Topics in Time Series (2)

Yisen Du (Eason)

1 Question1

1. A noisy measurement device is being examined for understanding the distribution of the errors that are being produced by it. Suppose that ten measurements led to the following observations on the errors made by the device:

$$-0.69, -4.26, 0.14, -0.86, 0.42, 24.21, 0.51, -1.23, 2.30, 4.15$$

Consider the following three models for the distribution of these errors:

- Model 1: $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$

- Model 2: $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} \operatorname{Lap}(0, \sigma)$. Recall that the $\operatorname{Lap}(0, \sigma)$ density is given by $x \mapsto \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$

- Model 3: $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} \operatorname{Cauchy}(0, \sigma)$. Recall that the Cauchy $(0, \sigma)$ density is given by $x \mapsto \frac{1}{\pi} \frac{\sigma}{x^2+\sigma^2}$.

a) Under the prior $\log \sigma \sim \operatorname{Unif}(-15, 15)$, calculate the evidences of each of the above three models, given the observed data. Use a numerical approximation method (by gridding the set of possible σ values) for evaluating the integral over σ. (6 points)

b) Normalize the three evidences to obtain posterior probabilities for the three models. Which model has the highest posterior probability? Comment on whether your results seem intutively sensible. (2 points).

1.1 Answer of Q1(a)

By definition of evidence,

$$Evi_M(\varepsilon) = f_{\varepsilon|M}(\varepsilon) = \int f_{\varepsilon|\sigma}(\varepsilon) f_\sigma(\sigma) d\sigma$$

for model1,

$$Evi_{M_1}(\varepsilon) = \int_{e^{-15}}^{e^{15}} (2\pi)^{-\frac{n}{2}} \sigma^{-n} exp(-\frac{\sum \varepsilon_i^2}{2\sigma^2}) \frac{1}{30\sigma} d\sigma$$

Similarly, for model2 and model3,

$$Evi_{M_2}(\varepsilon) = \int_{e^{-15}}^{e^{15}} (2\sigma)^{-n} exp(-\frac{1}{\sigma} |\varepsilon_i|) \frac{1}{30\sigma} d\sigma$$

$$Evi_{M_3}(\varepsilon) = \int_{e^{-15}}^{e^{15}} \pi^{-n} \sigma \prod_{i=1}^{n} \frac{1}{\varepsilon^2 + \sigma^2} \frac{1}{30\sigma} d\sigma$$

To numerically approximate these integrals, we grid $\log \sigma$ values from -15 to 15 and compute the mean of these probability density values. The results are

Evi_1	Evi_2	Evi_3
1.236900e-28	1.511484e-15	1.026217e-10

Table 1: Evidences of three models

Since we prefer the model with higher evidence, model3 is the best among them.

1.2 Answer of Q1(b)

From the hierarchical modelling view,

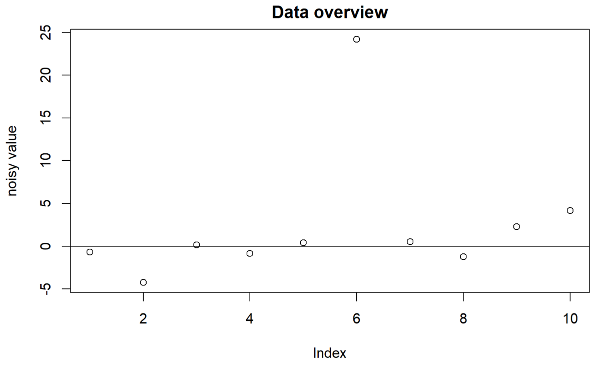
$$Pr\{M = i|Y = \varepsilon\} = \frac{f_{Y|M=i}(\varepsilon)}{\sum_{i=1}^{3} f_{Y|M=i}(\varepsilon)}$$

By normalizing the evidences, we get the posteriors of three models

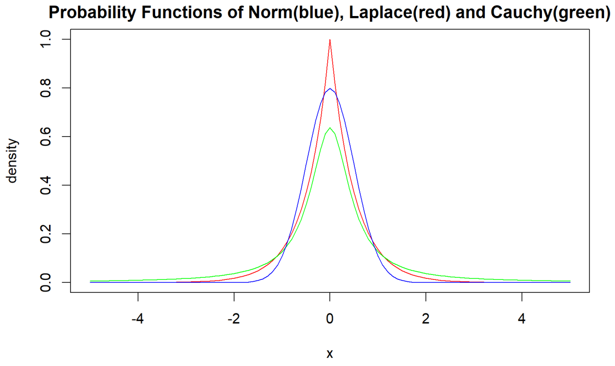
posterior of model1 (Norm)	posterior of model2 (Laplace)	posterior of model3 (Cauchy)
1.2e-18	1.5e-05	0.99

Table 2: posteriors of three models

We found that model3 has the highest posterior probability.
Comment: The reason lies behind the fact that Cauchy distribution has fatter tails comparing to Normal and Laplace. From the data, we find there's an outlier which has a very large value. Due to this point, fat-tail distribution is preferred when using evidence to select models.



(a) figure1



(b) figure2

2 Question2

2. In the file "HW3Data153Fall2022.csv", you will find data on two variables $(x_1, y_1), \ldots, (x_n, y_n)$. Consider the following four models for this dataset:

- Model 1: $Y_i = \beta_0 + \beta_1 x_i + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} C(0,\sigma)(C(0,\sigma)$ has the density $x \mapsto \frac{1}{\pi}\frac{\sigma}{x^2+\sigma^2})$

- Model 2: $Y_i = \beta_0 + \beta_1 x_i + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} \text{Lap}(0,\sigma)(\text{Lap}(0,\sigma)$ has the density $x \mapsto \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right))$

- Model 3: $Y_i = \beta_0 + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} C(0,\sigma)$

- Model 4: $Y_i = \beta_0 + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} \text{Lap}(0,\sigma)$.

a) Numerically calculate the Evidence for each of these models given the observed data. As priors, assume that β_0, σ and β_1 (if β_1 exists in the model) are independent with

$$\begin{aligned}\beta_0 &\sim \text{unif}\{-10, -9.9, -9.8, \ldots, 9.8, 9.9, 10\}\\ \beta_1 &\sim \text{unif}\{-10, -9.9, -9.8, \ldots, 9.8, 9.9, 10\}\\ \log\sigma &\sim \text{unif}\{-10, -9.9, -9.8, \ldots, 9.8, 9.9, 10\}\end{aligned}$$

Report the normalized evidences. Which model has the highest evidence and what is the value of this highest evidence? (6 points)

b) For your chosen model, describe your best estimates of β_0, σ and β_1 (if β_1 exists in the model) along with appropriate uncertainty quantification. Plot your best estimate of $\beta_0 + \beta_1 x$ on a plot of the data. (4 points)

2.1 Answer of Q2(a)

For model 1, if $Z_i \overset{\text{i.i.d}}{\sim} C(0,\sigma)$, then by transformation property of Z_i, $Y_i \overset{\text{i.i.d}}{\sim} Cauchy(\beta_0 + \beta_1 x_u, \sigma)$. Therefore, the likelihood of Y is

$$\prod_{i=1}^{n} \frac{1}{\pi}\frac{\sigma}{\sigma^2 + (y_i - (\beta_0 + \beta_1 x_i))^2}$$

After generating the grid of β_0, β_1, and σ, we can use this likelihood to numerically compute the evidence of model 1, similar to Q1. For model 2, 3, and 4, we found the likelihoods are,

$$(2\sigma)^{-n} exp(-\frac{1}{\sigma}\sum_{i=1}^{n}|y_i - (\beta_0 + \beta_1 x_i)|)$$

$$\prod_{i=1}^{n} \frac{1}{\pi}\frac{\sigma}{\sigma^2 + (y_i - (\beta_0))^2}$$

$$(2\sigma)^{-n} exp(-\frac{1}{\sigma}\sum_{i=1}^{n}|y_i - \beta_0|)$$

After the numerical computation, the normalized evidences are

	Normalized Evidence
model 1	1 (approximately)
model 2	1.66e-56
model 3	2.64e-97
model 4	6.50e-110

Table 3: Normalized evidences of four models

The result shows model 1 has the highest evidence. It intuitively makes sense. From the data, we could find there's increasing trend. Also, there are some outliers far away from the main data. Therefore, the model with slope parameter and fat-tailed residuals is preferred.

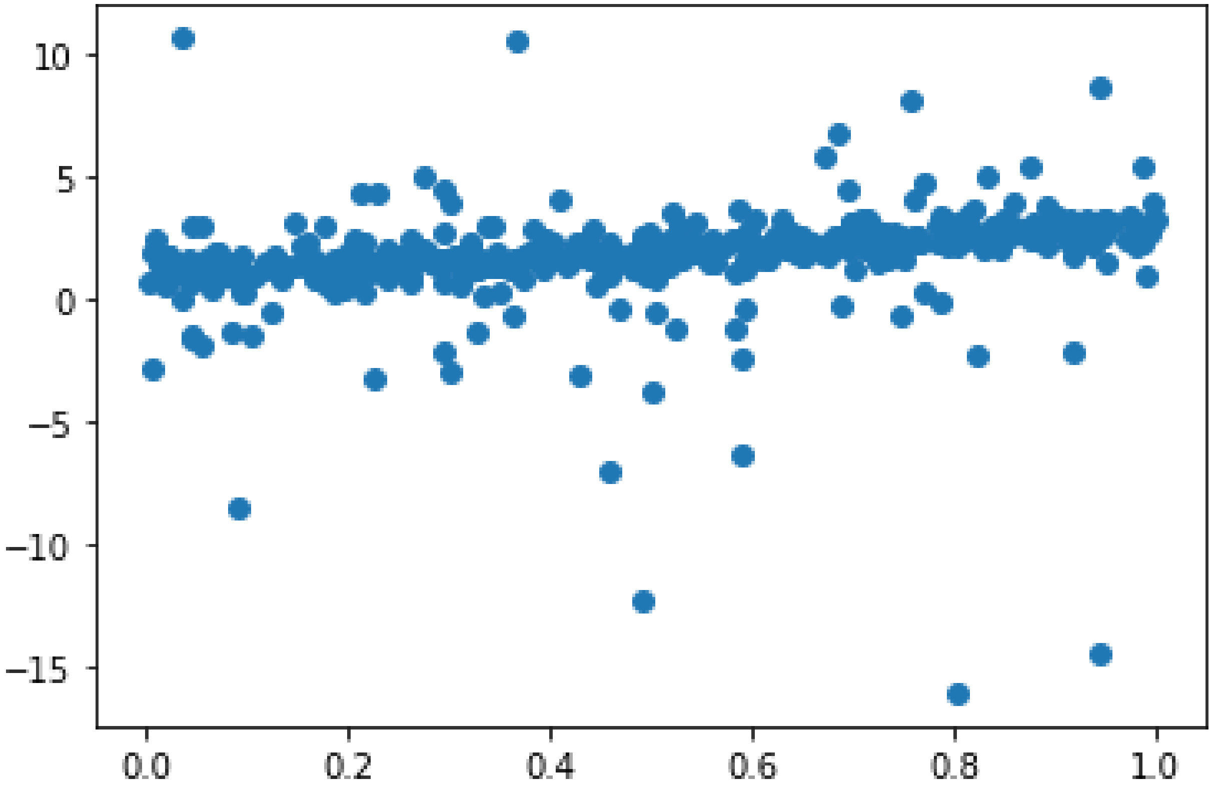
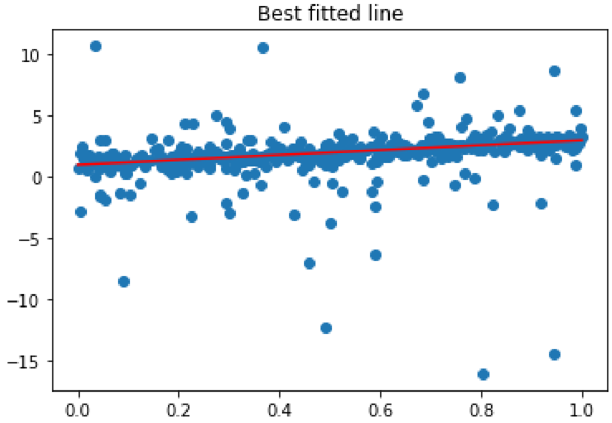


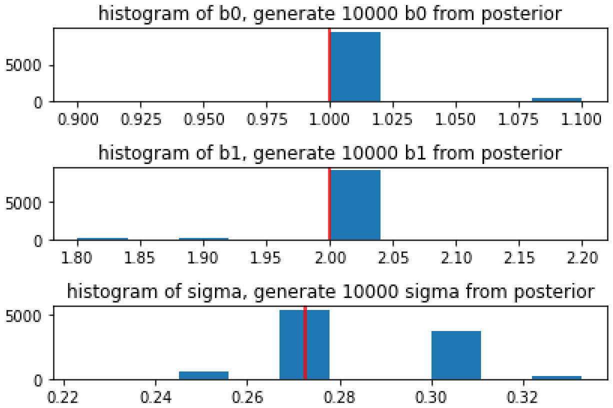
Figure 2: Q2 dataset

2.2 Answer of Q2(b)

The parameters with the highest posterior probabilities are shown below. To get uncertainty quantification, we generate 10000 points from posterior distributions of β_0, β_1, and σ. The histograms and 95% uncertainty intervals are shown below.



(a) Best fitted line



(b) Histograms of parameters

	β_0	β_1	σ
Best parameters	1.00	1.99	0.27

Table 4: Best parameters

	95% interval
β_0	[0.999, 1.099]
β_1	[1.799, 1.999]
σ	[0.246, 0.301]

Table 5: 95% Interval

3 Question3

3. R has an inbuilt dataset called state which gives some data on 50 states in America from the 1970s. You can access this dataset via (see help(state) for information about the data)

data (state); dt = data.frame (state.$x77$, row. names = state.abb)

We want to fit a linear model to this dataset with life expectancy as the response variable and some subset of the remaining seven explanatory variables (including the intercept term). Your goal is to figure out which subset of the explanatory variables provides the best explanation for the response variable in a linear model.

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the $2^7 = 128$ models (obtaining by taking all possible subsets of the explanatory variables along with the intercept term). How many of the 128 models get nontrivial Evidences (after normalization so that the Evidences sum to one)? Describe the models getting high evidences? Do the results of your analysis seem sensible? (6 points)

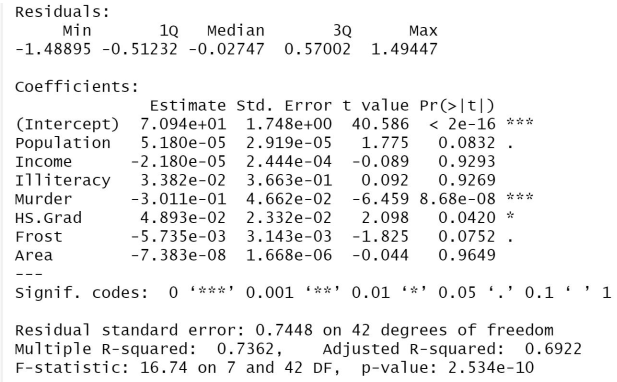
b) Calculate the best model (among the 128 possible models) using the BIC. Compare this model with the models obtaining high evidence from part (a). (3 points)

c) Calculate the best model (among the 128 possible models) using the AIC. Compare this model with the models obtaining high evidence from part (a). (3 points)

3.1 Answer of Q3(a)

Firstly, we use all the variables except Life.Exp to do the regression. The result shows Murder, HS.Grad, Frost, and Population are significant regressors.

Then, we compute evidences of 128 models and select models with highest evidences. The model using Intercept and Murder has the highest evidence. The below table shows the four best models.

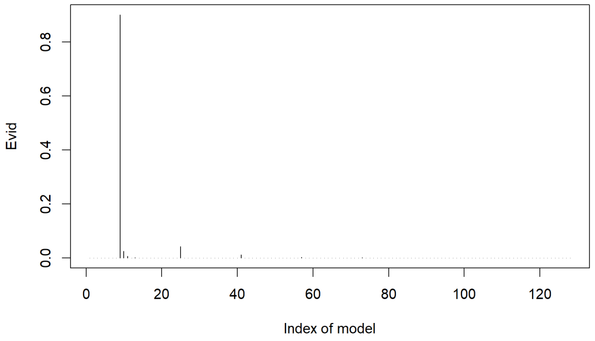


```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.48895 -0.51232 -0.02747  0.57002  1.49447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .  
Income      -2.180e-05  2.444e-04  -0.089   0.9293    
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269    
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *  
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .  
Area        -7.383e-08  1.668e-06  -0.044   0.9649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,	Adjusted R-squared:  0.6922 
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

(a) Regression using all the variables



(b) Evidence values of 128 models

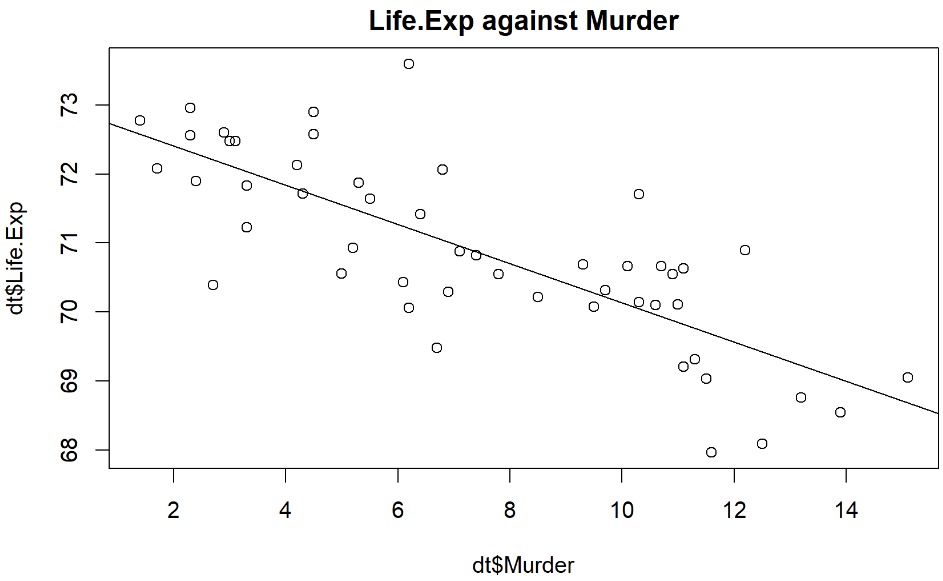


Figure 5: Murder is a great explanatory variable

Evidence	Regressors
0.9013	Intercept, Murder
0.0255	Intercept, Murder, Population
0.0432	Intercept, Murder, HS.Grad
0.0130	Intercept, Murder, Frost

Table 6: Models with highest evidences

Comment: The results shows **Life.Exp** (life expectancy in years (1969–71)) are highly correlated with **Murder** (murder and non-negligent manslaughter rate per 100,000 population (1976)), **Population** (population estimate as of July 1, 1975), **HS.Grad** (percent high-school graduates (1970)), and **Frost** (mean number of days with minimum temperature below freezing (1931–1960) in capital or large city).
It makes sense that higher murder rate leads to lower life expectancy. The higher percent of high-school graduates shows great education resources of the state. We might assume these states are richer and thus have high life expectancy. The higher mean number of days below freezing means more bad weathers, which may lead to more death and accidents.

3.2 Answer of Q3(b)

According to the derivation of AIC and BIC for Linear Model (Lecture Note 10), we have,

$$AIC(M_k) = n + nlog(\frac{2\pi}{n}||Y - X\hat{\beta}||^2) + 2(p+1)$$

$$BIC(M_k) = n + nlog(\frac{2\pi}{n}||Y - X\hat{\beta}||^2) + (logn)(p+1)$$

Using these fomulas, we found best models by AIC are,

AIC	Regressors
117.7196	Intercept, Murder, Population, Income, HS.Grad, Frost
117.7242	Intercept, Murder, Population, Illiteracy, HS.Grad, Frost
117.7309	Intercept, Murder, Population, HS.Grad, Frost, Area

Table 7: Models with lowest AIC

Comment: Compared with models selected by evidence, these models are more complex.

3.3 Answer of Q3(c)

Using the fomulas in Q3(b), we found best models by BIC are,

BIC	Regressors
127.2048	Intercept, Murder, Population, HS.Grad, Frost
127.5344	Intercept, Murder, HS.Grad, Frost
129.5775	Intercept, Population, Murder, HS.Grad

Table 8: Models with lowest BIC

Comment: Compared with models selected by AIC, these models are less complex since BIC uses a more stringent penalty for model complexity.

4 Question4

4. Download the google trends time series dataset for the query yahoo. This should be a monthly time series dataset that indicates the search popularity of this query from January 2004 to August 2022. The goal of this exercise is to use model selection to figure out the best polynomial trend model

$$Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k + Z_i \quad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$$

among the values $k = 1, 2, 3, 4, 5, 6, 7, 8$. To prevent numerical instability issues, take x_i to be some scaled version of time (for example, take $x_i = i/n$).

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the above 8 models. Report the normalized evidences of each of the 8 models. Which model has the highese evidence? Does this model selection method select models that seemings overfit? (6 points)

b) Calculate the best model using the BIC. Compare this model with the models obtaining high evidence from part (a). (3 points) c) Calculate the best model using the AIC. Compare this model with the models obtaining high evidence from part (a). (3 points)

4.1 Answer of Q4(a)

From Lecture 11, we derived that the evident of linear model with normal prior for β and uniform prior for σ is

$$Evi(M) \propto \frac{\Gamma(\frac{p}{x})}{||X\hat{\beta}||^p} \frac{\Gamma(\frac{n-p-1}{2})}{||Y - X\hat{\beta}||^{n-p}}$$

Due to this formula, we compute evidences for the above 8 models are

k	1	2	3	4	5	6	7	8
Evi	1.77e-172	1.15e-126	9.6e-68	2.99e-46	2.3e-27	5.8e-04	5.8e-03	9.9e-01

Table 9: Evidences

Therefore, the eighth model is chosen. It seems over-fit since the polynomial degree is too large. Essentially, it's not necessary to use such a complex model to capture the trend. The reason might be evidence using Zellner prior does not punish complex model as much as evidence using uniform prior.

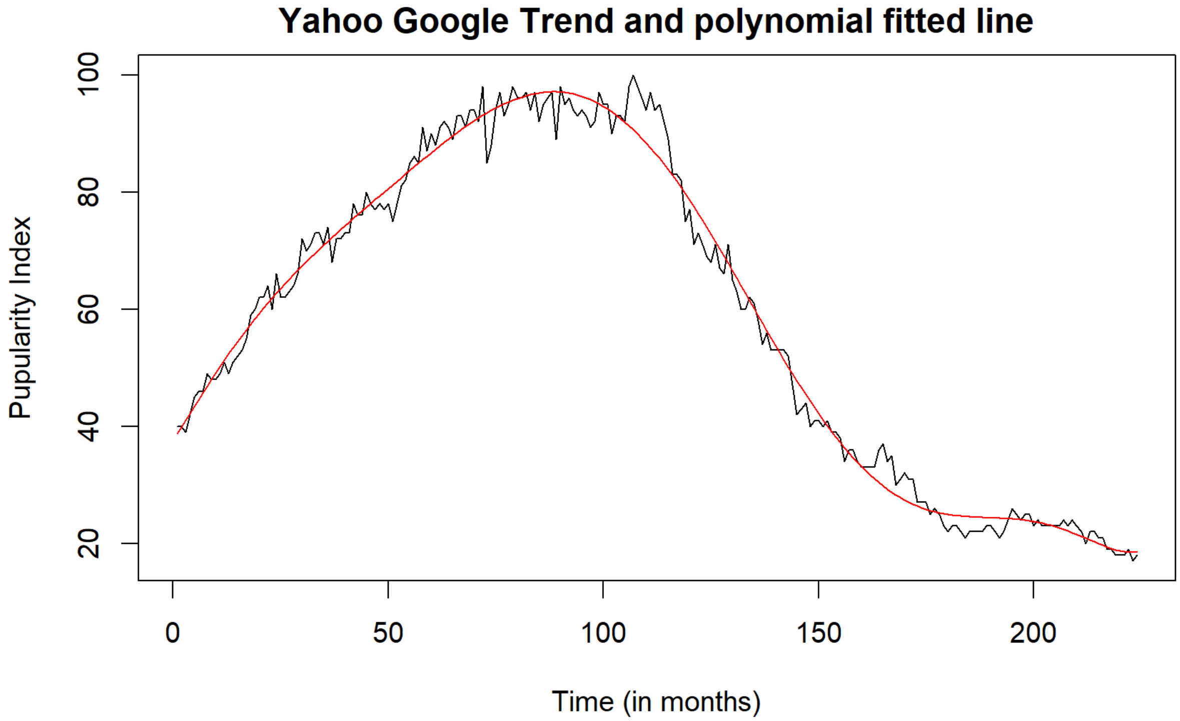


Figure 6: Fitted line

4.2 Answer of Q4(b)

Similarly to Q3, We compute AIC and the eighth model is selected. Here AIC gets the same result as evidence.

4.3 Answer of Q4(c)

Similarly to Q3, We compute BIC and the eighth model is selected. The result of BIC is the same as selection by evidence, which because BIC is an approximation of evidence.

5 Question5

5. Download the FRED dataset on ”Retail Sales: Beer, Wine, and Liquor Stores” from https://fred.stlouisfed.org/series/MrTSSM4453USN. This is a monthly dataset (the units are millions of dollars) and is not seasonally adjusted. To this data, we want to fit one of the models $M_{k,l}$ where k ranges in $0,1,2,3,4,5$ and l ranges in $0,1,2,3,4,5$. The model M_{kl} is given by:

$Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j \left(\frac{t}{n}\right)^j + \sum_{u=1}^{l} \left\{\alpha_{1u} \cos\left(\frac{2\pi ut}{12}\right) + \alpha_{2u} \sin\left(\frac{2\pi ut}{12}\right)\right\} + Z_t$ with $Z_t \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$

for $t = 1, \ldots, n$. Note that $k = 0$ means that the term $\sum_{j=1}^{k} \beta_j \left(\frac{t}{n}\right)^j$ is just missing (similarly for $l = 0$). For example, M_{00} corresponds to just fitting the constant β_0. The total number of models considered is $6^2 = 36$ corresponding to all pairs of choices $(k,l) \in \{0,1,2,3,4,5\} \times \{0,1,2,3,4,5\}$.

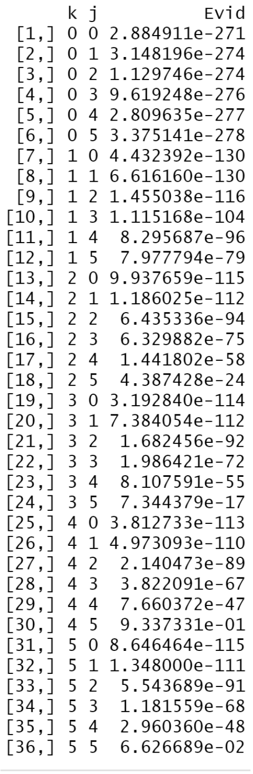
a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the above 36 models. Report the normalized evidences of each of the 36 models. Which models have high evidence? Does this model selection method favors models that seemingly overfit? (6 points)

b) Calculate the best model using the BIC. Compare this model with the models obtaining high evidence from part (a). (3 points)

c) Calculate the best model using the AIC. Compare this model with the models obtaining high evidence from part (a). (3 points)

5.1 Answer of Q5(a)

By the similar method in previous questions, the normalized evidences are shown below,



```
       k j          Evid
 [1,] 0 0 2.884911e-271
 [2,] 0 1 3.148196e-274
 [3,] 0 2 1.129746e-274
 [4,] 0 3 9.619248e-276
 [5,] 0 4 2.809635e-277
 [6,] 0 5 3.375141e-278
 [7,] 1 0 4.432392e-130
 [8,] 1 1 6.616160e-130
 [9,] 1 2 1.455038e-116
[10,] 1 3 1.115168e-104
[11,] 1 4  8.295687e-96
[12,] 1 5  7.977794e-79
[13,] 2 0 9.937659e-115
[14,] 2 1 1.186025e-112
[15,] 2 2  6.435336e-94
[16,] 2 3  6.329882e-75
[17,] 2 4  1.441802e-58
[18,] 2 5  4.387428e-24
[19,] 3 0 3.192840e-114
[20,] 3 1 7.384054e-112
[21,] 3 2  1.682456e-92
[22,] 3 3  1.986421e-72
[23,] 3 4  8.107591e-55
[24,] 3 5  7.344379e-17
[25,] 4 0 3.812733e-113
[26,] 4 1 4.973093e-110
[27,] 4 2  2.140473e-89
[28,] 4 3  3.822091e-67
[29,] 4 4  7.660372e-47
[30,] 4 5  9.337331e-01
[31,] 5 0 8.646464e-115
[32,] 5 1 1.348000e-111
[33,] 5 2  5.543689e-91
[34,] 5 3  1.181559e-68
[35,] 5 4  2.960360e-48
[36,] 5 5  6.626689e-02
```

Figure 7: Normalized Evidence

The model (k=4, j=5) has the highest evidence and the model (k=5, j=5) has the second highest evidence. Although the line of model (k=4, j=5) fits the data well, it seems to be over-fitted since it's very complex.

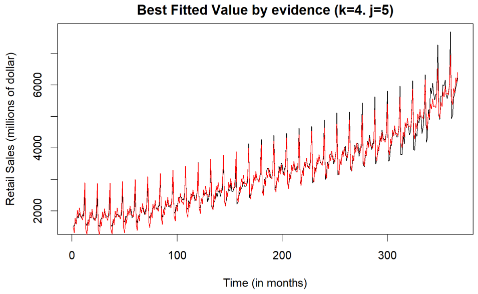


Figure 8: Fitted line with the highest evidence

5.2 Answer of Q5(b)

The model selected by BIC is the same as evidence, which because BIC is an approximation of evidence.

5.3 Answer of Q5(c)

Note that AIC does not punish model complexity as heacy as BIC, so it tends to choose complex model. Here the model (k=5, j=5) is selected by AIC, which is the most complex model among all.

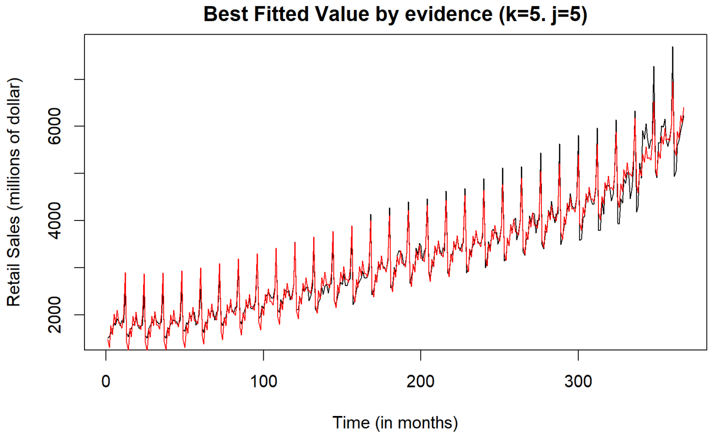


Figure 9: Fitted line with the highest AIC

6 Question6

6. A classic time series dataset is the Box and Jenkins airline passenger data (can be accessed in R via data(AirPassengers)). This gives monthly totals of international airline passengers from 1949 to 1960 . There are $n = 144$ observations in total corresponding to 12 years. To this dataset, consider fitting the models:

$$Y_t = a + bt + \sum_{f \in S} \left[\beta_{1f} \cos(2\pi f t) + \beta_{2f} \sin(2\pi f t)\right] + Z_t \quad \text{ with } Z_t \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$$

for some subset

$$S \subseteq \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{18}{n}\right\}$$

There are 2^{18} models here in total and we shall denote them by M_S as S ranges over all subsets of $\{1/n, \ldots, 18/n\}$. In addition to these 2^{18} models, also consider the models:

$$\log Y_t = a + bt + \sum_{f \in S} \left[\beta_{1f} \cos(2\pi f t) + \beta_{2f} \sin(2\pi f t)\right] + Z_t \quad \text{ with } Z_t \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$$

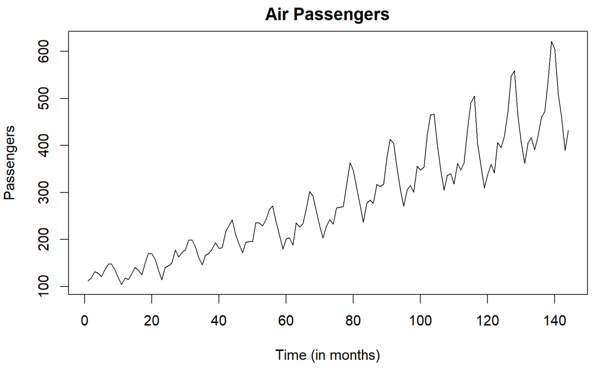
again as S ranges over all subsets of $\{1/n, \ldots, 18/n\}$ (note that the response variable above is $\log Y_t$ as opposed to Y_t). Let us denote these models by LM_S (L standing for logarithm). Consider all these $2^{18} + 2^{18} = 2^{19}$ models together.

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the above 2^{19} models. Which models have high evidences? Do the frequencies k/n appearing in the high evidence models have any intuitive meaning? (8 points)

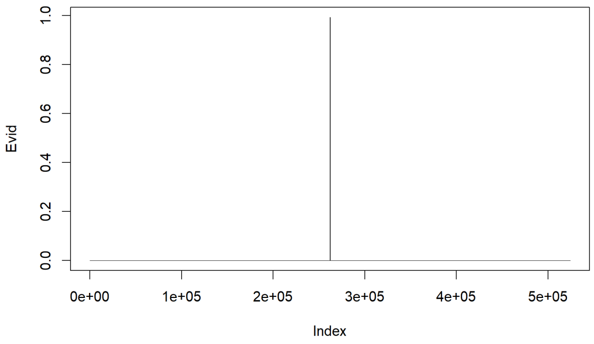
b) For fixed S, which of the two models M_S and LM_S generally has higher evidence? (3 points)

6.1 Answer of Q6(a)

By the similar method in previous questions, the models with highest evidences are $logY_i$ with frequencies $\frac{12}{n}$. It makes sense since it shows the yearly periodicity of data, which is obvious for the raw data.



(a) figure1



(b) figure2

6.2 Answer of Q6(b)

The result shows LM_s generally has higher evidence for fixed S. Note that in our model, we use a straight line $a + bt$ to describe the overall increasing trend. Log data seems to be more approperaite to be fitted by a straight line. Therefore, LM_s has higher evidence.