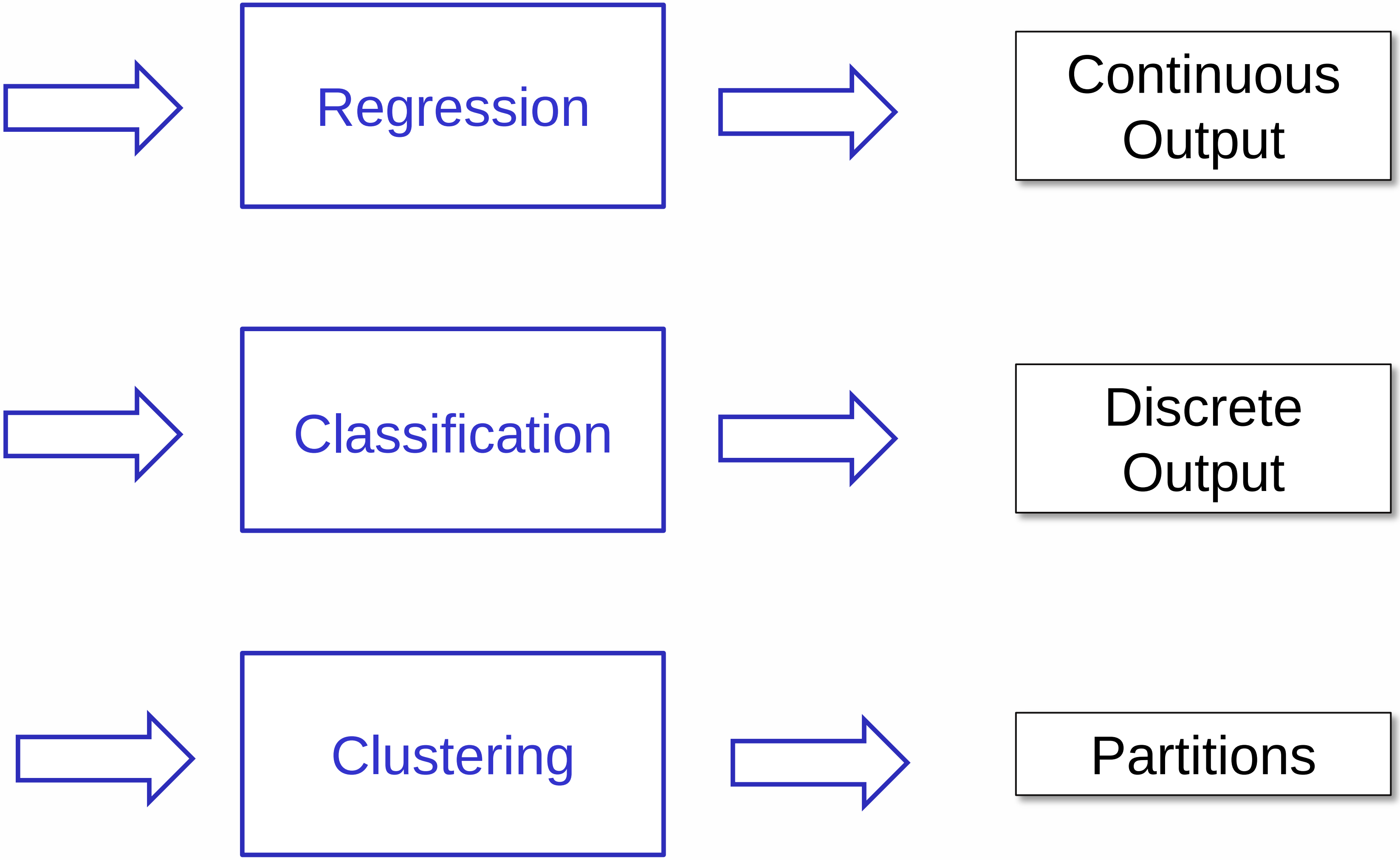
POLITECNICO DI MILANO

Pattern Analysis and Machine Intelligence

Linear Classification

Prof. Matteo Matteucci

Recall from the first lecture …



Example: Default dataset

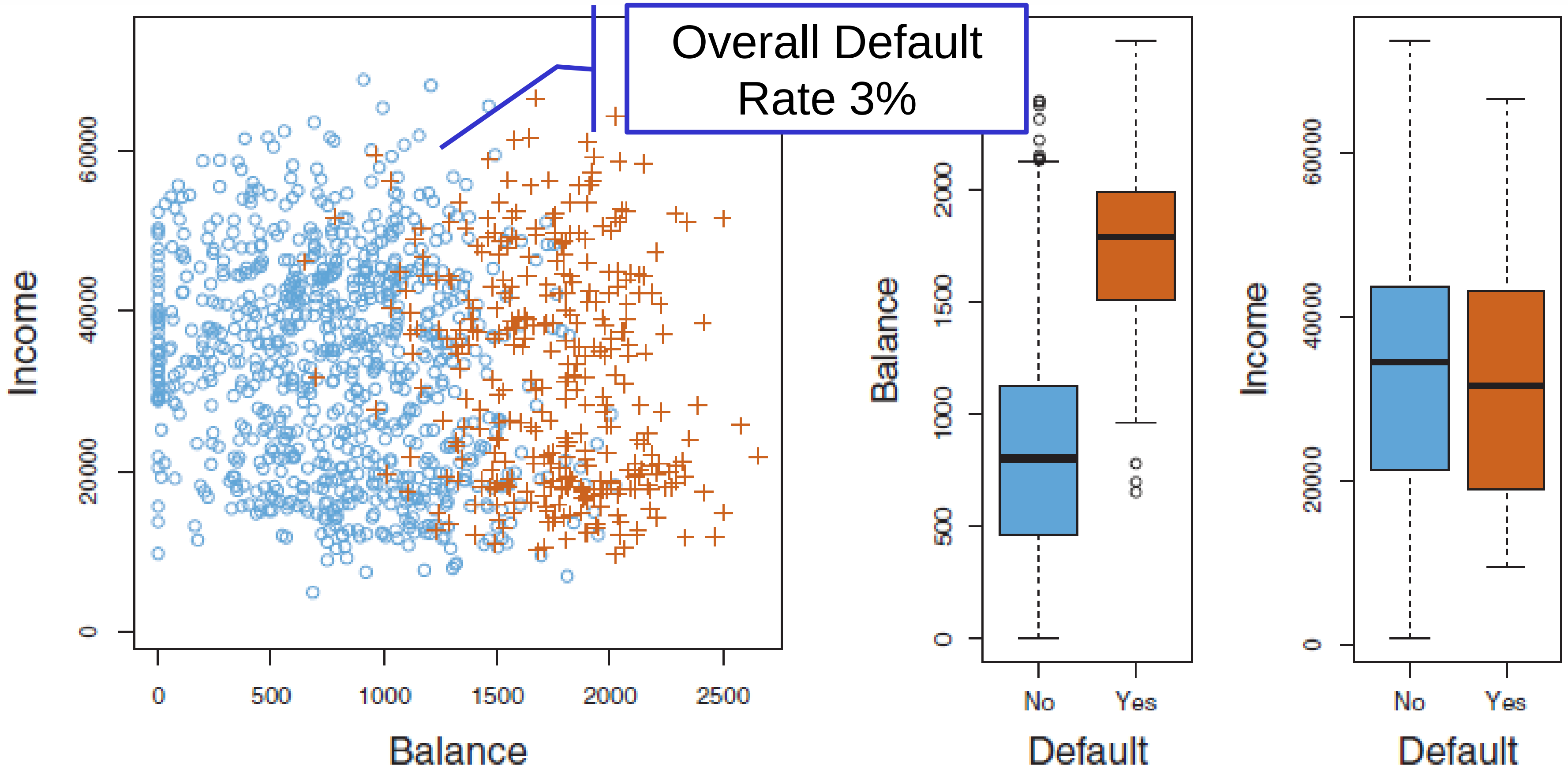


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

- Suppose to predict the medical condition of a patient. How should this be encoded?

  - We could use dummy variables in case of binary output

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

  but how to deal with multiple output?

  - Different encodings could result in different models

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases} \qquad Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}. \end{cases}$$

Recall here the Bayesian Classifier

- For a classification problem we can use the error rate i.e.

$$Error \quad Rate = \sum_{i=1}^{n} I(y_i \neq \hat{y}_i) / n$$

  - Where $I(y_i \neq \hat{y}_i)$ is an indicator function, which will give 1 if the condition $(y_i \neq \hat{y}_i)$ is correct, otherwise
  - The error rate represents the fraction of inc classifications, or misclassifications

The best classifier possible estimates the class posterior probability!!

- The Bayes Classifier minimizes the Average Test Error Rate

$$\max_j P(Y = j \mid X = x_0)$$

- The **Bayes error rate** refers to the lowest possible Error Rate achievable knowing the "true" distribution of the data

$$1 - E\left(\max_j \Pr(Y = j|X)\right)$$

Logistic Regression

- We want to model the probability of the class given the input

$$p(X) = \Pr(Y = 1|X)$$

$$p(X) = \beta_0 + \beta_1 X$$

but a linear model has some drawbacks

Example: Default data & Linear Regression

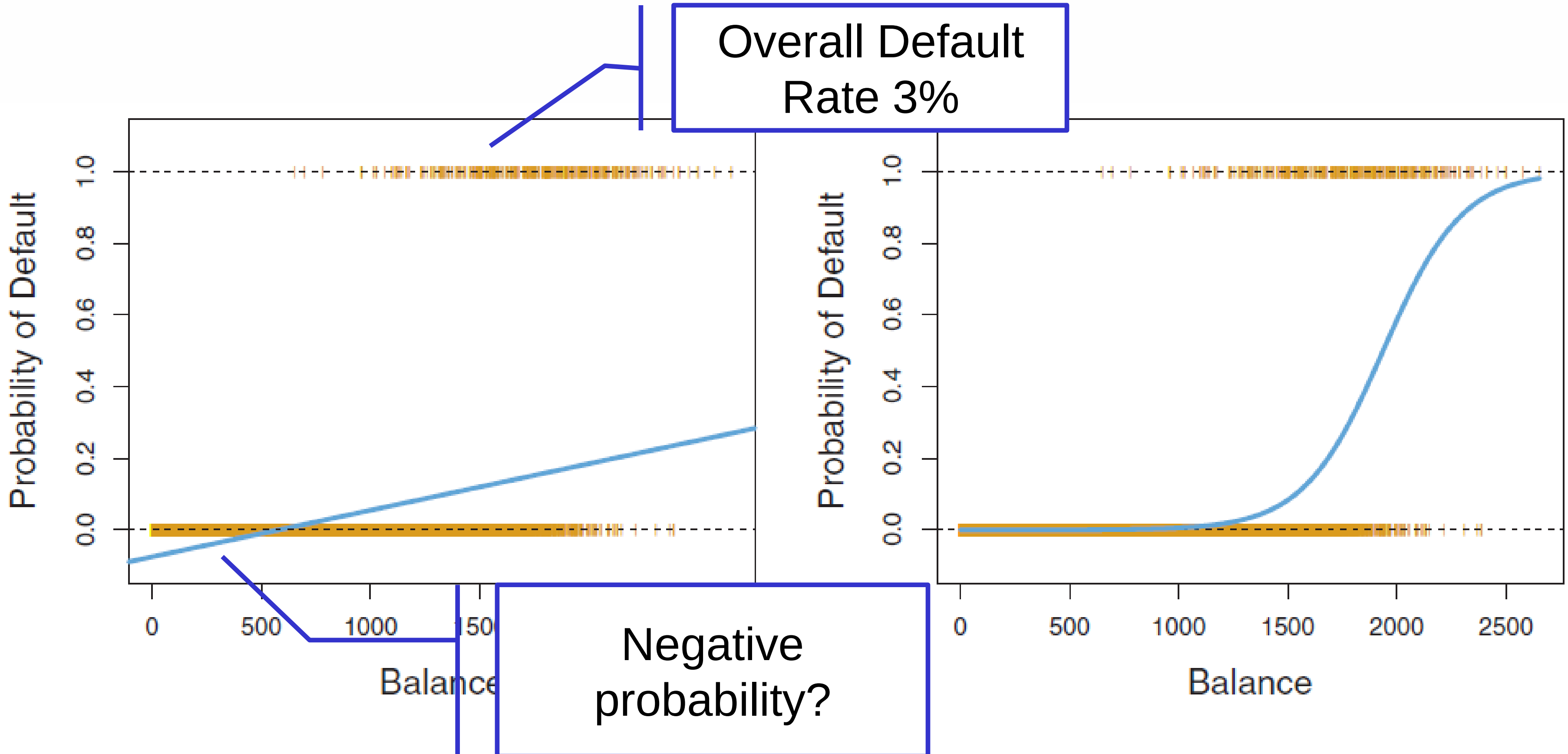


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic Regression

- We want to model the probability of the class given the input

$$p(X) = \Pr(Y = 1|X)$$
$$p(X) = \beta_0 + \beta_1 X$$

  but a linear model has some drawbacks

- Logistic regression solves the negative probability (ad other issues as well) by regressing the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Example: Default data & Logistic Regression

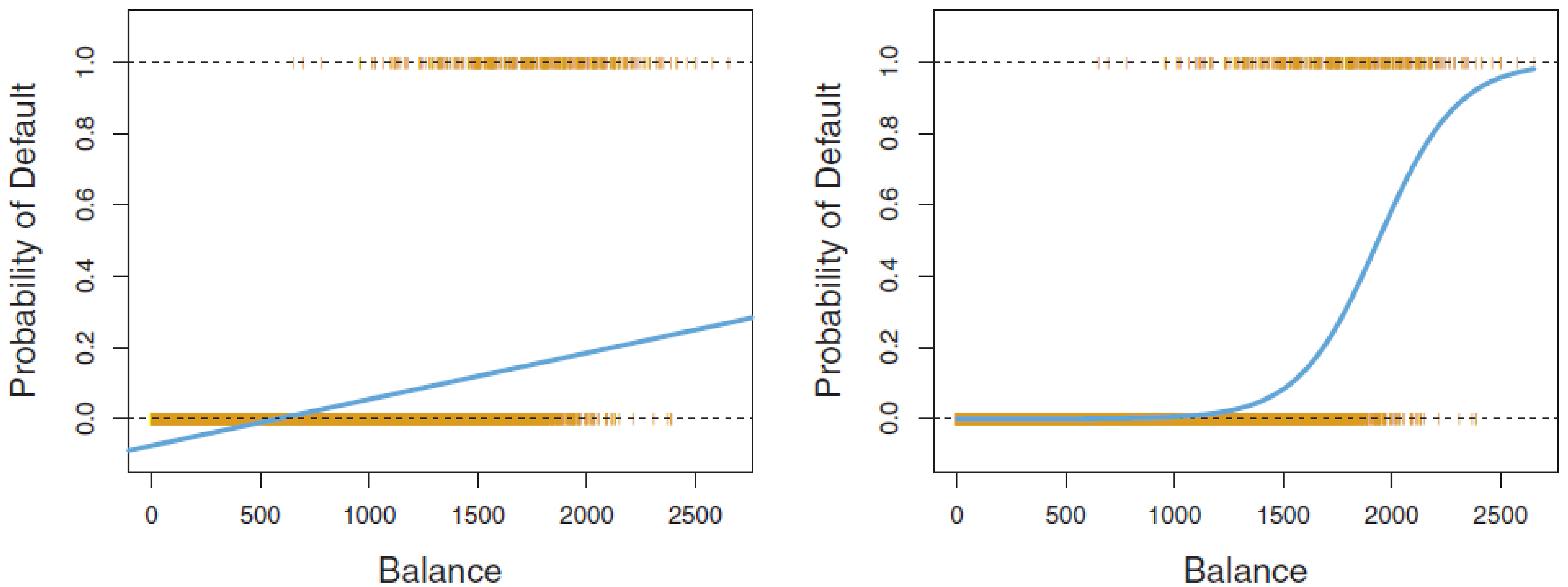


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic Regression

- We want to model the probability of the class given the input

$$p(X) = \Pr(Y = 1|X)$$

$$p(X) = \beta_0 + \beta_1 X$$

Linear Regression

but a linear model has some drawbacks (see later slide)

- Logistic regression solves the negative probability (ad other issues as well) by regressing the logistic function

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

Logistic Regression

from this we derive

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

This is called *odds*

and taking logarithms

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

This is called *log-odds or logit*

Coefficient interpretation

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting P(Y) and not Y.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

  - If $\beta_1 = 0$, this means that there is no relationship between Y and X
  - If $\beta_1 > 0$, this means that when X gets larger so does the probability that Y = 1
  - If $\beta_1 < 0$, this means that when X gets larger, the probability that Y = 1 gets smaller.
- But how much bigger or smaller depends on where we are on the slope, i.e., it is not linear

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Training Logistic Regression (1/4)

- For the basic logistic regression wee need two parameters

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- In principle we could use (non linear) Least Squares fitting on the observed data the corresponding model

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

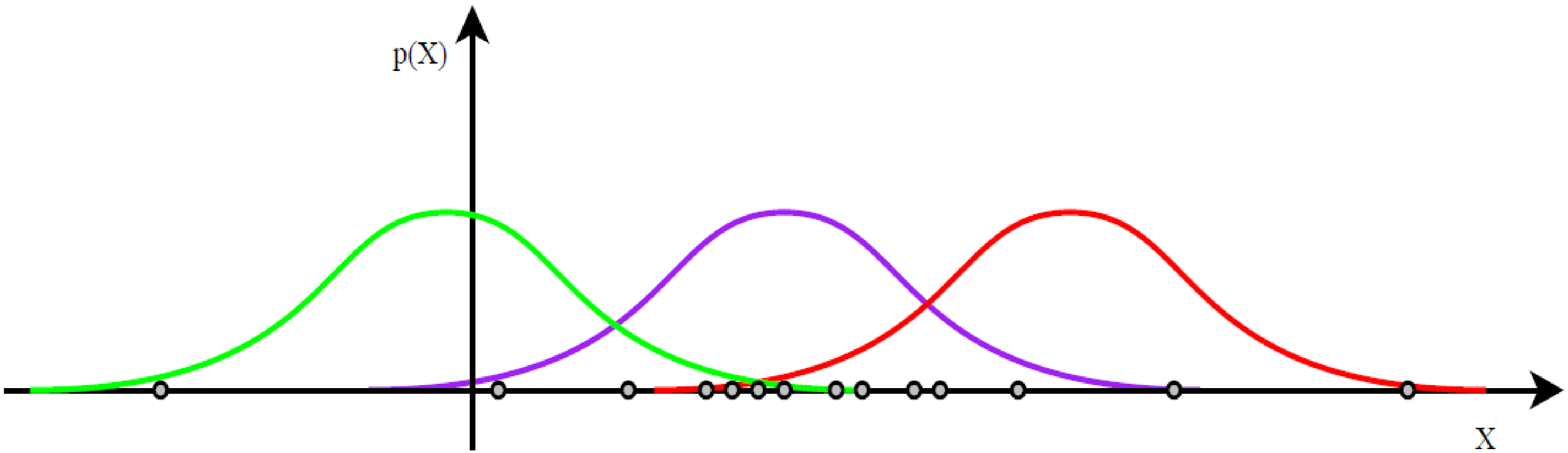
- But a more principled approach for training in classification problems is based on Maximum Likelihood
  - We want to find the parameters which maximize the likelihood function

$$\ell(\beta_0,\beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'}))$$

Maximum Likelihood flash-back (1/6)

- Suppose we observe some i.i.d. samples coming from a Gaussian distribution with known variance:

$$x_1, x_2, \ldots, x_K \sim N(\mu, \sigma^2) \qquad p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$



<u>Which distribution do you prefer?</u>

Maximum Likelihood flash-back (2/6)

- There is a simple recipe for Maximum Likelihood estimation

1. Write the likelihood $L = P(Data|\theta)$ for the data
2. (Take the logarithm of likelihood $\mathcal{L} = \log P(Data|\theta)$)
3. Work out $\partial L/\partial\theta$ or $\partial\mathcal{L}/\partial\theta$ using high-school calculus
4. Solve the set of simultaneous equations $\partial\mathcal{L}/\partial\theta_i = 0$
5. Check that θ^{mle} is a maximum

- Let's try to apply it to our example

$$x_1, x_2, \ldots, x_K \sim N(\mu, \sigma^2) \qquad p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

Maximum Likelihood flash-back (3/6)

- Let's try to apply it to our example

$$x_1, x_2, \ldots, x_K \sim N(\mu, \sigma^2) \qquad p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

1. Write le likelihood for the data

$$\begin{aligned} L(\mu) &= p(x_1, x_2, \ldots, x_N|\mu,\sigma^2) = \prod_{n=1}^{N} p(x_n|\mu,\sigma^2) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \end{aligned}$$

Maximum Likelihood flash-back (4/6)

- Let's try to apply it to our example

$$x_1, x_2, \dots, x_K \sim N(\mu, \sigma^2) \qquad p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

2. (Take the logarithm of the likelihood -> log-likelihood)

$$\begin{aligned}
\mathcal{L} &= log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x_n-\mu)^2}{2\sigma^2}) \\
&= N(\log \frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n-\mu)^2
\end{aligned}$$

Maximum Likelihood flash-back (5/6)

- Let's try to apply it to our example

$$x_1, x_2, \ldots, x_K \sim N(\mu, \sigma^2) \qquad p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

3. Work out the derivatives using high-school calculus

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= \frac{\partial}{\partial \mu} N (\log \frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \\
&= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \mu} \sum_{n=1}^{N} (x_n - \mu)^2 = \\
&= \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(x_n - \mu)
\end{aligned}$$

Maximum Likelihood flash-back (6/6)

- Let's try to apply it to our example

$$x_1, x_2, \ldots, x_K \sim N(\mu, \sigma^2) \qquad p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

4. Solve the unconstrained equations $\partial \mathcal{L} / \partial \theta_i = 0$

$$\frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(x_n - \mu) = 0$$

$$\sum_{n=1}^{N} (x_n - \mu) = 0$$

$$\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} \mu$$

$$\boxed{\mu^{mle} = \frac{1}{R} \sum_{r=1}^{R} x_r}$$

Training Logistic Regression (1/4)

- For the basic logistic regression wee need two parameters

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- In principle we could use (non linear) Least Squares fitting on the observed data the corresponding model

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- But a more principled approach for training in classification problems is based on Maximum Likelihood
  - We want to find the parameters which maximize the likelihood function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Training Logistic Regression (2/4)

- Let's find the parameters which maximize the likelihood function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- If we compute the log-likelihood for N observations

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)$$

Taken from ESL

where $p_k(x_i;\theta) = \Pr(G = k|X = x_i;\theta)$

- We obtain a log-likelihood in the form of

Can you derive it?

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{N} \left\{ y_i \log p(x_i;\beta) + (1 - y_i) \log(1 - p(x_i;\beta)) \right\} \\ &= \sum_{i=1}^{N} \left\{ y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right\}. \end{aligned}$$

Training Logistic Regression (3/4)

- Let's find the parameters which maximize the likelihood function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

  - Z-statistics has the same role of the regression t-statistics, a large value means the parameter is not null
  - Intercept does not have a particular meaning is used to adjust the probability to class proportions

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

TABLE 4.1. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`*. A one-unit increase in* `balance` *is associated with an increase in the log odds of* `default` *by* 0.0055 *units.*

Example: Default data & Logistic Regression

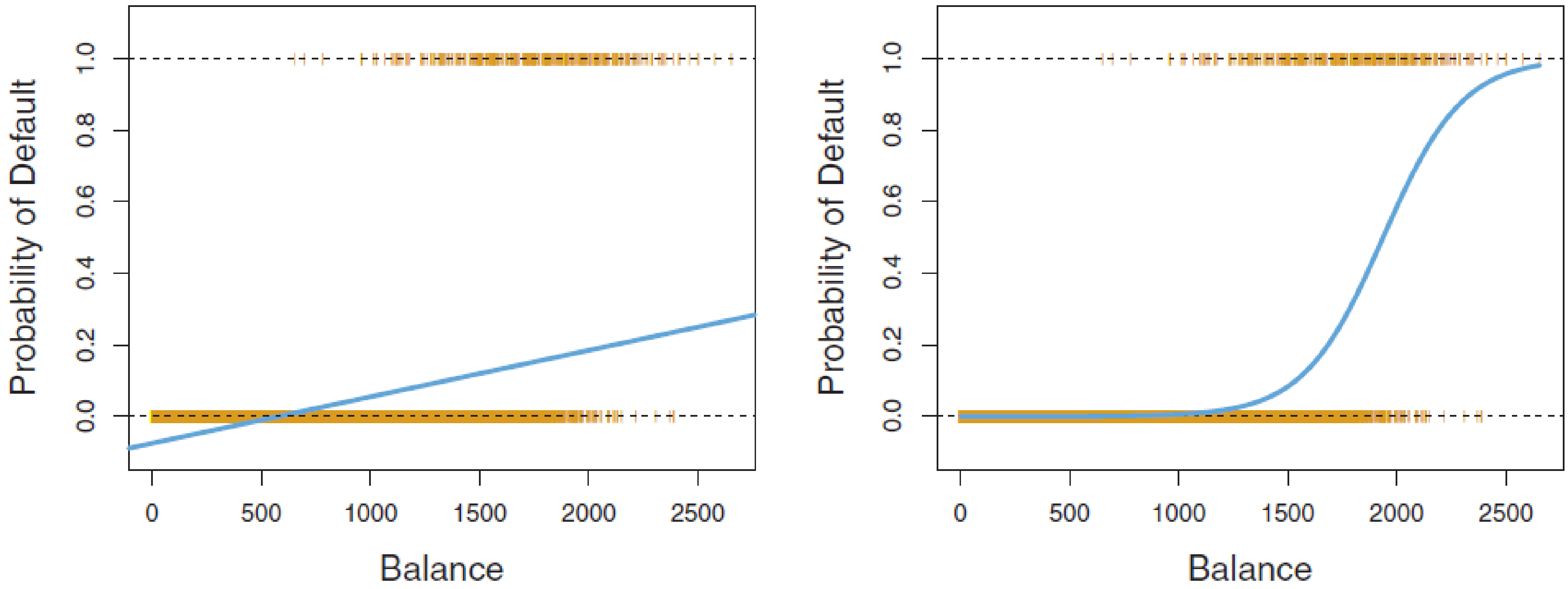


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Training Logistic Regression (4/4)

- Let's find the parameters which maximize the likelihood function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

  - We can train the model using qualitative variables through the use of binary (dummy) variables

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

TABLE 4.2. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable* student[Yes] *in the table.*

Making predictions with Logistic Regression

- Once we have the model parameters we can predict the class

- The Default probability having 1000$ balance is <1%

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

  while with a balance of 2000$ this becomes 58.6%

- With qualitative variables, i.e., dummy variables, we get that being a students results in

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Multiple Logistic Regression

- So far we have considered only one predictor, but we can extend the approach to multiple regressors

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- By maximum likelihood we learn the corresponding parameters

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

What about this?

TABLE 4.3. *For the* Default *data, e… the logistic regression model that predicts the probability … nce,* income, *and* student *status. Student status is encoded as a dummy variable* student[Yes], *with a value of 1 for a student and a value of 0 for a non-student. In fitting this model,* income *was measured in thousands of dollars.*

An apparent contradiction

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Positive

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Negative!!!

Example: Confounding in Default data set

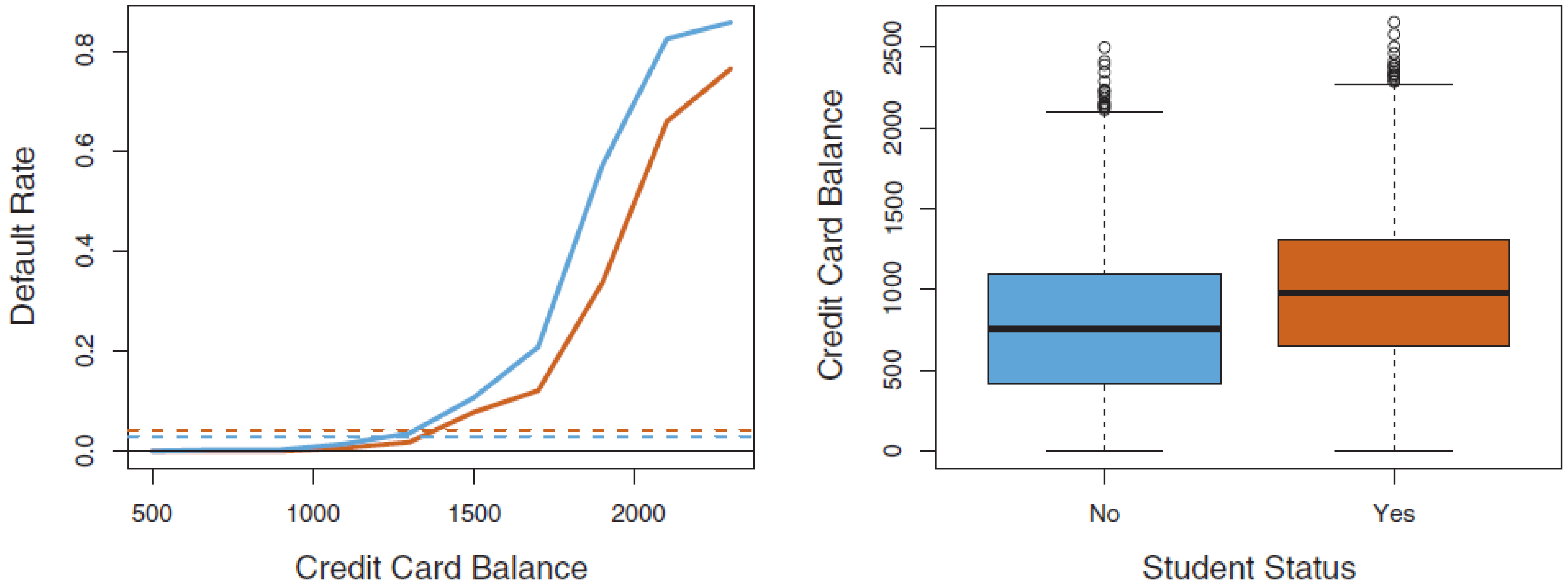


FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

Example: South African Heart Disease

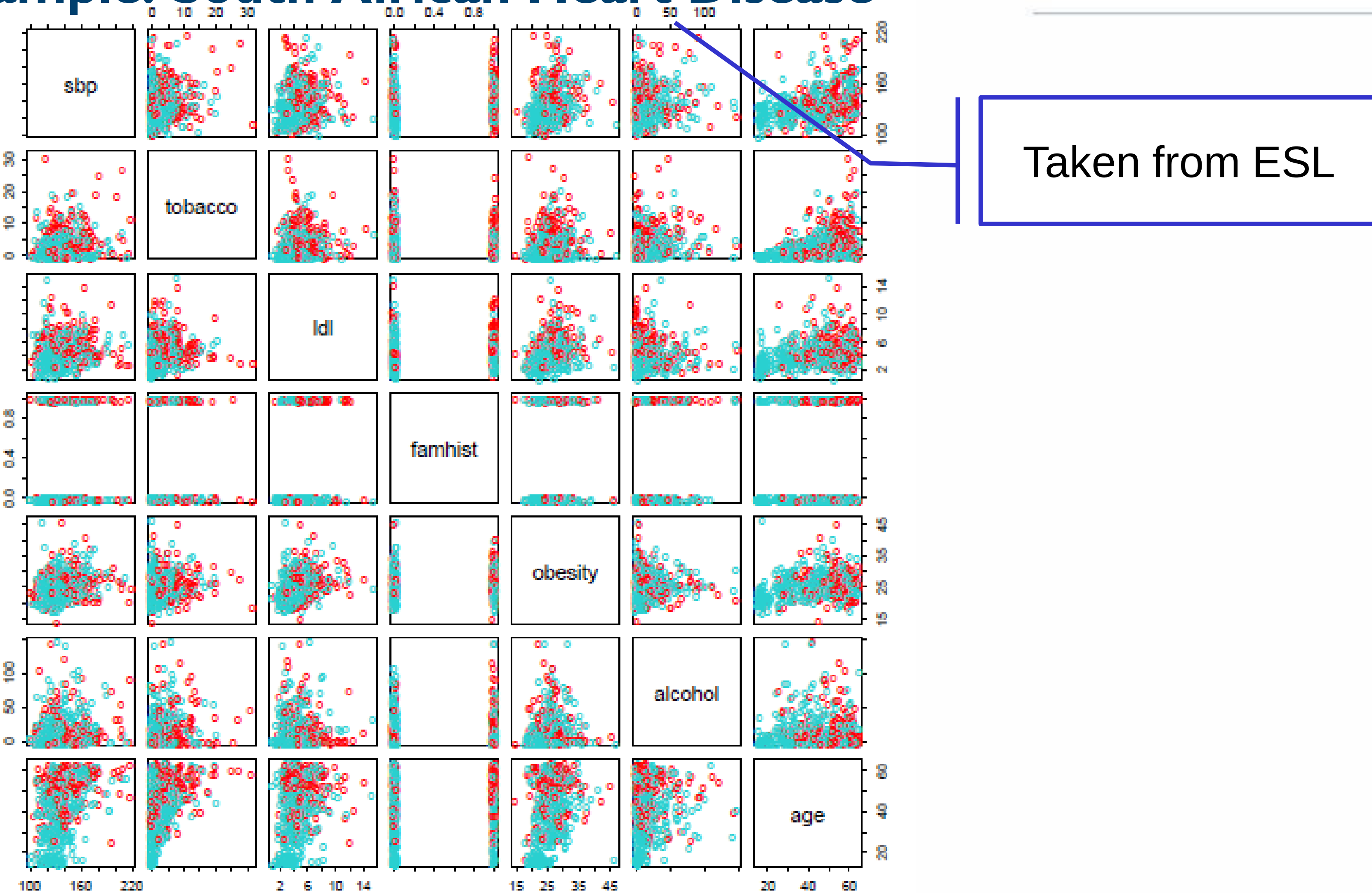


FIGURE 4.12. *A scatterplot matrix of the South African heart disease data. Each plot shows a pair of risk factors, and the cases and controls are color coded (red is a case). The variable* family history of heart disease (famhist) *is binary (yes or no).*

Logistic Regression for Feature Selection

- If we fit the complete model on these data we ge

Taken from ESL

TABLE 4.2. *Results from a logistic regression fit to the South Af disease data.*

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

- While if we use stepwise Logistic Regression

Taken from ESL

TABLE 4.3. *Results from stepwise logistic regression fit to South African heart disease data.*

	Coefficient	Std. Error	Z score
(Intercept)	−4.204	0.498	−8.45
tobacco	0.081	0.026	3.16
ldl	0.168	0.054	3.09
famhist	0.924	0.223	4.14
age	0.044	0.010	4.52

Logistic Regression parameters meaning

- Regression parameters represent the increment on the logit of probability given by a unitary increment of a variable

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Let consider the increase of *tobacco* consumption in life od 1Kg, this count for an increase in log-odds of exp(0.081)=1.084 which means an overall increase of 8.4%
- With a 95% confidence interval $\exp(0.081 \pm 2 \times 0.026) = (1.03, 1.14)$

TABLE 4.3. *Results from stepwise logistic regression fit to South African heart disease data.*

Taken from ESL

	Coefficient	Std. Error	Z score
(Intercept)	−4.204	0.498	−8.45
tobacco	0.081	0.026	3.16
ldl	0.168	0.054	3.09
famhist	0.924	0.223	4.14
age	0.044	0.010	4.52

Regularized Logistic Regression

- As for Linear Regression we can compute a "Lasso" version

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

- As for Linear Regression we can compute a "Lasso" version

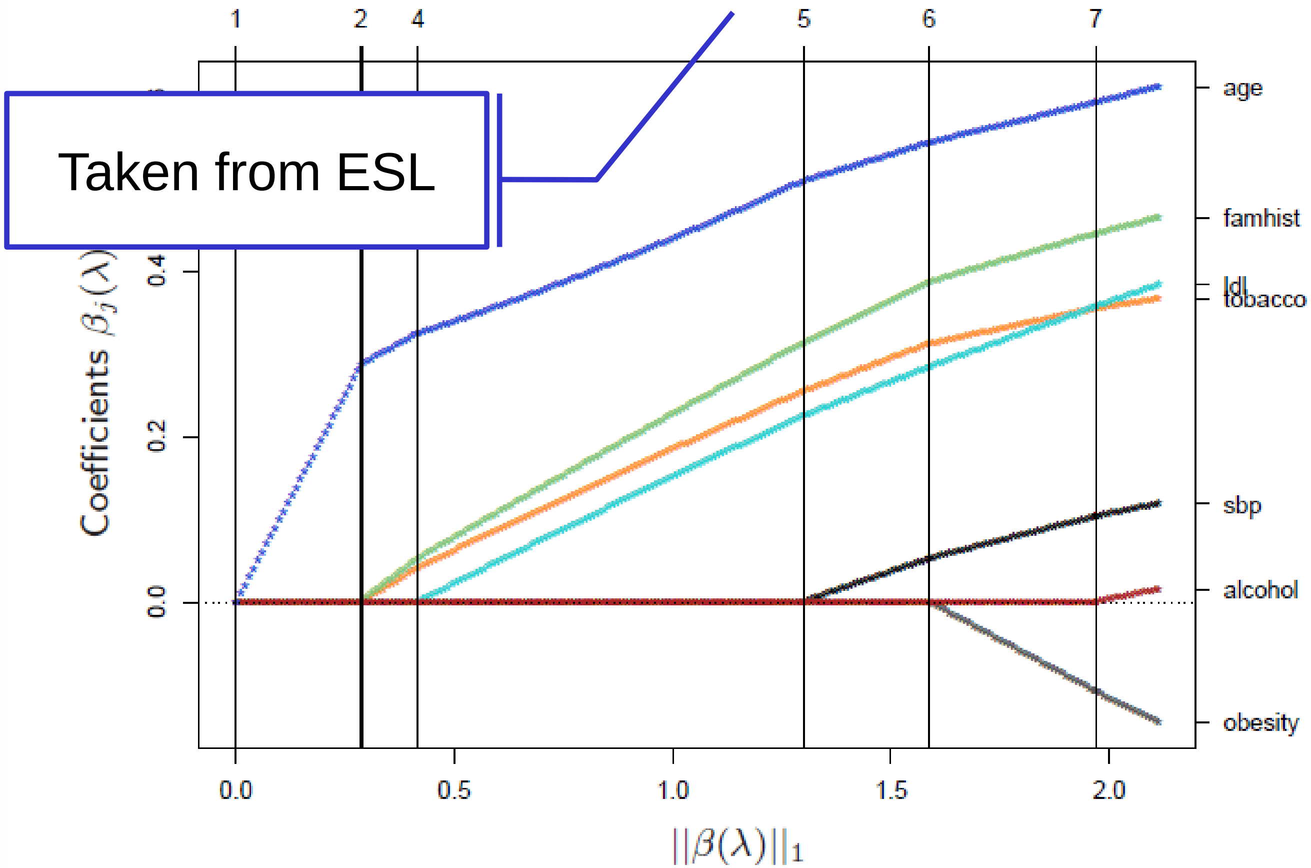


FIGURE 4.13. *L_1 regularized logistic regression coefficients for the South African heart disease data, plotted as a function of the L_1 norm. The variables were all standardized to have unit variance. The profiles are computed exactly at each of the plotted points.*

Multiclass Logistic Regression

- Logistic Regression extends naturally to multiclass problems by computing the log-odds w.r.t. the K[th] class

$$\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$$
$$\log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$$
$$\vdots$$
$$\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

Comes from ESL, but it's worth knowing!!!

Notation different because it comes from ESL

- This is equivalent to

$$\begin{aligned}
\Pr(G=k|X=x) &= \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T x)}, \quad k = 1, \ldots, K-1, \\
\Pr(G=K|X=x) &= \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T x)},
\end{aligned} \qquad (4.18)$$

- Can you prove it ?!?!?

Wrap-up on Logistic Regression

- We model the log-odds as a linear regression model

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- This means the posterior probability becomes

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- Parameters represent log-odds increase per variable unit increment keeping fixed the others
- We can use it to perform feature selection using z-scores and forward stepwise selection
- The class decision boundary is linear, but points close to the boundary count more … this will be discussed later