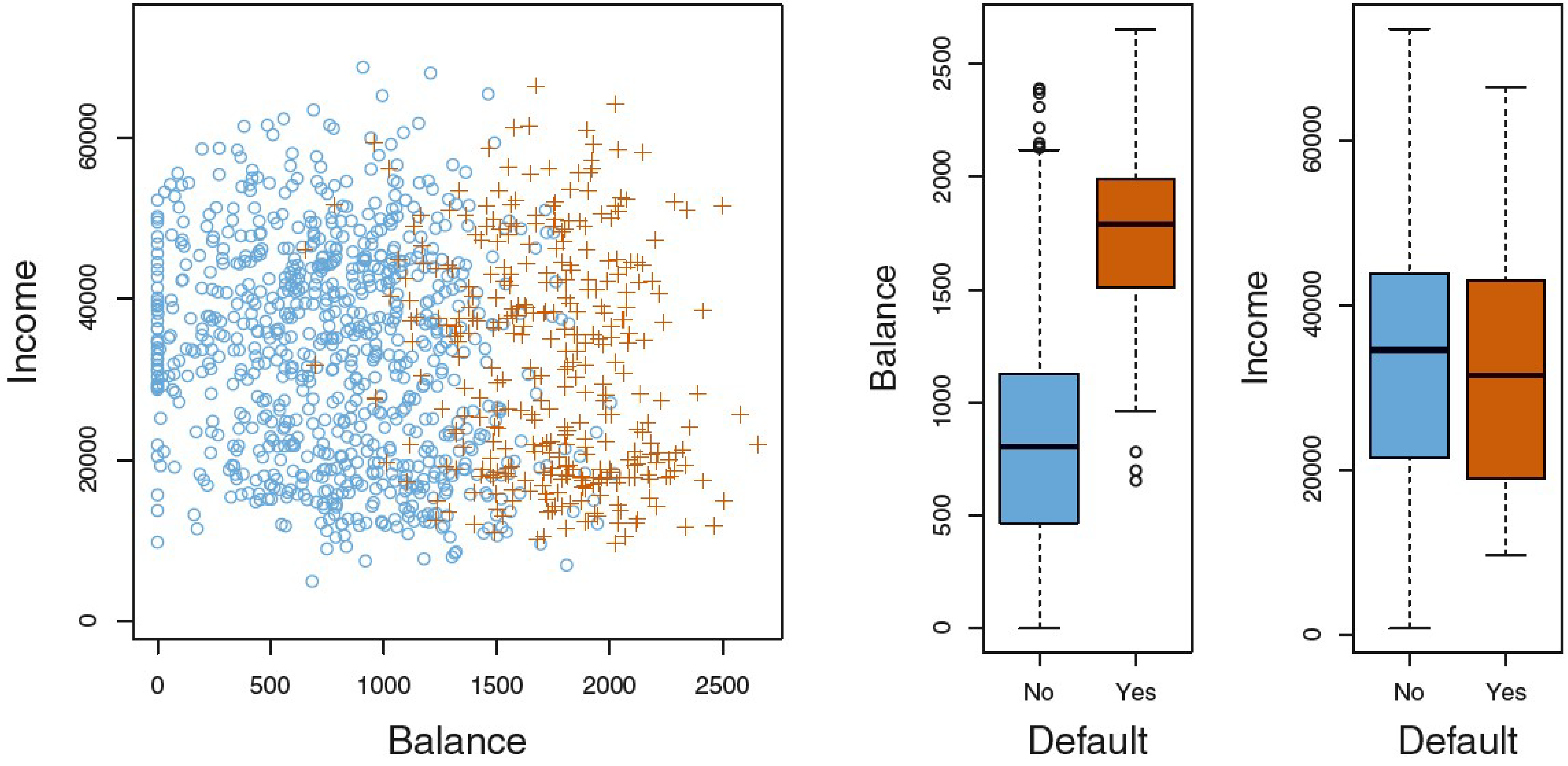
Chapter 4: Classification

- Response variable is qualitative
  1. Logistic regression
  2. Linear discriminant analysis
  3. *K*-nearest neighbors
  4. Poisson regression

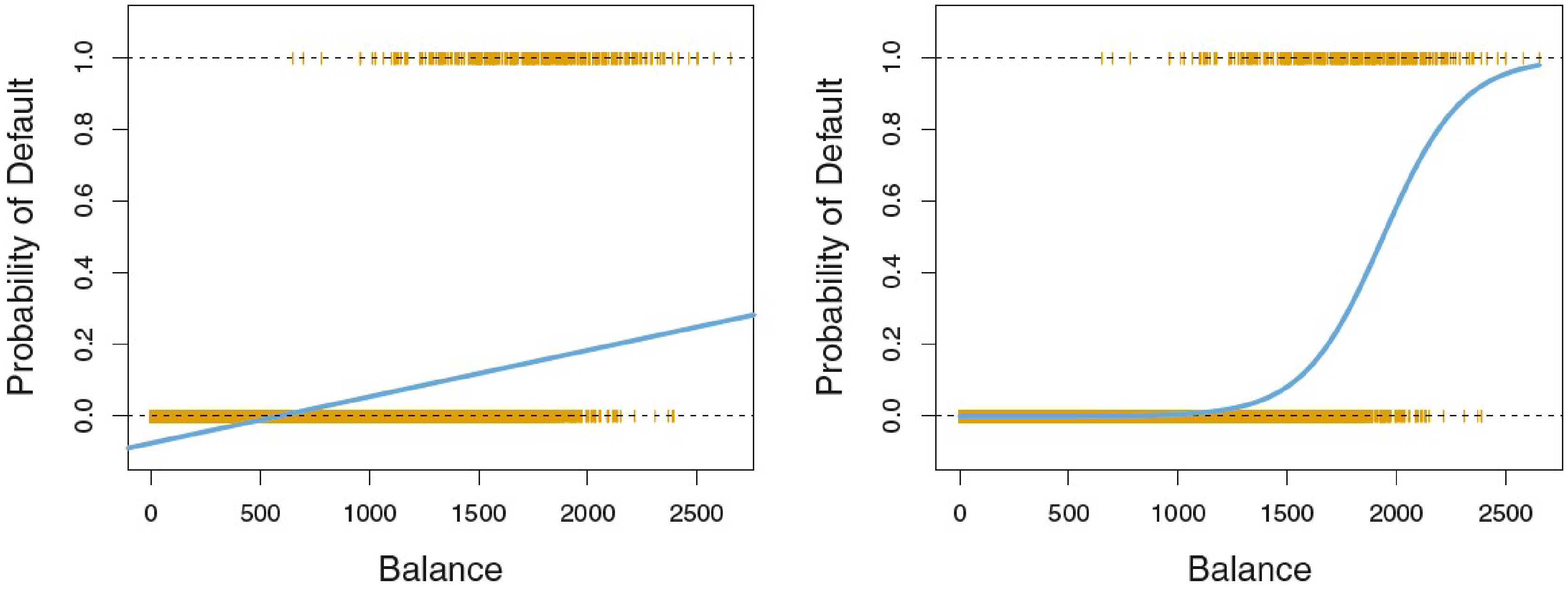
Overview

- Use phenotypic of genetic data to classify species or populations
- Can default be predicted from credit card balance and income?



Logistic Regression

- When there are two outcomes, like alive (0) or dead (1), we can predict the probability of either outcome as , $p(X) = Pr(Y = 1|X)$.



The blue line on the left is a straight line and on the right a logistic equation

Logistic Regression

- $p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1+e^{\beta_0 + \beta_1 X}}$ and $1 - p(X) = \frac{1}{1+e^{\beta_0 + \beta_1 X}}$
- The odds, $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$
- Log-odds, or logit, $log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$
- This model can be expanded to include multiple predictors logit=$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Logistic Regression (cont.)

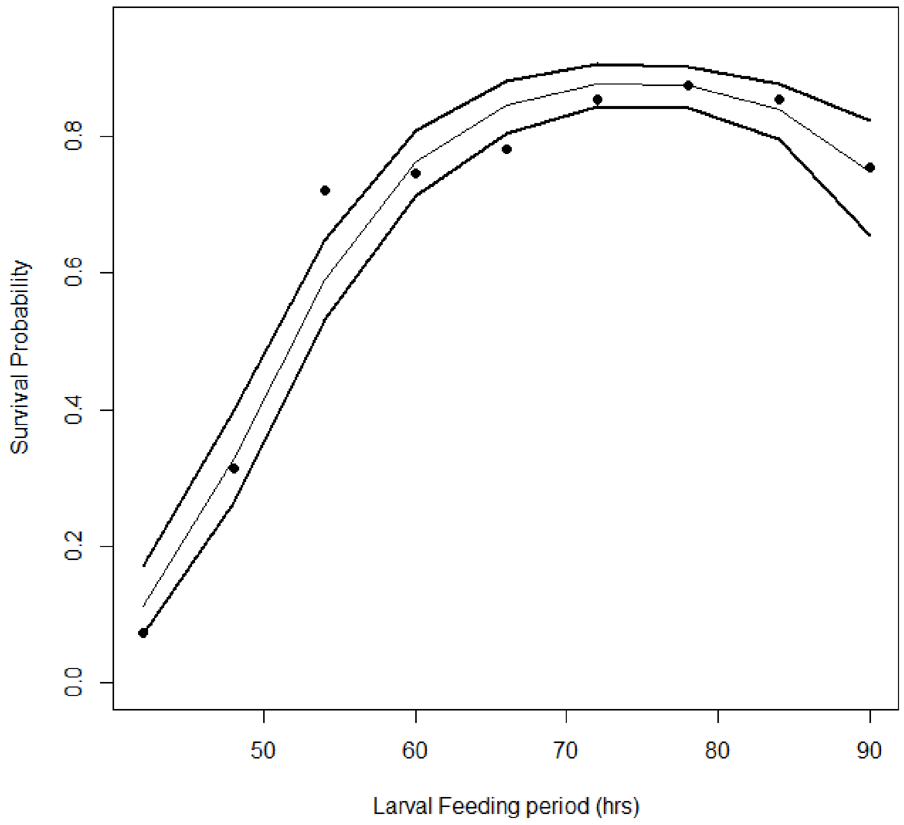
- These coefficients will be estimated by maximum likelihood
- Likelihood function $=l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0}(1 - p(x_i))$,which is the probability of observing the sample. β's are chosen to maximize the likelihood function. This can be done taking derivatives as with the least squares estimates or if there are constraints on the parameter values a technique like Lagrange multipliers can be used.

Logistic Regression: Example

Drosophila larvae are allowed to feed on excess yeast past for various periods.

At each sample time larvae are removed and placed in vials with agar only (no food).

The number of adult survivors at each time sample are recorded.



Logistic Regression: Example

selection	replicate	age	alive	dead
utb	1	12	0	30
utb	1	24	0	30
utb	1	30	0	30
utb	1	36	0	30
utb	1	42	2	28
utb	1	48	2	28
utb	1	54	9	21
utb	1	60	25	5
utb	1	66	25	5
utb	1	72	24	6

etc

There is a second selection treatment “tb” a control for “utb” -> larvae raised in urea food.

This data file (survival.data) is read into R,

```
viability.data<- read.table("survival.data ",header=TRUE)
```

Logistic Regression: Example

A linear model will not do well with these data so a quadratic model is used

```
#So let’s also try the analysis on hours 42 and above
viability.data2<- viability.data[viability.data$age>36,]

viability.data3<- cbind(viability.data2,viability.data2$age^2)

dead.data2<- as.matrix(viability.data2[,4:5])

names(viability.data3)<- c("selection", "replicate", "age", "alive", "dead","age2")

viability.glm3<- glm(dead.data2~ age*selection+age2*selection,data=viability.data3,
                     family=binomial)
```

Logistic Regression: Example

```
> summary(viability.glm3)
Call:
glm(formula = dead.data2 ~ age * selection + age2 * selection,
    family = binomial, data = viability.data3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3493  -1.5558  -0.1557   1.7634   4.3268

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)$\beta_0$       -1.919e+01  1.505e+00 -12.754  < 2e-16 ***
age $\beta_1$               5.673e-01  4.663e-02  12.166  < 2e-16 ***
selectionutb $\alpha_0$      3.969e+00  2.160e+00   1.838  0.06613 .
age2 $\beta_2$             -3.800e-03  3.452e-04 -11.009  < 2e-16 ***
age:selectionutb $\alpha_1$ -1.514e-01  6.695e-02  -2.261  0.02373 *
selectionutb:age2 $\alpha_2$ 1.284e-03  4.983e-04   2.577  0.00997 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1234.99  on 89  degrees of freedom
Residual deviance:  365.04  on 84  degrees of freedom
AIC: 635.29

Number of Fisher Scoring iterations: 5
```

$$log\left(\frac{p_i(t)}{1-p_i(t)}\right) = \beta_0 + \delta_i\alpha_0 + (\beta_1+\delta_i\alpha_1)t + (\beta_2+\delta_i\alpha_2)t^2$$

where tb (*i*=1) and utb (*i*=2) and δ_i=0 if *i*=1 1 otherwise.

To see all the side results that you can use write,
attributes(viability.glm3)
This may reveal stuff not documented in the help page.

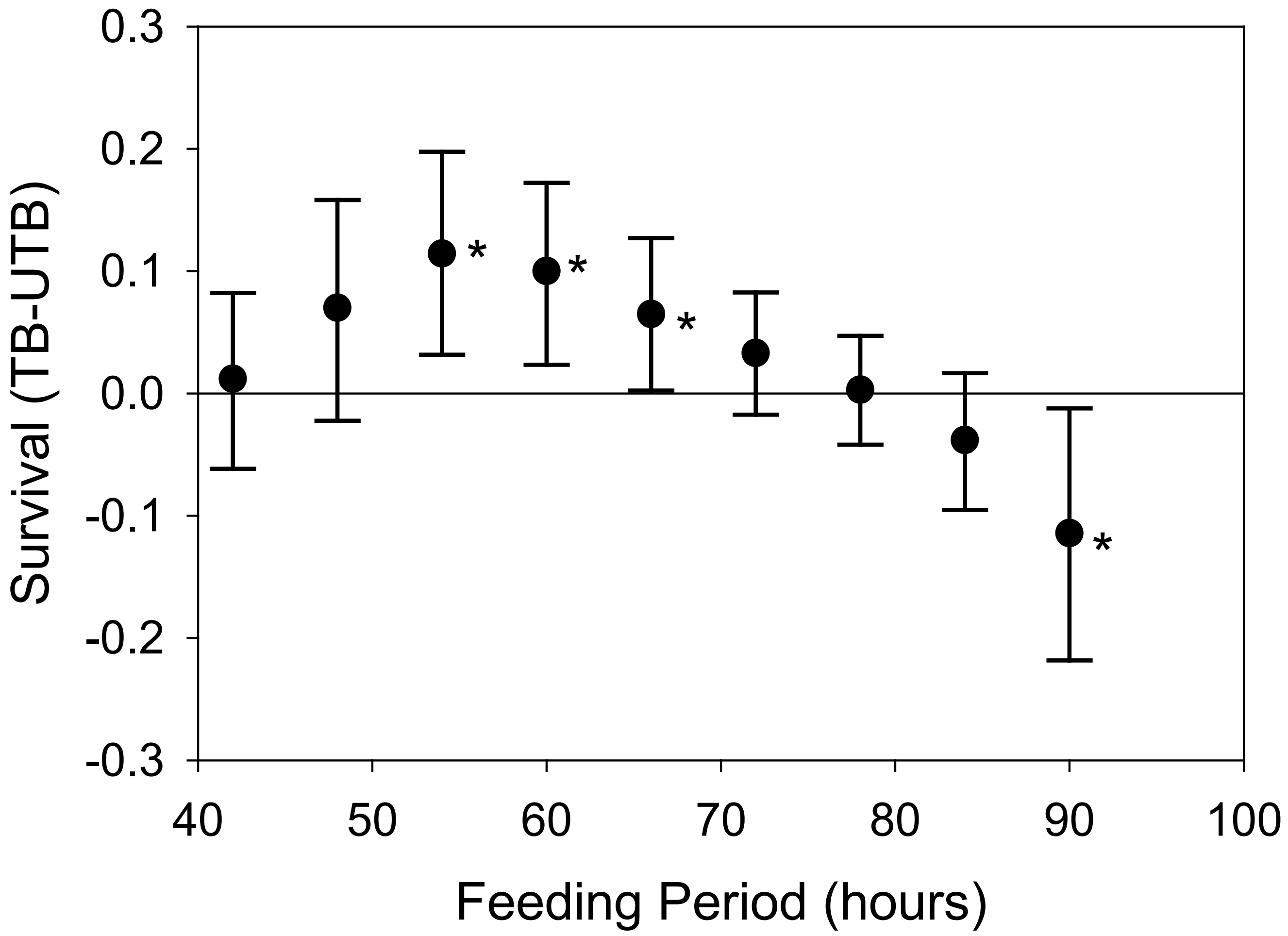
Logistic Regression: Example

- An important question here is if the predicted survival probabilities in the TB and UTB populations are significantly different.
- Use predictions rather than individual observations at each time interval since these are based on all the data.
- The R predict function can generate standard errors for the logit function but won’t generate tests between different predictions.
- However, we can generate random samples of the regression parameters, make predictions for TB and UTB and save the differences.

Logistic Regression: Example

```
library(mvtnorm)

logit2.ftn<- function(t,a0,a1,a2){
  y<- a0+a1*t+a2*t^2
  exp(y)/(1+exp(y))
}

cov.b<- summary(viability.glm3)$cov.unscaled
mean.b<- coefficients(viability.glm3)
age.range<- c( 42, 48, 54, 60, 66, 72,78,84,90)
sim.num<-5000
conf.band<- sapply(1:sim.num, function(x) {
  #generate random parameters and make sure they are all >0
  b.x<- rmvnorm(1,mean= mean.b,sigma=cov.b)
  tb.x<- c(b.x[1], b.x[2], b.x[4])
  utb.x<- c(b.x[1]+ b.x[3], b.x[2]+ b.x[5], b.x[4]+ b.x[6])
      logit2.ftn(age.range,tb.x[1], tb.x[2],tb.x[3])- logit2.ftn(age.range,utb.x[1],
                 utb.x[2],utb.x[3])
})
```

Logistic Regression: Example



Multinomial Logistic Regression

- If there are now $k>2$ categories we can extend the logistic equation method in two ways.
- The first requires that you pick one category as the baseline and then estimate coefficients for the remaining k-1 categories. The complication is that all odds ratios are defined relative to the baseline category.
- The *softmax coding* treats all categories symmetrically and is used in other statistical learning methods.

Multinomial Logistic Regression, softmax coding

- Assume we have *K* categories, then the Pr(*Y*=*k*|*X*=*x*) is

$$\frac{e^{\left(\beta_{k0}+\beta_{k1}x_1\ +\cdots+\beta_{kp}x_p\ \right)}}{\sum_{i=1}^{K} e^{\left(\beta_{i0}+\beta_{i1}x_1\ +\cdots+\beta_{ip}x_p\ \right)}}$$

- Now the odds ratio between the *kth* and *jth* class is,

$$log\left(\frac{Pr(Y=k|X=x)}{Pr(Y=j|X=x)}\right)=\left(\beta_{k0}-\beta_{j0}\right)+\left(\beta_{k1}-\beta_{j1}\right)x_1+..+\left(\beta_{kp}-\beta_{jp}\right)x_p$$

- In R use the *nnet* package, and the *multinom* function.

Linear Discriminant Analysis

- LDA will outperform logistic regression when, (i) classes are well separated and (ii) *n* is small and the distribution of the predictors is approximately normal.
- LDA can also handle multiple response classes
- Three ways to find LDA predictors, (i) Bayes classifiers, (ii) find a scaling that maximizes the mean differences between response classes, and (iii) use the Mahalonobis distance.

Bayes Theorem

- Suppose we have K response classes, $K \geq 2$. π_k is the prior distribution of class k. May be uniform, or estimated from sample.
- Let $f_k(X)$ be the probability density or mass function of X or $\Pr(X=x|Y=k)$
- From Bayes Theorem we have, $\Pr(Y=k|X=x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$
- If we assume the predictors have a normal distribution then we can get some specific results, assuming all $f_k(X)$ have a common variance (σ^2) and each has mean, μ_k.
- Assign observation to class-k if, $x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$ is largest.

Maximize distance between distributions

- Suppose $\boldsymbol{X}_1 \sim \mathrm{MVN}(\mu_1,\Sigma)$ and $\boldsymbol{X}_2 \sim \mathrm{MVN}(\mu_2,\Sigma)$.
- The sample means, $\bar{X}_1$ and $\bar{X}_2$ have sample variances, S/N_1 and S/N_2. We will do a linear transformation (e.g $\mathbf{a}\boldsymbol{X}_1$) of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to separate their distributions.
- Find $\mathbf{a}^T=(a_1, \ldots,a_p)$ such that $t^2(\mathbf{a})$ is maximized, where

$$t^2(a) = \left[\frac{a^T(\bar{X}_1 - \bar{X}_2)}{\sqrt{a^T S a}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}\right]^2$$

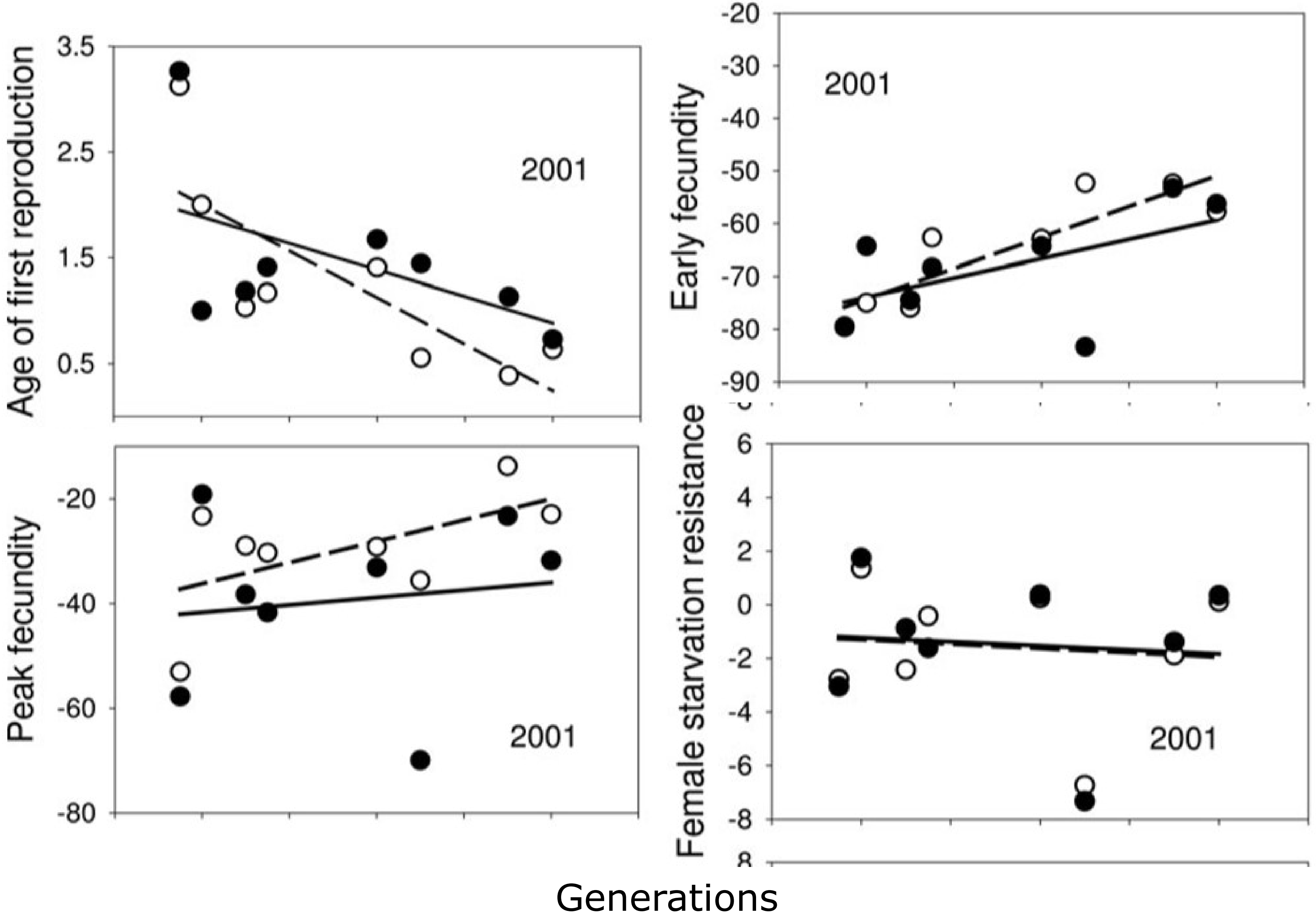
- To do this maximization use Lagrangian multipliers with the solution that $\mathbf{a}= S^{-1}(\bar{X}_1 - \bar{X}_2)$
- Then for a set of features $\mathbf{X}$, classify according to whether $\mathbf{a}^T\mathbf{X}$ is closer $a^T\bar{X}_1$ or $a^T\bar{X}_2$.

Mahalonobis distance

- Another way to derive the linear discriminant function.
- Find the Mahalonobis distance between an unknown (***X***) and the mean of every group and assign ***X*** to the group ***X*** is closest to, e.g. find the min over all *i*,

$$D_i^2 = (X - \bar{X}_i)^T S^{-1} (X - \bar{X}_i)$$

Linear Discriminant Analysis: example



Wild populations of *Drosophila subobscura* brought into the lab and they adapt over 22 generations.

Four different phenotypes are monitored and are believed to be related to fitness.

Can these component fitness measures be used to distinguish flies at the start of selection and flies at the end?

Linear Discriminant Analysis: example

generation	age	earlyfec	peakfec	rf	rm	location	replicate	subpop	year
4	0.550	-100.050	-100.450	-10.500	-2.100	sintra	1	a	1998
4	0.550	-42.050	-20.450	-10.500	-2.100	sintra	1	a	1998
4	-0.450	-50.050	-23.450	1.500	3.900	sintra	1	a	1998
4	1.550	-86.050	-52.450	-4.500	-2.100	sintra	1	a	1998
4	-0.450	-69.050	-79.450	-10.500	9.900	sintra	1	a	1998
4	1.550	-77.050	-41.450	-10.500	-2.100	sintra	1	a	1998
4	0.550	-55.050	-46.450	-10.500	3.900	sintra	1	a	1998
4	1.550	-72.050	-54.450	1.500	-2.100	sintra	1	a	1998
4	0.550	-53.050	-36.450	-4.500	3.900	sintra	1	a	1998
4	0.550	-46.050	-46.450	1.500	33.900	sintra	1	a	1998
4	-0.450	-50.050	-83.450	7.500	3.900	sintra	1	a	1998
4	-0.450	30.950	-5.450	1.500	-2.100	sintra	1	a	1998
4	1.550	-51.050	-4.450	-16.500	-8.100	sintra	1	a	1998
4	-0.450	-25.050	-44.450	-4.500	-2.100	sintra	1	a	1998
4	0.550	-33.050	-18.450	-16.500	-8.100	sintra	1	a	1998
4	0.550	-58.050	-28.450	1.500	-2.100	sintra	1	a	1998
4	1.550	-86.050	-114.450	-16.500	-8.100	sintra	1	a	1998

Etc

Age = age at first reproduction; earlyfec= early fecundity; peakfec = peak fecundity; rf= female starvation resistance.

All values are shown relative to a lab adapted control run at the same time.

Linear Discriminant Analysis: example

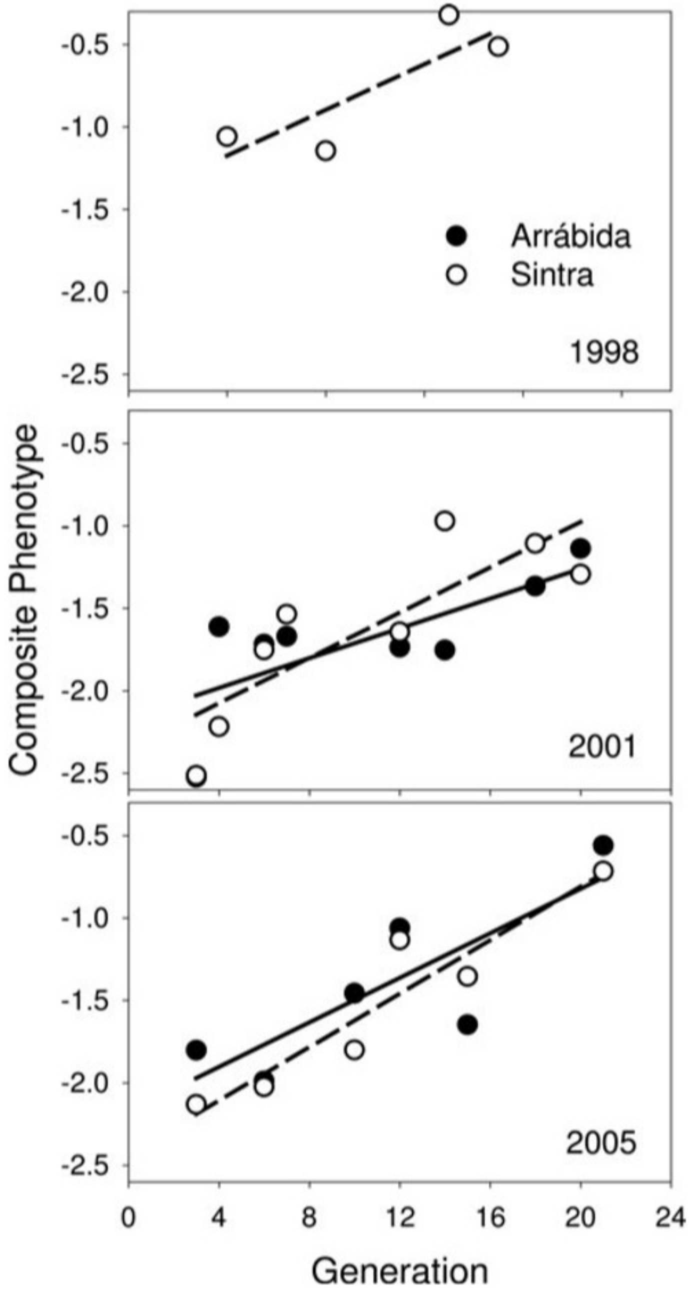
```
> port.lda <- lda(na.omit(port),port.name)
> port.lda
Call:
lda(na.omit(port), port.name)

Prior probabilities of groups:
        b         f
0.5352697 0.4647303

Group means:
        age  earlyfec   peakfec         rf
b 2.0833152 -73.04842 -51.57066 -0.1002842
f 0.3036518 -45.21231 -34.16832 -0.3851994

Coefficients of linear discriminants:
                   LD1
age      -0.4087049732
earlyfec  0.0214200071
peakfec  -0.0088328690
rf       -0.0005101683
```

With no formula in lda, the first entry are the features and the second vector are the groupings or class membership for each observation (b=begin, f=final).

To assess the importance of each factor the variable should be scaled first, e.g. divide each observation by the standard error of that variable , e.g. port$age<-port$age/sd(port$age).

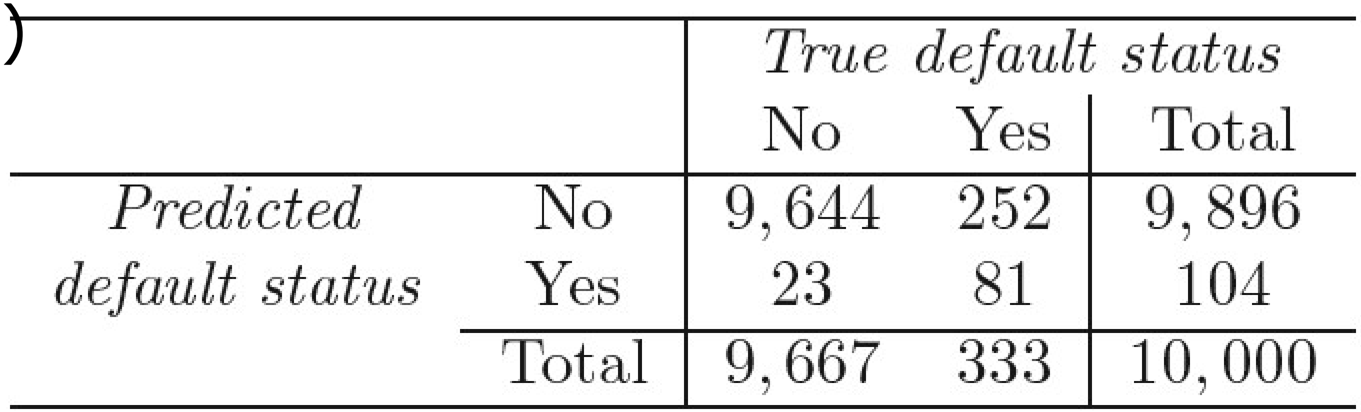
Linear Discriminant Analysis: example

- The discriminant function has given an objective measure of how the combined phenotypes change from the start to the end of the period of adaptation. Perhaps this is a way of summarizing the increase in fitness.
- While there is no evolutionary theory that arrives at the lda weightings it is clear that natural selection weights fitness components differently.



LDA Error Analysis

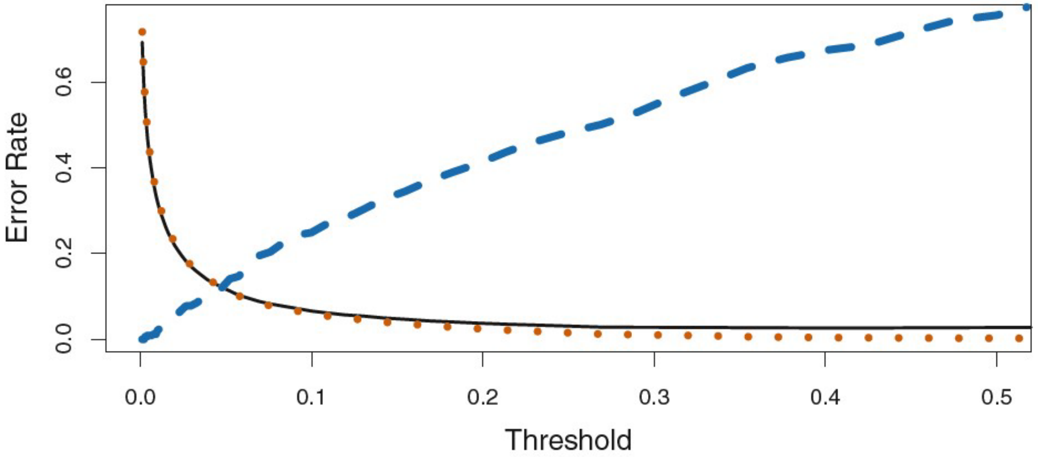
- A confusion matrix shows the true vs. the predicted status of samples in an LDA analysis
- Assume the user is interested in predicting only one of the two states (default).
- % of true defaulters identified= 81/333=24% (sensitivity, power, 1-type II error)
- % of true non-defaulters identified= 9644/9667=99.8% (specificity, 1-type I error)

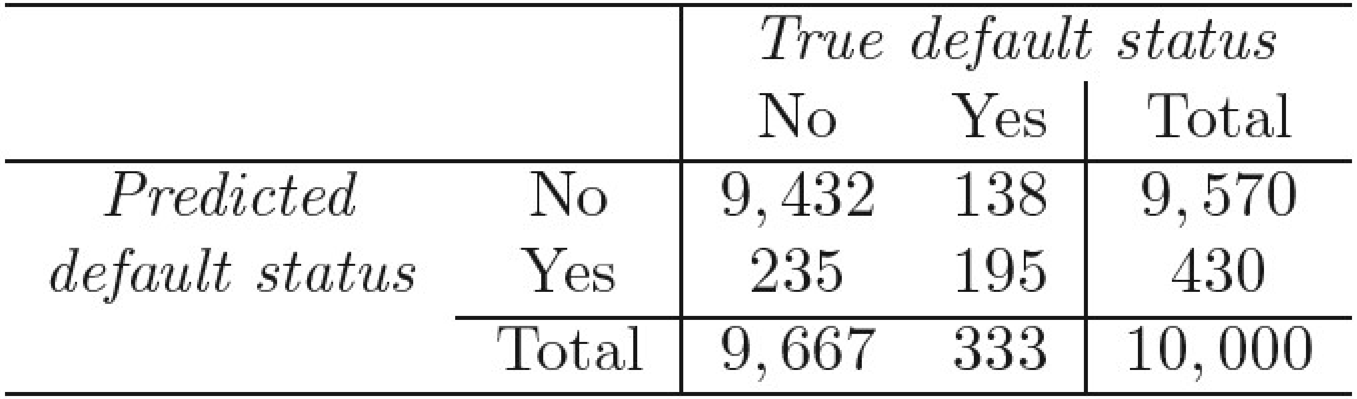


		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

LDA Error Analysis

- How to improve the prediction of defaulters?
- Lower your criteria!
- Standard is for two choices, Pr(default=yes|X=x)>0.5. Lower this cutoff to say 0.2.
- Now the sensitivity has increased to 195/333= 41%
- But specificity has decreased to 9432/9667= 97.6%

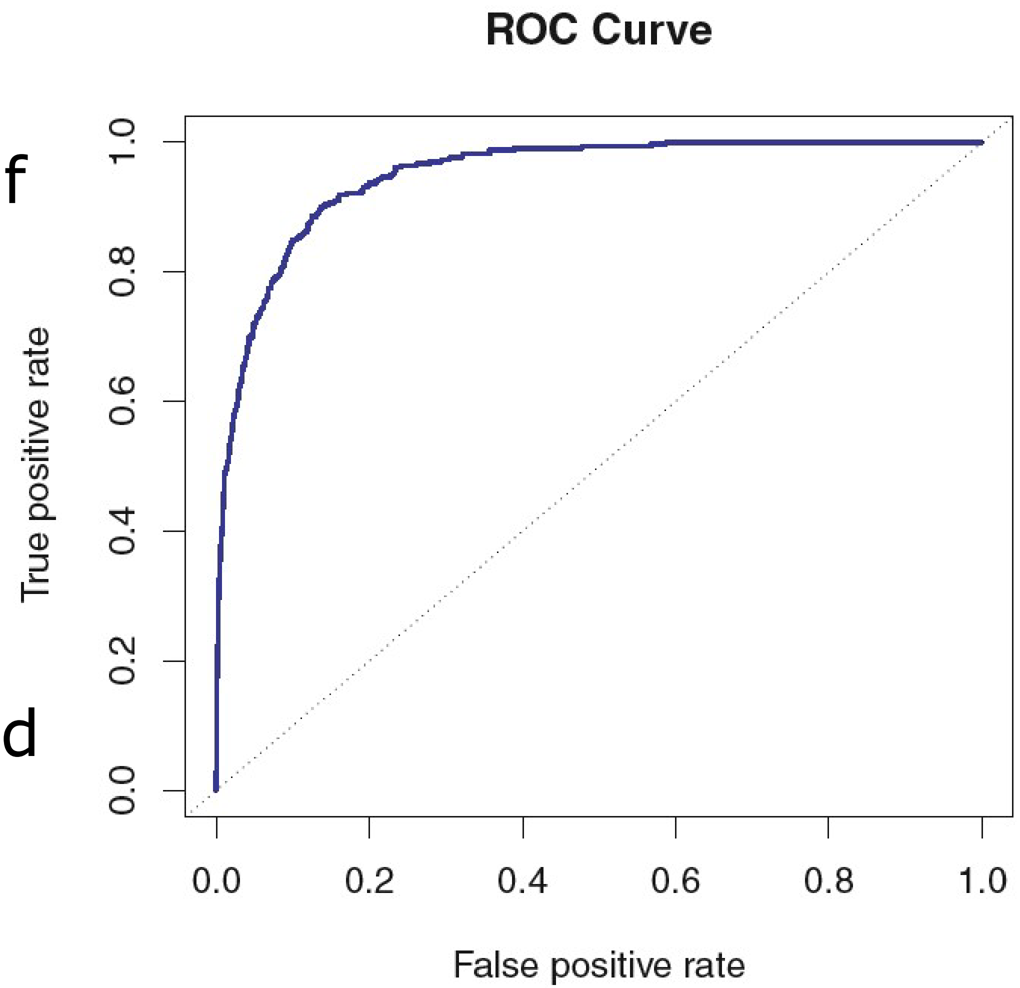




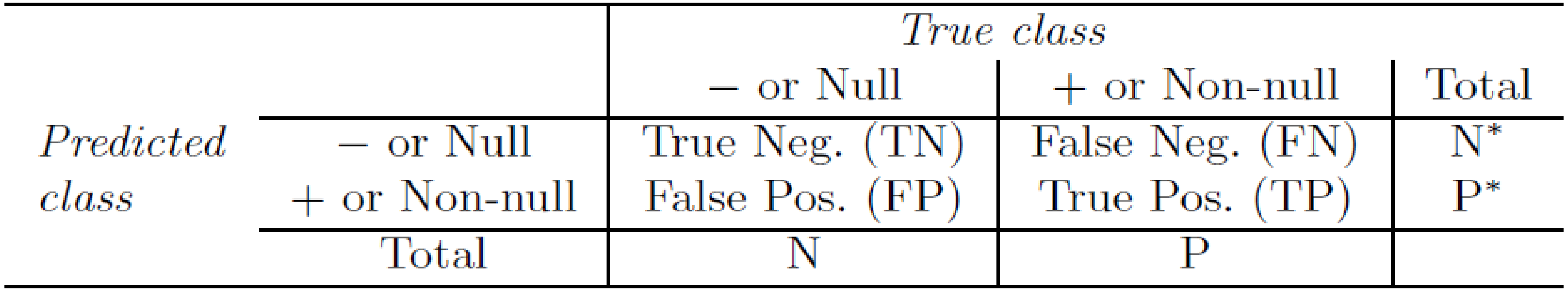
		True default status		
		No	Yes	Total
Predicted	No	9,432	138	9,570
default status	Yes	235	195	430
	Total	9,667	333	10,000

ROC Curve

- ROC stands for receiver operating characteristics.
- This is a plot of the true positive rate (sensitivity) vs false positive rate (=#of incorrect predicted defaults/total non-defaults)
- Ideally you would like the true positive rate to be very high at very low false positive rates. Thus, the best methods would have an area under the curve (AUC) close to 1.0. Just guessing should get an AUC of 0.5.

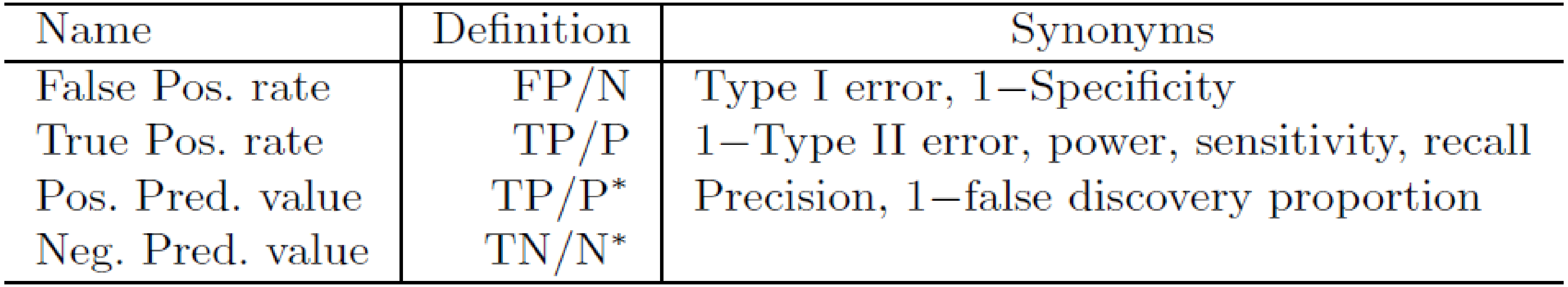


Statistical Nomenclature about errors



		True class		
		– or Null	+ or Non-null	Total
Predicted	– or Null	True Neg. (TN)	False Neg. (FN)	N^*
class	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P^*
	Total	N	P	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*



Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1–Specificity
True Pos. rate	TP/P	1–Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1–false discovery proportion
Neg. Pred. value	TN/N^*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Quadratic Discriminant Analysis

- With this type of analysis we allow the variance-covariance matrix of each class to be different.
- This will reduce the bias of the LDA predictor but increase the variance due to the great increase in the number of parameters that must be estimated.
- The Bayesian classifier assuming normally distributed predictor variables yields, $\delta_k = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}log|\Sigma_k| + log\pi_k$

Quadratic Discriminant Analysis

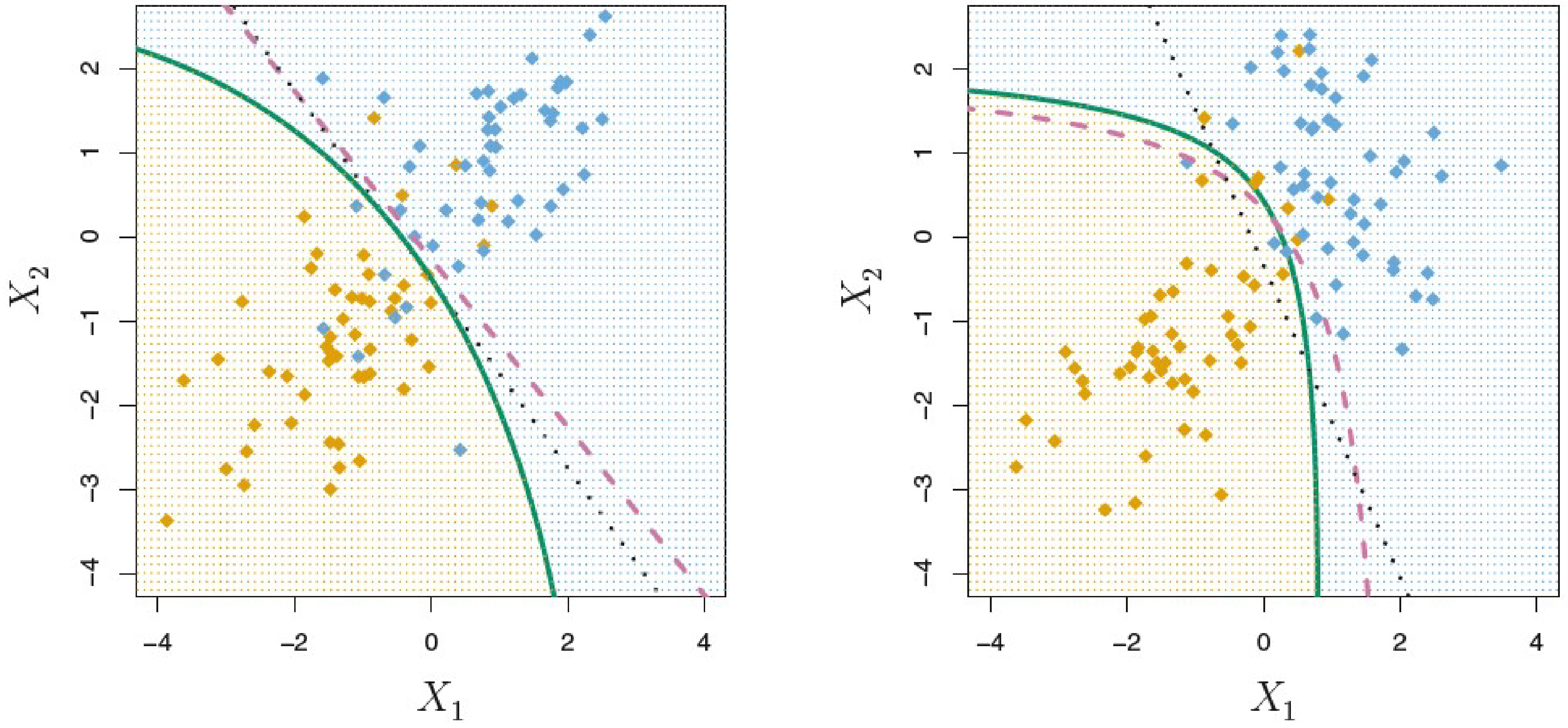
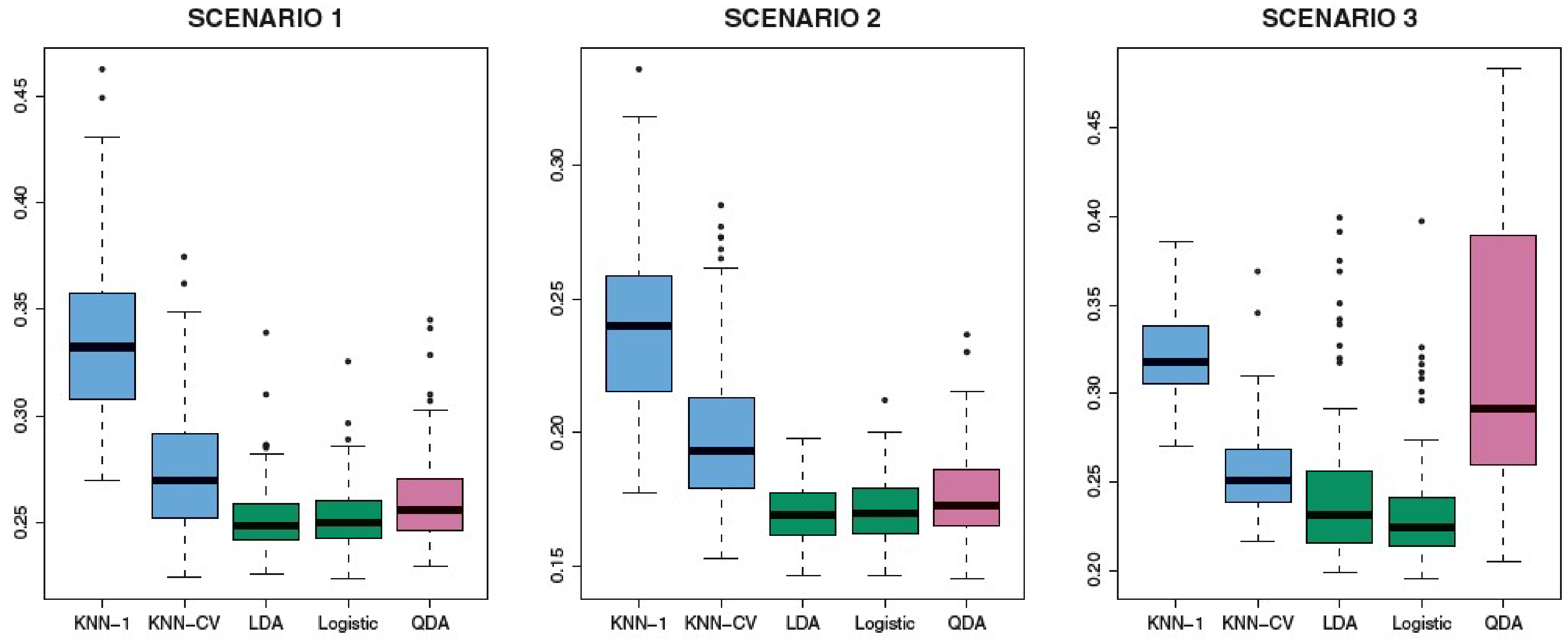


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Comparing Methods

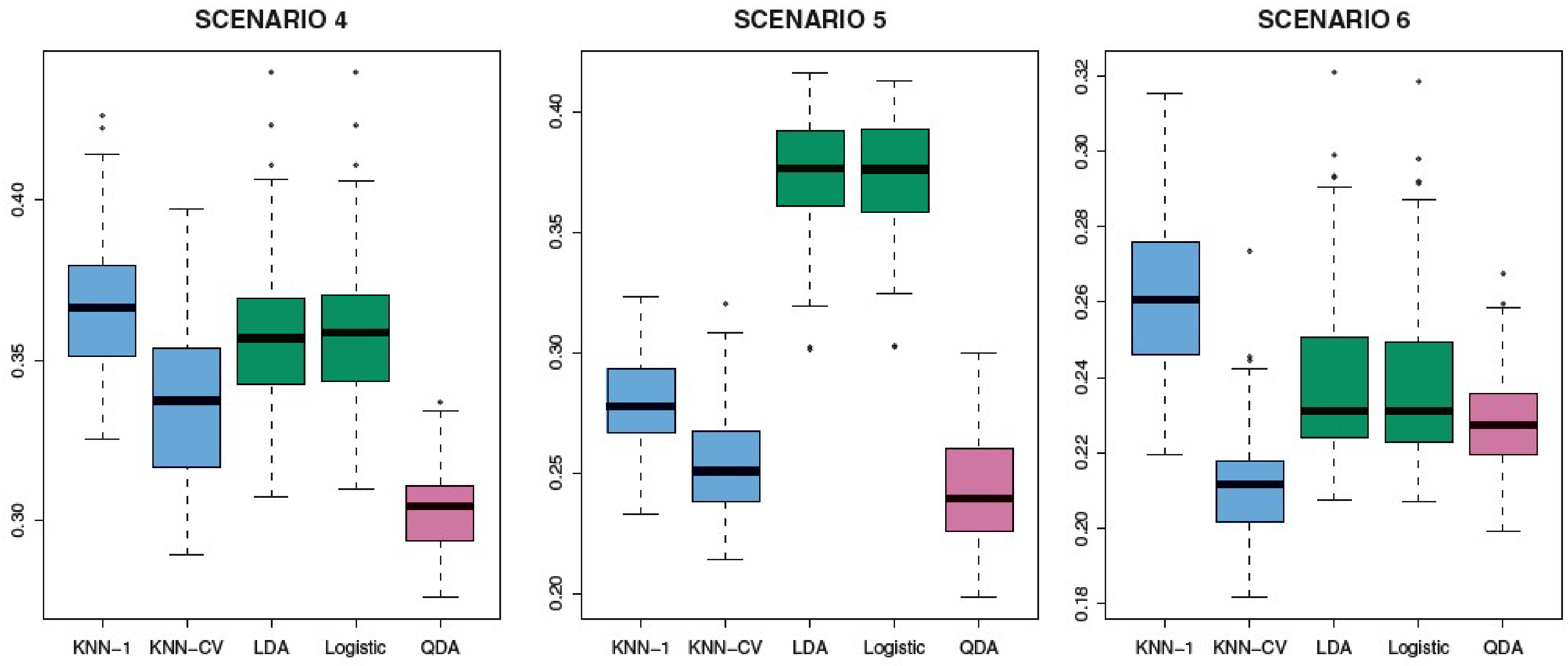
- LDA and logistic regression produce linear decision boundaries but estimate model parameters differently.
- If observations are truly normal then LDA may outperform logistic regression. However, if this the distribution is not normal logistic regression may be better.
- KNN is non-parametric so it should do well when the decision boundary is non-linear. But you can’t weight the importance of the predictor variables.
- QDA may be viewed as a compromise between LDA, logistic regression and KNN.

Linear Simulations (p=2)



Scenario 1: uncorrelated predictors from a normal distribution, common covariance
Scenario 2: Same as 1 except predictors have a correlation of -0.5
Scenario 3: Samples drawn from a *t*-distribution, now logistic regression does better
KNN-CV, cross-validation with the training set to help choose neighborhood size.

Non-linear Simulations (p=2)



Scenario 4: class 1 predictors normal distribution with correlation 0.5, second class predictors normal distribution with a correlation of -0.5. Now QDA does best.
Scenario 5: Observation are from a normal distribution with uncorrelated predictors. However, the response is generated from logistic equation with terms $X_1^2, X_2^2, and\ X_1 \times X_2$. This generates a quadratic decision boundary.
Scenario 6: Same as 5 except an even more complicated non-linear function.

Poisson Regression

- Suppose the response variable you want to predict is a non-negative integer, and its variance increases with the mean value.
- The increasing variance violates the assumptions of linear regression but also violates the binomial/multinomial assumptions of logistic regression.
- These response variables may have a Poisson distribution where,

$$Pr(Y=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}\ for\ k=0,1,2,\dots$$

- The mean and the variance of the Poisson distribution is λ.
- Example: bikeshare data in the textbook.

Poisson Regression

- Rather than model the response variables the mean, λ, is modeled eg. $\lambda(X_1, X_2, \ldots, X_p)$.
- The linear model is, $log(\lambda(X_1, X_2, \ldots, X_p)) = \beta_0 + \beta_0 X_1 + \cdots + \beta_p X_p$which implies that, $\lambda(X_1, X_2, \ldots, X_p) = e^{\beta_0 + \beta_0 X_1 + \cdots + \beta_p X_p}$
- These regression coefficients will be estimated using maximum likelihood as was done with logistic regression.
- The likelihood of the observations, $(y_1,\mathbf{x}_1)$, $(y_2,\mathbf{x}_2)$,..,$(y_n,\mathbf{x}_n)$ is

$$\prod_{i=1}^{n} \frac{e^{-\lambda(\boldsymbol{x}_i\ )}\lambda(\boldsymbol{x}_i)^{y_i}}{y_i!}$$

Generalized Linear Models

- The R glm function has several options,

  ```
  glm(dead.data2~
  age*selection+age2*selection,data=viability.data3,
                       family=binomial)
  ```

- family has several options:
  family=gaussian -> linear regression
  family= binomial -> logistic regression
  family=poisson -> Poisson regression
  other options -> Gamma, inverse.gaussian, quasi, quasibinomial, quasipoisson

Comments on the Quasi-Binomial Distribution

- Binomial: $g(x) = \binom{n}{x} p^x (1-p)^{n-x}$

  Quasi-binomial: $g(x) = \binom{n}{x} p(p + x\varphi)^{x-1}(1 - p - x\varphi)^{n-x}$
- The extra parameter φ can inflate (or deflate) the variance relative to the binomial distribution ($\varphi = 0$).
- A population contaminated with individuals that show two (or more) different binomial probabilities will have a variance greater than the binomial.
- Example: niche overlap measures for two species contaminated with genotypic variation.