Chapter 5

Linear Methods for Prediction

Today we describe three specific algorithms useful for classification problems: linear regression, linear discriminant analysis, and logistic regression.

5.1 Introduction

We now revisit the classification problem and focus on linear methods.

Since our prediction $\hat{G}(x)$ will always take values in the discrete set $\mathcal{G}$ we can always divide the input space into a collection of regions taking the same predicted values.

The question is: What is the best sub-division of this space?

We saw previously that the boundaries can be smooth or rough depending on the prediction function.

For an important class of procedures these *decision boundaries* are linear, this is what we will mean by linear methods for classification. We will see that these can be quite flexible (much more than linear regression).

Suppose we have K classes labeled $1, \dots, K$. We can define a 0-1 indicator for each class k and for each of these perform regression. We would end-up with a regression function $\hat{f}_k(x) = \hat{\beta}_{0k} + \hat{\beta}'_{1k}x$ for each class k.

The decision boundary between class k and l is simply $\hat{f}_k(x) = \hat{f}_l(x)$ which is the set $\{x : (\hat{\beta}_{0k} - \hat{\beta}_{0l}) + (\hat{\beta}_{1k} - \hat{\beta}_{1l})'x = 0\}$ which is on a plane.

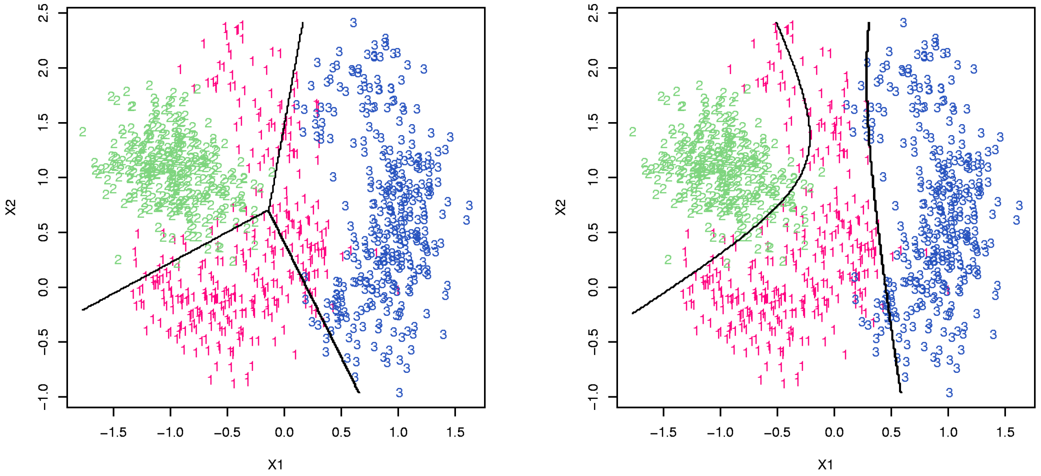
Since the same is true for any pair of classes the division of the space of inputs are piecewise planes.

This regression approach is a member of a set of methods that model *discriminant function* $\delta_k(x)$ for each class, and then classify X to the class with the largest value for its discriminant function.

Methods that model the posterior probability $\Pr(G = k|X = x)$ are also in this class. If this is a linear function of x then we say its linear. Furthermore if its a monotone transformation of a linear function x we will sometimes say its linear. An example is *logistic regression*. More on that later.

Both decision boundaries shown in Figure 5.1 are linear:

Figure 5.1: Two linear decision boundaries. One obtained with straight regression, the other using the quadratic terms.



5.2 Linear Regression of an Indicator Matrix

Each response is coded as a vector of 0-1 entries. If $\mathcal{G}$ has K classes then $Y_k, k = 1, \ldots, K$ with $Y_k = 1$, if $G = k$ and 0 otherwise.

These are collected in a vector $Y = (Y_1, \ldots, Y_K)$ and the N training vectors are collected into a $N \times K$ matrix we denote with $\mathbf{Y}$.

For example, if we have $K = 5$ classes

$$\begin{array}{ccccccc} 1 & & 0 & 0 & 0 & 0 & 1 \\ 2 & & 0 & 0 & 0 & 1 & 0 \\ 3 & \rightarrow & 0 & 0 & 1 & 0 & 0 \\ 4 & & 0 & 1 & 0 & 0 & 0 \\ 5 & & 1 & 0 & 0 & 0 & 0 \end{array}$$

On the left is the original data and the left the coded data.

To fit regression to each class simultaneously we can simply use the same matrix multiplication trick.

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Notice the matrix

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

has the coefficients for each regression in the columns. So for any point x we can get the prediction by

- compute the fitted output $\hat{f}(x) = [(1, x)\hat{\mathbf{B}}]'$, a K vector
- identify the largest component and classify accordingly:

$$\hat{G}(x) = \arg\max_{k \in \mathcal{G}} \hat{f}_k(x)$$

What is the rational for this approach?

A formal justification is that we are trying to estimate $\mathrm{E}(Y_k|X = x) = \Pr(G = k|X = x)$. The real issue is: How good is the linear approximation here? Alternatively, are the $\hat{f}_k(x)$ good estimates of $\Pr(G = k|X = x)$.

We know they are not great because we know $\hat{f}_k(x)$ can be larger than 1 and smaller than 0. However, as we have discussed this does not necessarily matter if we get good predictions.

A more conventional way of fitting this model is by defining t_k as the vector with a 1 in the k entry and $y_i = t_k$ if $g_i = k$. Then the above approach is equivalent to minimizing

$$\min_{\mathbf{B}} \sum_{i=1}^{N} ||y_i - \{(1, x_i)\mathbf{B}\}'||^2.$$

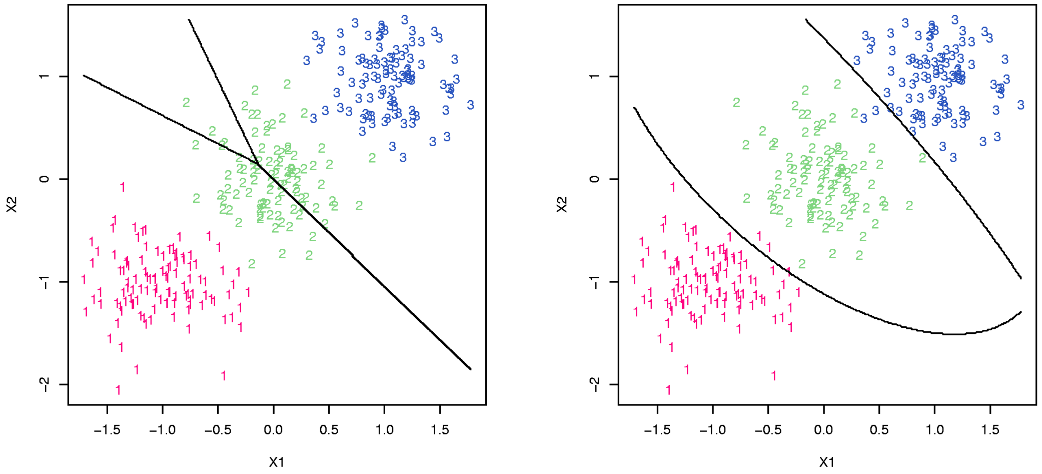
A new observation is classified by computing $\hat{f}(x)$ and choosing the closest target

$$\arg\min_{k} ||\hat{f}(x) - t_k||^2$$

Because of the rigid nature of regression, a problem arises for linear regression when $K \geq 3$. Figure 5.2 shows an extreme situation. Notice that the boundaries can easily be formed by eye to perfectly discriminate the three classes. However, the regression discriminatory does not do well.

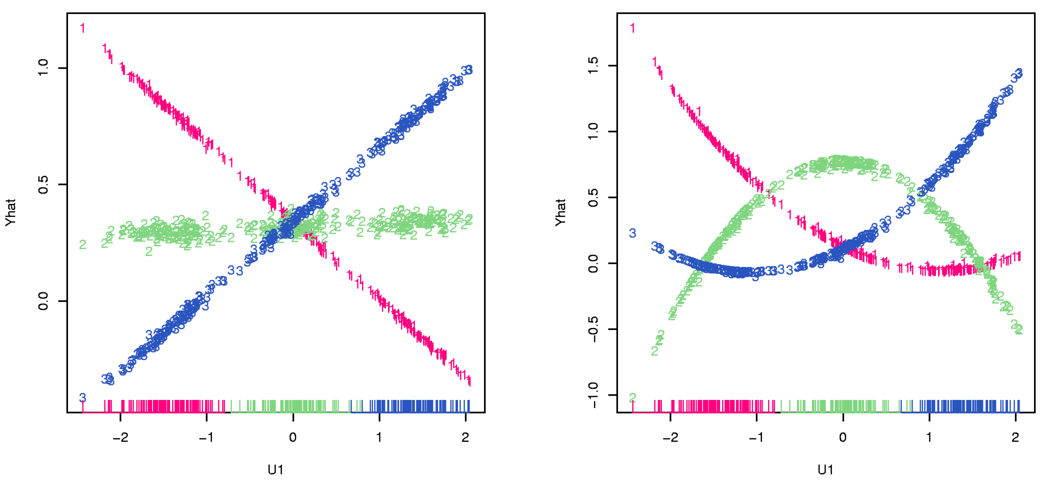
Why does linear regression miss the classification in Figure 5.2? To see what is going on the ideas we learned in the last chapter are useful. Notice that the best direction to separate these data is a line going through the centroids of the data. Notice there is no information in the projection orthogonal to this one.

Figure 5.2: Two linear decision boundaries. One obtained with straight regression, the other using the quadratic terms.



If we then regress the Y on the transformed X we see there is barely any information about the the second class. These is clearly seen in the left side of Figure 5.2.

Figure 5.3: Projection to best direction and regression of Y onto this projection



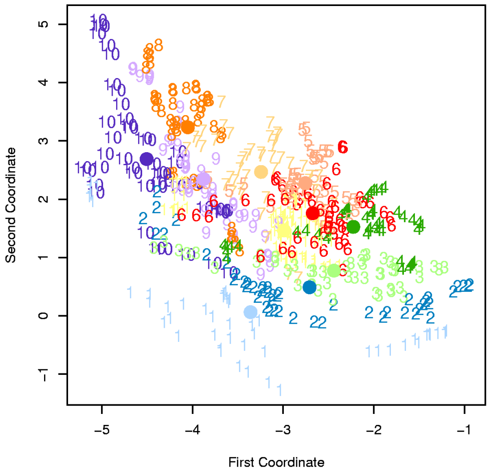
However, by making the regression function a bit more flexible, we can do a bit better. One way to do this is to use the quadratic terms (there are 3, which are they?) In Figures 5.2 and 5.2 this linear regression version that includes the quadratic term does much better.

However, if we increase the number of classes to $K = 4$ we would then need to start adding the cubic terms and now we are dealing with lots of variables.

A Data set that we may be working on later will be the vowel sound data. Figure

5.2 contains a plot of the first 2 coordinates and the classes.

Figure 5.4: Vowel sound data



5.3 Linear Discriminant Analysis

Decision theory for classification tells us what we need to know the class posteriors $\Pr(G|X)$ for optimal classification.

Suppose $f_k(x)$ is the class conditional density of X in a class $G = k$, and let π_k be the prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$. A simple application of

Bayes theorem gives us

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}.$$

Notice that having the quantities $\mathbf{f}_k(x)$ is almost equivalent to having the $\Pr(G = k|X = x)$ that provide Bayes rule.

Suppose we model each class density as a multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\{-\frac{1}{2}(x - \mu_k)'\mathbf{\Sigma}_k^{-1}(x - \mu_k)\}$$

Figure 5.3 shows regions that contain 95% of the data for three bivariate distributions with different means but the same covariance structure. The covariance structure makes these be ellipses instead of circles.

The lines are Bayes rule.

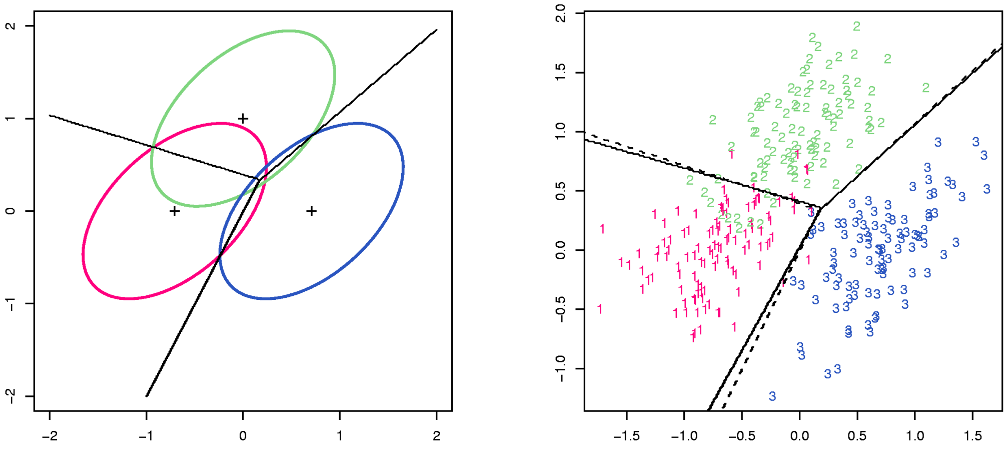
Linear discriminant analysis (LDA) arises when we assume the covariance structure is the same for all classes. In this case we can see that discrimination function is simply

$$\delta_k(x) = x'\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k\mathbf{\Sigma}^{-1}\mu_k + \log\pi_k$$

Notice: if we assume $\pi_k = 1/K$ then the last term is not needed. In any case, notice that this is a linear function of x!

In practice we do not have the means μ and the covariance structure Σ. The strategy is to use the training data to estimate these. In Figure 5.3 we see some outcomes a three class simulation. The distributions used to create them where those shown in the left side of 5.3.

Figure 5.5: Bivariate normal distributions, outcomes of these, and the Bayes and LDA prediction rules



To estimate the parameters we simply

- $\hat{\pi}_k = N_k/N$, where N_k is the observed number of subjects in class k.
- $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$
- $\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)'/(N - K)$

Figure 5.3 also show both the Bayes rule (dashed) and the estimated LDA decision boundary.

Technical Note: For two classes LDA is the same as regression.

Now if we assume that each class has its own correlation structure then we no longer get a linear estimate. Instead we have that the decision boundary is simply:

$$\delta_k(x) = -\frac{1}{2}|\mathbf{\Sigma}_k| - \frac{1}{2}(x - \mu_k)'\mathbf{\Sigma}^{-1}(x - \mu_k) + \log \pi_k$$

The decision boundary is now described with a quadratic function. This is therefore called quadratic discriminant analysis (QDA).

QDA and LDA decision boundaries are shown in Figure 5.3 for the same data.

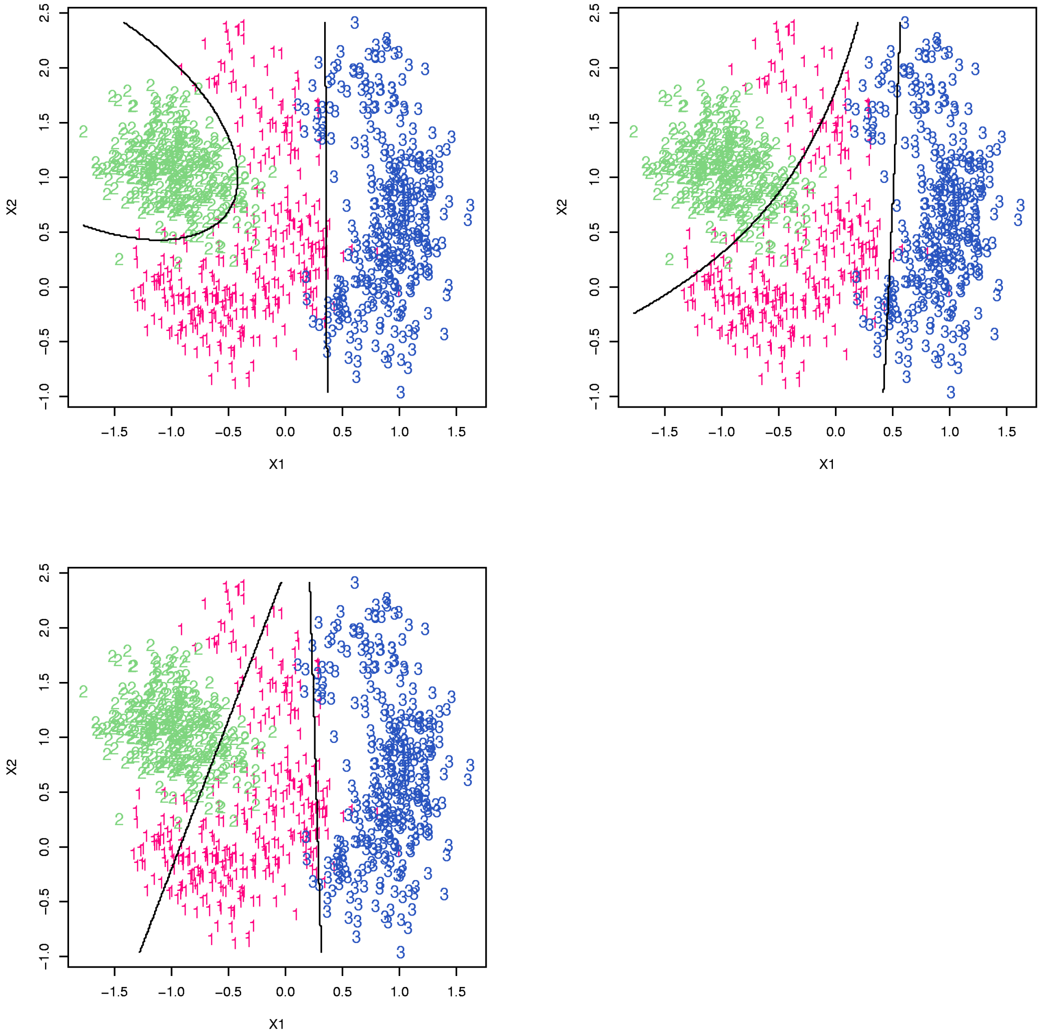
Notice that both LDA and QDA are finding the centroid of classes and then finding the closest centroid to the new data point. However, correlation structure is taken into consideration when defining distance.

Note: When the number of covariates grows the number of things to estimate in the covariance matrix gets very large. One needs to be careful.

There is a method that tries to find a happy medium called *Regularized Discriminant Analysis*. Basically it comes down using shrunken version of the class specific covariance structures.

$$\hat{\mathbf{\Sigma}}_k(\alpha) = \alpha\hat{\mathbf{\Sigma}}_k + (1 - \alpha)\hat{\mathbf{\Sigma}}$$

Figure 5.6: QDA and LDA with quadratic terms and straight LDA



5.3.1 Computations for LDA

Suppose we compute the eigen decomposition for each $\hat{\mathbf{\Sigma}}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{U}'_k$, where $\mathbf{U}_k$ is $p \times p$ orthonormal and $\mathbf{D}_k$ a diagonal matrix of positive eigenvalues d_{kl}. The ingredients for $\delta_k(x)$ are

- $(x - \hat{\mu}_k)' \hat{\mathbf{\Sigma}}_k^{-1} (x - \hat{\mu}_k) = [\mathbf{U}'_k (x - \hat{\mu}_k)]' \mathbf{D}_k^{-1} [\mathbf{U}'_k (x - \hat{\mu}_k)]$
- $\log |\hat{\mathbf{\Sigma}}_k| = \sum_l \log d_{kl}$.

Notice this is much easier to compute because the $\mathbf{D}$ are diagonal!

Given this we can now compute and interpret the LDA classifier as follows:

- *Sphere* the data with respect to the common covariance estimate $\hat{\mathbf{\Sigma}}$: $X^* = \mathbf{D}^{-\frac{1}{2}} \mathbf{U}' X$ where $\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}'_k$. The common covariance estimate of X^* will no be the identity matrix!
- Classify to the closest class centroid in the transformed space, modulo the effect of the class prior probability π_k.

Note: Section 4.3.3 of the book has a nice description of how LDA provides the solution as what one obtains when asking for the linear combination $Z = a'X$ that provides the biggest ration of within

5.4 Logistic Regression

Assume

$$\begin{aligned}
\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} &= \beta_{01} + \beta_1' x \\
\log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} &= \beta_{01} + \beta_2' x \\
&\vdots \\
\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} &= \beta_{01} + \beta_{K-1}' x
\end{aligned}$$

Notice $g(p) = \log(\frac{p}{1-p})$ is called the logistic link and is $g : (0,1) \rightarrow \mathbf{R}$

When $K = 2$ this has a very simple form (only one set of covariates) and is a very popular model used in biostatistical applications.

With this probability model we can now write the log-likelihood...

$$l(\beta_0, \beta; y) = \sum_{i=1}^{N} \{ y_i \log p_i + (1 - y_i) \log(1 - p_i) \}$$

- Out estimate will be the MLE: $\max_{\beta_0, \beta} l(\beta_0, \beta; y)$.
- In practice how do we find this maximum? We will see this later.

The rest of this sections presents some technical details about maximum likelihood estimates.

5.4.1 Elementary Likelihood Theory

In linear model theory we found: $\mathrm{E}(\hat{\boldsymbol{\beta}})$, $\mathrm{var}(\hat{\boldsymbol{\beta}})$, and the asymptotic properties of $\hat{\boldsymbol{\beta}}$. To do this for the MLE's we need to develop some asymptotic theory.

We will derive some properties that will be useful. We will give some of the main results without presenting proofs. If you are interested in the details lookup the following books:

- Bickel P. and Doksum K. *Mathematical Statistics: Basic Ideas and Topics.* (1977) Holden-Day
- Lehman, E. *Theory of Point Estimation* (1983) Wadsworth & Brooks/Cole.

Say Y has density function $f_Y(y;\beta)$ then remember

$$\int_{y \in A} f_Y(y;\beta)\, dy = 1$$

here A is the range of Y.

First we will assume β is 1 dimensional.

We define the log-likelihood function $l(\beta; y) = \log f_Y(y;\beta)$.

In GLM we will usually assume that the data are independent observations coming from a density determined by the parameter β. If $\mathbf{Y}$ is a vector having n

independent components then the log-likelihood is a sum of n independent terms

$$l(\beta; \mathbf{y}) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \beta)$$

Notice the following

- We consider the log-likelihood as a function of β given the observed data y.
- The log-likelihood is always a real number.
- We allow β to "move" so we assume $\beta \in \beta$ a parameter space, so we are actually defining a family of distributions: $\{f(\mathbf{y}; \beta) : \beta \in \beta\}$
- We will specify a true distribution with a true parameter β_0. The expected value and probability measure with respect to that density will be denoted E_{β_0} and Pr_{β_0}.

For the results described in this sections to hold, the family of densities must satisfy some regularity condition. We will call the family of densities ML-regular if they satisfied these conditions. See L&C or B & D.

If $\mathbf{Y}$ is a vector having n independent components with joint distribution that is ML-regular, then

- $\mathrm{E}_{\beta_0}\{l(\beta_0; \mathbf{Y})\} > E_{\beta_0}\{l(\beta; \mathbf{Y})\}$
- $\lim_{n\to\infty} P_{\beta_0}[l(\beta_0; \mathbf{Y}) > l(\beta; \mathbf{Y})] = 1$

for all $\beta \neq \beta_0$.

The first part of this result says that the value considered to be the *true* value of the parameter maximizes the expected log-likelihood. The second part says that the chance of the log-likelihood being bigger at the true value is VERY likely for large values of n.

This properties motivate the use of ML-estimation. Given the observed data we estimate the true parameter with the maximum likelihood estimate (MLE) is defined as

$$\hat{\beta} = \max_{\beta \in B} l(\beta; \mathbf{y}) \tag{5.1}$$

A consequence of Theorem 1 is the consistency of the MLE.

Theorem 2. Define $\hat{\beta}_n$ as the MLE of β_0 when n observations are used. Then $\hat{\beta}_n$ converges in probability to β_0.

Define the *score statistic* as:

$$U(\beta; \mathbf{y}) = \frac{\partial l(\beta; \mathbf{y})}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \log f_{Y_i}(y_i; \beta)}{\partial \beta}$$

Notice that, just like the log-likelihood, this is a function of β once the random variable $\mathbf{Y}$ is observed to be $\mathbf{y}$.

Why are we defining this ?

The maximum likelihood estimate (MLE) defined by (5.1), under the regularity conditions, is equivalent to finding the β for which $U(\beta;\mathbf{y}) = 0$.

Notice that the score statistic may be considered to be a random variable since it is a function of the data. Once we obtain the data we find the value of β that has 0 score, the MLE.

How good of a statistic is the score statistic?

First notice that

$$\mathrm{E}_{\beta_0}\{U(\beta_0;\mathbf{Y})\} = \mathrm{E}_{\beta}\left(\left. \frac{\partial l(\beta;\mathbf{Y})}{\partial \beta} \right|_{\beta=\beta_0} \right) = 0$$

This shows that the score has expected value of 0 at the true β_0, which is what we want.

The next thing we would like is to see how variable it is... how precise is our estimate? Look at the variance. One important property is that

$$\mathrm{var}_{\beta_0}\{U(\beta_0;\mathbf{Y})\} = \mathrm{var}_{\beta_0}\left(\left. \frac{\partial l(\beta;\mathbf{Y})}{\partial \beta} \right|_{\beta=\beta_0} \right) = -\mathrm{E}_{\beta_0}\left(\left. \frac{\partial^2 l(\beta;\mathbf{Y})}{\partial \beta^2} \right|_{\beta=\beta_0} \right) = I_n(\beta)$$

This last quantity is called Fisher's information quantity. It is telling us how much "information" about the true parameter is given by the data.

5.4.2 Asymptotic results

Under regularity conditions

$$I_n(\beta)^{-\frac{1}{2}}U(\beta;\mathbf{y}_n) \sim N(0,1) + O_p(n^{-\frac{1}{2}})$$

Example 3.1 I.I.D Normal we have distribution

$$f_Y(\mathbf{y}_n;\beta_0) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0)^2\right\}$$

with σ^2 known.

For data $\mathbf{y}_n$ we have log-likelihood function

$$l(\beta;\mathbf{y}_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)$$

Notice that we can calculate the expected value of the log-likelihood for any β.

$$\begin{aligned}\mathrm{E}[2l(\beta;\mathbf{Y}_n)] &= \mathrm{E}[-n\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}[(y_i-\beta_0)-(\beta-\beta_0)]^2] \\ &= -N\left[\log(2\pi\sigma^2)+1+\frac{(\beta-\beta_0)}{\sigma^2}\right]\end{aligned}$$

Notice that the maximum occurs at β_0 as Theorem 1 implies.

We can similarly show that

$$\begin{aligned}\Pr[l(\beta_0;\mathbf{y}_n) > l(\beta;\mathbf{y}_n)] &= \Pr[2\{l(\beta_0;\mathbf{y}_n)-l(\beta;\mathbf{y}_n)\}>0] \\ &\geq \Pr\left[|\bar{Y}-\beta_0| > \frac{N}{2}|\beta-\beta_0|\right] \to 1\end{aligned}$$

The score statistic is

$$U(\beta;\mathbf{y}_n) = \frac{1}{\sigma^2}\sum(y_i - \beta)$$

From here it is obvious that the MLE is $\hat{\beta}_n = \bar{y}$, that $\mathrm{E}_{\beta_0}[U(\beta_0;\mathbf{Y}_n)] = 0$, and

$$\mathrm{var}_{\beta_0}[U(\beta_0;\mathbf{Y}_n)] = n/\sigma^2 = -\mathrm{E}_{\beta_0}\left[\sum -\frac{1}{\sigma^2}\right] = I_n(\beta_0)$$

It is known that

$$\sqrt{N}(\bar{Y} - \beta)/\sigma^2 \rightarrow N(0,1)$$

here the convergence is in distribution. Notice that this is $I_n(\beta)^{1/2}U(\beta;\mathbf{y}_n)$.

What we really want to know is the asymptotic distribution of MLE not the score equation.

Use usual trick... since Theorem 2 tell us that $\hat{\beta}_n$ is close to β_0 we can use a Taylor expansion to write the MLE as a linear combination of the score equation.

$$U(\beta_0;\mathbf{y}_n) \approx U(\hat{\beta}_n;\mathbf{y}_n) + \left.\frac{\partial U(\beta;\mathbf{y}_n)}{\partial \beta}\right|_{\beta=\hat{\beta}_n}(\beta_0 - \hat{\beta}_n)$$

Notice $U(\hat{\beta}_n;\mathbf{y}_n) = 0$.

$$\begin{aligned}\left.\frac{\partial U(\beta;\mathbf{y}_n)}{\partial \beta}\right|_{\beta=\hat{\beta}_n} &\approx \mathrm{E}\left[\left.\frac{\partial U(\beta;\mathbf{Y}_n)}{\partial \beta}\right|_{\beta=\hat{\beta}_n}\right] \\ &\approx \mathrm{E}\left[\left.\frac{\partial U(\beta;\mathbf{Y}_n)}{\partial \beta}\right|_{\beta=\beta_0}\right] = -I_n(\beta_0)\end{aligned}$$

So we have

$$U(\beta_0; \mathbf{y}_n) \approx -I_n(\beta_0)(\beta_0 - \hat{\beta}_n)$$

which implies

$$I_n(\beta_0)^{-1/2} U(\beta_0; \mathbf{y}_n) \approx I_n(\beta_0)^{1/2}(\hat{\beta}_n - \beta_0)$$

So $I_n(\beta_0)^{1/2}(\hat{\beta}_n - \beta_0)$ is approximately standard normal when n is big.

5.4.3 Algorithm for fitting GLMs

We want to obtain the MLE but in general there is no way of getting it in closed form. Instead we must use some minimization procedure. For GLMs there is a nice trick that reduces the minimization procedure to a iterative re-weighted least squares procedure

Say we have some data and we have specified all the necessary elements to define a GLM

Let $g(\cdot)$ be the link function. Notice that for all i, y_i should be somewhat close to μ_i so we can write

$$\begin{aligned} g(y_i) &\approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i) \\ &= \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i} \end{aligned}$$

Notice hat we have a linear model with a deterministic and random part.

Letting $Z_i = g(Y_i)$ and $\epsilon_i = (Y_i - \mu_i)\frac{d\eta_i}{d\mu_i}$ we have

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where the ϵ_i's are IID with variance $V(\mu_i)\frac{\phi}{w_i}\left(\frac{d\eta_i}{d\mu_i}\right)^2$

Notice that with this regression model the MLE is obtained by minimizing the weighted least squares with quadratic weights

$$W_i = \left(V(\mu_i)\frac{\phi}{w_i}\left(\frac{d\eta_i}{d\mu_i}\right)^2 \right)^{-1} \tag{5.2}$$

In this case the solution is $(\mathbf{X'WX})^{-\mathbf{1}}\mathbf{X'Wz}$. *Problem the weights depend on $\boldsymbol{\beta}$...*

In this case we may use iteratively re-weighted least squares.

The procedure is as follows

1. Let $\boldsymbol{\beta}^{(0)}$ be the current estimate and $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{\mu}^{(0)}$ be the values derived from it.

2. Define
$$\mathbf{z}^{(0)} = \boldsymbol{\eta}^{(0)} + (\mathbf{y} - \boldsymbol{\mu}^{(0)}) \left(\left. \frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}} \right|_{\boldsymbol{\eta}=\boldsymbol{\eta}^{(0)}} \right)$$

3. Define the weights $W_i^{(0)}$ by evaluating equation (5.2) with $\boldsymbol{\eta}^{(0)}$.

4. Obtain new estimates $\boldsymbol{\beta}^{(1)}$ using weighted least squares. From this from new estimate $\boldsymbol{\eta}^{(1)}$ and from this from $\mathbf{z}^{(1)}$.

5. Iterate until $||\mathbf{z}^{(1)} - \mathbf{z}^{(0)}||$ is small

5.4.4 Justification for the fitting procedure

How do we know the estimate obtained from the procedure described above is equivalent to the MLE.

Once we specify a model we find the MLE by maximizing the log-likelihood or equivalently

Find MLE by solving the system of equations $\mathbf{U}(\boldsymbol{\beta}; \mathbf{y}) = 0$ or

$$\frac{\partial l}{\partial \beta_j} = 0, \text{ for } j = 1, \ldots, p$$

But we don't necessarily know how the log-likelihood depends on $\boldsymbol{\beta}$.

We are going to abuse notation for a bit and stop writing the index i:

For each observation, we know how:

- the log-likelihood depends on θ (by the way we define the distributions)

- θ depends on the mean μ (from the exponential family property $\mu = b'(\theta)$)
- μ depends on the linear predictor η (the link function specifies this)
- η depends on β (η is a linear combination of the components of $\boldsymbol{\beta}$)

To find $\frac{\partial l}{\partial \beta_j}$ simply use the chain rule.

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j}$$

Lets see what each one of these pieces is... We know that

For each observation $l = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)$ so

$$\frac{\partial l}{\partial \theta} = (y - \mu)/a(\phi)$$

From $b'(\theta) = \mu$ and $b''(\theta) = V(\mu)$ we derive $d\mu/d\theta = V(\mu)$.

Finally from $\eta = \sum \beta_j x_{\cdot j}$ we obtain $\frac{\partial \eta}{\partial \beta_j} = x_{\cdot j}$

Therefore The contribution from observation i from the j-th equation of the score statistic is

$$\frac{y_i - \mu_i}{a(\phi)}\frac{1}{V(\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ij}$$

Notice that $(a(\phi)V(\mu_i))^{-1}d\eta_i x_{ij}$ is $W_i\frac{d\eta_i}{d\mu_i}$.

By adding up the contribution to the likelihood of each observation we can know write

The system of equation involving the score $\mathbf{U}$ can be written as

$$u_j = \sum_{i=1}^{n} W_i (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} x_{ij} \text{ for } j = 1, \dots, p$$

with W_i as in equation (5.2).

Newton-Rapson's method to find the the solution to $\mathbf{U} = 0$ consist on iterating using

$$\mathbf{A}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}) = \mathbf{U}^{(0)}$$

with

$$\mathbf{A} = - \left. \frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} \right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}}$$

to obtain the next "estimate" of $\boldsymbol{\beta}$

Fisher's score method *works similarly but instead uses*

$$\mathbf{A} = \mathrm{E} \left(- \left. \frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} \right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}} \right)$$

Takes expectations simplifies the computations for the procedure and statistically ends up being the same.

Notice that in many instances $\partial^2 l / \partial \boldsymbol{\beta}^2$ is a constant and Fisher's method is equivalent to Newton-Rapson's

Notice that the j, k entry of $\mathbf{A}$ is

$$\mathrm{E}\left(\frac{\partial u_j}{\partial \beta_k}\right) = \mathrm{E}\left(\sum_{i=1}^{n}(y_i - \mu_i)\frac{\partial}{\partial \beta_k}\left\{W_i \frac{d\eta_i}{d\mu_i} x_{ij}\right\} + W_i \frac{d\eta_i}{d\mu_i}\frac{\partial}{\partial \beta_k}(y_i - \mu_i)\right)$$

Notice that the first term has expectation 0 and that the second term is

$$\sum_{i=1}^{n} W_i \frac{d\eta_i}{d\mu_i} x_{ij} \frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{n} W_i x_{ij} x_{ik}$$

So $\mathbf{A} = (\mathbf{X}' W \mathbf{X})$, with $W = \text{diag}[W_i] = \text{diag}[\text{var}^{-1}(Z_i)]$.

The new estimate satisfies

$$\mathbf{A}\boldsymbol{\beta}^{(1)} = \mathbf{A}\boldsymbol{\beta}^{(0)} + \mathbf{U}^{(0)}$$

and notice that $j - th$ entry of the vector $\mathbf{A}\boldsymbol{\beta}^{(0)}$ is

$$[\mathbf{A}\boldsymbol{\beta}^{(0)}]_j = \mathbf{X}_j' W \mathbf{X}_j \beta_j = \sum_{i=1}^{n} W_i x_{ij} \eta_i$$

This means that the j-th entry of

$$\mathbf{A}\boldsymbol{\beta}_j^{(1)} = \sum_{i=1}^{n} W_i x_{ij} \left\{\eta_i - (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}\right\}$$

So

$$\mathbf{A}\boldsymbol{\beta}^{(1)} = \mathbf{X}' W \mathbf{z}^{(0)}$$

And the solution to this equation is the weighted regression estimate

$$\boldsymbol{\beta}^{(1)} = (\mathbf{X}' W \mathbf{X})^{-1} \mathbf{X}' W \mathbf{z}^{(0)}$$

Which is what the suggested procedure does.

5.4.5 Initial values

One advantage of this procedure is that we can use the use the transformed data as initial values for the $\mathbf{z}$. This sometimes presents a problem. For example when modeling with Poisson and we obtain a 0 then the starting value will be $\log(0)$. Problems like these need to be addressed. Notice, R does it for you.

5.4.6 Summary

- The MLE parameter $\hat{\beta}$ satisfy a self-consistency relationship: they are the coefficients of a weighted least squares fit, where the responses are

  $$z_i = x_i'\hat{\beta} + \frac{y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)},$$

  and the weights are $\hat{p}_i(1-\hat{p}_i)$. This connection implies the following

- The weighted residuals of sum-of-squares is the familiar Pearson chi-square statistic.

  $$\frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)},$$

  a quadratic approximation of the variance

- Asymptotic results give tell us the coefficients converge to $N(\beta, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$.

5.5 Logistic Regression versus LDA

Notice logistic regression provides a similar solution to LDA.

Using Bayes theory we can show that

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k+\mu_K)'\mathbf{\Sigma}^{-1}(\mu_k-\mu_K) + x'\mathbf{\Sigma}^{-1}(\mu_k - mu_K)$$

which can be re-written as

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \alpha_{0k} + \alpha_k' x.$$

For logistic regression we explicitly write

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \beta_{0k} + \beta_k' x.$$

The difference comes from how the parameters are estimated. Notice that the estimates of the αs use the parameters of the conditional distribution of x which was assumed to be normal.

Thus, the main difference is that LDA imposes a distribution assumption on X which, if a good, will result in more efficient estimates. If the assumption is in fact true the estimates will be 30% more efficient. Logistic regression is conditional methodology. We condition on X and do not specify any distribution for it. This presents a big advantage in cases were we know the X can't be normal, e.g. categorical variables.

However, in practice the two methods perform similarly. Even in extreme cases with categorical covariates.