1.3 Fundamental definitions

The purpose of regression analysis

Regression analysis is performed for two reasons: **prediction** or **inference**.

Prediction

In many situations X is available but the output Y cannot be easily obtained. Since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{1.3.1}$$

where $\hat{f}$ is the estimate for f, and $\hat{Y}$ *represents the resulting prediction for* Y.

The estimate $\hat{f}$ is characterised by a **reducible** error and by an **irreducible** error.

$$\begin{aligned} \mathbb{E}[Y - \hat{Y}]^2 &= \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + \mathbb{V}ar(\varepsilon) \end{aligned} \tag{1.3.2}$$

Inference

The goal of inference is to understand the **relationship** between X and Y. How Y changes as a function of $X_1, X_2, \ldots, X_p$.

1. Which predictors are associated with the response?
2. What is the relationship between the response and each predictor?
3. Is the relationship between Y and each predictor linear or more complex?

Depending on the purpose of the analysis, various methods of estimating f may be more appropriate. For **prediction**, use of *non-linear models* may be more appropriate (in some cases overfitting causes problems), while for **inference**, *linear models* may be more useful.

Methods of estimation

Two main types of approaches

- **parametric**
    1. Assumption on the functional form of f

$$f(\mathbf{x}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{1.3.3}$$

    2. Method to fit the model (estimating $\beta_0, \beta_1, \ldots, \beta_p$)
- **nonparametric:** no explicit assumptions about the functional form of f at the price of not having a small and fixed amount of parameters (i.e. larger sample sizes are needed)

Maximum likelihood estimation

$\mathbf{y} = [Y_1, Y_2, \ldots, Y_n]^\top$ denotes a random vector.

$f(\boldsymbol{y}; \boldsymbol{\theta})$ - joint probability density function of Y_i, which depends on the vector of parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^\top$.

The **likelihood function** $L(\boldsymbol{\theta}; \boldsymbol{y})$ is algebraically the same as $f(\boldsymbol{y}; \boldsymbol{\theta})$ but the emphasis shifts to parameters $\boldsymbol{\theta}$ while $\boldsymbol{y}$ stays fixed.

Remark: L is itself a random variable!

⚠ **Warning:** Do not mix up $f(\cdot; \boldsymbol{\theta})$ above with the function $f(\cdot)$ of the model $\mathrm{y} = f(\mathbf{x}) + \epsilon$!!!

Definition:

> The **maximum likelihood estimator** of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood function, that is
>
> $$L(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \geq L(\boldsymbol{\theta}, \boldsymbol{y}) \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta \quad (\text{parameter space}). \qquad (1.3.4)$$

Equivalently, $\hat{\boldsymbol{\theta}}$ is the value that maximizes the log-likelihood function

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \ln L(\boldsymbol{\theta}, \boldsymbol{y}), \qquad (1.3.5)$$

since the logarithmic function is monotonic.

The estimate $\hat{\boldsymbol{\theta}}$ is obtained by *differentiating the log-likelihood function with respect to each element θ_j of $\boldsymbol{\theta}$* and *solving the simultaneous equations*

$$\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \theta_j} = 0 \qquad j = 1, \ldots, p \qquad (1.3.6)$$

If the matrix of second derivatives

$$\frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \theta_j \partial \theta_k} \qquad (1.3.7)$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is negative definite, then $\hat{\boldsymbol{\theta}}$ *maximizes the log-likelihood function in the interior of* Θ.

Note that it is also necessary to *check if there are any values of $\boldsymbol{\theta}$ at the edges of the parameter space* Θ

that give local maxima of $l(\boldsymbol{\theta}, \mathbf{y})$: when all local maxima have been identified, the value of $\hat{\boldsymbol{\theta}}$ corresponding to the largest one is the MLE.

Invariance property of MLE

Invariance: If $g(\boldsymbol{\theta})$ is any function of the parameters $\boldsymbol{\theta}$, then the maximum likelihood estimator of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$.

Other properties: consistency, sufficiency, asymptotic efficiency, asymptotic normality.

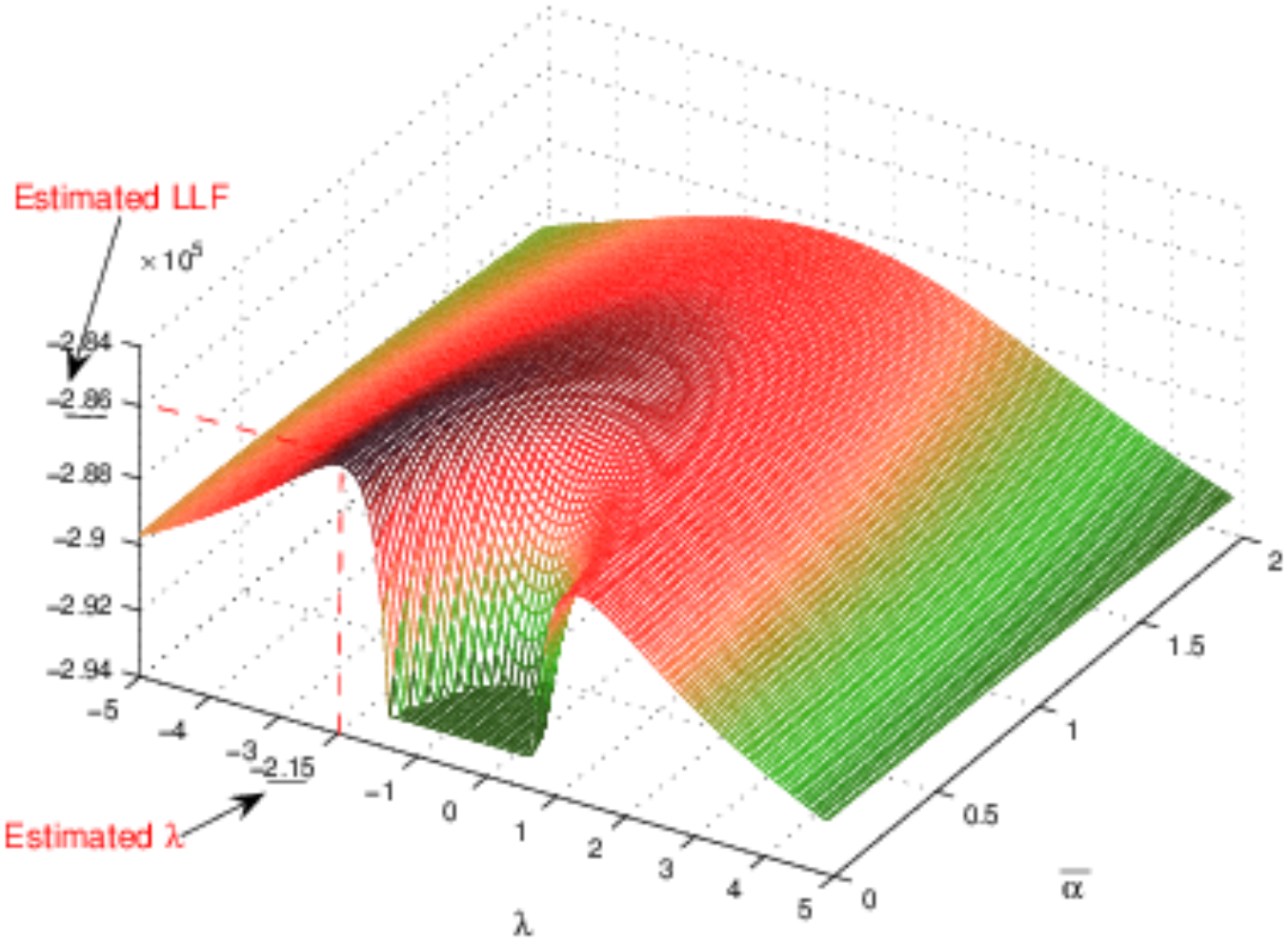


Figure 1.3.1: Example of a log-likelihood function. Source: Platen & Rendek (2008)

Example: Poisson distribution

Let $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_n$ be independent random variables with Poisson distribution

$$f(y_i, \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!},$$

$y_i = 0, 1, \ldots$, with the same parameter θ.

Find: The MLE of θ.

Least squares estimation

Let $\mathrm{Y}_1, \ldots, \mathrm{Y}_n$ be independent r.v. with expected values $\mu_1, \ldots, \mu_n$ which are functions of the parameter vector (that we want to estimate) $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$, $p < n$. Thus

$$\mathbb{E}(Y_i) = \mu_i(\boldsymbol{\beta}).$$

The simplest **method of least squares** consists of finding the estimator $\hat{\boldsymbol{\beta}}$ that *minimizes the sum of squares of the differences between* Y_i*'s and their expected values*

$$SS = \sum_{i=1}^{n} [\mathrm{Y}_i - \mu_i(\boldsymbol{\beta})]^2. \tag{1.3.8}$$

and $\hat{\boldsymbol{\beta}}$ is obtained by differentiating SS with respect to the elements β_j

$$\frac{dSS}{d\beta_j} = 0 \qquad j = 1, \ldots, p \tag{1.3.9}$$

Weighted least squares

If Y_i's have variances σ_i^2 that are not all equal it may be desirable to *minimize the weighted sum of squared differences*

$$WSS = \sum_{i=1}^{n} w_i[\mathrm{Y}_i - \mu_i(\boldsymbol{\beta})]^2. \tag{1.3.10}$$

where the weights are $w_i = (\sigma_i^2)^{-1}$. In this way the observations that are less reliable will have less influence on the estimates.

More generally, if $\mathbf{y} = [\mathrm{Y}_1, \ldots, \mathrm{Y}_n]^\top$ is a random vector with mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\top$ and variance-covariance matrix $\mathbf{V}$, then

$$WSS = (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \tag{1.3.11}$$

Comments

1. Method of least squares can be used without making assumptions about the distribution of the response variables Y_i in contrast to the maximum likelihood estimation.
2. For many situations maximum likelihood and least squares estimates are identical.
3. In many cases **numerical methods** are used for parameter estimation.

Recommended reading: Dobson pages 13–16.

Model fitting

Model fitting involves four steps

- **Model specification**: an equation linking the response and the explanatory variables and a probability distribution for the response variable
- **Estimation** of the parameter of the model
- **Checking the adequacy of the model**
- **Inference**: calculating confidence interval, testing hypotheses

Australian longitudinal study on women's health, Lee et al. (2005)

1. **Observation:** women living in country areas tend to have fewer consultations with GPs than women who live near a wide range of health services
2. **Hypotheses:** is this because they are healthier or because of structural factors?

Table 2.1 *Number of chronic medical conditions of 26 town women and 23 country women with similar use of general practitioner services.*

Town	
0 1 1 0 2 3 0 1 1 1 1 2 0 1 3 0 1 2 1 3 3 4 1 3 2 0	
$n = 26$, mean = 1.423, standard deviation = 1.172, variance = 1.374	
Country	
2 0 3 0 0 1 1 1 1 0 0 2 2 0 1 2 0 0 1 1 1 0 2	
$n = 23$, mean = 0.913, standard deviation = 0.900, variance = 0.810	

Group of study:

- Women living in country towns (*town group*) or in rural areas (*country group*) in NSW
- Women aged 70-75 years
- Same socio-economic status
- ≤ 3 GP visits in 1996

Think: Which distribution is suitable for these data?

Let Y_{jk} be a r.v. representing the number of conditions for woman k in group j ($j = 1$ town group, $j = 2$ country group).

$$\mathrm{Y}_{jk} \sim \mathcal{P}ois(\theta_j) \qquad k = 1, \ldots, K_j$$

The **question of interest** can be formalised as

$$\begin{aligned} H_0 &: \theta_1 = \theta_2 = \theta \longrightarrow \mathbb{E}[Y_{jk}] = \theta \\ H_1 &: \theta_1 \neq \theta_2 \longrightarrow \mathbb{E}[Y_{jk}] = \theta_j \end{aligned}$$

If H_0 is true:

$$l_0(\theta; \mathbf{y}) = \sum_{j=1}^{2} \sum_{k=1}^{K_j} [y_{jk} \log \theta - \theta - \log y_{jk}!]$$

and the MLE is

$$\hat{\theta} = \sum_{j=1}^{2} \sum_{k=1}^{K_j} \frac{y_{jk}}{N}$$

where $N = \sum_{j=1}^{2} K_j$, $\hat{\theta} = 1.184$, with $\hat{l}_0 = -68.3868$.

Practice: Implement these calculations in R

If H_1 is true:

$$l_1(\theta_1, \theta_2; \mathbf{y}) = \sum_{k=1}^{K_1} [y_{1k} \log \theta_1 - \theta_1 - \log y_{1k}!] + \sum_{k=1}^{K_2} [y_{2k} \log \theta_2 - \theta_2 - \log y_{2k}!]$$

and the MLE is

$$\hat{\theta}_1 = \sum_{k=1}^{K_1} \frac{y_{1k}}{K_1}, \qquad \hat{\theta}_2 = \sum_{k=1}^{K_2} \frac{y_{2k}}{K_2}$$

with $\hat{\theta}_1 = 1.423$ and $\hat{\theta}_2 = 0.913$, with $\hat{l}_1 = -67.0230$.

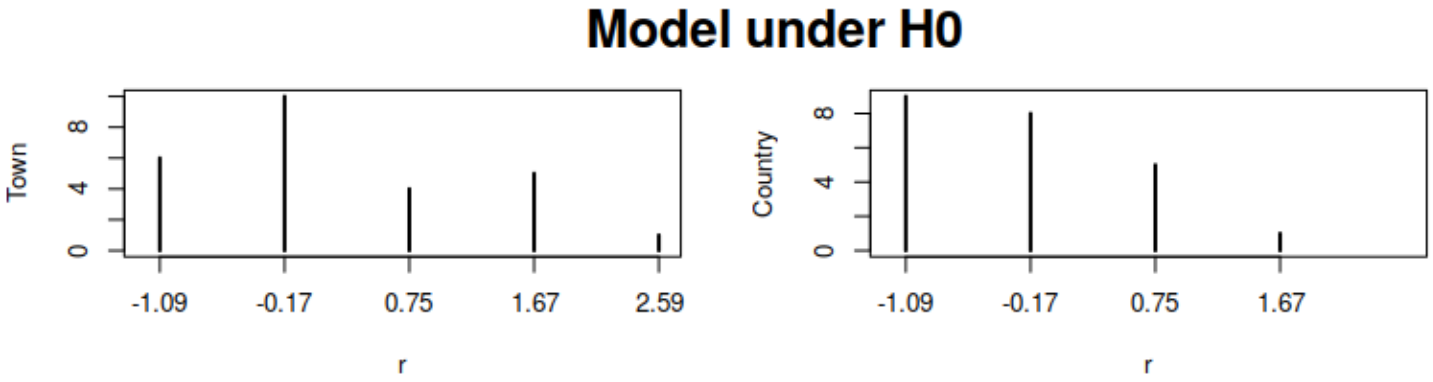
Practice: Implement these calculations in R

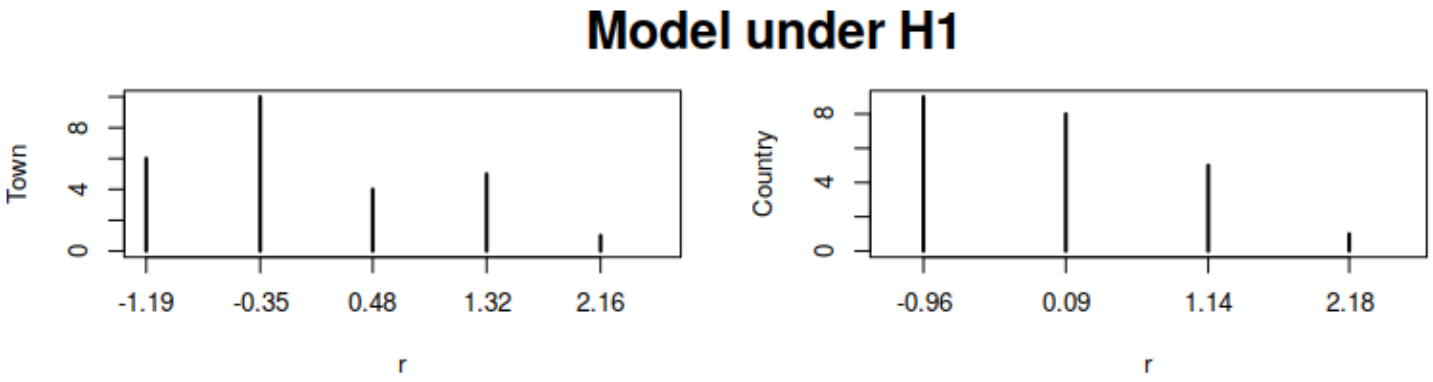
Think: Is the difference statistically significant?

Remark: $l_1 \geq l_0$ because one more parameter has been fitted.

The **adequacy** of the model is usually evaluated by looking at the **(standardised) residuals**

$$r = \frac{\mathrm{Y} - \hat{\theta}}{\sqrt{\hat{\theta}}} \tag{1.3.12}$$





Practice: Reproduce these plots in R

Assuming Poisson independent data, the *standardised residuals are approximately distributed as a standard normal distribution* $\mathcal{N}(0,1)$, then $r_1^2 \sim \chi_1^2$ and

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\theta}_i)^2}{\hat{\theta}_i} \sim \chi_m^2 \tag{1.3.13}$$

where χ_m^2 is a χ^2 distribution with m degrees of freedom and $m = n - p$

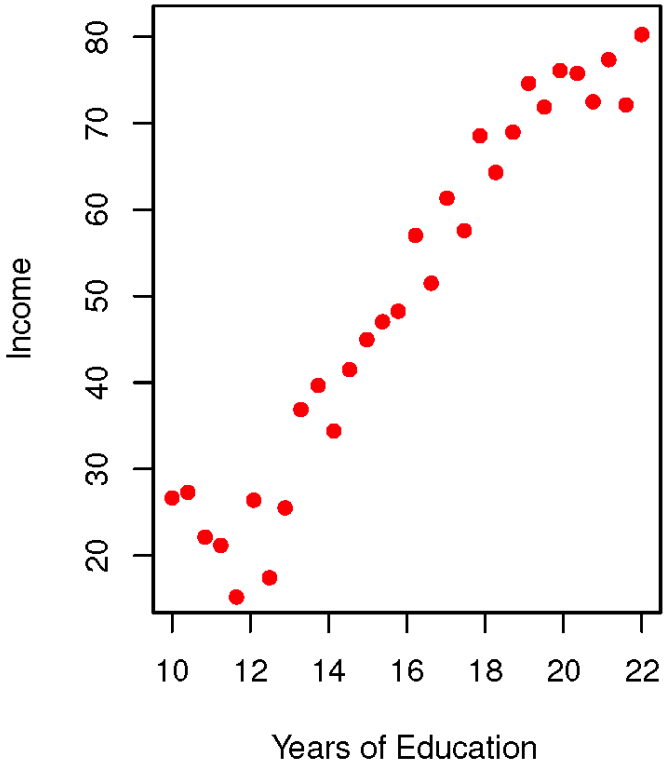
Under H_0, $\sum_{i=1}^{N} r_{0i}^2 = 46.8457$ and $m = K_1 + K_2 - 1 = 26 + 23 - 1 = 48$.

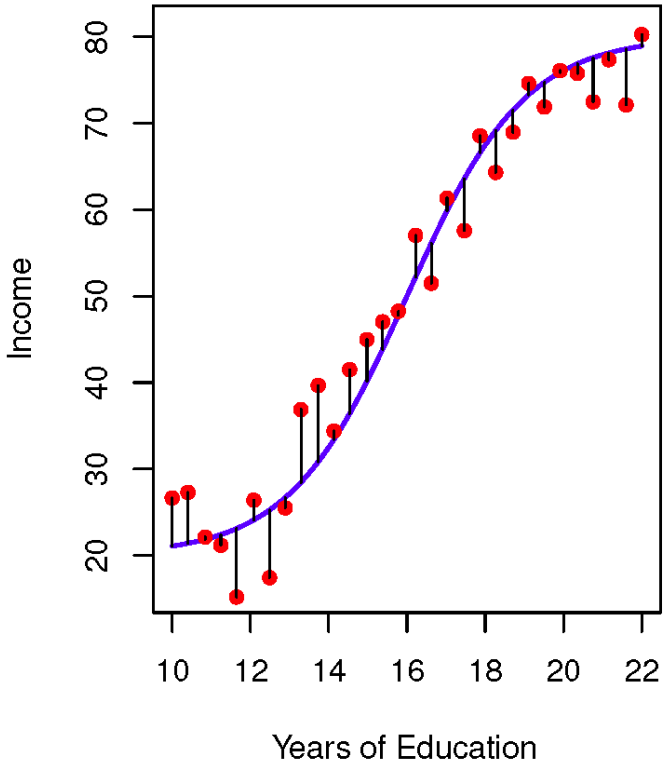
Under H_1, $\sum_{i=1}^{N} r_{1i}^2 = 43.6304$ with $m = K_1 + K_2 - 2 = 26 + 23 - 2 = 47$.

For now, it is enough to say that $46.8457 - 43.6304 = 3.2153$ seems small, but later we will quantify the interpretation on it.

Relating income to years of education

Let's analyse the dataset **Income** available in the R package **ISLR**





Predicting

- **Input:** years of education (X_1) and seniority (X_2)
- **Output:** income (Y)

Hypothesis: there is some relationship between Y and $\mathbf{x} = (\mathrm{X}_1, \mathrm{X}_2)$

$$\mathrm{Y} = f(\mathbf{x}) + \varepsilon \tag{1.3.14}$$

- f is an unknown function (systematic part)
- ε is an error term

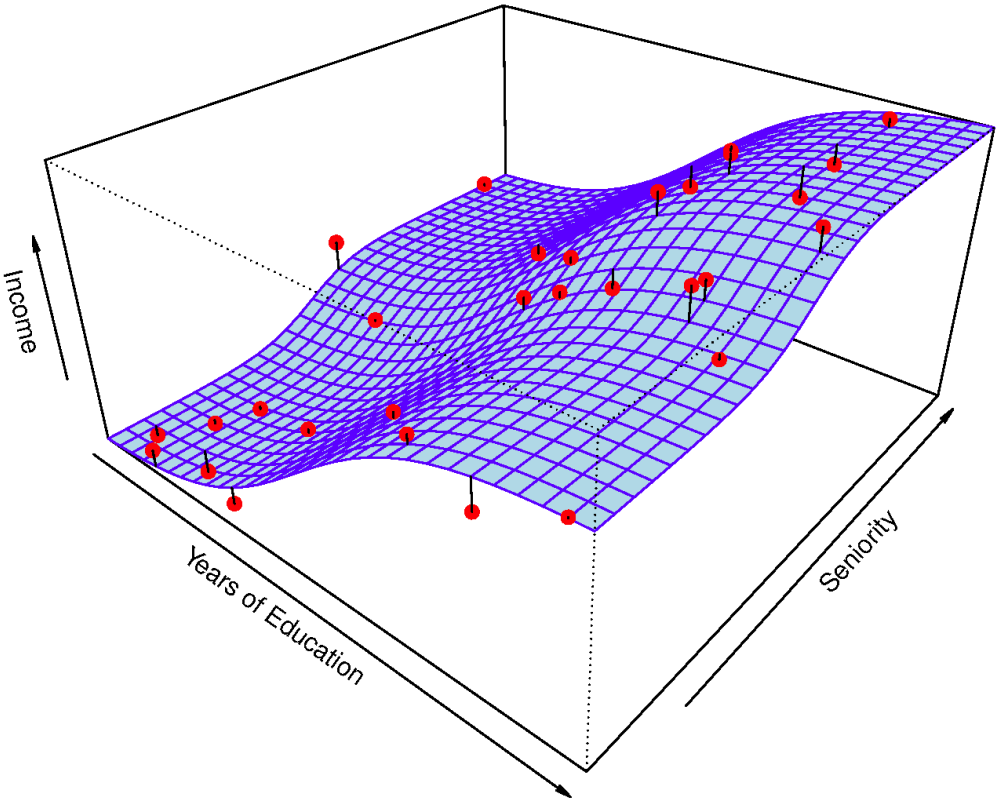


Figure 1.3.2: Plot of `income` as a function of `years of education` and `seniority`. The *blue* surface represents the true underlying relationship (simulated data).

We can fit the **linear model**:

$$\mathbf{income} = \beta_0 + \beta_1 \times \mathbf{education} + \beta_2 \times \mathbf{seniority} + \varepsilon$$

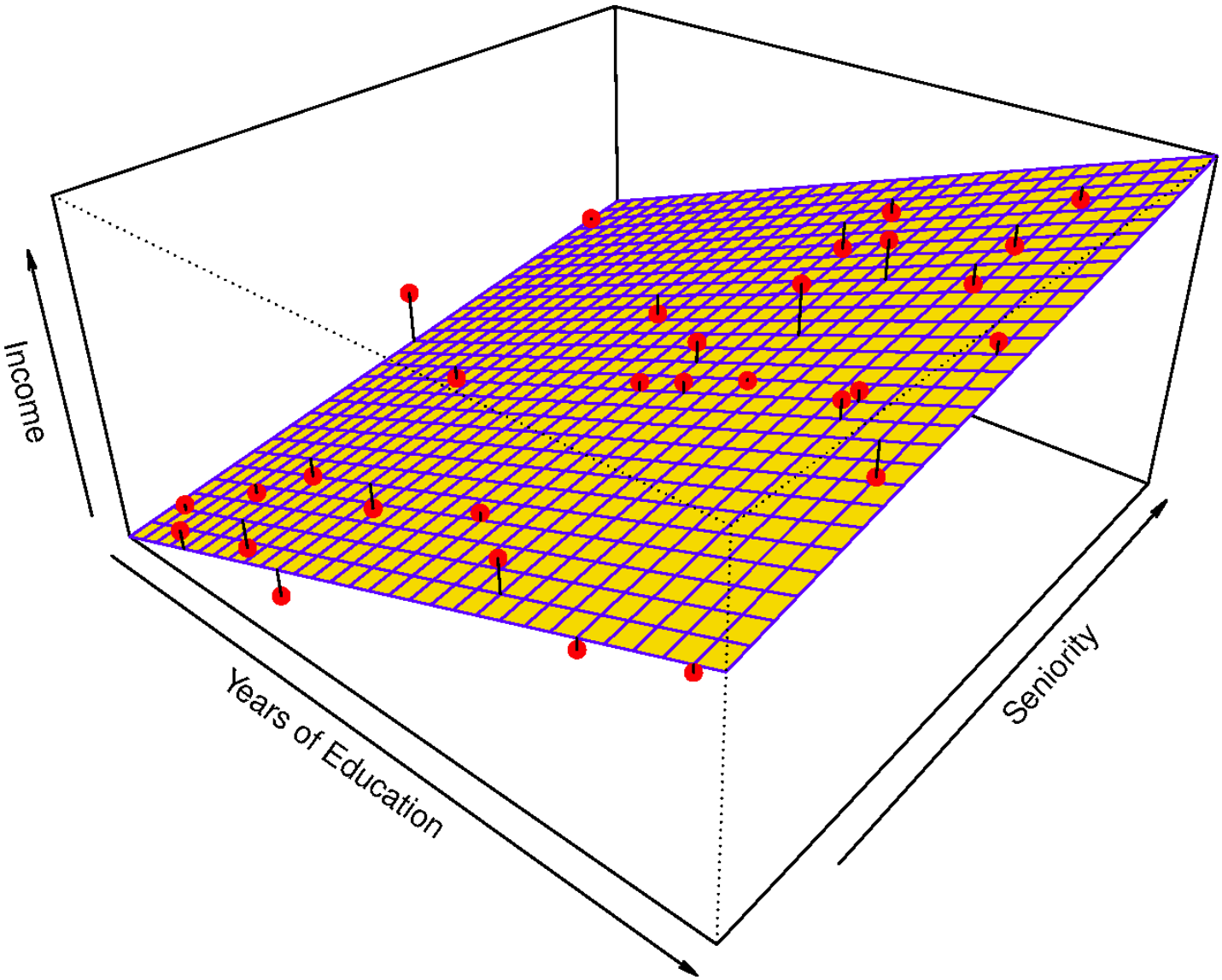
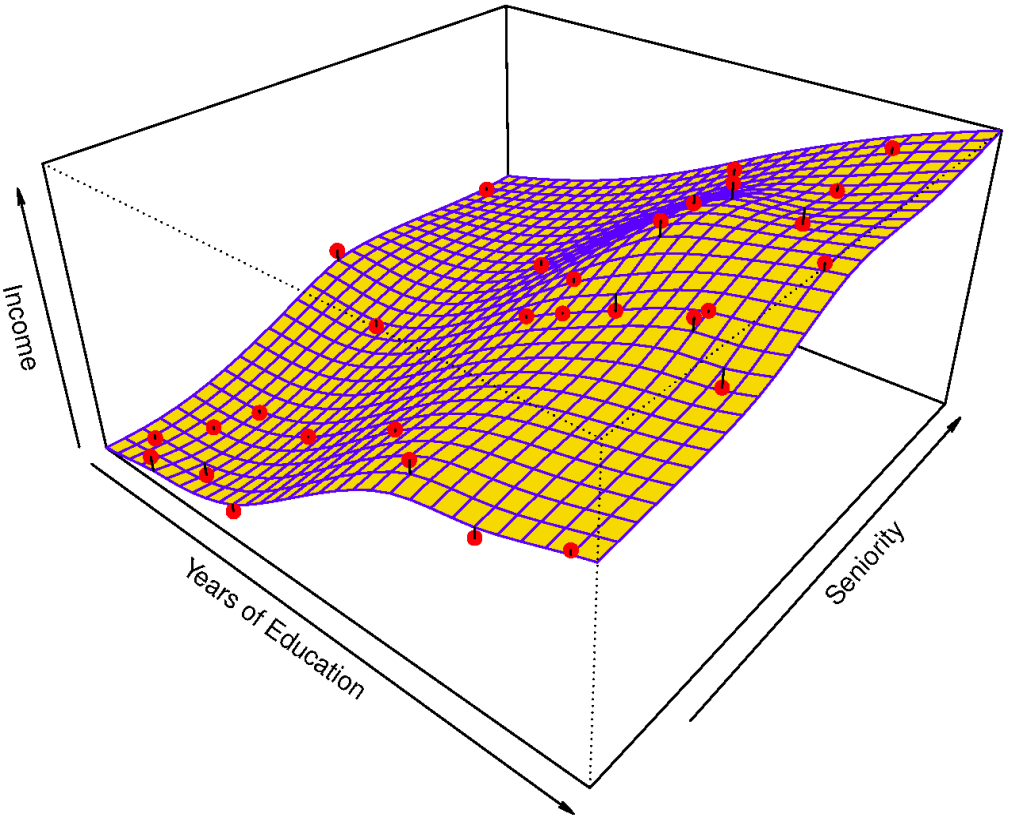
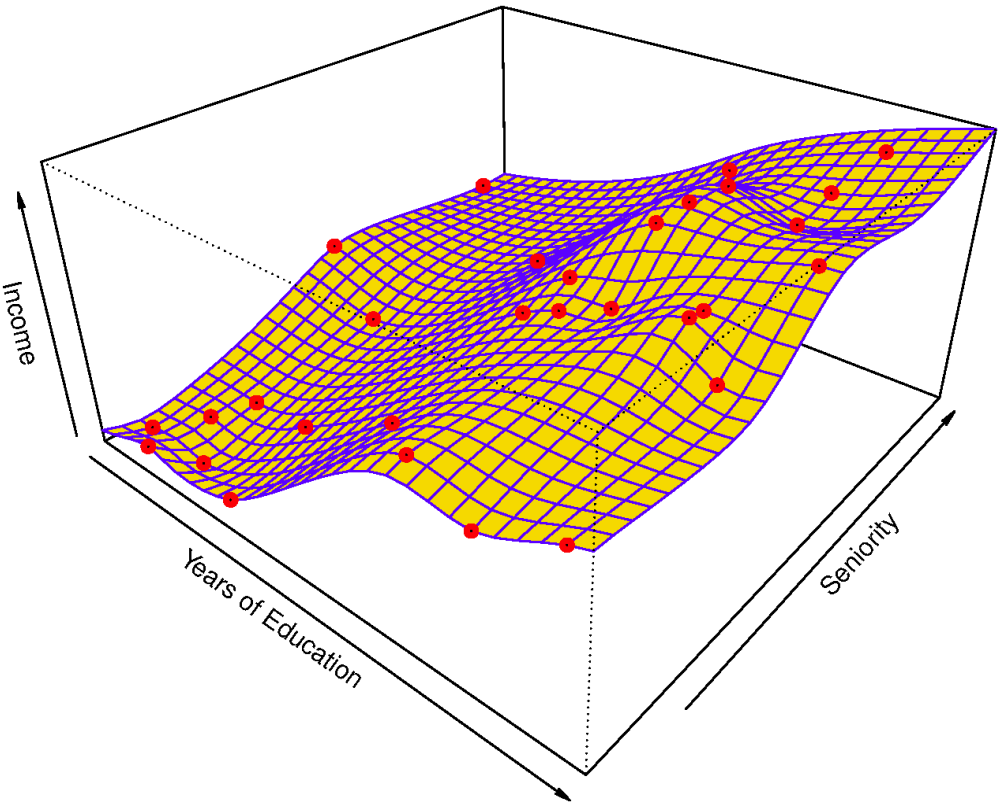


Figure 1.3.3: The yellow surface represents $\hat{\beta}_0 + \hat{\beta}_1 \times \textbf{education} + \hat{\beta}_2 \times \textbf{seniority}$

Nonparametric Model: Thin-Plate Spline

Think: How to choose the level of smoothness?


Income
Years of Education
Seniority


Income
Years of Education
Seniority

What have we learnt from these examples?

1. What is the scale of measurement?
2. What is a reasonable **distribution** to model the data?
3. What is the **relationship** with other variables?

$$\mathbb{E}[\mathrm{Y}] = \alpha + \beta x$$
$$\log[\mathbb{E}(\mathrm{Y})] = \alpha + \beta \sin(\gamma x)$$

4. What is the **best parameter estimation process**? MLE, Least Squares, Bayesian methods
5. Why choosing a restrictive (parametric method) instead of a very flexible (nonparametric) approach?
6. **Model checking**: residual checking, plots, checking of the assumptions

Don't under-evaluate exploratory statistics!

Wage dataset: wages for a group of males from the Atlantic region of the US (available in the `R` package `ISLR`)

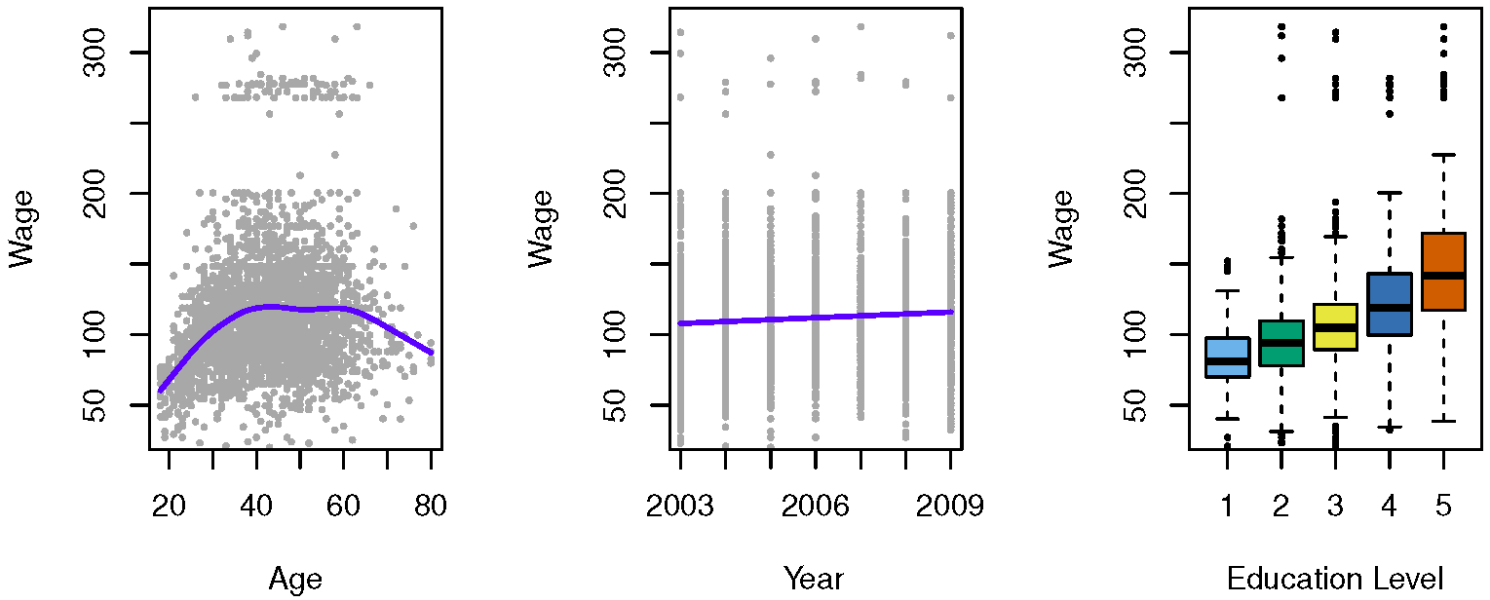


Figure 1.3.4: Graphical representations of `wage` as a function of `age`, `year` and `education`

Comments

1. Many statistical learning methods are relevant and useful in a wide range of disciplines, beyond just statistical sciences.
2. Statistical learning should not be viewed as a series of black boxes.
3. While it is important to know what job is performed by each tool, it is not necessary to have the skills to construct the machine inside the box.
4. We will work on real-world problems.

Measuring the quality of a fit

The goal: We need to measure how well the model predictions match the data.

In a regression setting, this is done by the **mean squared error (MSE)**

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{1.3.15}$$

However, a low MSE can hide problems of overfitting on the dataset at hand. Then, what we really want to have is *the accuracy of the predictions when we apply the method on unseen data.*

Suppose we have (**training**) observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ on which we estimate $f(\boldsymbol{x})$; then we obtain estimates $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$.

We want to measure the accuracy of the model on new input (*test*) variables x_0, to obtain the **test MSE**

$$\mathrm{Ave}(\hat{f}(x_0) - y_0)^2$$

which is the average squared prediction error on the test observations (x_0, y_0)

Then we can

- select the model which minimises the **test MSE**, if test observations are available
- select the model which minimises the **training MSE**, if test observations are not available
- use estimation method for the test MSE, like **cross-validation**

The bias-variance trade-off

There are **two competing properties** of statistical learning methods.

The expected test MSE, for a given value x_0 can be decomposed into **three quantities**:

$$\begin{aligned}
\mathrm{MSE}(x_0) &= \mathbb{E}\left[y_0 - \hat{f}(x_0)\right]^2 \\
&= \mathbb{E}\left[f(x_0) + \varepsilon - \hat{f}(x_0)\right]^2 \\
&= \mathbb{E}\left[f(x_0) - \mathbb{E}(\hat{f}(x_0)) + \mathbb{E}(\hat{f}(x_0)) - \hat{f}(x_0)\right]^2 + \mathbb{V}\mathrm{ar}[\varepsilon] \\
&= \left[f(x_0) - \mathbb{E}(\hat{f}(x_0))\right]^2 + 2\left[f(x_0) - \mathbb{E}(\hat{f}(x_0))\right]\mathbb{E}\left[\hat{f}(x_0) - \mathbb{E}(\hat{f}(x_0))\right] \\
&\quad + \mathbb{E}\left[\hat{f}(x_0) - \mathbb{E}(\hat{f}(x_0))\right]^2 + \mathbb{V}\mathrm{ar}[\varepsilon] \\
&= \mathrm{Bias}^2\left(\hat{f}(x_0)\right) + \mathbb{V}\mathrm{ar}\left(\hat{f}(x_0)\right) + \mathbb{V}\mathrm{ar}[\varepsilon]
\end{aligned}$$

Comments

- The test MSE can never be lower than $\mathbb{V}\mathrm{ar}[\varepsilon]$
- **Variance:** the amount by which $\hat{f}$ would change when changing the training dataset (in general, flexible methods have larger variance)
- **Bias:** the error that is introduced by approximating a potentially complicated relationship between Y and X with a simpler model (in general, restrictive methods have a larger bias)
- In practice, as we increase the flexibility of the model, the bias tends to decrease faster than the variance increases; however, at some point, increasing flexibility has little impact on the bias and significantly increases the variance
- In practice, we cannot explicitly compute the test MSE

The classification setting

Model accuracy can be also applied for **categorical outputs**, with slight modifications.

In a **classification** setting, model accuracy is measured by the (*training*) **error rate**

$$\mathrm{ER} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i \neq \hat{y}_i) \tag{1.3.16}$$

where $\hat{y}_i$ is the predicted label for the i-th observation, using the estimate $\hat{f}$ and $\mathbb{I}(y_i \neq \hat{y}_i)$ is an indicator function which is equal to

- 1 if $y_i \neq \hat{y}_i$ (miss-classification)
- 0 if $y_i = \hat{y}_i$ (correct classification)

As in the case of regression, we are more interested in the **test error rate**, which is the error rate that results from applying the classifier to test observations, not available in the training dataset

$$\mathrm{Ave}\left(\mathbb{I}(y_0 \neq \hat{y}_0)\right)$$

where $\hat{y}_0$ is the prediction obtained by applying the classifier on input x_0.

The Bayes classifier

The **test error rate** is minimised, on average, by a simple classifier - **the Bayes classifier** - that *assigns each observation to the most likely class given its predictor values*, i.e. we should simply assign a test observation with predictor x_0 to the class j for which

$$\Pr(Y = j|X = x_0)$$

is maximised.

If there are only 2 classes, the Bayes classifier predicts class 0 if

$$\Pr(Y = 0|X = x_0) > 0.5$$

and class 1 otherwise.

The **expected prediction error** is

$$\begin{aligned}\text{EPE} &= \mathbb{E}\left[\mathbb{I}(y_0 \neq \hat{y}_0)\right] \\ &= \mathbb{E}_X \sum_{j=0}^{1} \left[\mathbb{I}(y_0 \neq \hat{y}_0)\right] \Pr(y_0 = j|X = x)\end{aligned}$$

and to minimise the expected error, you need to minimize the probability of being wrong, or

$$\hat{y}_0 = 1 \qquad \text{if} \quad \Pr(y_0 = 1|X = x_0) = \max_{j \in \{0,1\}} \Pr(y_0 = j|X = x_0)$$

The **expected Bayes error rate** is then

$$1 - \mathbb{E}_X[\max_{j \in \{0,1\}} \Pr(Y_0 = j|X)]$$

where the expectation is with respect to the probability over all possible values of X.

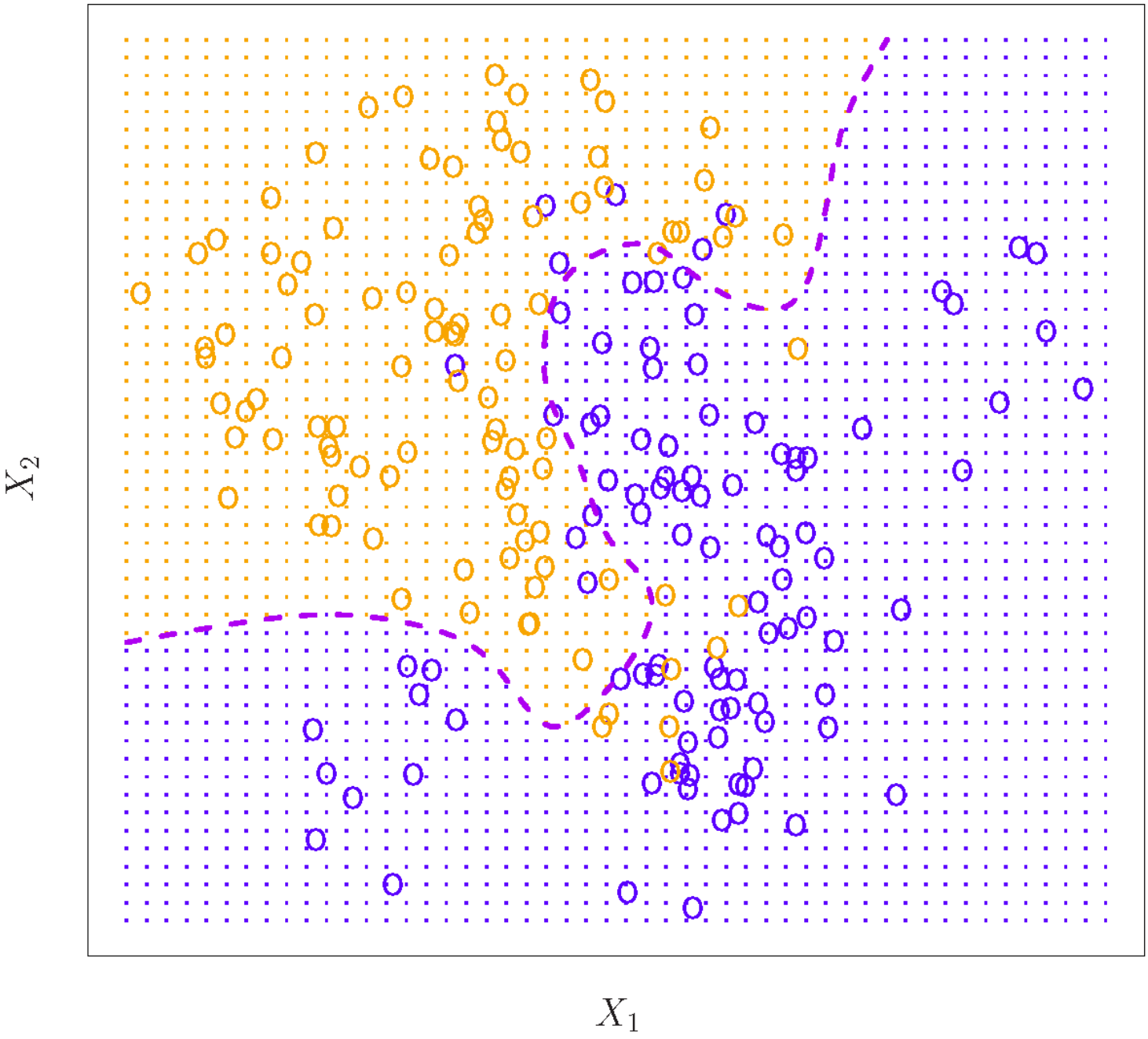


Figure 1.3.5: 100 observations from two groups. The purple line indicates the Bayes decision boundary. The grid colour indicate the group to which a test observation will be allocated to.

K-nearest neighbours

In practice, we do not know the conditional distribution of Y given X, so we can't compute the Bayes classifier. The Bayes classifier is an **unattainable gold standard**.

The **K-nearest neighbours (KNN)** classifier estimates the conditional probability of Y given X and classifies a given observation to the class with the highest estimated probability.

Despite its simplicity, the KNN classifier produces classifications that are often close to the Bayes classifier.

Given a positive integer K (**the choice of K is essential!**) and a test input observation x_0, the KNN classifier

- Identifies the K points in the training dataset closest to x_0 - a neighbourhood of x_0, N_0
- Computes

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} \mathbb{I}(y_i = j)$$

- Applies the Bayes rule

$$\Pr(X = x_0|Y = j) = \frac{\Pr(Y = j|X = x_0)}{\Pr(Y = j)}$$

- Classifies the test observation x_0 to the class with the largest probability

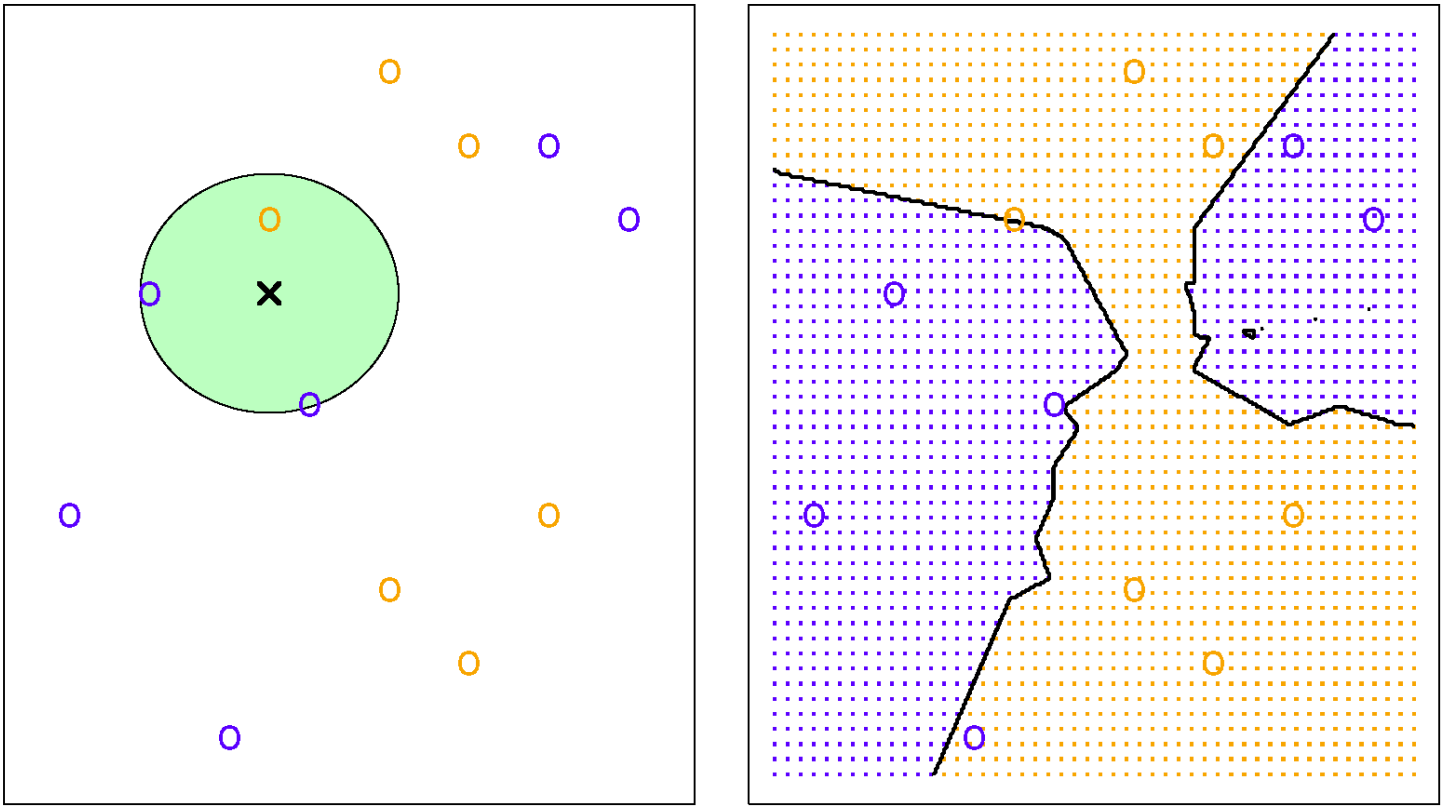


Figure 1.3.6: KNN approach using $K = 3$. Left: a test observation. Right: decision boundary.

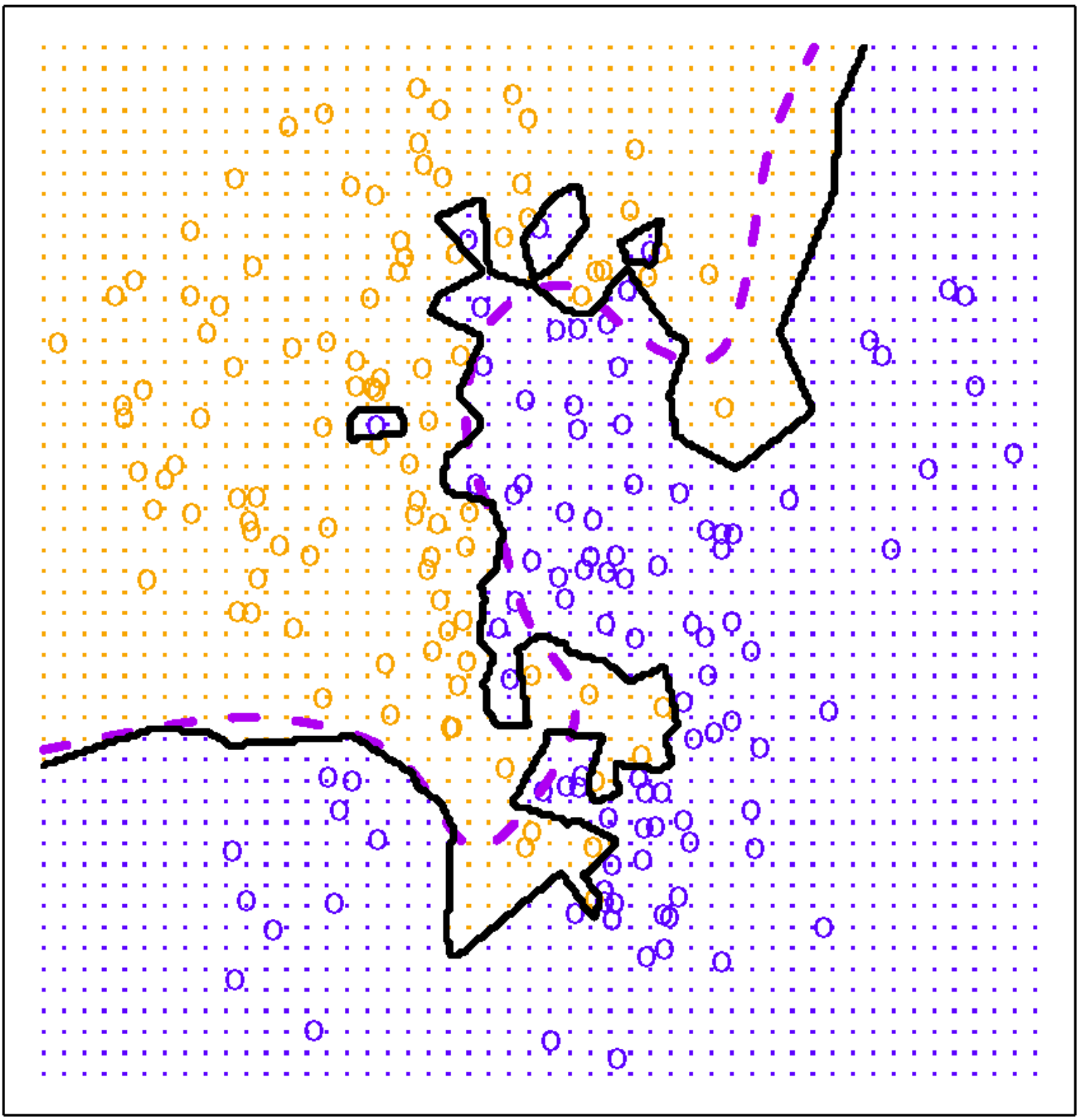


Figure 1.3.7: KNN with $K = 1$. KNN decision boundary compared with Bayes decision boundary.

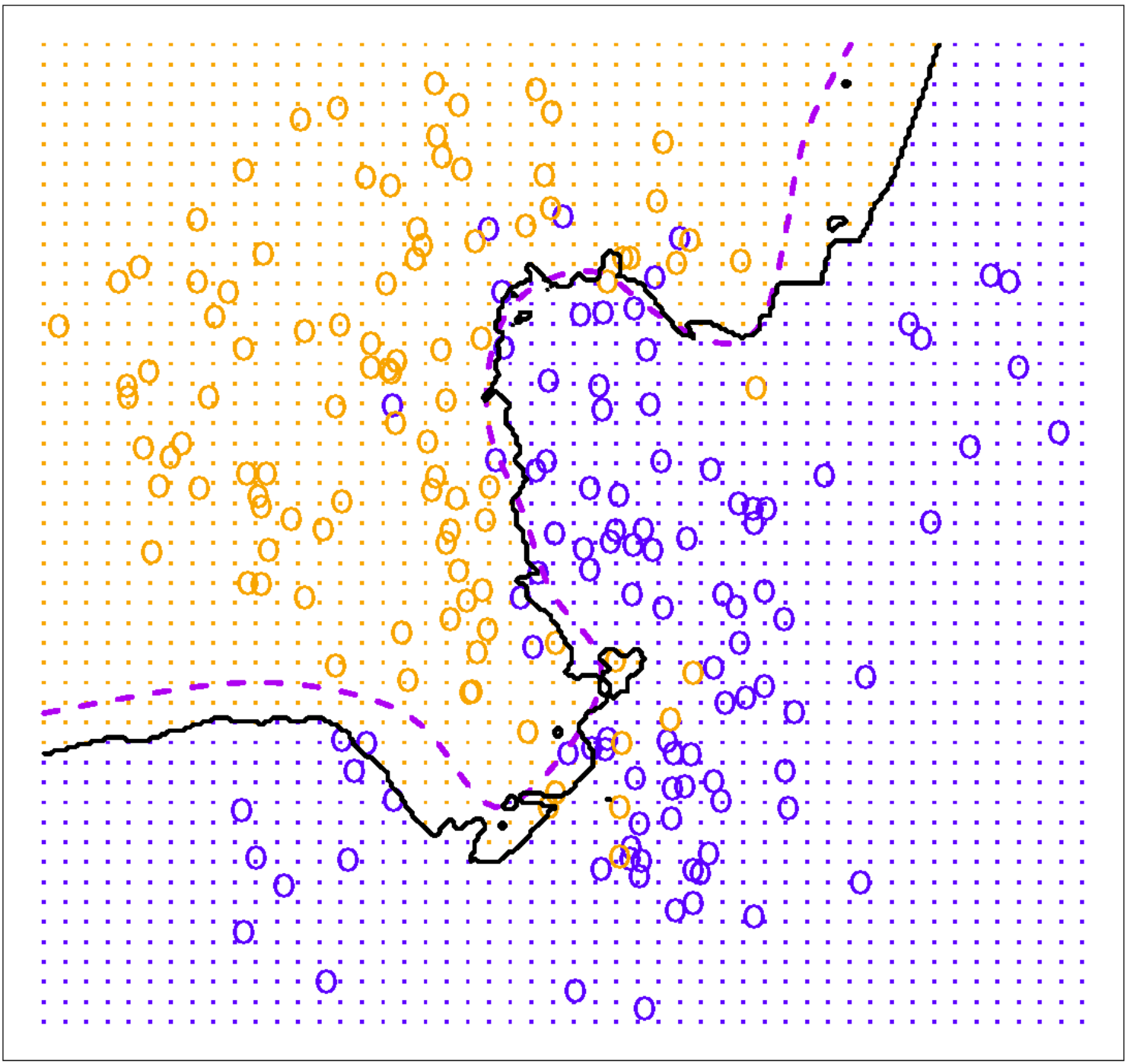


Figure 1.3.8: KNN with $K = 10$. KNN decision boundary compared with Bayes decision boundary.

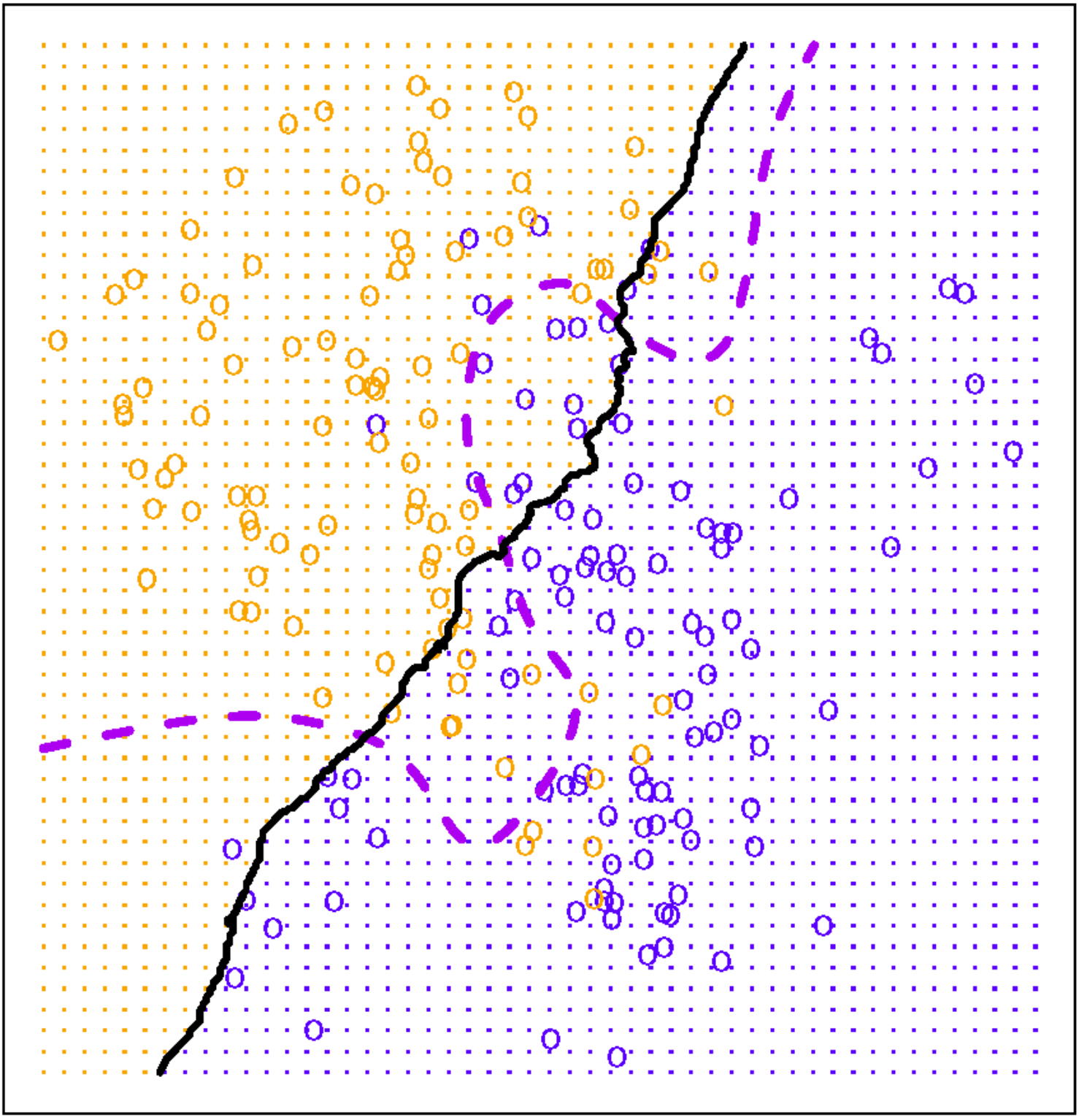


Figure 1.3.9: KNN with $K = 100$. KNN decision boundary compared with Bayes decision boundary.

References

Sources and recommended reading:

1. A. J. Dobson & A. G. Barnett (2018) An introduction to generalised linear models, Chapman and Hall. Chapters 1-2.
2. G. James, D. Witten, T. Hastie & R. Tibshirani (2013) An Introduction to Statistical Learning with Applications in R, Springer. Chapters 1-2.

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani