depict the relationship between ldl and these new indicator variables using boxplots. Boxplots are excellent graphical tools for examining the relationship between numerical and categorical variables. The boxes represent the 50% most central data, the horizontal bar represents the median value. The fences show the 1.5 times Inter-quartile-range, and is generally used to denote the "well-behaved" range of the data set. The circles denote outliers that exceed this "well-behaved" range. In the top panel of 2 we illustrate the relationship between ldl and the two heart disease indicators (self and family history). From figure 2 we deduce that a) both disease indicators show that disease status is associated with ldl levels, b) tobacco usage as a binary variable is associated with ldl level, while c) alcohol usage as a binary variable exhibits a very weak relationship with ldl.

From Figure 1 we concluded that non-constant variance could be a problem. We applied a log-transform to the ldl data to stabilize the variance. Furthermore, we applied a log transform to some of the numerical explanatory variables to enhance their relationship with ldl. Obesity and age were log transformed because the correlation between these variables and ldl increased after taking logs (see Figure 3).

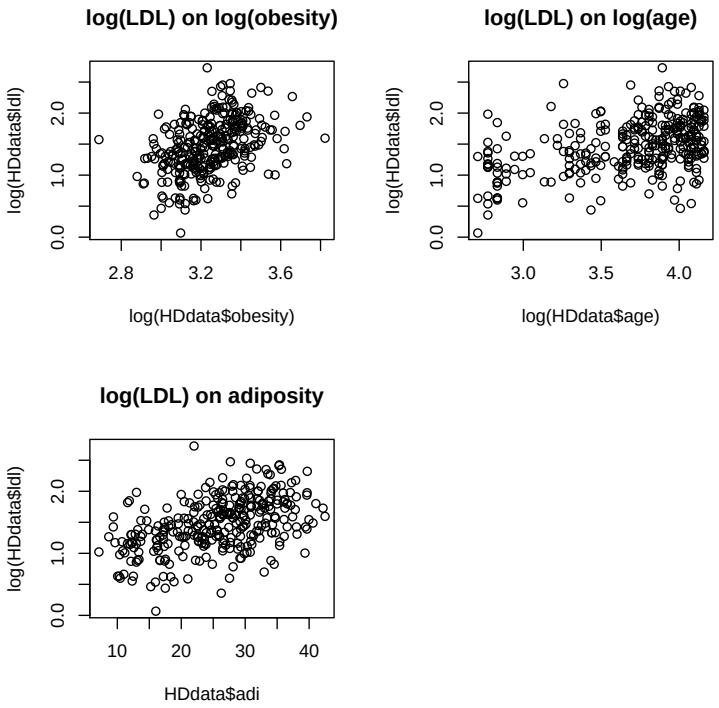


Figure 3: After log-transformation of ldl, obesity and age, the variance of ldl appears to be constant across the range of the attributes, and the dependency appears to be reasonable well characterized with a linear model.

The graphical exploration of the data suggests that there are some linear relationships between the explanatory variables and ldl that can be used to formulate a predictive model of the ldl level of a patient. In the next section we will attempt to use ordinary least squares to fit a linear regression model to the heart disease data.

4 Ordinary Least Squares

The linear regression modeling assumption states the following;

$$y = X\beta + \epsilon,\ E(y) = X\beta, E(\epsilon\epsilon') = \sigma^2 I,$$

where y is a $n \times 1$ vector, X is a $n \times p$ design matrix, and ϵ is a $n \times 1$ error vector. The errors are assumed to be uncorrelated, and have constant variance σ^2. With the exception of the first column, the columns of the design matrix X correspond to the measurements of each explanatory variable. The first column is all 1's, and corresponds to the intercept of the mode.

The most common way to fit a regression line to the data (X, y) is via least squares:

$$\hat{\beta} = argmin\ ||y - X\beta||^2 = argmin\ \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_{p-1} x_{ip-1})^2.$$

The least squares criterion weighs the contribution of each observation i equally. In addition, positive and negative residuals $y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_{p-1} x_{ip-1}$ contribute equally as well. Therefore, it is essential that the following basic assumptions are satisfied:

1. Errors are uncorrelated
2. No outliers
3. Constant error variance
4. Symmetric errors
5. The linear model $X\beta$ is sufficient

Assumptions 1-3 guarantee that an equal contribution from each observation i is sensible. Assumption 4 validates the use of the squared residuals in the fitting criterion. Assumption 5 guarantees that there is an optimal model $X\beta$ to estimate (that a regression line makes sense - in class we discussed some problems that could be present, e.g. groups in the data).

In class, we derived a closed-form solution for the unbiased and minimum variance coefficient estimates:

$$\hat{\beta} = (X'X)^{-1}X'y,\ E(\hat{\beta}) = \beta,\ V(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

As can be seen from the above, the estimation variance of the regression coefficients have three sources of variability:

- σ^2 - the overall noise level of the data
- The sample size $V(\hat{\beta}) \simeq \sigma^2/n$

- The variance of the explanatory variables.

The larger the spread of the individual x-variables, the more of the relationship between y and x we get to observe, and therefore we can pin down the true value of β with more accuracy. However, there is a source for concern. $X'X$, the covariance of the x's, can be close to non-singular if some of the explanatory variables are highly correlated. This is called the multi-collinearity problem. If some x's are correlated, the estimation variance of the $\hat{\beta}$ can get very large, and be correlated. That is, the sign and magnitude of individual coefficient estimates will be near impossible to interpret.

The fitted values from an OLS fit can also be obtained in closed-form:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy,$$

where $H = X(X'X)^{-1}X'$ is called the hat-matrix. We can write

$$\hat{y}_i = \sum_j h_{ij} y_j,$$

where h_{ij} is the $i-th$ row of matrix H. Thus, each observation y_j contributes to the fitted value $\hat{y}_i$. The diagonal elements of H, h_{ii} are called "leverage". h_{ii} measures how much y_i contributes to its own prediction. If h_{ii} is very large, i is called a high-leverage observation. Such observations can pose a problem. If y_i takes on an unusual value at the same time as h_{ii} is large, the regression line will be pulled towards y_i and can create a poor fit elsewhere in the data. An observation y_i can be problematic even if its leverage is not large. If y_i is unusual, but with low leverage, its residual e_i will be large. When we estimate the noise level of the data, $\hat{\sigma}^2 = \sum_i e_i^2/n - p$ (where p is the number of coefficients estimated), a large residual can lead to an overestimate of σ^2.

Let us proceed with a basic linear model fit applied to the heart disease data set. We will use ordinary least squares, and then carefully check the assumptions. In figure 4 we examine the residuals from the first model fit, where all explanatory variables are included. The residual plots do not exhibit any patterns (curve-linear, or skewed errors), and we conclude that the overall linear model assumption seems to hold. However, the normal QQ-plot does bring one problem to our attention. Observation 11 is an unusually large residual. We will examine this residual in more detail in the discussion to follow. In addition, there are 3-4 very high leverage observations. We will return to a discussion of potential outliers in the data set below. The residual standard error $\hat{\sigma} = 0.3632$. In addition, the R-Squared is 0.31, indicating that about 30% of the variability of ldl is explained by the other patient attributes. (Note when you run the demos you might get a different set of outliers - the training and test sets are randomized.)

Basic inference of the regression model is based on t and F-tests. These tests both rely on one additional assumption, namely that the errors are normally distributed. We check this assumption with the residual QQ plot in figure 4. The residuals do seem to follow a normal distribution fairly closely, though observation 11 deviates somewhat and should perhaps be

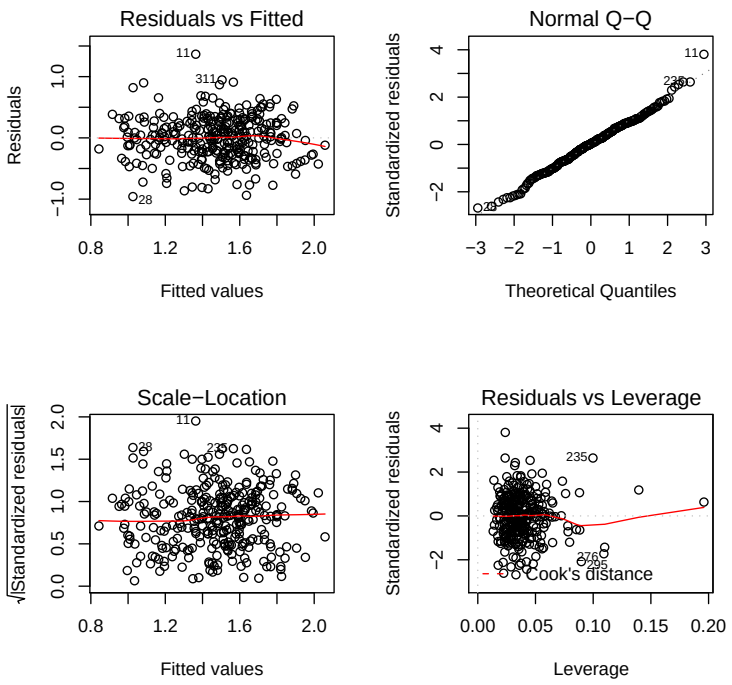


Figure 4: Basic diagnostic plots of a model fit. We observe at least one outlier that deviates from the overall data set, and at least a few high-leverage observations. There appears to be no patterns in the residual plots however, so the overall sufficiency of the linear model is not called into question.

removed prior to performing these tests. We will return to this discussion below.

If we assume that $\epsilon \sim N(0, \sigma^2)$ it follows that $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. However, we don't know σ^2 and instead estimate it by the MSE $\sum_i e_i^2/(n-p)$. Now, if the errors are normally distributed, the residual sum of squares is χ^2-distributed, and

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{jj}}} \sim t_{n-p}.$$

We can thus perform hypothesis testing on β_j using the t-test. In table 1 we examine the t-values for each coefficient, where $t_j = \hat{\beta}_j / SE(\hat{\beta}_j)$, $SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X'X)^{-1}_j j}$ (corresponding to testing the null $\beta_j = 0$). As we can see from the table, the t-based inference identify only 3 variables as significantly related to ldl (at the 5% level): tobacco usage indictor, alcohol usage and adiposity.

To validate the overall fit of the model we use the F-test. We compare the residual sum of square under the full model, to the residual sum of square under the intercept model. If the intercept model is true (the null), then the normalized ratio of these RSS is F-distributed. For this data set our F-statistic is 12.44 on 11 and 300 degrees of freedom, corresponding

Coeff.	Estimate	Std. Error	t value	p-value
(Intercept)	-0.9455862	0.7035160	-1.344	0.17994
age	0.1182099	0.0803304	1.472	0.14219
obesity	0.4150529	0.2140178	1.939	0.05340
chd	0.0841454	0.0487453	1.726	0.08534
famhist	0.0666580	0.0447531	1.489	0.13742
tobind	0.1692175	0.0593127	2.853	0.00463 ***
alcind	-0.0032777	0.0536708	-0.061	0.95134
tobacco	-0.0033543	0.0054797	-0.612	0.54092
alcohol	-0.0022336	0.0009915	-2.253	0.02499 **
adiposity	0.0116767	0.0053684	2.175	0.03040 *
typea	0.0006552	0.0022024	0.297	0.76630
sbp	0.0012328	0.0011870	1.039	0.29985

Table 1: The initial full model fit. Only 3 coefficients are significant at the 5% level. These are: tobacco usage (binary), alcohol usage and adiposity.

to a p-value $< 2.2e-16$. Thus, we reject the null (intercept model), and conclude that the overall relationship between the explanatory variables and ldl is stronger than would be expected by chance alone.

4.1 Outliers

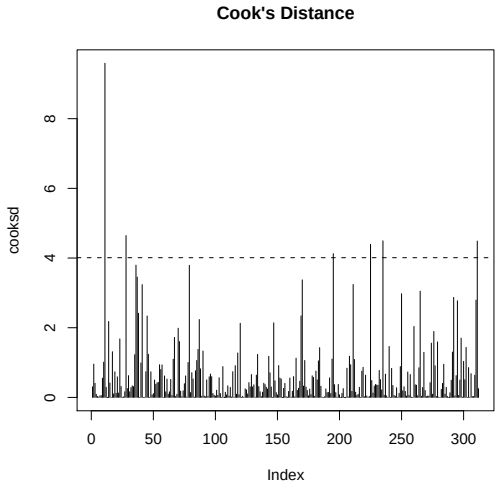


Figure 5: Cook's distance: observation 11 clearly stands out.

As discussed above, observations that deviate from the overall structure of the data can cause problems. These "outliers" come in two forms: influential outliers with high leverage that have an impact on the coefficient estimates, and "pure" outliers that do not have high leverage but impact the MSE (estimate of σ^2). Both of these outliers cause problems in terms of inference. Influential outliers can generate a model that doesn't fit the data as a whole. Pure outliers can lead to a large $\hat{\sigma}^2$, which will then impact the test-based validation. There are many tools for identifying outliers. Here, we will use the Cook's distance: $C_i = e^2_{(i)}/(p\hat{\sigma}^2)$, where $e_{(i)}$ denotes the studentized residual, i.e. the residual for observation i if i is not contributing to the fit. We will also examine the leverage h_{ii}, and the change in coefficient estimates that result from dropping an observation i from the data. In figure 5 we depict the Cook's distance for all observations. Clearly, observation

11 deviates from the overall data set. We will drop observation 11 from the data set and refit without it.

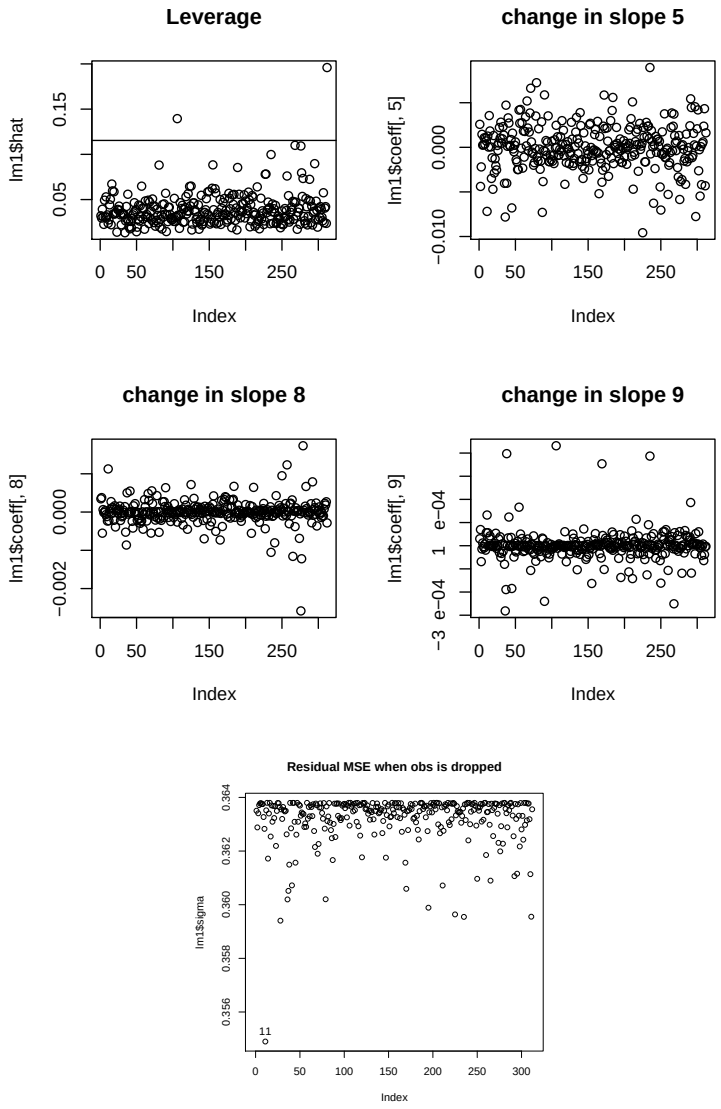


Figure 6: Top: The leverage of observations i, as well as observation i's impact on the 3 significant slope coefficients. Observation 243 may constitute a problem. Bottom: The impact on $\hat{\sigma}^2$ when an observation is dropped.

In figure 6 we examine the impact of observations on the coefficient estimates. While a few observations did seem to affect the coefficient estimates, we did not deem the effect large enough to be a concern. The R-squared values did not chance much after we dropped one of these observations, nor did the sign or significance of the coefficient estimates. When observation 11 was dropped, the residual MSE decreased by a large amount. In conclusion, we identified observation 11 as a pure outlier, whose impact on the MSE estimate was severe. We will drop observation 11 from the analysis and refit the model without it.

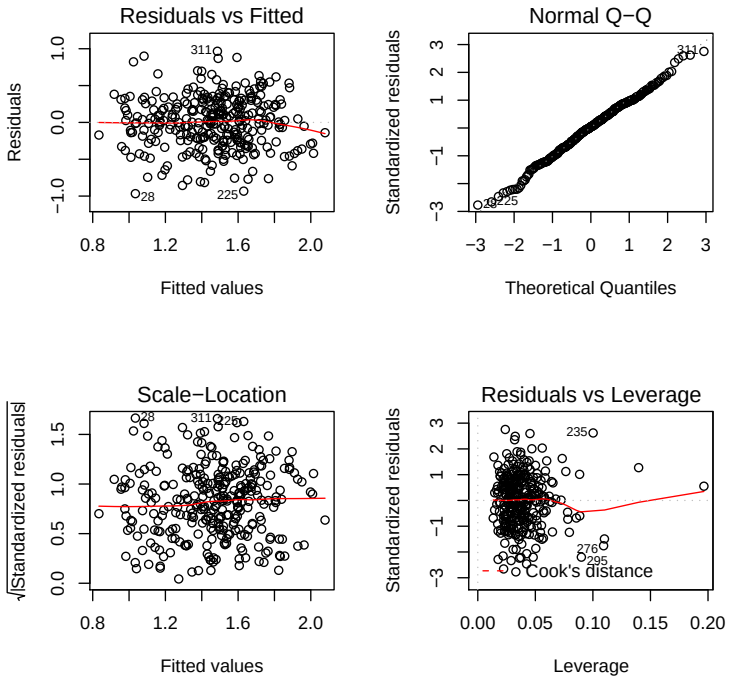


Figure 7: Basic diagnostic plots after observation 11 is dropped. The residuals follow a normal distribution quite closely, and not extreme outliers are detected.

In figure 7 we review the basic diagnostic plots once the outlier (observation 11) was removed from the data set. There are no patterns in the residuals, and the residuals seem to follow a normal distribution closely. The sign and significance of the estimated coefficients did not change much (results omitted) after the outliers was dropped. The residual standard error decreased to 0.3549, and the multiple R-Squared increased to 0.3279.

4.2 Interactions

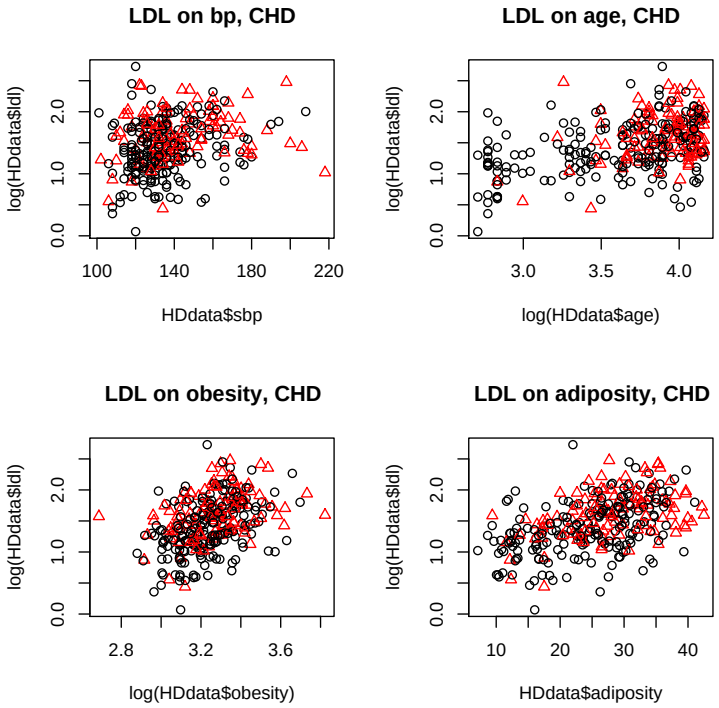


Figure 8: Interactions: No strong interactions were detected via graphical analysis. In the figure, we see some potential for an interaction between obesity and chd, and adiposity and chd.

Since many categorial variables were included in the data set, and health status and ldl most likely follow a complex relationship, it was necessary to examine the data set for potential interaction effects. In figure 8 we investigate the potential interactions between heart disease status (chd) and some numerical variables. This figure constitutes a small set of investigations only. Our full graphical analysis of the data set did not lead to the detection of any strong interaction effects. While figure 8 does hint to a potential interaction effect between chd and obesity, and between chd and adiposity, the effects are deemed too weak to model (they may be spurious). Therefore, we will use our full additive model as a starting point for model selection and predictive model building.

5 Model Selection

There are two main goals behind model building: data interpretation and prediction. For both goals we are better off if we keep the model as simple as possible. Model selection is the common name given to statistical procedures that try to identify the most simple model to interpret or predict an outcome using explanatory variables.

The F-test can be extended to perform model selection. If we compare the residual sum of square of a complex model RSS_c and a simple model RSS_s, the F-ratio (under the simple model - null)

$$F = \frac{(RSS_s - RSS_c)/(p_c - p_s)}{RSS_c/(n - p_c)} \sim F_{p_c - p_s, n - p_c}$$

where p_c and p_s refer to the number of coefficients in each model respectively. To use the F-test for model selection we need to specify the subset models (simple and complex) we want to compare. In general, these models are not pre-specified, and we have to search for the best subset model.

The stepwise F-test is often used to simplify a model. We start with the full model, and then drop one variable at a time. Each time we drop a variable we perform an F-test. If we don't reject the simple model (with one variable dropped), we take the simple model as our current model and consider dropping one more variable. As soon as the F-test leads to a rejection we halt our search. This is called a backward search. We can also use the stepwise F-test to build a model from the intercept up. In this forward search we start with the intercept and add one variable at a time. As soon as the F-test leads to a non-rejection we halt our search. The problem with stepwise searches is that they are greedy - once a variable is dropped(added) we can't go back on the decision. However, we can improve on the stepwise searches if we also allow for moves in both direction - at each step you can choose to go forward or backward.

Applying a stepwise backward search to the heart disease data we eliminate many of the variables. The final set of explanatory variables selected are sbp, obesity, famhist, chd, alcohol, adiposity, and tobind. That is, we dropped age, alcind, tobacco and typea. However, we cannot conclude from this stepwise search that e.g. age is unrelated to cholesterol. Age, adiposity and obesity are correlated ($\sim$ correlation 0.4), and our regression model search cannot tease out the contribution of each individual predictor. Similarly, blood pressure is correlated with adiposity and age. The most difficult case of collinearity exists between obesity and adiposity (correlation .77). Of course, the indicator variables we constructed are correlated with the consumption measures.

As an alternative to stepwise F-tests, we discussed a few other selection criteria in class. The Cp and AIC criteria estimates the expected size of the gap between the MSE and the prediction MSE:

$$Cp = RSS + 2 * p\sigma^2$$

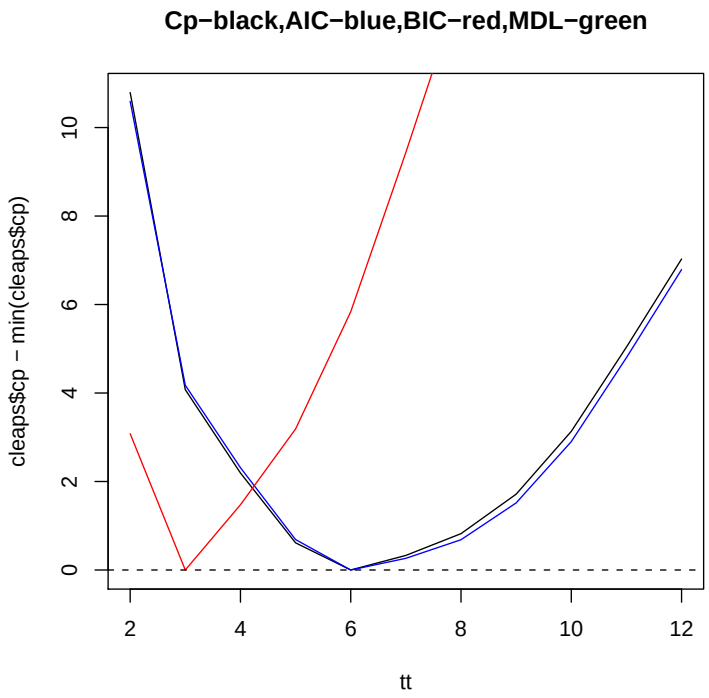


Figure 9: AIC (blue), Cp (black) and BIC (red) selection curves. BIC selects the smallest model (only 3 coefficients).

$$AIC = n\log(RSS) + 2 * p.$$

The BIC criterion penalizes the model size more than either Cp and AIC:

$$BIC = n\log(RSS) + p * \log(n).$$

We split the training data set into 2/3 for model selection and 1/3 for prediction error estimation. In figure 9 we depict the AIC, Cp and BIC curves obtained on the 2/3 training data set. We can clearly see that BIC is more conservative than AIC and Cp. BIC selects a model with 3 coefficients only: intercept, adiposity and chd. AIC and Cp both select a model of size 6: intercept, adiposity, age, typea, alcohol, chd and tobind. For this particular random split, the Cp and AIC prediction errors (on the 1/3 left out) was 0.1519, whereas the more simple BIC model prediction error was 0.1522.

Repeating the split one more time, we obtain a different set of models. BIC now selects a model with 4 variables: adiposity, famhist, alcohol and tobind. AIC and Cp again agree and select a model with 6 variables: adiposity, famhist, obesity, alcohol, age and tobind. The BIC model prediction error is 0.1066 and the Cp/AIC prediction error is 0.1074.

If we repeat the split again, we obtain yet another model. Clearly, the multi-collinearity and overall noise level of the data makes model selection a difficult task. We will examine the instability of model selection in more detail in the next section.

6 Cross-validation

To break away from using model selection criteria like AIC and Cp, we can also let the data decide on the actual model to select. This is called cross-validation. We split the data into K parts, letting each $1/K$-fraction of the data take turns to be the test set. When part k is the test set, all other parts of the data come together to form the training set. Each model is fit to the training set, and then evaluated on the test set. The final prediction error for each model is averaged across all K test sets.

1. Split the data set into K parts
2. Enumerate models to compare: $M_1, \cdots, M_J$
   - for $k = 1, \cdots, K$, fit the model M_j to the training data $data[-k]$ where component k is not included in the fitting.
   - For each component k and model M_j, compute the prediction error $PE_{k,j} = \sum_{i \in k}(y_i - \hat{y}_i[M_j, -k])^2 / \sum_{i \in k} 1$, where $\hat{y}_i[M_j, -k]$ is the prediction of observation i with model M_j fit to the training data excluding part k.
   - Final prediction error for model M_j is $PE_j = \sum_k PE_{k,j}/K$
3. Select model j^* where $PE_{j^*} < PE_j \; \forall j \neq j^*$, i.e. the model with the smallest prediction error

7 Model Selection Instabilities - Random Splits

Model selection can be difficult if there are many variables to choose from, and these variables are related (collinearity problems). We can use re-sampling techniques to examine the selection instability in more detail and also to evaluate a model's predictive performance.

A simple re-sampling scheme is the following:

1. Randomly split the data into training and testing
2. Apply a model selection criterion like AIC or BIC to the training data to identify a subset model
3. Record which variables were selected
4. Apply the selected model to the test data and record the prediction error
5. Repeat

The results can be tabulated to identify informative variables (that are frequently selected), and models that are frequently selected.

Cross-validation is just another selection criterion, and will also produce different models if we perturb the data (i.e. start with a different training data set altogether). To investigate model selection instability we can thus perform another random split outside the cross-validation selection:

1. Randomly split the data into training and testing
2. Take the training data as your full data, and perform cross-validation to select a subset model (i.e. split the training data to a smaller training and a test set).
3. Record which variables were selected
4. Apply the selected model to the test data and record the prediction error
5. Repeat

PE	sbp	tobacco	adipo.	famhist	typea	obesity	alco.	age	chd	tobind	alcind
0.3891	F	F	T	T	F	T	T	F	T	T	F
0.3895	T	F	T	T	F	T	T	F	T	T	F
0.3895	F	F	T	T	F	F	T	F	T	T	F
0.3901	T	F	T	T	F	F	T	F	T	T	F
0.3911	F	F	T	F	F	T	T	F	T	T	F
0.3912	F	F	T	T	F	T	T	T	T	T	F
0.3914	F	F	T	F	F	F	T	F	T	T	F
0.3914	T	F	T	T	F	F	T	F	F	T	F
0.3916	F	F	T	T	T	T	T	F	T	T	F
0.3916	T	F	T	T	F	T	T	F	F	T	F

Table 2: 100 random splits into training and testing. 5-fold CV is applied to each training set, and selection results tabulated. Shown here: the top 10 models and their corresponding prediction error based on the average across the 100 splits.

In table 2 we randomly split the data set into training and testing 100 times. For each split we applied CV to the training data and recorded which variables/models were selected, and the corresponding prediction error. As can be seen from the table, many models are equally competitive. A few decisions stand out; tobind is always selected, but tobacco is not. It appears that the fact that you smoke is more predictive of ldl than the amount you smoke. In contrast, alcohol is always selected, but alcind is not. The amount of alcohol you drink is more predictive of ldl than the indicator if you drink or not. The family history of heart disease, and the patient's history of heart disease are also selected as predictors.

Out of the 101 models we considered (essentially 10 of each model size), the top 10 models were selected 21,18,12,6,5,4,3,3,3 and 3 times respective. While there is no clear majority, the top 3 models seem to be selected much more frequently than the rest. If we look at the prediction error curve, we see that model selection is not an easy problem for this data set.

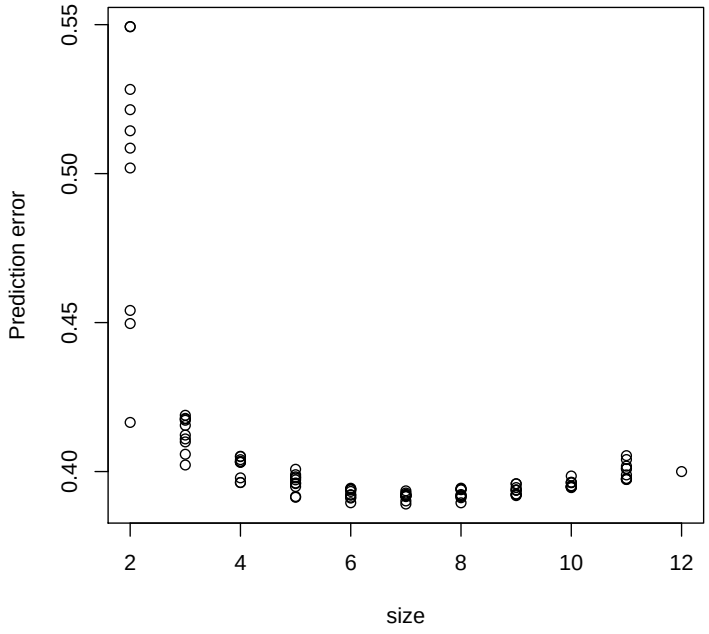


Figure 10: Prediction errors for selected models. Notice that the prediction error curve is "flat", indicating that there is a lot of uncertainty in terms of selecting an optimal model for prediction - many models perform near equally well.

Many models provide similar prediction error results, and these models may differ in terms of variables included as well as in model size.

It is also of interest to rank variables based on their marginal counts: how often is a particular variable selected to be in the predictive model? In table 3 we record these statistics for the 100 5-fold CV splits we performed. Clearly, some variables stand out. Adiposity is always included in the model, as is tobind. That is, body fat and smoking status are important predictors for predicting cholesterol levels. In the second tier of important variables we find disease history parameters, where family history and patient history are equally important. Alcohol consumption is also an important factor. Notice that the behavioral variable, type-A, is never selected, and tobacco usage is rarely selected. Age may be a factor, but is not selected in more than 23/100 CV runs.

We also computed the prediction errors of the top 5 models on the original test set. This test set was removed from the analysis prior to graphical exploration and selection. The prediction errors of the top 5 models are 0.1291, 0.1333, 0.1319, 0.1327 and 0.138. We can compare this to the full model prediction error 0.135. The top model does in fact provide a smaller prediction error than the full model, but the 5th best model selected is worse. This is again supporting the conclusion that model selection for optimal prediction is difficult for this data set.

sbp	41
tobacco	8
adiposity	100
famhist	84
typea	0
obesity	74
alcohol	94
age	23
chd	88
tobind	100
alcind	0

Table 3: Marginal comparison of variable importance: how many times is a variable selected in the cross-validation scheme? Note that adiposity, alcohol and tobind are almost always included in a model. Heart disease history of the patient and/or family members as well as obesity are also important variables, but not always included.

8 A Strategy for Modeling

If you have enough data I advocate the following approach;

1. Split the data into TRAINING and TESTING.
2. On the TRAINING data, perform
   - Graphical exploration/Transformations
   - Full model fitting
   - Diagnostics and Outlier detection
   - Iterate these steps until the full model is adequate
3. Perform model selection via Cp, AIC, BIC, F-test, CV, and identify a set of top models.
4. Examine selection stability via random splits. That
   - Split the TRAINING into a smaller Train and Test set.
   - Do model selection on the Train data.
   - Compute the prediction error on the Test data.
   - Repeat many times.
   - Summarize the results in terms of e.g. how often a variable is included in a selected model, how stable the model size is, how much the prediction errors vary, how often a particular model is selected, etc.
   - From the random splits, identify a few candidate models.

5. Apply the candidate models from steps 3 and 4 to the TESTING data and record the corresponding prediction errors.

6. Repeat steps 1-5 at least once.

7. Step back and try to interpret the results you've found.

Please note, the prediction errors in 4 are most likely underestimating the true prediction errors because you have used the full data in step 2 to clean up the data and choose a good modeling format. The prediction error in step 5 are fair estimates of future performance of the selected models.

What if you have a small sample size? The easiest fix is to use the same strategy as above BUT use a small TESTING and repeat steps 1-5 many times. The reason is that a small TESTING cannot provide you with a reliable estimate of prediction performance, while if you repeat the steps 1-5 several times you may get some sense of the expected performance. However, a small sample size *will* make model selection a difficult task.

9 Regularized Regression

In class we talked about the multi-collinearity problem. The effect of multi-collinearity is a large estimation variance for the coefficients corresponding to correlated explanatory variables. We can address this problem in several ways. Above, we applied model selection to reduce the impact of multi-collinearity. However, there is no guarantee that we select the "true" model, but we reduce the set of variables in our model, decreasing the estimation uncertainty in the process.

An alternative to selection is Principal Component Regression. We create a new set of variables consisting of the principal components of the design matrix X. Thus, if x_1 and x_2 are highly correlated, it may suffice to use an artificial variable $x_{12} = ax_1 + (1-a)x_2$ ($\sim$ average of the two variables) in the regression. Often, the principal components will take on the form of averages of the correlated variables, or differences of averages. Sometimes it's difficult to interpret the meaning of each component. In general, we only use the components with maximum variance in our regression. From section 3 we know that high variance explanatory variables have low estimation variance.

Ridge regression is related to PC regression. In essence, ridge regression shrinks coefficients corresponding to variables aligned with principal components with low variance, and leaves the coefficients corresponding to variables aligned with the major principal components unaffected. We can write ridge regression as the following penalized least squares problem:

$$\hat{\beta} = argmin\ ||y - X\beta||^2 + \lambda||\beta||^2,$$

which has a closed form solution

$$\hat{\beta}_R = (X'X + \lambda I)^{-1} X'y.$$

Now, the λI will only really impact the small diagonal elements, i.e. x's with small variance, and those correlated with these x's. When $X'X = D$ (diagonal) it is easy to see that we can write $\hat{\beta}_R = \gamma \hat{\beta}_{OLS}$, where γ is a *shrinkage factor* between 0 and 1. We can pick γ (λ) to minimize prediction errors etc. The logic behind ridge regression is simply that it's better to set $\hat{\beta}$ to a small value if the estimation variance is high - otherwise we might make a large error that will hurt us when it's time for prediction.

The problem with ridge regression is that no variables are ever completely dropped from the model. The shrinkage factor can be decreased toward 0, but is never actually 0 (then all variables are dropped). An alternative to ridge regression is the LASSO:

$$\hat{\beta} = argmin\ ||y - X\beta||^2 + \lambda ||\beta||,$$

which penalizes the least squares criterion with the L1 norm of the coefficients rather than the L2. This more moderate penalty can be shown to correspond to a different form of shrinkage (again when $X'X$ is diagonal):

$$\hat{\beta}_{LASSO} = (|\hat{\beta}| - \delta)_+ sign(\hat{\beta}),$$

where $(z)_+$ is 0 if $z < 0$ and z otherwise. That means, a coefficient is set to 0 if it's below a certain threshold δ (a function of λ), and shrunk by a constant amount δ otherwise. In figure 11 we show the $\hat{\beta}_{LASSO}$ estimates for the heart disease data. Variable 3 (adiposity) is the first one to cross a δ threshold. This coefficient quickly attains its optimal value and remains steady. The next variabel to enter is age, followed by obesity and tobind. The corresponding prediction errors for each variable subset are summarized in table 4. The smallest prediction

Prediction errors - Lasso										
adipo.	age	obesity	tobind	chd	famhist	alco.	sbp	tobacco	alcind	typea
0.1491	0.1417	0.1414	0.1410	0.1357	0.1350	0.1335	0.1335	0.1345	0.1345	0.1350

Table 4: LASSO prediction errors. Each prediction error corresponds to the error obtained after the corresponding column variable is added to the model.

error is obtained after 7 variables have entered the model. Note, LASSO was only applied to the whole training data. One could now apply the LASSO to random splits of the data to examine the selection stability of the regularized fit.

10 CART

If it's difficult to find a linear model that fits the data, or a graphical exploration to identify interactions is very cumbersome, CART models (Classification and Regression Trees) can be

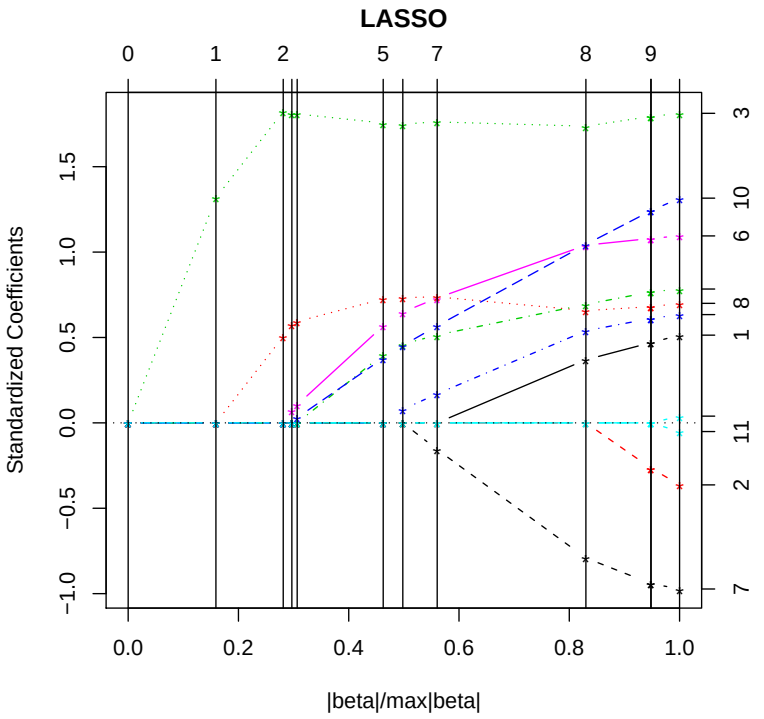


Figure 11: LASSO regularization paths.

an easy and intuitive way to summarize the data. CART is like a parlour game of "Twenty Questions". The questions you can ask are related to thresholds on the variables: e.g. "Is the observed value of variable x_j less than or equal to 2, or greater than 2?". This kind of question is called a "split". Depending on which side of the split an observation i falls, we will attribute a constant fitted value. There will be one fitted value for all the observations for which the answer to the above question is "yes", and similarly for all the observations for which the answer is "no". The gain of the split is measured in terms of RSS. Here, before the split the RSS is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

After the split, let's say n_1 observations answered "yes", and n_2 observations answered "no". The RSS has now been reduced to

$$\sum_{i:x_{ij}\leq 2}(y_i - \hat{\mu}_1)^2 + \sum_{i:x_{ij}>2}(y_i - \hat{\mu}_2)^2,$$

where

$$\hat{\mu}_1 = \sum_{i:x_{ij}\leq 2} y_i/n_1, \hat{\mu}_2 = \sum_{i:x_{ij}>2} y_i/n2.$$

For each branch we can now ask a new question to refine the predictions. Again, this leads to a reduction in RSS.
The combination of splits, e.g.

$$\{X_1 \leq 2\} \cap \{X_3 > 5\} \cap \{X_8 \leq 2.3\},$$

constitute a rectangle in, here, 3 dimensional space. This example rectangle corresponds to asking 3 questions. For each rectangle we calculate the mean value of the observations for which the answer to each question is "yes". That is the corresponding prediction.

The results of the splits are illustrated using a "Tree", where each split is a node and the branches correspond to the different answers. The length of the branches are drawn proportional to the reduction in RSS. At the bottom of the tree we have the so-called "leaves" (a high-dimensional rectangle) for which we calculate the predicted value from the mean of the observations corresponding to "yes" answers.

Let us now discuss how to choose the split. There are two choices to be made for each split: the variable j to split on, and where to split (2 in the above example). For each question you intend to ask, you only worry about the current question, not the future ones. Thus, you need to identify which of the total of p variables to ask about, and then perform a search for the optimal threshold. Optimality here refers to minimizing the RSS we obtain after the split. This is a greedy search since we only look at the optimal decision locally (the current question). Once we make a decision on a split we focus on the next question. We will ask one question for the "yes" from the first question, and a potentially different question for the "no". We search for the optimal variable and threshold separately for each.

How many questions should we ask? If we ask questions until every observation has its own rectangle we are clearly over-fitting. On future data, it is not plausible that the same detailed questions will summarize the data well. To evaluate this we perform cross-validation. We cut the bottom branches of the tree (pruning) until the cross-validation error increases. In Figure 12 we illustrate a full tree (a), the cross-validation error (b), and the tree which minimized the cross-validation error (c). The winning tree asks only two questions, related to adiposity.

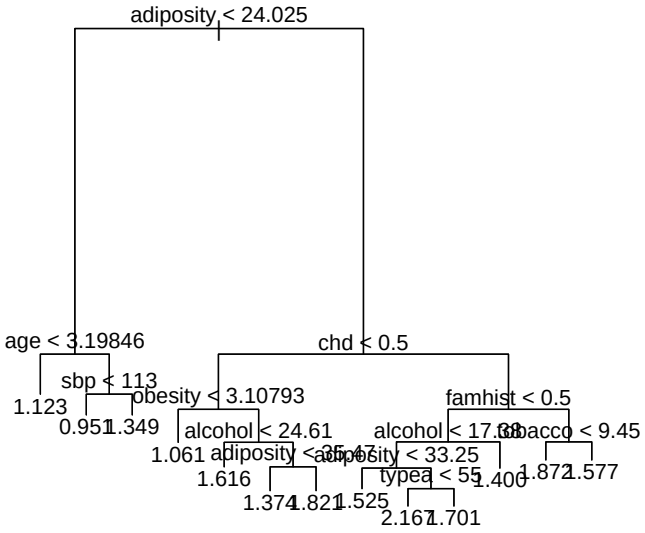
I repeatedly split the data and ran the same analysis. For all the trees, adiposity was the first question. On average 2.34 questions were asked. The second and sometimes third question were related to either obesity, alcohol, age or heart disease (chd).

10.1 CART for classification

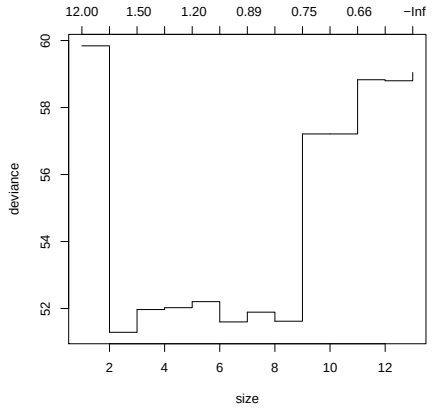
CART can also be used for categorical outcome data. Instead of letting each leaf represent a mean value for $y_i : i$ in leaf, we apply a majority voting scheme. If the majority of the observations in a leaf correspond to category A, we vote A.

I demonstrate CART classification using the South African heart disease data. I switch the outcome variable to heart disease status (chd), and make ldl one of the predictors.

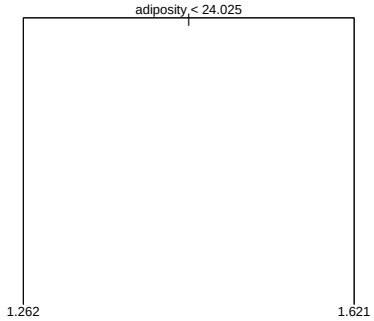
In Figure 13 you see the full tree which separates heart disease cases from healthy patients. The cross-validation error is minimized in a range from size 2 to 6. The winning tree involves



(a)



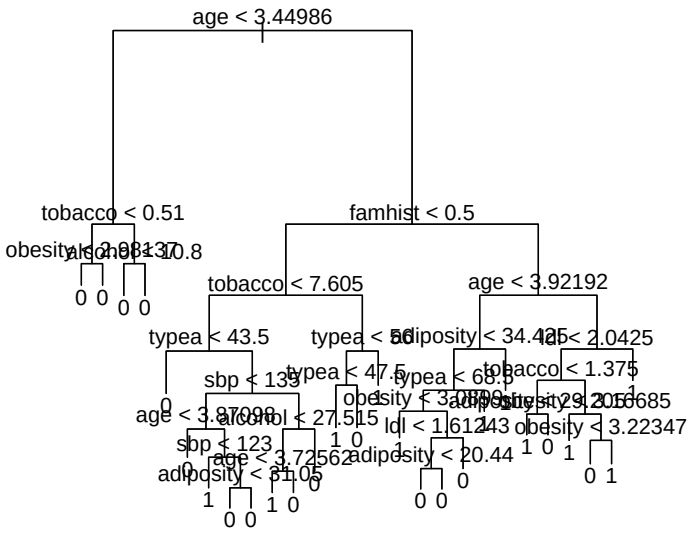
(b)



(c)

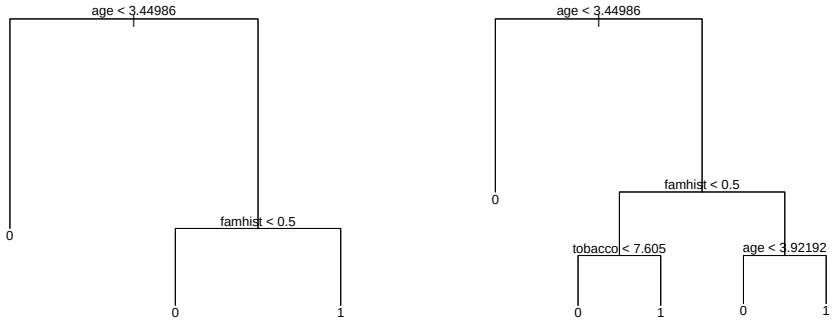
Figure 12: CART: (a) The full CART tree, (b) Cross-validation error as a function of the size of the tree (number of questions), (c) the pruned tree.

only two questions: age and family history of heart disease (Fig. 14 (a)). I also provide a slightly larger tree in Figure 14 (b) for reference.



(a)

Figure 13: CART classification



(b) (c)

Figure 14: CART classification: (a) The full CART tree, (b) the cross-validation tree, (c) a tree of size 5.

11 Generalized Linear Models

In the last class, we touched briefly on the topic of generalized linear models. Up to that point we had focused on the normal error models. What do we do when the error are clearly not normal, such as when we observe counts or categorical outcomes?

We revisit the heart disease data, but this time we try to predict the disease status (0 or 1) from the other attributes. We will model this as a binomial model:

$$y_i \sim Bin(1, \pi_i),\ E(y_i) = \pi_i,\ log(\frac{\pi_i}{1-\pi_i}) = X_i\beta,$$

where y_i is the disease outcome for observation i, π_i is the probability of disease for observation i given the explanatory variables. Here we use the so-called logit-link, $log(\frac{\pi_i}{1-\pi_i}) = X_i\beta$, which relates the explanatory variables (linear model) to the parameter of interest (π_i) in a non-linear fashion. That is, we can write $\pi_i = e^{X_i\beta}/(1 + e^{X_i\beta})$.

This link-function is the key to generalized linear models. In general we write

$$E(y_i) = \mu_i,\ g(\mu_i) = X_i\beta,$$

where the link function $g(.)$ links the linear predictor $X_i\beta$ to the expected value of the outcome μ_i. Note, by choosing an appropriate link, we are guaranteed that our predictions make sense, i.e. positive if we are modeling counts, or between 0 and 1 if we are modeling probabilities.

The sampling distribution of coefficient estimates in general don't have a simple form, and we resort to using approximations. Appealing to the central limit theorem, we assume that coefficient estimates are approximately normally distributed (review the class notes on non-linear models).

We apply a binomial GLM fit to the disease data and obtain and the following result. Only 4 coefficients are significant in this fit: ldl, famhist, typea and age. Using a stepwise selection procedure, we can eliminate all variables except these. Note that the behavior variable is predictive of disease status, while it was not informative for predicting ldl. We can take our selected model and apply it for predicting disease data on the test set. Here, the full model resulted in an error rate of 25%, whereas model selection reduced this to 23.3%. In figure 15 we use a boxplot to summarize the prediction results. We plot each $\hat{\pi}_i$ (i.e. estimated probability of disease) against the true disease status. While the prediction performance is not perfect, we do see a strong association between $\hat{\pi}_i$ and the true outcome.

This is just a skim on the surface of GLM modeling. Please take either 586, 587 or a 600-level class to learn more about GLMs.

12 Conclusion

What is the take-home message of this class? There are a few things I hope you will remember:

Coeff.	Estimate	Std. Error	z-value	p-value
(Intercept)	-8.013681	5.083791	-1.576	0.11495
sbp	0.010391	0.007479	1.389	0.16473
tobacco	0.056261	0.034278	1.641	0.10073
ldl	0.812569	0.395012	2.057	0.03968*
adiposity	0.016157	0.036922	0.438	0.66168
famhist	0.920864	0.281493	3.271	0.00107***
typea	0.048424	0.015123	3.202	0.00136***
obesity	-1.553562	1.422733	-1.092	0.27485
alcohol	-0.001205	0.006291	-0.191	0.84816
age	1.507834	0.612842	2.460	0.01388*
tobind	0.789599	0.443453	1.781	0.07498
alcind	-0.196312	0.364551	-0.539	0.59023

Table 5: Coefficients of the binomial model fit. The p-values are approximate only, since the coefficient estimates are no longer linear in the data, and the errors are not normally distributed.

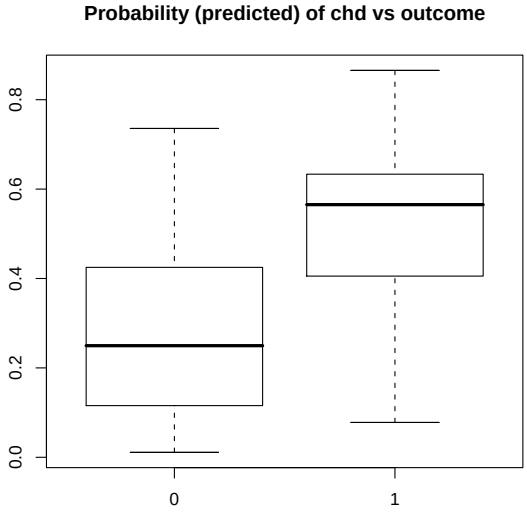


Figure 15: $\hat{\pi}_i$ on the test set compared with true disease status.

- The basic assumptions
  - if these do not hold, the least squares approach makes no sense!

- Unusual observations
  - sometimes outliers are outliers, sometimes they are important subsets of the data.
  - always report results with and without the outliers

- Dare to simulate
  - while statistical testing is handy, always remember that the tests you use are only valid if the underlying assumptions are true (e.g. normal errors).
  - let the computer do the work! Just try it: simulate some data sets with and without normal errors - how does your data set compare?
  - use bootstrap to investigate the modeling without as many assumptions on the error distribution

- There may not exist a best model
  - most of the time, it is very difficult to identify the "best" model.
  - if variables are correlated, you won't be able to tell which variable has a direct influence on the outcome. Either, find a data expert (the biologist, or engineer who generated the data) and get their help, OR report the model selection uncertainty.
  - report a set of candidate models and leave the final decision to the data expert.

- Transformation, non-linear or polynomial?
  - if you can get away with a transformation, that's great!
  - remember to keep your polynomial models simple.
  - non-linear modeling can be tricky! be careful with CI, convergence issues

- .... and finally, how you present your results is almost as important as the results themselves!