Computer Lab 2

Andrea Bruzzone, Thomas Zhang

2016-04-20

Assignment 1

The multinomial model is:

$p(y_i|\theta_i) \propto \prod_{k=1}^{K} \theta_{ik}^{y_{ik}}$

The prior is:

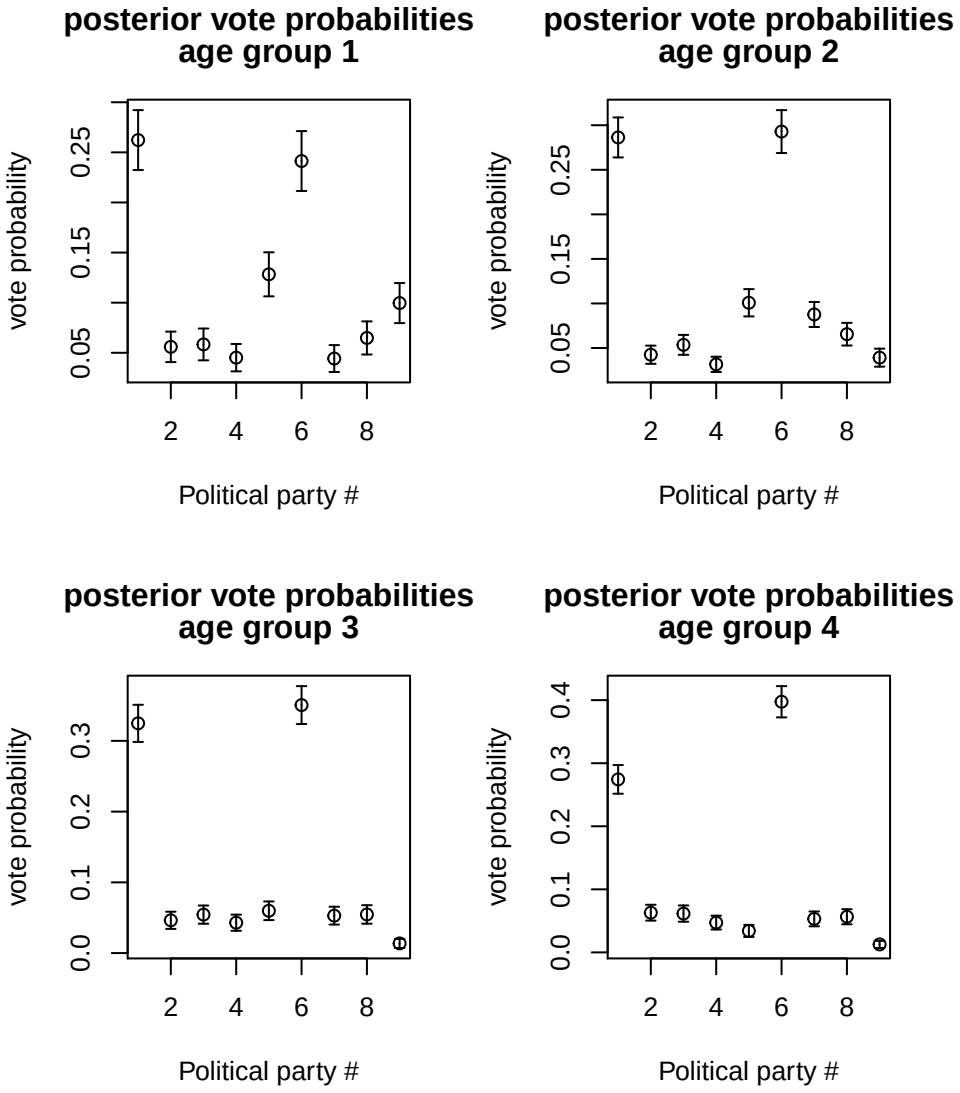
$p(\theta_i) \propto \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1}$

So the posterior is the product of the liklihood and the prior, so the posterior is:

$p(\theta_i|y_i) \propto \prod_{k=1}^{K} \theta_{ik}^{y_{ik}} \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_{ik}^{(\alpha_k + y_{ik}) - 1}$

Or in other words, the posterior is $Dirichlet(\alpha_1 + y_{i1}, \ldots \alpha_k + y_{ik})$-distributed.

We show the posterior means and two standard deviation error bars for θ_{ik}, the probability that a voter in age group i votes for party k, in the plots below. 1000 draws were simulated.



We see that the youngest age group have a propensity to vote for party 5 (Miljöpartiet) and party 9 (Others) while the second youngest group vote more than average for party 5 and party 7 (Vänsterpartiet). The oldest age group tends to vote for party 6 (Socialdemokraterna) more than any other age group.

We determine given these 1000 posterior draws of θ_{ik} the probability that, in each of the four different age groups, the Red-Green bloc will win over the Alliance bloc in a general election.

```
## [1] "Posterior probability of Red-Greens winning in age group 1 is 0.399 percent"
```

```
## [1] "Posterior probability of Red-Greens winning in age group 2 is 0.998 percent"
```

```
## [1] "Posterior probability of Red-Greens winning in age group 3 is 0.424 percent"
```

```
## [1] "Posterior probability of Red-Greens winning in age group 4 is 0.94 percent"
```

Finally, we wish to find the probability that the Red-green bloc will win over the Alliance bloc in a general election. We first find the likely number of eligible voters Y_i in each age group i using the given multinomial

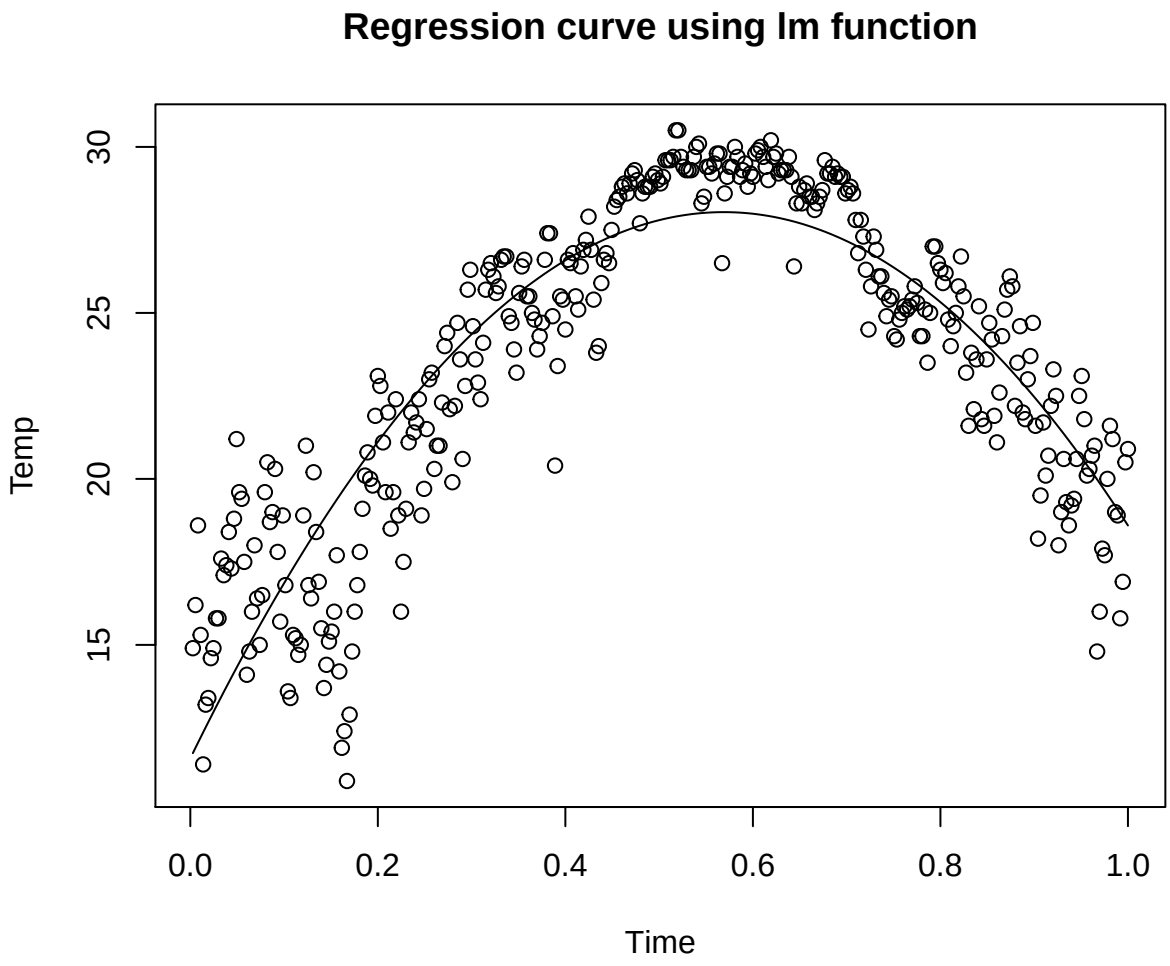
distribution, and then we draw θ_{ik} from the posterior distribution and then we count the votes across all age groups. 1000 repetitions of this procedure were used.

```
## [1] "posterior probability of red-green election win: 0.973"
```

Since we now know that the red-green bloc did indeed beat the alliance bloc in the 2014 general election, we are happy to have found this result from the posterior probability.

Assignment 2

We load data from `JapanTemp.dat` and try to fit the data using a quadratic regression model using the `lm()` function.

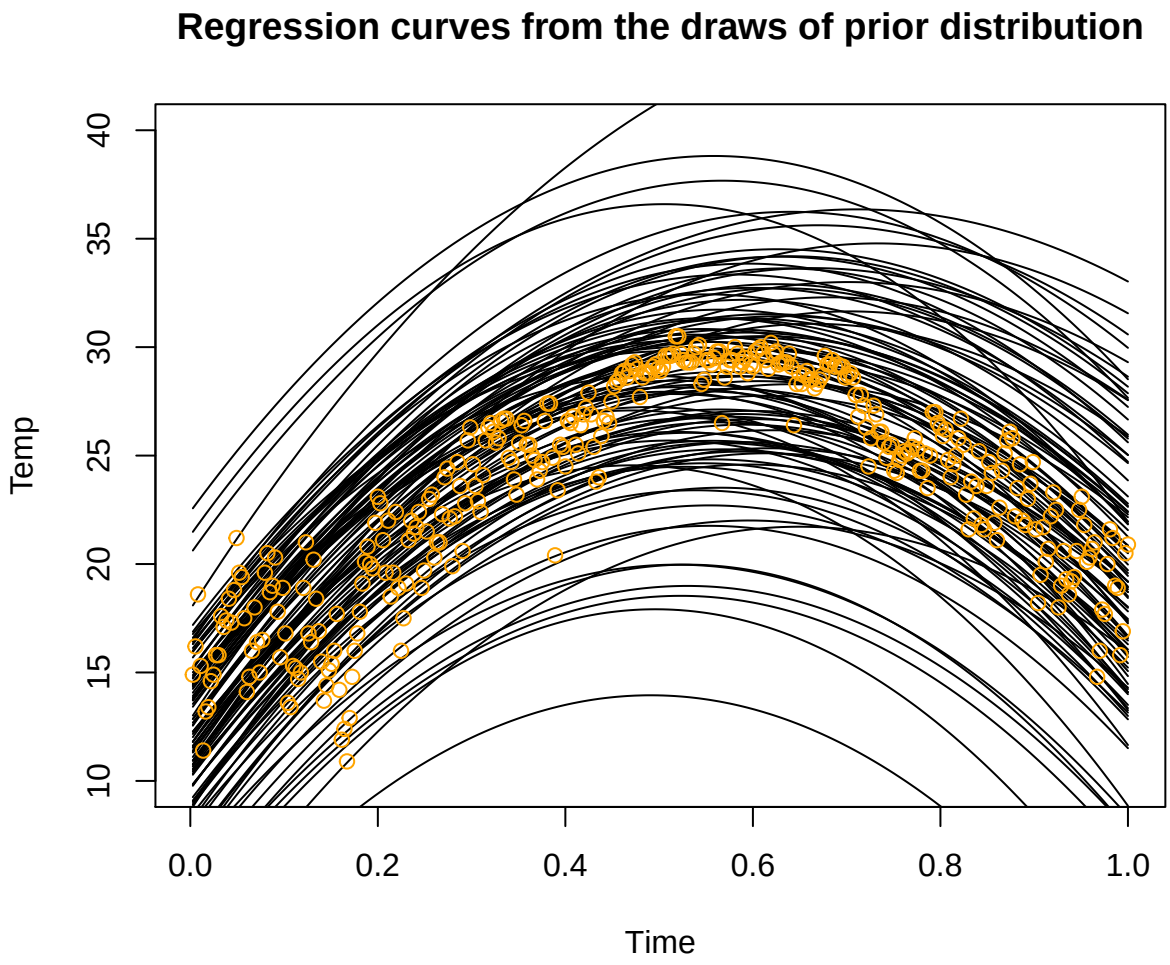


We see that the data is a little noisy, and we calculate the standard deviation of the difference between the curve and the data. it is around 2.17.

we now try to find hyperparameters β_0,Ω_0, ν_0 and s_0^2 and λ for the conjugate prior of a linear regression model. These hyperparameters should 1: simulate the data noise correctly and 2: Create reasonable regression curves during draws.

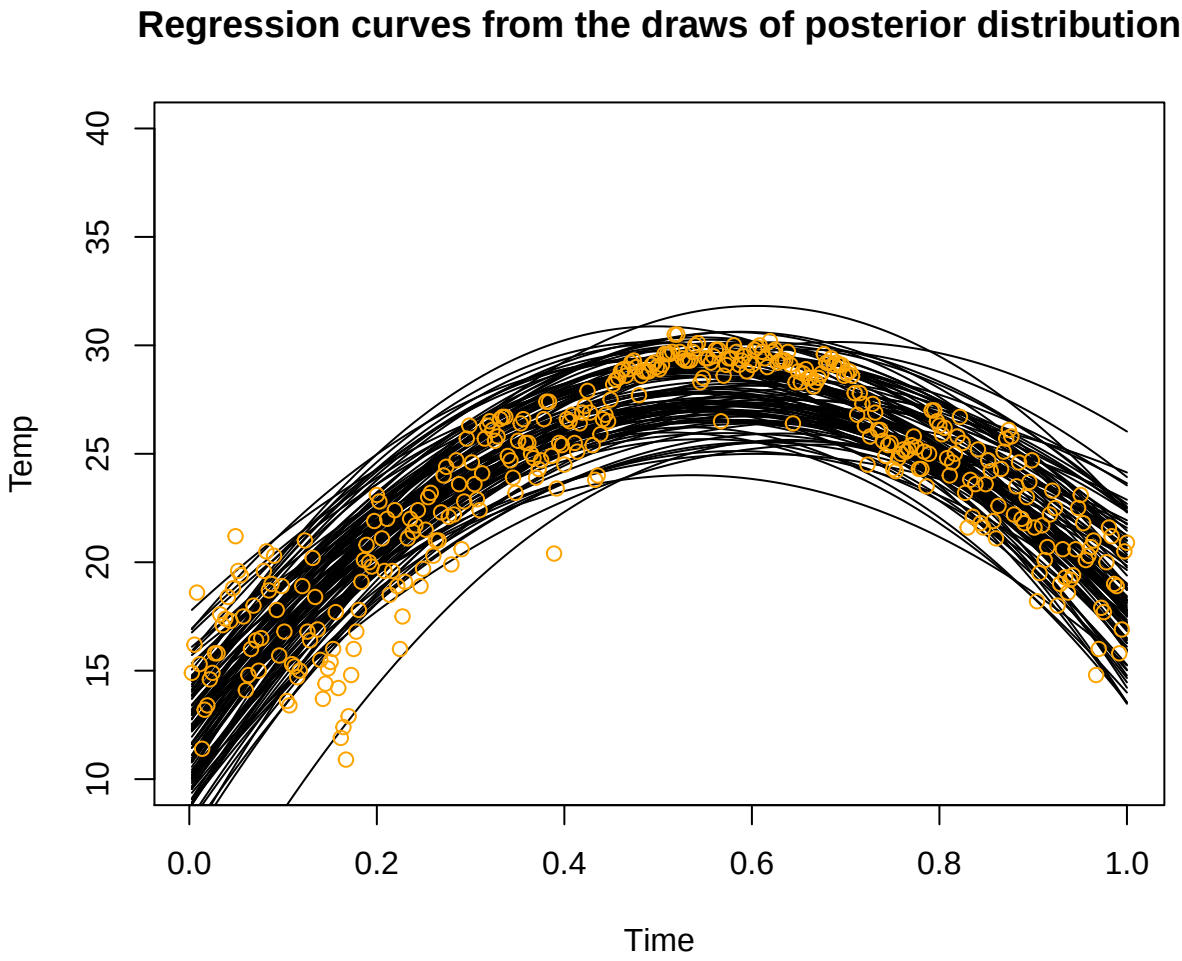
We started with trying to find values of ν_0 and s_0^2 which give us a σ^2 around 4-5. After some trial and error, we decided upon $\nu_0 = 50$ and $s_0^2 = 25$. We further set $\Omega_0 = I_3$ and $\lambda = 2$. The β_0 we get from the `lm()` quadratic model, where $\beta_0 = (11.58, 57.83, -50.82)$.

We simulate draws from the joint prior of all parameters and for every draw we compute the regression curve and we plot it together with the data points.



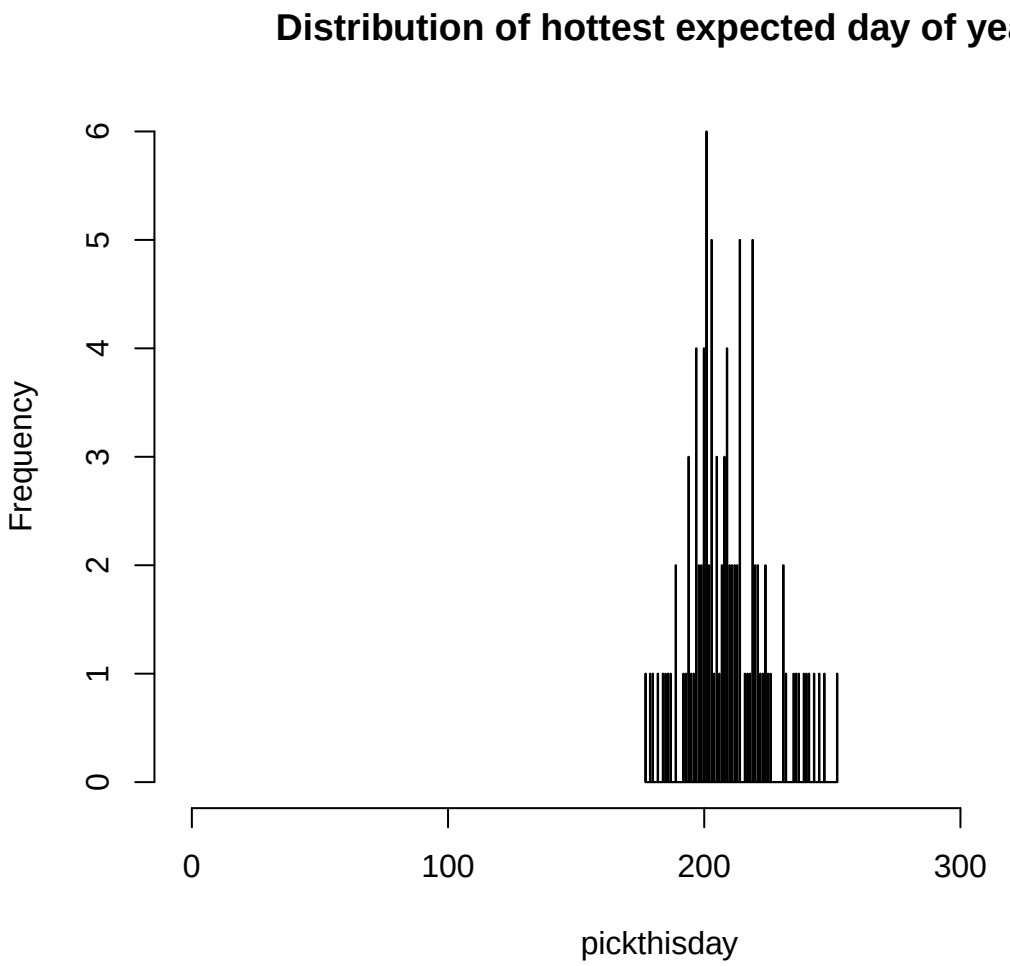
From the plot, looking at the shape of the curves, it can be seen that our prior seems to be sensible.

Then we write a program to simulate from the joint posterior distribution of the parameters and using 100 draws obtained the parameters and then we plot the regression curves resulting from these draws.



We can see that the posterior is less spread out so it can be said that we have a better fit.

Using the values simulated previously we want to simulate from the posterior distribution of the day with the highest temperature. We plot it with an histogram:



It can be seen that the hottest day is around day 200 that is in July, and all the other high values are in the summer period.

As for eliminating the higher order variables in the polynomial model, we suggest a Laplace prior, since then many β_k are close to zero, a situation reminiscent of the LASSO variable selection method.