STAT40180 — Stochastic Models

Brendan Murphy

Week 1

Introduction

Module Details

Lectures:

- Online videos
- Discussion forum
- Problem sheets
- Code examples

Assesment:

- Homework – 20%.
- Final Exam – 80%.

Grading:

http://mathsci.ucd.ie/tl/grading/en06

Lecturing Format

- Lecture notes on blackboard.
- Videos available to stream (backup of videos available to download)
- Mixture of theory and practical examples.
- **Please** ask questions, answer questions and discuss material on blackboard.

Assessment Format

- The module will have a number of assignments:
  Week 2, Week 5, Week 8, Week 11.
- The assignments will look at modeling problems of the type covered in class.
- The assignments will consist of a mixture of calculations, code and application.
- There will be two weeks to complete each assignment.

Module Overview

- The course topics will include:
    - Statistical Models
    - Inference
    - Developing Models
    - Smoothing & Flexible Regression Models
    - Time To Event Data
    - Stochastic Processes
    - ...

STAT40180 — Stochastic Models

Brendan Murphy

Week 1

Motivating Examples

Crime Statistics

- Suppose you work in the police station of a city.
- Suppose that a certain number of crimes are reported per week.
- However, you don't know how many crimes happen.
- You also don't know what percentage of crimes are reported.
- You get data of the number of crimes that are reported for the last ten weeks:
  38 34 32 34 32 27 28 36 37 33
- Can you estimate the number of crimes that happen per week?
- Can you estimate the percentage of crimes reported?

Taxi Cab Problem

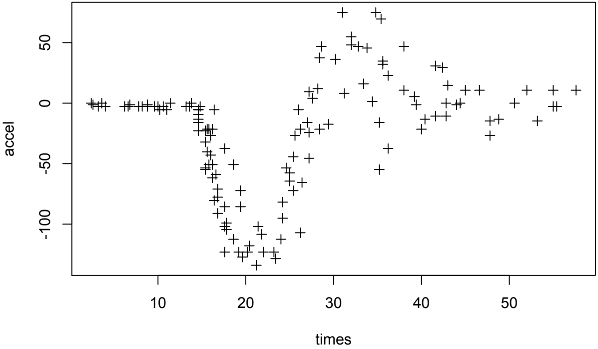
- Suppose you land into a new city that you've never been to before.
- You walk out of the airport and you see that some taxis.
- You are wondering how many taxis there are in the city.
- You see eight taxis parked outside the airport.
  You notice that they are numbered and their numbers are:
  `127 469 404 148 315 170 271 131`
- Can you estimate how many taxis there are in the city?

Employee Retention

- You are working for a company where they are concerned about employee retention.
- The company has collected sample data on a number of employees (former and current) and how long they worked in the company.
- The following data were recorded (the time units are weeks):
  40* 94 83 88 13 70* 49 130* 55 100*
  31 79 17 162 76 2* 11 97 30* 77
  Employees marked with a * are still with the firm.
- How long do employees stay working with the company?
- Suppose we had other covariates about the employees.
  How could we use this in studying retention?

Motorcycle Crash

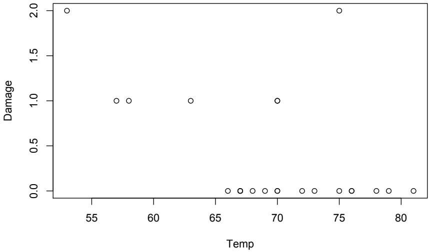
- The head acceleration of a motorcyclist was recorded in a simulated crash.
- A plot of acceleration versus time is given as follows:



- How do we model the relationship between time and acceleration?

Space Shuttle

- On January 20th, 1986 the space shuttle Challenger exploded shortly after it was launched.
- The root cause of the explosion was the failure of rubber O-rings on the fuel tanks (there were six O-rings).
- Data on launch temperature (in degrees Fahrenheit) and the number of failed O-rings are given below:



- The temperature on January 20th, 1986 was 32 degrees. What was the probability of O-ring failure on that day?

Summary

- For each of the above problems, we have:
  - Scenario
  - Data
  - Question
- We need to:
  - Develop a model
  - Infer the model unknowns
  - Use the model to answer the question of interest.

STAT40180 — Stochastic Models

Brendan Murphy

Week 1

Models

Models

- We now consider potential models for each of the motivating examples:
  - Crime Statistics
  - Taxi Cab Problem
  - Employee Retention
  - Motorcycle Crash
  - Space Shuttle

Crime Statistics

- Let's assume that:
  - the data from each week are independent.
  - the number of crimes happening per week is constant over the data collection period.
  - the probability of a crime being reported is the same for all weeks and crimes.
- What model does this suggest?
- It suggests that the data can be modeled by a Binomial(n, p), where n and p are both unknown.

Taxi Cab Problem

- Let's assume that:
  - the taxis are numbered consecutively.
  - the taxi number doesn't affect it being observed outside the airport.
- What model does this suggest?
- We can assume that each number observed is a draw from a uniform distribution on the numbers $1, 2, \ldots, N$ where N is the unknown number of taxis in the city.

Employee Retention

- Let's assume that:
  - the employees retention times are independent.
  - the times are non-negative.
- What model does this suggest?
- We could use any probability distribution which accommodates positive values:
  - exponential
  - gamma
  - Weibull
  - log-normal
- We would need to allow the parameters to depend on the covariates, if these are available.

Motorcycle Crash

- The relationship between acceleration and time is clear, but it is complex.
- Standard linear regression models won't fit very well.
- If we could change the regression assumption from

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

to

$$Y_i = s(x_i) + \epsilon_i,$$

where $s(\cdot)$ is a "smooth" function, then we may be able to better model the crash.

Space Shuttle

- Let's assume that:
  - the launches are independent.
  - the number of failed O-rings is

$$\text{Binomial}(6, p),$$

    where p depends on the launch temperature.
- What model does this suggest?
- We could fit a binomial regression model:

$$Y_i \sim \text{Binomial}(6, p(x_i))$$

  where

$$p(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

  .

Famous Quote

- For each scenario being modeled, we made a series of assumptions.
- This allowed us to posit a stochastic model for the scenario.
- It could be argued that some of the assumptions are unrealistic.
- However, we may still gain useful information from the modeling excercise.
- George Box once said,

  All models are wrong, but some models are useful

STAT40180 — Stochastic Models

Brendan Murphy

Week 1

Inference Example

Models

- We will consider the crime statistics example.
- We will establish how we can fit the posited model to the data.

Crime Statistics

- We have ten observation $x_1, x_2, \ldots, x_{10}$ that we are proposing to model as Binomial(n, p) with n and p unknown.
- This is an unusual binomial problem because n is unknown.
- Further, n is a discrete quantity so we can't use calculus based methods.
- Also, $0 < p < 1$ which means it is bounded; this may (or may not) be problematic.

Crime Statistics: Method of Moments

- We could try to use method of moments to estimate the model.
- We have two unknown parameters (n, p), so we will need two equations to uniquely identify them.
- We know that under the binomial model

$$\mathbb{E}(X_i) = np \text{ and } \mathbb{V}\text{ar}(X_i) = np(1-p).$$

- If we replace the expected values by the sample moments, we get

$$np = \overline{x} \text{ and } np(1-p) = s^2.$$

Crime Statistics: Method of Moments

- Thus,

$$\begin{aligned} & \overline{x}(1-p) = s^2 \\ \Rightarrow \quad & (1-p) = \frac{s^2}{\overline{x}} \\ \Rightarrow \quad & p = 1 - \frac{s^2}{\overline{x}} \end{aligned}$$

and

$$n = \frac{\overline{x}}{p}$$

Crime Statistics: Estimates

- For the given data, we get:
  $\hat{p} = 0.61$ and $\hat{n} = 54$.
- Thus, we estimate that there are 54 crimes per week and 61% of crimes occuring are reported.
- The following code can be used:

```
x <- scan()
38 34 32 34 32 27 28 36 37 33

xbar <- mean(x)
s2 <- var(x)

phat <- 1-s2/xbar
nhat <- xbar/phat

phat
nhat
```

Crime Statistics: Likelihood

- We could try to estimate (n, p) using maximum likelihood.
- It turns out to be non-trivial, but it is perfectly managable.
- For the observed data, we get the following likelihood function:

$$L(n, p) = \prod_{i=1}^{m} \left(\begin{array}{c} n \\ x_i \end{array} \right) p^{x_i} (1-p)^{n-x_i}.$$

- The log-likelihood is:

$$\ell(n, p) = \sum_{i=1}^{m} \log \left(\begin{array}{c} n \\ x_i \end{array} \right) + \left(\sum_{i=1}^{m} x_i \right) \log p + \left(nm - \sum_{i=1}^{m} x_i \right) \log(1-p).$$

- We want to maximize this, with respect to (n, p).

Crime Statistics: Likelihood

- Suppose, for a moment, that n is known.
- We could maximize the likelihood with respect to p to find that

$$\hat{p}(n) = \frac{\sum_{i=1}^{m} x_i}{nm} \qquad \text{Check!}$$

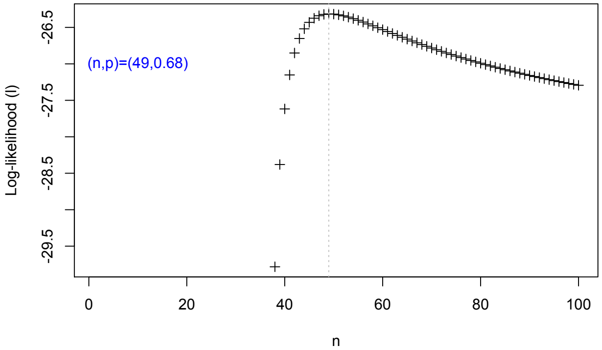
- *I have written it as a function of* n *because the calculation assumed* n *known*.
- We could replace p in the likelihood by $\hat{p}(n)$ to get

$$\ell(n, \hat{p}(n)) = \sum_{i=1}^{m} \log \binom{n}{x_i} + \left(\sum_{i=1}^{m} x_i\right) \log \hat{p}(n) + \left(nm - \sum_{i=1}^{m} x_i\right) \log(1 - \hat{p}(n))$$

- There is no straightforward way to maximize the resulting function with respect to n.

Crime Statistics: Likelihood

- However, because n is a whole number, we can evaluate $\ell(n, \hat{p}(n))$ for a range of values of n.
- The resulting plot is as follows:



- In this case, we got a lower value for the number of crimes but a higher percentage being reported.

Crime Statistics: Likelihood Code

- The code for doing the maximum likelihood estimation.

```
l<-function(n,p,x)
{
sum(dbinom(x,n,p,log=TRUE))
}

phat<-function(x,n)
{
m<-length(x)
sum(x)/(n*m)
}

l2<-function(n,x)
{
l(n,phat(x,n),x)
}

nvec<-1:100
lvec<-rep(NA,length(nvec))

for (n in nvec)
{
lvec[n]<-l2(nvec[n],x)
}
plot(nvec,lvec,pch=3,xlab="n",ylab="Log-likelihood (l)")

abline(v=49,col="gray",lty=3)
text(10,-27,"(n,p)=(49,0.68)",col="blue")
```