Chapter 19

Generalised linear models and other animals

19.1 Seatbelts

The file `glm_seatbelts.csv` contains data on the monthly total of car drivers killed (on a log 10 scale) in Great Britain between January 1969 and December 1984, see:

```
https://stat.ethz.ch/R–manual/R–devel/library/datasets/html/
   UKDriverDeaths.html
```

It also contains a measure of petrol prices over the same period, as well as a variable that represents the month on a scale of 1-12.

During the period for which the data runs there was a change in the law that meant it became a legal requirement to wear seatbelts in cars. In this question we are going to estimate when this event occurred by examining the data.

Problem 19.1.1. Plot the data. Can you see by eye when the legislation was likely enacted?

It looks like there is a structural break in the series around 1983 (see Figure 19.1), which happens to be when the law was enacted (at the end of January that year).

Problem 19.1.2. A model is proposed of the form,

$$deaths(t) \sim \mathcal{N}\left(\alpha + \beta petrol(t) + \sum_{i=1}^{11} \delta_i D(i,t) + Gamma(t,s), \sigma\right) \tag{19.1}$$

where,

$$\gamma = \begin{cases} 0, & \text{if } t < s \\ \gamma_0, & \text{if } t \geq s, \end{cases} \tag{19.2}$$

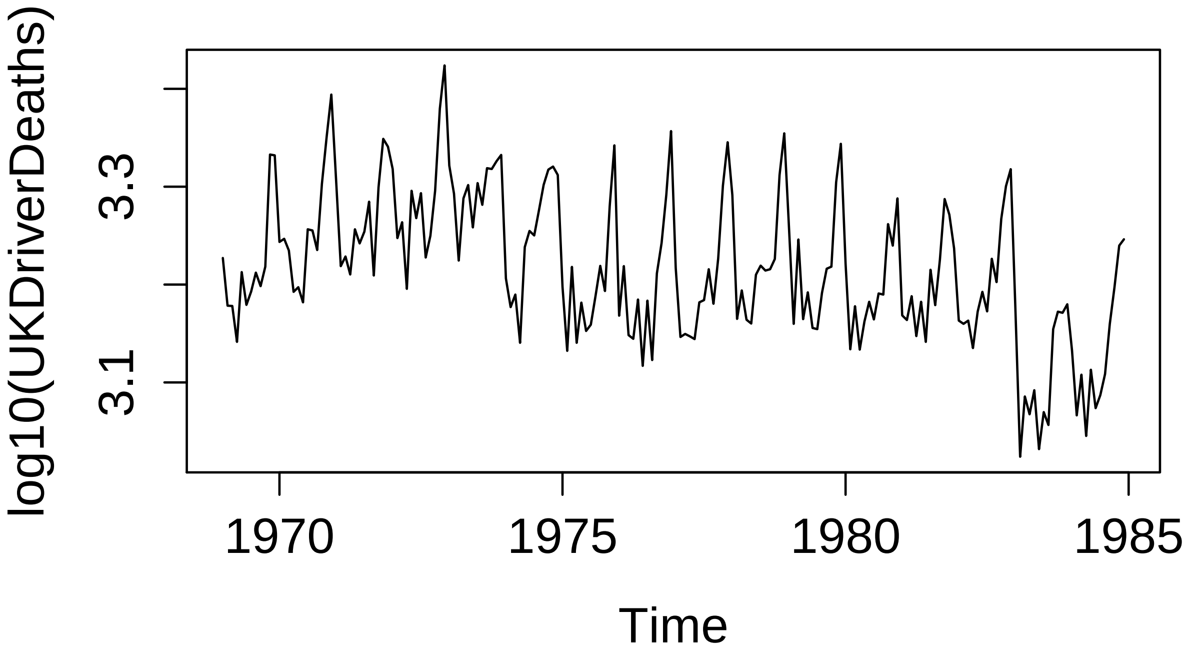

Figure 19.1: The number of car drivers killed in car accidents in Great Britain.

and $\gamma_0 < 0$ represents the effect of the seatbelt legislation on the numbers of car drivers killed after some implementation date s; $D(i,t)$ is a dummy variable for month i equal to 1 if and only if the date t corresponds to that month, and is zero otherwise.

Implement the above model in Stan, and hence estimate the effect that the seatbelt legislation had on car driver deaths.

This can be done with the following code,

```
functions{
  // function that returns a dummy variable from 1-11 if in that month.
  // For december return zero to avoid falling into the dummy variable trap
  real dummySelector(int aMonth, real[] dummies){
    real aDummy;
    if(aMonth < 12){
      return(dummies[aMonth]);
    }else{
      return(0.0);
    }
  }
}
```

```
data{
  int N;
  real deaths[N];
  real petrol[N];
  int month[N];
}

transformed data{
  real log_unif;
  log_unif = - log(N);
}

parameters{
  real beta;
  real delta[11];
  real alpha;
  real<upper=0> gamma;
  real<lower=0> sigma;
}

transformed parameters{
  vector[N] lp;

  // discrete uniform prior on s
  lp = rep_vector(log_unif, N);
  for(s in 1:N)
    for(t in 1:N)
      lp[s] = lp[s] + normal_lpdf(deaths[t] | t < s ? (alpha +
                beta * petrol[t] + dummySelector(month[t], delta)) :
                (alpha + beta * petrol[t] +
                 dummySelector(month[t], delta) + gamma), sigma);
}

model{
  alpha ~ normal(0, 1);
  beta ~ normal(0, 1);
  delta ~ normal(0, 1);
  gamma ~ normal(0, 1);
  sigma ~ normal(0, 1);

  // marginalise out s
  target += log_sum_exp(lp);
}
```

The median estimate of the effect size (gamma) is around an 8% reduction in car driver deaths.

Problem 19.1.3. Using the `generated quantities` block estimate the date when the legislation was enacted.

This can be done using the softmax function along with a random sample from a categorical distribution,

```
generated quantities {
  int<lower=1, upper=N> s;
  s = categorical_rng(softmax(lp));
}
```

If we run the code we get an output of the form shown in Figure 19.2, which has a median of January 1983 as hoped.

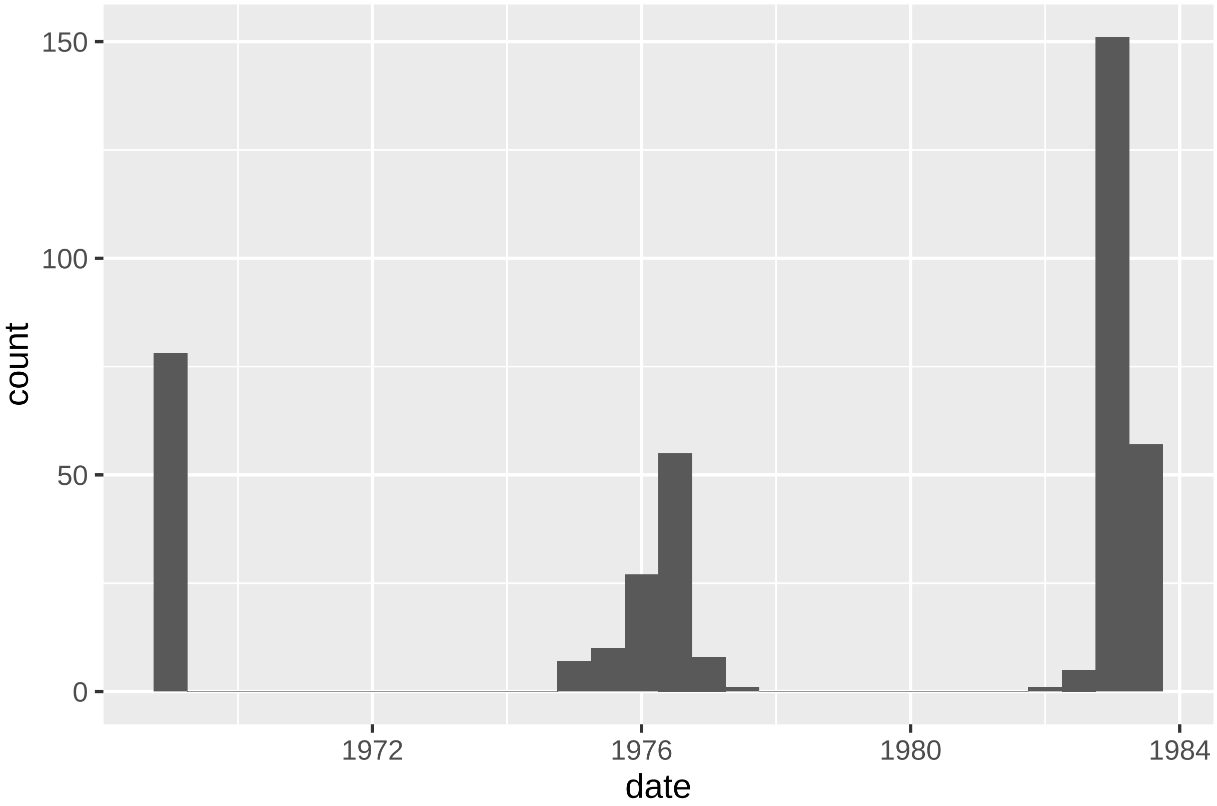

Figure 19.2: The model estimated date when the seatbelt legislation was enacted.

19.2 Model choice for a meta-analysis

Suppose that the data contained in `glm_metaAnalysis.csv` contains the (fictitious) result of 20 trials of a new drug. In each trial 10 patients with a particular disorder treated with the drug, and the data records the number of individuals cured in each trial.

Problem 19.2.1. Graph the data across all 20 trials. What does this suggest about a potential model to explain the data?

There is much more variability in the data than could be explained using a binomial model with a single θ value.

Problem 19.2.2. Suppose that we have two models that we could use to describe the data,

$$X_i \sim \mathcal{B}(10, \theta), \tag{19.3}$$

or alternatively,

$$X_i \sim beta - binomial(10, a, b), \tag{19.4}$$

where X_i is the number of successes in trial $i \in [1, 20]$. Write two Stan programs to fit each of the above models to the data, and use the estimated LOO-CV (use the “loo” package for R) to choose between the above models. (Assign $\theta \sim beta(1, 1)$ and $a, b \sim \mathcal{N}(2, 5)$ for priors for each model, where a and b are constrained to be positive.)

The code to estimate each model is shown below,

```
data{
  int N;
  int n;
  int X[N];
}

parameters{
  real<lower=0> a;
  real<lower=0> b;
}

model{
  X ~ beta_binomial(n, a, b);
  a ~ normal(2, 5);
  b ~ normal(2, 5);
}

generated quantities{
  vector[N] logLikelihood;
  for(i in 1:N)
    logLikelihood[i] = beta_binomial_lpmf(X[i] | n, a, b);
}
```

and,

```
data{
  int N;
  int n;
  int X[N];
}

parameters{
  real<lower=0,upper=1> theta;
}

model{
  X ~ binomial(n, theta);
  theta ~ beta(1, 1);
}

generated quantities{
  vector[N] logLikelihood;
  for(i in 1:N)
    logLikelihood[i] = binomial_lpmf(X[i] | n, theta);
}
```

If we estimate the above we obtain estimates of the elpd of -49.6 and -45.4 for the binomial and beta-binomial models respectively. Using the “compare” function from “loo” we obtain a difference of 4.2 with a standard error of 3. This has a p value well above the threshold for statistical significance. Therefore By this criterion there is nothing to choose between these models.

Problem 19.2.3. An alternative way to choose between these models is to use Bayes factors. Rather than determine the marginal likelihoods explicitly, this can actually be done in Stan by allowing a discrete model choice parameter $s \in \{1, 2\}$ that dictates which model to use. Code up this model in Stan, and by examining the posterior distribution for $Pr(s)$ determine which sampling distribution fits the data best. (Hint: assign equal probability to each model *a priori* and marginalise out s to obtain the log-probability.)

The following code estimates this model,

```
data{
  int N;
  int n;
  int X[N];
}

parameters{
  real<lower=0> a;
  real<lower=0> b;
  real<lower=0, upper=1> theta;
}
```

```
transformed parameters{
  vector[2] lp;
  for(s in 1:2){
    if(s==1)
      lp[s] = log(0.5) + binomial_lpmf(X | n, theta);
    else
      lp[s] = log(0.5) + beta_binomial_lpmf(X | n, a, b);
  }
}

model{
  target += log_sum_exp(lp);
  theta ~ beta(1, 1);
  a ~ normal(2, 5);
  b ~ normal(2, 5);
}

generated quantities{
  vector[2] lProbs;
  lProbs = exp(lp - log_sum_exp(lp));
}
```

The posterior distribution for the $Pr(s = 2|X)$ has a mean of 0.99. In this case we strongly prefer the beta-binomial model. The two approaches use different criteria to choose. They both tend towards the same answer, that the beta-binomial model is better.

Problem 19.2.4. An alternative approach is to use the binomial likelihood, but use a hierarchical model where each θ_i is drawn from some population-level distribution. Comment on whether you would prefer this approach or the beta-binomial model. (Hint: do not estimate the hierarchical model.)

The hierarchical model is really the same as the beta-binomial case, since the latter is essentially, $X_i \sim \mathcal{B}(10, \theta_i)$ where $\theta_i \sim beta(a, b)$. This is the same as the hierarchical model.

19.3 Terrorism

In this question we are going to investigate the link between the incidence of terrorism and a country's level of income. The data in `glm_terrorism.csv` contains for one hundred countries (those for which the latest data was available) the following series,

- **count**: the number of acts of terrorism perpetrated in each country from 2012 to 2015, as compiled by START [1].
- **gdp**: the gross domestic product of each country in 2015 as compiled by the World Bank.
- **population**: the population of each country in 2015 as compiled by the World Bank.

- **gdpPerCapita**: the GDP per capita in each country.
- **religion**, **ethnic**, **language**: measures of fractionalisation with respect to each of these measures, obtained from: `http://www.anderson.ucla.edu/faculty_pages/romain.wacziarg`.
- **law** and **corruption**: measures of the rule of law and corruption (actually an inverse measure) as compiled by the World Bank in their 2016 World Governance Indicators report.
- **democracy** and **autocracy**: indicators of democracy and autocracy respectively from the polity4 database.
- **region** and **region_numeric**: the region to which a country belongs out of Asia, Europe, Middle East and North Africa, Sub-Saharan Africa, South America, and North America.

Problem 19.3.1. Graph the data. What does this tell you about the processes?

Using the `pairs` plotting function in R where we have logged the count, GDP, and population we obtain Figure 19.3. From this plot it is clear that there is a fairly strong relationship between terror count and population size. Otherwise it is hard to see any strong associations with the terror variable although perhaps there is a negative correlation between terrorism and rule of law and the corruption variable (actually an inverse measure of corruption, meaning there is a positive association between corruption and terrorism). Otherwise there is some covariance between GDP per capita and rule of law and corruption (a higher GDP per capita means a lower corruption level). There is similarly strong associations between linguistic and ethnic fractionalisation, and also between rule of law and corruption.

Problem 19.3.2. A simple model for the terrorism count is the following,

$$count_i \sim Poisson(\alpha + \beta_1 population_i + \beta_2 gdpPerCapita_i), \tag{19.5}$$

where i corresponds to one of the countries in our dataset. Code up this model in Stan, and use it to obtain estimates of the effect of a country's income level on the incidence of terrorism.

Problem 19.3.3. Now include corruption, religion and ethnic as further variables in the above generalised linear model. What is the impact of each of these variables on the terrorism count?

Problem 19.3.4. Conduct posterior predictive checks to assess the applicability of the model to the data. What do these tests suggest?

19.4 Eurovision

The data in `Eurovision.csv` contains historical data of the outcome of the Eurovision song contest from 1976 to 2015 for the twenty countries who have featured most consistently in the finals throughout the years. Along with the results from the contest we also include data on the distance between pairs of countries, whether those countries share a common language, and if one was ever colonised by the other. In this question we ask you to develop a model to help explain the way in which countries award points to one another.

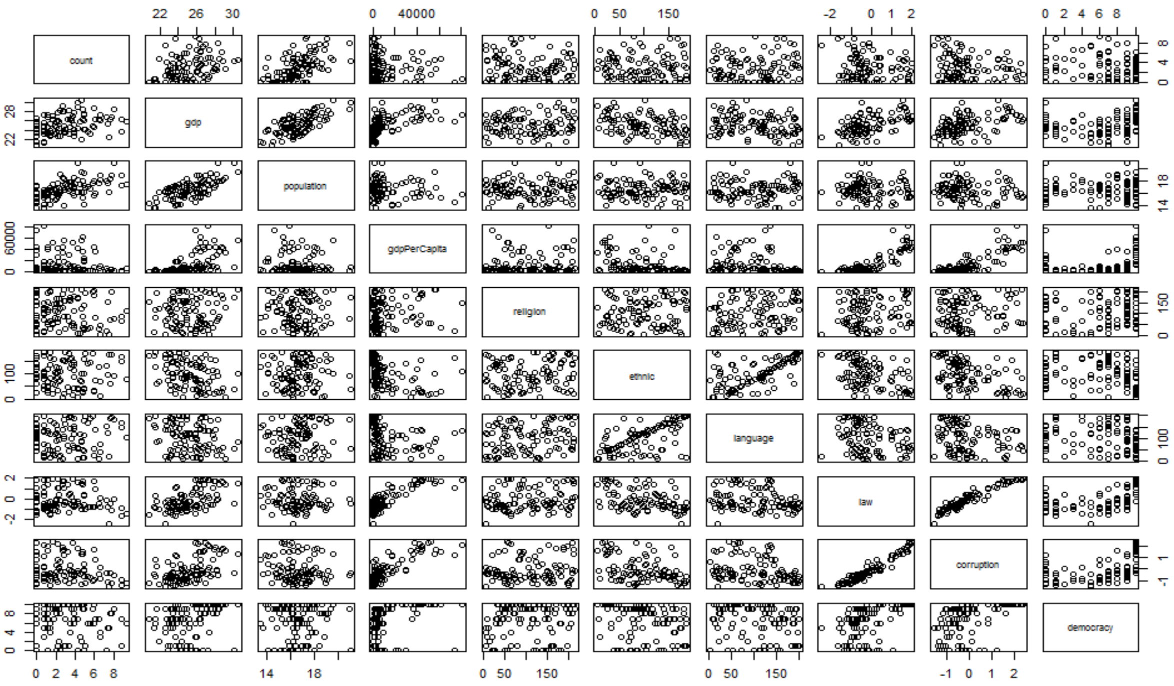

Figure 19.3: Scatter plots of the data for the terrorism example. Note that the count, GDP, and population are logged.

19.5 More terrorism (harder)

The data in `terrorism.csv` contains historical pairwise counts of terrorist attacks perpetrated by citizens of an origin country against a target country, compiled by Alan Krueger see:

```
http://krueger.princeton.edu/pages/
```

assembled from the U.S. State Department's annual list of significant international terrorist incidences (PGT). In this question we ask students to develop a model to explain the incidence of such attacks using data on the attributes of each country (the origin and target).

Bibliography

[1] National Consortium for the Study of Terrorism and Responses to Terrorism (START). Global terrorism database. 2016.