Latent Class Analysis

Mark Andrews
Psychology Department, Nottingham Trent University

✉ mark.andrews@ntu.ac.uk

Parametric model fitting

- So far, we have encountered some probability distributions used to model the conditional distribution of outcome variables in regression model:
  - Normal distributions. These are used to model the outcome variable in standard linear regression models.
  - Bernoulli distributions. These are models of binary outcome variables, and are used in binary logistic regression as elsewhere.
  - Poisson distributions. These are used to model count variables, and are used in Poisson regression.

In any cases, we can fit these models to data by modifying their parameters to achieve the best fit, often done by maximum likelihood estimation.

Fitting parametric models

- Assume our data is n observations $y_1, y_2 \dots y_n$.
- If we assume that

$$y_i \sim N(\mu, \sigma^2), \quad \text{for } i \in 1 \dots n,$$

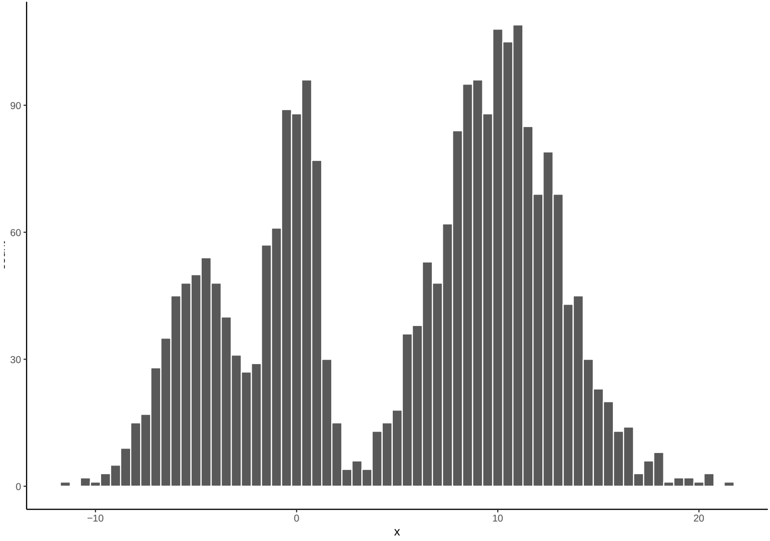
then we can calculate the likelihood function for μ and σ^2, i.e.

$$L(\mu, \sigma^2 | y_1 \dots y_n) \propto \prod_{i=1}^{n} P(y_i | \mu, \sigma^2),$$

and maximize this function for μ and σ^2.

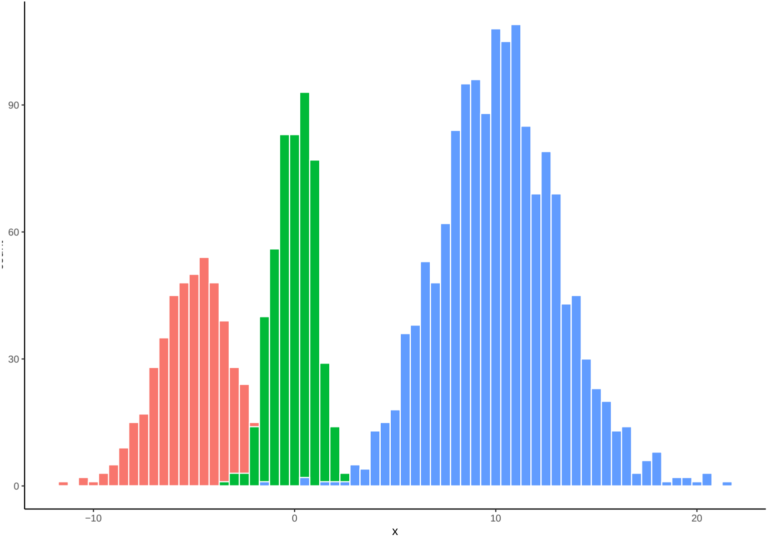
Irregular distributions

- What should we do when encounter data of the following form?



Mixture model

- A mixture model assumes that the data is sampled from independent component distributions, each of which can be modelled by parametric distributions.



Latent variables

- With irregular data, even if assume it is derived from a mixture of independent distributions, we do not know which data point came from which distributions.
- In other words, we have a set of data $y_1, y_2 \ldots y_n$, but we don't know which distribution each data point came from or even how many distributions there are.
- In this situation, we assume that for each y_i data point, there is an z_i that tells us which distribution y_i came from.
- This z_i is a *latent* variable. It has some value, but we don't or can't observe it directly.
- Another name for a model of this kind is a *latent class model*. We assume each y_i belongs to some class, but we just don't or can't observe what that class is.

Mixture models: The probabilistic generative model

- We start by assuming that there are K distinct hidden classes, e.g. $K = 3$.
- So each $z_i \in \{1, 2, 3\}$.
- Then, our model is

$$y_i \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } z_i = 1 \\ N(\mu_2, \sigma_2^2), & \text{if } z_i = 2 \\ N(\mu_3, \sigma_3^2), & \text{if } z_i = 3 \end{cases},$$

$$z_i \sim P(\pi),$$

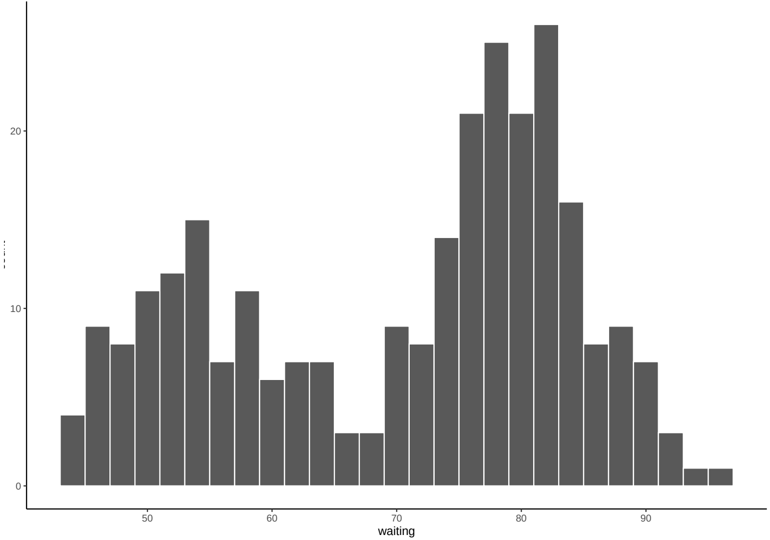
where $\pi = [\pi_1, \pi_2, \pi_3]$ is a probability distribution of $\{1, 2, 3\}$, i.e. π_1 gives the probability that the latent's class's value is class 1, π_2 gives the probability that the latent's class's value is class 2, π_3 gives the probability that the latent's class's value is class 3.

Mixture models: Inference

- In a normal mixture model with $K = 3$ components, we have 9 parameters:
  - μ_1, σ_1^2: The parameters of component distribution 1.
  - μ_2, σ_2^2: The parameters of component distribution 2.
  - μ_3, σ_3^2: The parameters of component distribution 2.
  - π_1, π_2, π_3: The relative probabilities of each component.
- In addition, we have the probability distribution over each value $x_1, x_2 \ldots x_n$.
- Inferring these values by maximum likelihood estimation is usually done by the *expectation-maximization* algorithm.

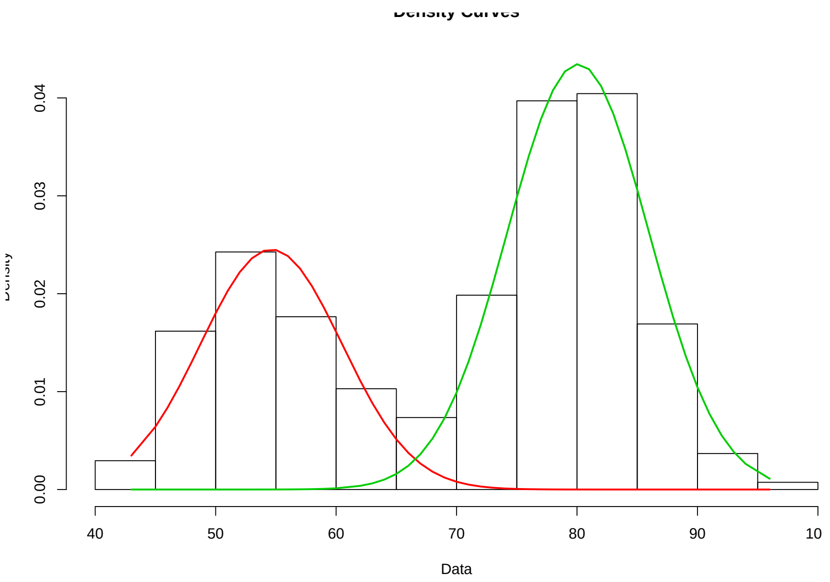
Example: Old faithful

- The distribution of waiting times.



Example: Old faithful

```
M <- normalmixEM(faithful$waiting, k=2)
```



Example: Old faithful

- The inferred means are

```
## [1] 54.61489 80.09109
```

- The inferred standard deviations

```
## [1] 5.871245 5.867716
```

- The relative probabilities of the two components

```
## [1] 0.360887 0.639113
```

Example: Old faithful

- The probabilities for each z_i (for first 10 values)

comp.1	comp.2
0	1
1	0
0.004	0.996
0.967	0.033
0	1
1	0
0	1
0	1
1	0
0	1

Mixture regression models

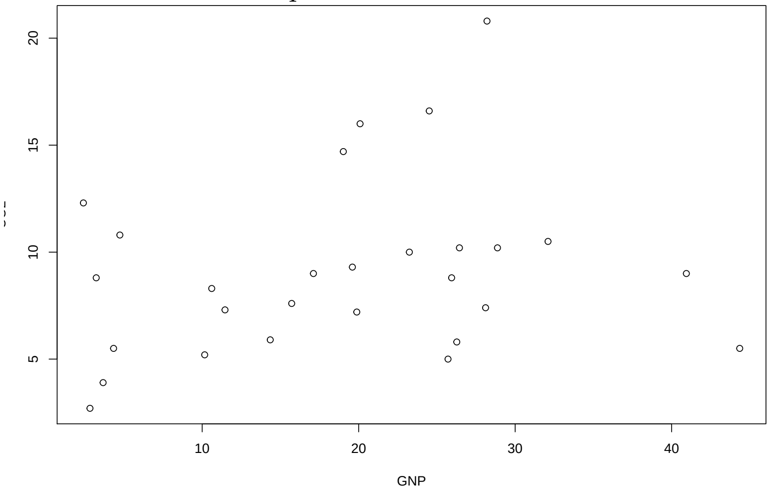
- In a mixture of regressions, we assume that there are K regression models.
- Each data point being associated with one of them.
- Again, we don't know which component it came from. This is given by a latent variable.

$$y_i \sim \begin{cases} N(\alpha_1 + \beta_1 x_i, \sigma_1^2), & \text{if } z_i = 1 \\ N(\alpha_2 + \beta_2 x_i, \sigma_2^2), & \text{if } z_i = 2 \end{cases},$$

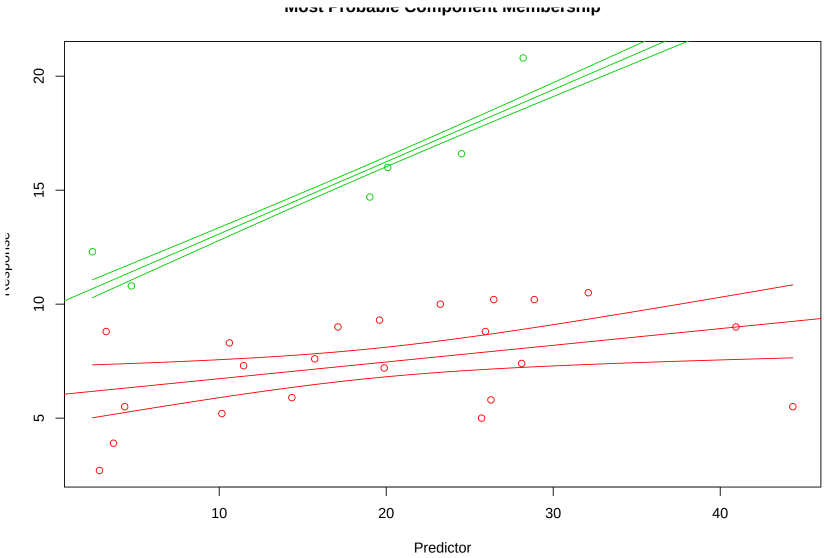
$$z_i \sim P(\pi),$$

Mixture regression models

- A mixture of two scatterplots?



Mixture regression models



Mixture regression models

- The inferred coefficients are

	comp.1	comp.2
beta.0	5.998	9.914
beta.1	0.07324	0.3166

- The inferred standard deviations

```
## [1] 2.025022 1.316107
```

- The relative probabilities of the two models

```
## [1] 0.7875034 0.2124966
```

Mixture of regressions

- The probabilities for each z_i (for first 10 values)

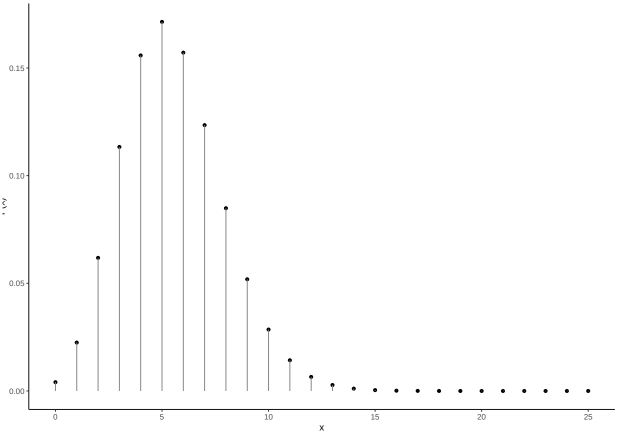
comp.1	comp.2
0.006	0.994
1	0
0	1
1	0
1	0
0	1
1	0
1	0
1	0
0.193	0.807

Zero-Inflated Poisson regression

- A zero inflated Poisson regression is $K = 2$ mixture regression model.
- There are two component models, so $K = 2$ and each latent variable $z_i \in \{0, \}$.
- The probability that $z_i = 1$ is a logistic regression function of the predictor(s) x_i.
- The two component of the zero-inflated Poisson model are:
  1. A Poisson distribution.
  2. A zero-valued point mass distribution (a probability distribution with all its mass at zero).

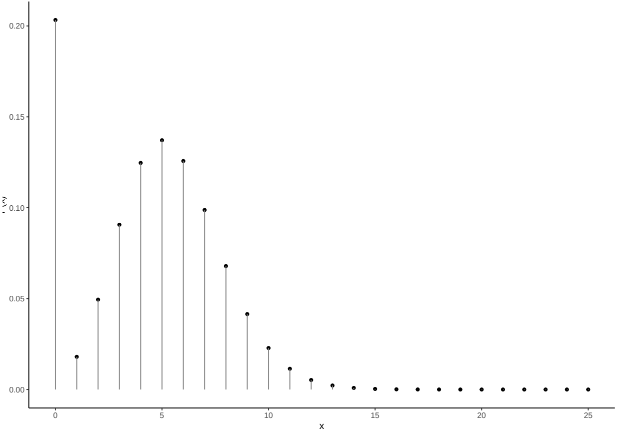
Poisson Distribution

A sample from a Poisson distribution with $\lambda = 5.5$.



Zero inflated Poisson Distribution

A sample from a zero inflated Poisson distribution with $\lambda = 5.5$, with probability of *zero-component* is 0.2.



Poisson regression to Zero-Inflated Poisson regression

- In Poisson regression (with a single predictor, for simplicity), we assume that each y_i is a Poisson random variable with rate λ_i that is a function of the predictor x_i.
- In Zero-Inflated Poisson regression, we assume that each y_i is distributed as a Zero-Inflated Poisson mixture model:

$$y_i \sim \begin{cases} \text{Poisson}(\lambda_i) & \text{if } z_i = 0, \\ 0, & \text{if } z_i = 1 \end{cases}$$

where rate λ_i and $P(z_i = 1)$ are functions of the predictor x_i.

Zero-Inflated Poisson regression

- Assuming data $\{(x_i, y_i), (x_2, y_2) \ldots (x_n, y_n)\}$, Poisson regression models this data as:

$$y_i \sim \begin{cases} \text{Poisson}(\lambda_i) & \text{if } z_i = 0, \\ 0, & \text{if } z_i = 1 \end{cases},$$

$$z_i \sim \text{Bernoulli}(\theta_i),$$

where θ_i and λ_i are functions of x_i.

Zero-Inflated Poisson regression

- The θ_i and λ_i variables are the usual suspects, i.e.

  $$\log(\lambda_i) = \alpha + \beta x_i,$$

  and

  $$\log\left(\frac{\theta_i}{1-\theta_i}\right) = a + b x_i.$$

- In other words, λ_i is modelled just as in ordinary Poisson regression and θ_i is modelled in logistic regression.

Ordinary Poisson

- Here we use a Poisson regression model to predict smoking frequency as function of education level.

```
Df <- read_csv('../data/smoking.csv')

M <- glm(cigs ~ educ, family = poisson, data=Df)
```

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  2.47235218 0.049273913 50.175682 0.000000e+00
## educ        -0.02514134 0.003906852 -6.435192 1.233176e-10
```

Ordinary Poisson

- The role of education according to the Poisson model is

$$\lambda = e^{2.472+-0.025x_i}.$$

So, for $x_i = 6$,

$$\lambda = e^{2.472+-0.15} = \lambda = e^{2.322} = 10.196,$$

and for $x_i = 18$,

$$\lambda = e^{2.472+-0.45} = \lambda = e^{2.022} = 7.553,$$

Zero-inflated Poisson

```r
library(pscl)
M <- zeroinfl(cigs ~ educ, data=Df)
```

- The coefficients for the (non-zero) Poisson model.

```
##              Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) 2.6978666 0.056778631 47.515527 0.000000e+00
## educ        0.0347184 0.004536394  7.653304 1.958797e-14
```

- The coefficients for the logistic regression predicting the zero component.

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.56273268 0.30605009 -1.838695 0.0659601033
## educ         0.08356933 0.02417456  3.456912 0.0005464025
```

Zero-inflated Poisson

- The zero-inflated Poisson provides us first with the probability of the zero-model as

$$\theta_i = \frac{1}{1 + e^{-(-0.563 + 0.084 x_i)}},$$

so, for $x_i = 6$,

$$\theta_i = \frac{1}{1 + e^{-(-0.563 + 0.504)}} = 0.485$$

and for $x_i = 18$,

$$\theta_i = \frac{1}{1 + e^{-(-0.563 + 1.512)}} = 0.721$$

Zero-inflated Poisson

- The role of education according to the Poisson component of the zero-inflated Poisson model is

$$\lambda = e^{2.698+0.035\cdot x_i}.$$

So, for $x_i = 6$,

$$\lambda = e^{2.698+0.21} = 18.32,$$

and for $x_i = 18$,

$$\lambda = e^{2.698+0.63} = 27.883,$$

- There is a substantial increase in smoking with education level.

Zero-inflated Poisson

- The expected number of cigarettes smoked by, for example, a person with 18 years of education is obtained by calculating the expected number according to the non-zero Poisson component and multiplying this by the probability of choosing this component.
- The expected number of cigarettes according to the Poisson component is 27.883.
- The probability of choosing this component is $1 - 0.721 = 0.279$.
- Therefore, the expected number is $27.883 \times 0.279 = 7.779$.
- Using similar procedure for the person with 6 years of education, the expected number of cigarettes is $18.32 \times 0.515 = 9.435$.