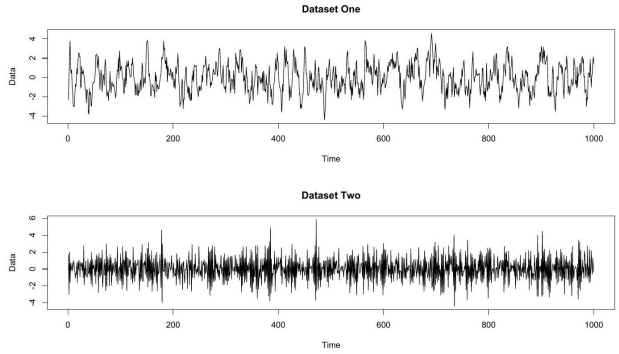
Topics in Time Series (1):

Yisen Du

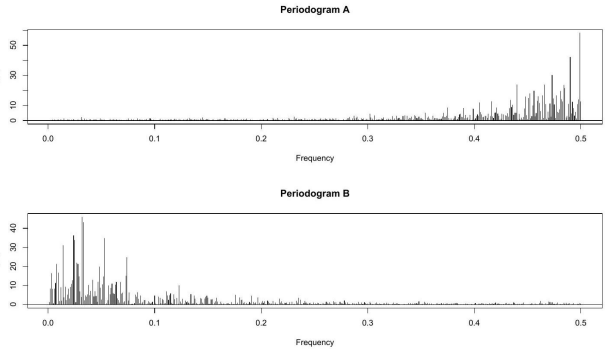
1 Problem1

1.1 Question

In Figure 1, you will find two different time series datasets (Dataset One and Dataset Two) of the same length $n = 1000$. In Figure 2, you will find two periodograms (Periodogram A and Periodogram B). One of these periodograms corresponds to Dataset One and the other to Dataset Two. Identify the correct periodograms giving reasons for your answer. (5 points)



(a) Two Time Series Datasets



(b) Two Periodograms

1.2 Answer

Note that the Dataset one looks sparser than Dataset two. Since $T = 1/f$, the larger the f the smaller the period. Figure 2 shows periodogram A has more spikes with larger frequencies, which means smaller periods are shown in time series data. Therefore, periodogram A is corresponded to Dataset two and periodogram B is corresponded to Dataset one.

2 Problem 2

2.1 Question

Consider the dataset lynx that is available in base R. This gives the annual numbers of lynx trappings for 1821-1934 in Canada. Type help(lynx) to learn more about the dataset.

a) Plot the periodogram of the data and comment on its notable features. points).

b) To this dataset, fit the model:

$$Y_t = \beta_0 + \beta_1 \cos(2\pi f t) + \beta_2 \sin(2\pi f t) + Z_t \quad \text{with } Z_t \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right) \tag{1}$$

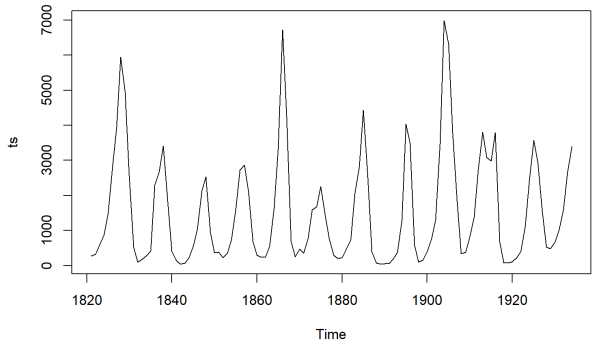
c) Provide point estimates and 95% marginal uncertainty intervals for $\beta_0, \beta_1, \beta_2$ and σ(6 points).

d) Comment on whether (1) is a good model for this dataset. (2 points)

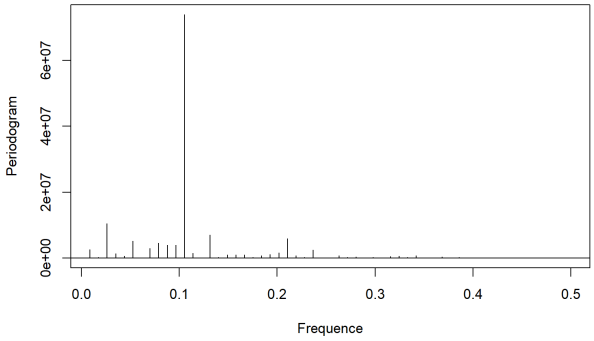
2.2 Answer

2.2.1 (a)

The periodogram shows there's strong spike at $f = 0.11$, which means the time series data has the period $T \approx 9.1$ (years). From the time series data, it seems that it's periodic with $T \approx 9.1$(years).



(a) Time Series



(b) Periodogram

2.2.2 (b)

Following the method mentioned in Lecture4, we can computer the posterior of f and find the point estimate and uncertainty interval. The point estimate is 0.104 and the 95% interval is $[0.103, 0.105]$.

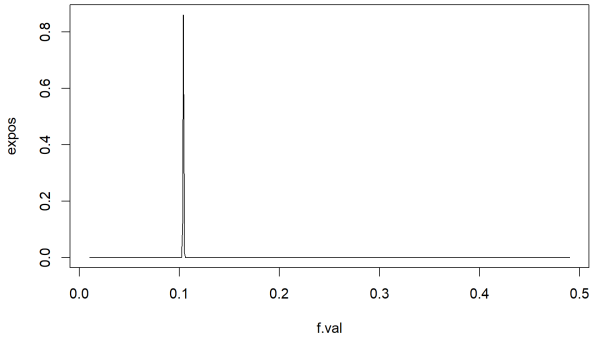
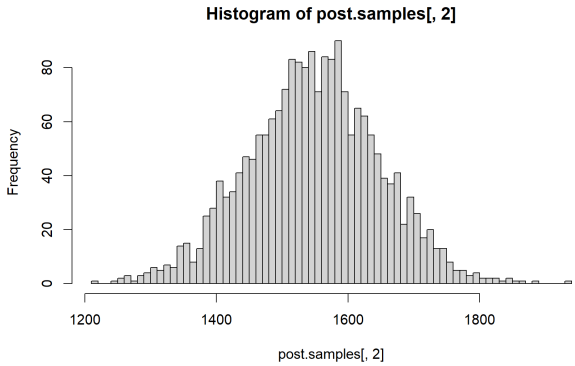


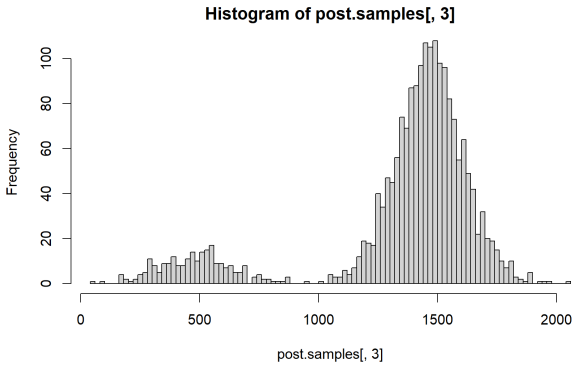
Figure 3: Posterior of f

2.2.3 (c)

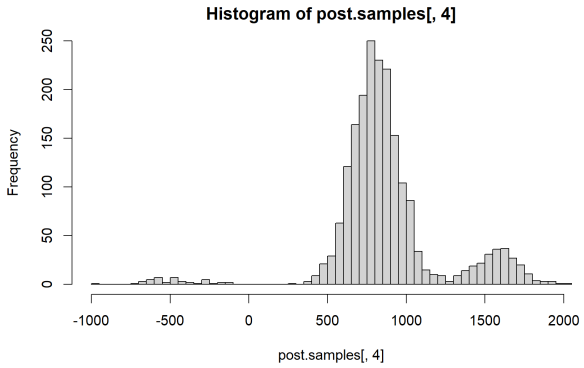
Conditional on a fixed f, we know β follows t-distribution and SSE/σ^2 follows a chi-square distribution. Therefore, we could generate samples of these parameters. According to the histograms, we could derive the point estimates and marginal uncertainty intervals.



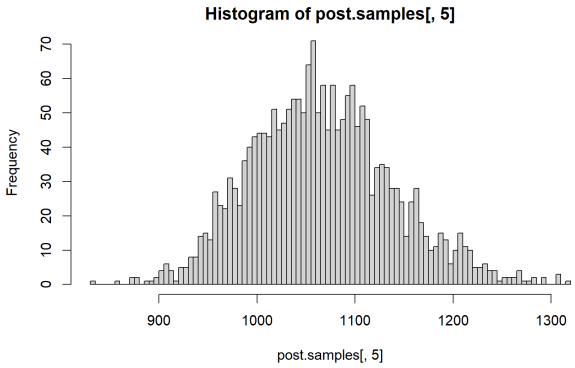
(a) Histogram of β_0



(b) Histogram of β_1



(a) Histogram of β_2



(b) Histogram of σ

	Point estimate	95% Uncertainty Interval
β_0	1547.4	[1347, 1740]
β_1	1349.97	[353.86, 1748.07]
β_2	869.11	[425.39, 1680.32]
σ	1065.88	[938.92, 1216.11]

Table 1: Uncertainty intervals

2.2.4 (d)

Basically, (1) is a good model since it captures the perodicity of the data. From the periodogram, we can clearly find a period of data. If we randomly choose some samples of estimated parameters and plot them, we can find the the fitted data captures the periodicity greatly.
However, this model is not perfect because it cannot capture the peaks in the data. It seems like there's a strong peak of data every four periods.

3 Problem 3

3.1 Question

3. Download the Google Trends Data (for the United States) for the query mask. This should be a monthly time series dataset that indicates the search popularity of this query from January 2004 to September 2022. To this data, fit the single change point model:

$$Y_t = \beta_0 + \beta_1 I\{t > c\} + Z_t \quad \text{with } Z_t \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right). \tag{2}$$

a) Provide a point estimate and 95% uncertainty interval for the changepoint parameter c. Explain whether your answers make intuitive sense in the context of this dataset (4 points).

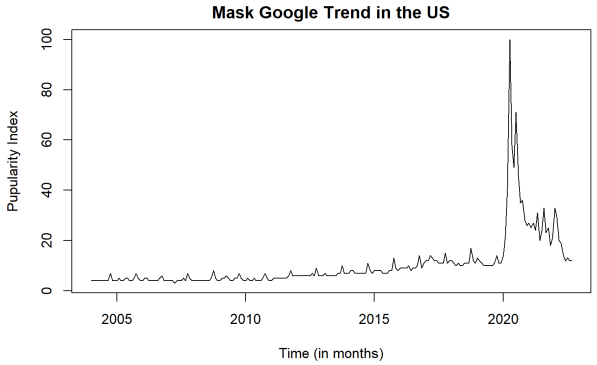
b) Provide point estimates and 95% marginal uncertainty intervals for the prechangepoint mean level $\mu_0 := \beta_0$ and the post-changepoint mean level $\mu_1 := \beta_0 + \beta_1$. (4 points).

c) Comment on whether (2) is a good model for this dataset. (2 points)

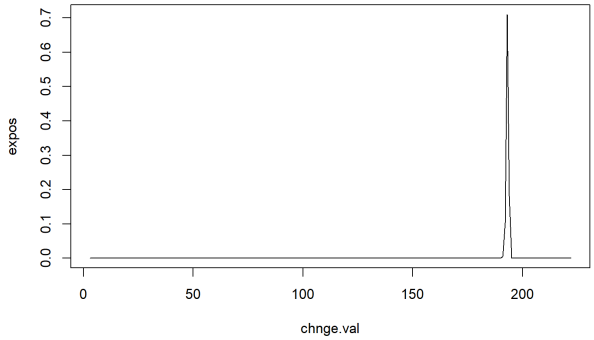
3.2 Answer

3.2.1 (a)

The point estimate of c is (1, 2020). The 95% interval is $[(12, 2019), (2, 2020)]$. Note that the probability of this interval is higher than 95% since the posterior of c is concentrated.
It makes sense since the pandemic was happened in 12,2019. After that, people began to wear mask so the search of 'mask' increased a lot since 1, 2020.



(a) Time Series



(b) Posterior of c

3.2.2 (b)

The point estimate of β_0 is 6.95, and the 95% interval is $[5.90, 8.03]$. The point estimate of β_1 is 30.47 and the 95% interval is $[27.84, 33.07]$.
To compute the estimate of $\mu_1 = \beta_0 + \beta_1$, we should sum the samples of β_0 and β_1. We find the mean of μ_1 is 37.35 and the 95% interval is $[34.50, 40.09]$.

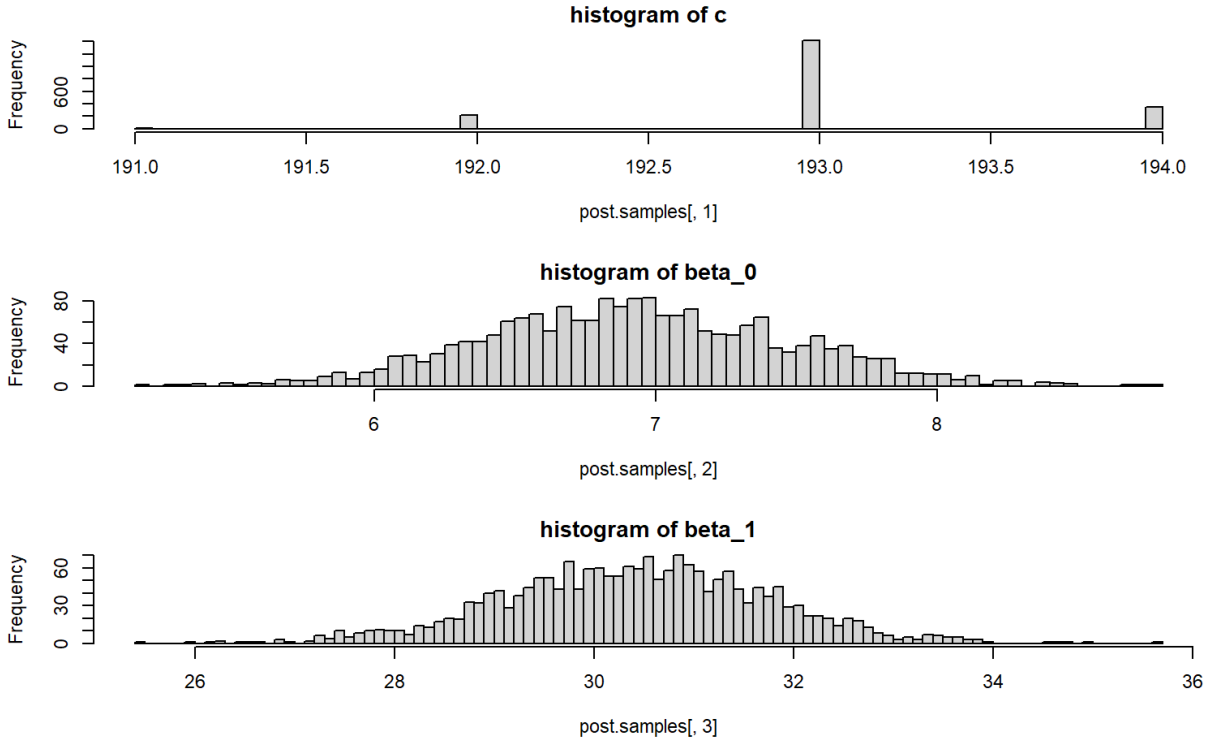


Figure 7: Histograms of estimated parameters

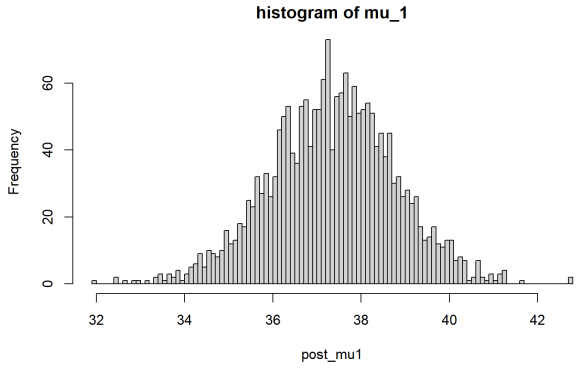


Figure 8: Histogram of μ_1

3.2.3 (c)

The model is great in the sense that it captures the change point precisely. Before the pandemic, most people don't care masks. After the pandemic, people pay more attention on it.
The model is insufficient in the sense that it does not capture **the decay of trend** after the pandemic. We could use the linear model or exponential model to depict the post-pandemic trend.

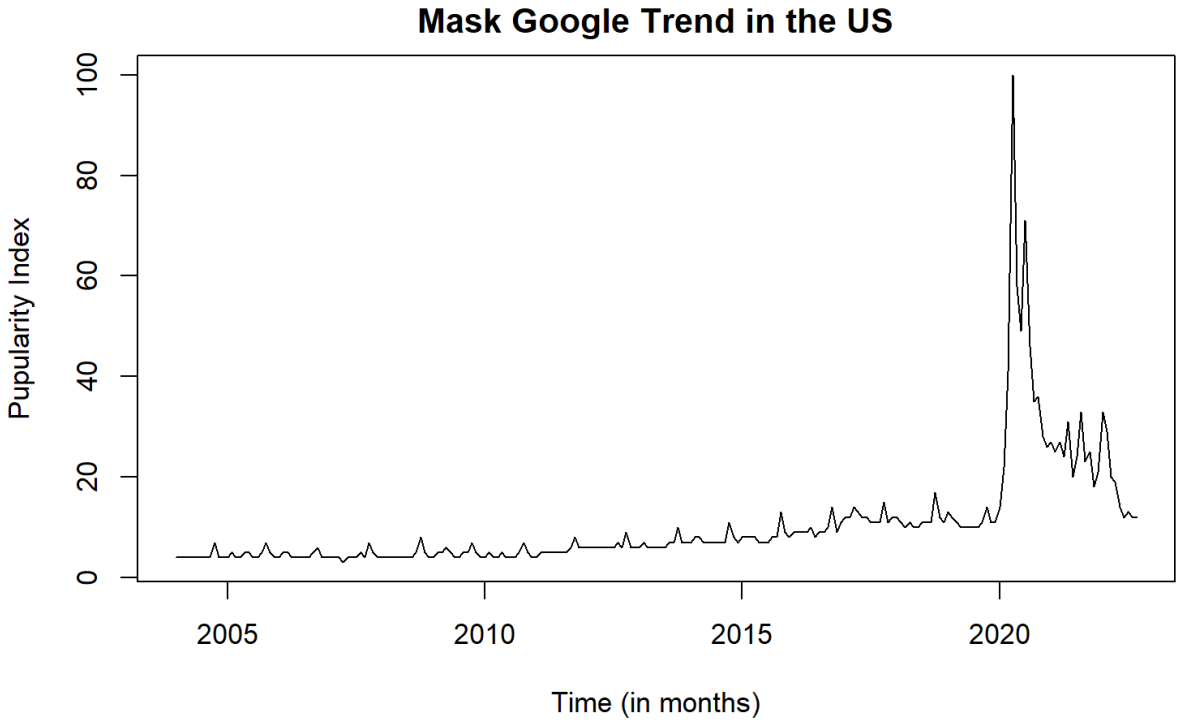


Figure 9: The decay of the trend

4 Problem 4

4.1 Question

4. Download the Google Trends Data (for the United States) for the query golf. This should be a monthly time series dataset that indicates the search popularity of this query from January 2004 to September 2022. To this data, fit the model:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \cos(2\pi f t) + \beta_4 \sin(2\pi f t) + Z_t \quad \text{with } Z_t \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right). \tag{3}$$

a) Provide a point estimate and a 95% uncertainty interval for the unknown frequency parameter f. (4 points) b) On a scatter plot of the data, plot your best estimate of the fitted function:

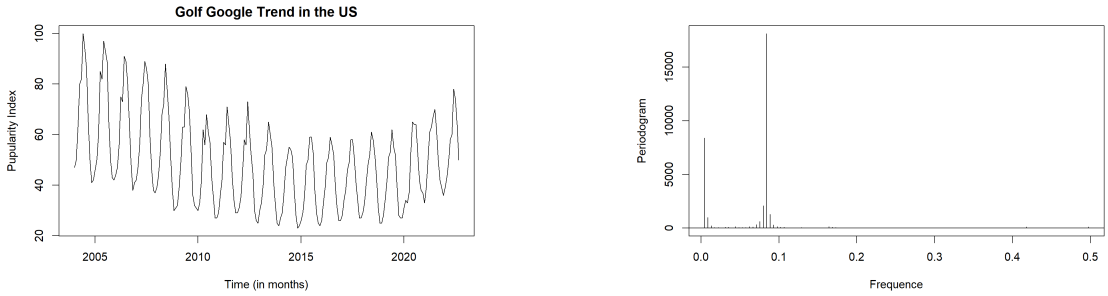
$$t \mapsto \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \cos(2\pi f t) + \beta_4 \sin(2\pi f t)$$

along with appropriate uncertainty quantification. (5 points).
c) Comment on whether model (3) is appropriate for this dataset. (2 points).

4.2 Answer

4.2.1 (a)

The point estimate of f is 0.082. The 95% interval is $[0.082, 0.084]$. Note that the probability of this interval is higher than 95% since the posterior of f is concentrated.



(a) Time Series

(b) Periodogram

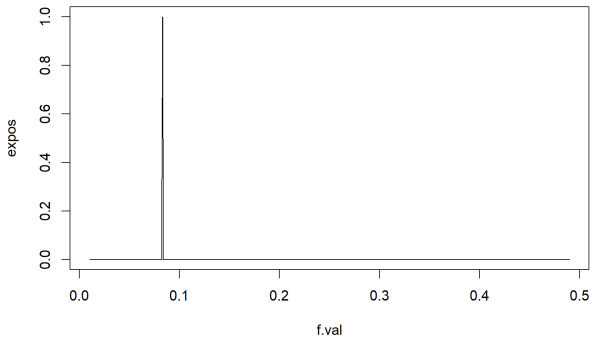


Figure 11: Posterior of f

4.2.2 (b)

To plot the best estimate, we firstly generate samples from posterior distributions of parameters. Then, we regard the means as the best point estimate. The fitted function is plotted in figure 9. The uncertainty quantification is 95% uncertainty interval of each parameters.

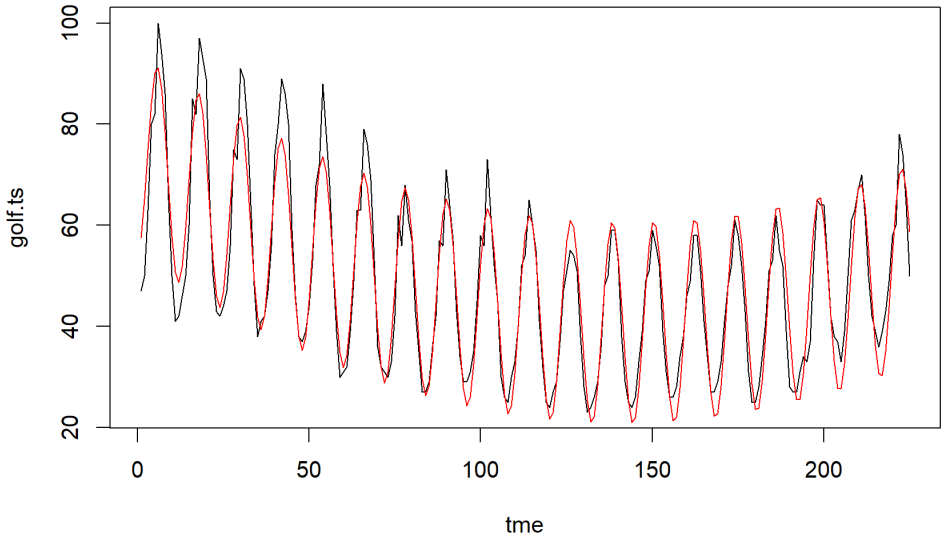


Figure 12: Best estimate of the fitted function

	Point estimate	95% Uncertainty Interval
f	0.083	0.083
β_0	74.02	[71.87574, 76.18146]
β_1	-0.47	[-0.5117891, -0.4246094]
β_2	0.0016	[0.001455316, 0.001832519]
β_3	-19.87	[-20.86085, -18.90471]
β_4	2.50	[1.492784, 3.533789]

Table 2: Uncertainty quantification

4.2.3 (c)

From figure 9, it seems that the model is appropriate for this dataset. It not only captures the periodicity of the data by using a sinusoidal function, but depicts the long-term trend by adopting parabola. The long-term trend is decreased at the first half and increased at the second half. From c), we found the point estimate of β_2 is positive. It coincides with our observation.

5 Problem 5

5.1 Question

5. Download the FRED dataset on Total Construction Spending in the United States from https://fred.stlouisfed.org/series/T
This gives monthly seasonally adjusted data on total construction spending in the United States in millions of dollars from January 1993 to July 2022. To this dataset, fit the model:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 (t - s_1)_+ + \beta_3 (t - s_2)_+ + Z_t$$

with $Z_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. The unknown parameters in this model are $\beta_0, \beta_1, \beta_2, \beta_3, s_1, s_2, \sigma$.

a) Provide point estimates and 95% uncertainty intervals for the change of slope parameters s_1 and s_2. (5 points)

b) On a scatter plot of the data, plot your best estimate of the fitted function:

$$t \mapsto \beta_0 + \beta_1 t + \beta_2 (t - s_1)_+ + \beta_3 (t - s_2)_+$$

along with appropriate uncertainty quantification. (5 points).

c) Comment on whether model (4) is appropriate for this dataset. (2 points).

5.2 Answer

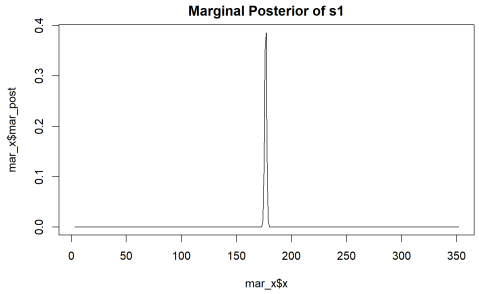
5.2.1 (a)

Similar to previous questions, it's a non-linear model with parameters β, $\mathbf{s}$, and σ. We could write the model in vector form:

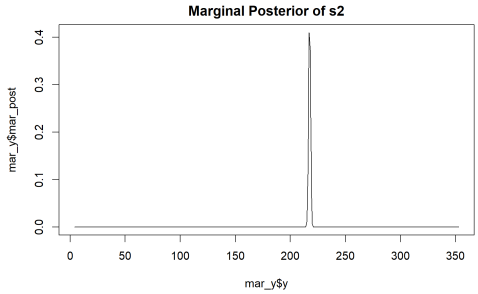
$$\mathbf{y} = X * \beta + \mathbf{z},$$

$$\text{where } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \mathbf{s} = \begin{bmatrix} s_0 \\ s_1 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & t_1 & (t_1 - s_1)_+ & (t_1 - s_2)_+ \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & (t_n - s_1)_+ & (t_n - s_2)_+ \end{bmatrix}$$

Firstly, we could get the joint posterior of $s1$ and $s2$. From the joint posterior, we could compute the marginal posteriors of them, respectively. The following figures and tables show the marginal posteriors and 95% intervals. From figure, we can conclude that these point estimates make sense.



(a) Marginal Posterior of s_1



(b) Marginal Posterior of s_2

	Point estimate	95% Uncertainty Interval
s_1	177	[175, 178]
s_2	217	[216, 219]

Table 3: Uncertainty quantification

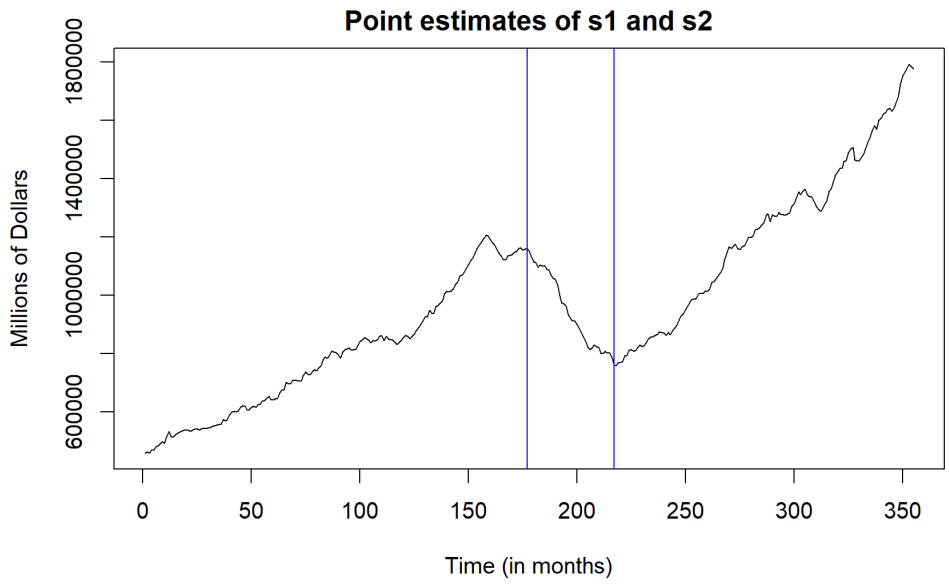
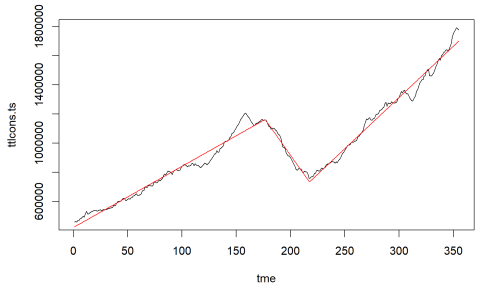
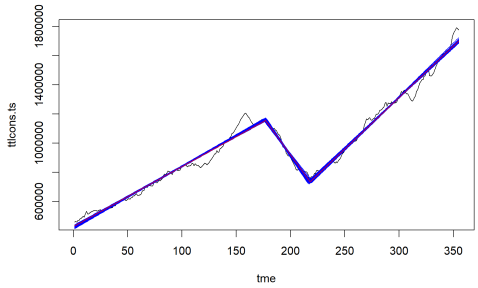


Figure 14: Point estimates of s_1 and s_2

5.2.2 (b)



(a) Best estimate of the fitted function



(b) Uncertainty Interval

	Point estimate	95% Uncertainty Interval
β_0	422277.6	[410797.2, 433513.6]
β_1	4198.07	[4091.47, 4308.35]
β_2	-14707	[-15794, -13679]
β_3	17538.9	[16519, 18625]

Table 4: Uncertainty quantification

5.2.3 (c)

I think this model is appropriate since it captures the three main trends from the data. Also, each part could be fitted by linear model.

6 Problem 6

6.1 Question

6. Download the google trends time series dataset for the query playoffs (download the trends for the United States and not worldwide). This should be a monthly time series dataset that indicates the search popularity of this query from January 2004 to September 2022.

a) Plot the data and observe that the scale of variability increases with time. To fix this, take the logarithm of the data. Plot the logarithm and comment on whether the variability can now be assumed to be constant over time. (2 points)

b) The logarithmed data have an increasing trend and a periodic component superimposed on the trend. To get an idea of the frequencies present in the periodic component, fit a linear trend model to the logarithmed data, compute the residuals and then plot the periodogram of the residuals. Comment on the location and size of the main spikes in the periodogram. (3 points)

c) To the logarithmed data, fit the model

$$y_i = g(t_i) + Z_i \quad \text{where } Z_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$$

and

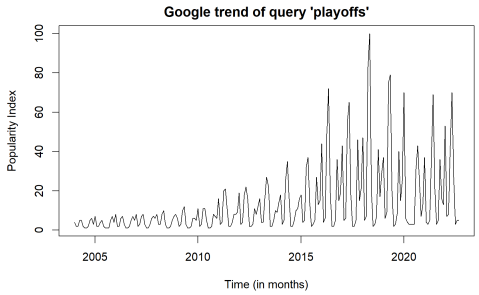
$$\begin{aligned} g(t) = \beta_0 + \beta_1 t \\ + \beta_2 \cos(2\pi f_1 t) + \beta_3 \sin(2\pi f_1 t) \\ + \beta_4 \cos(2\pi f_2 t) + \beta_5 \sin(2\pi f_2 t) \\ + \beta_6 \cos(2\pi f_3 t) + \beta_7 \sin(2\pi f_3 t). \end{aligned}$$

Treat $\beta_0, \beta_1, \ldots, \beta_7, f_1, f_2, f_3, \sigma$ as unknown parameters and estimate them from data. Plot your estimate of g (and appropriate uncertainty indicators) along with the actual data. (6 points)

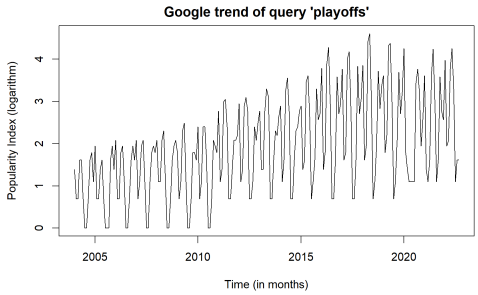
6.2 Answer

6.2.1 (a)

After the logarithmic transformation, the variability can be considered to be constant according to the figure (b).



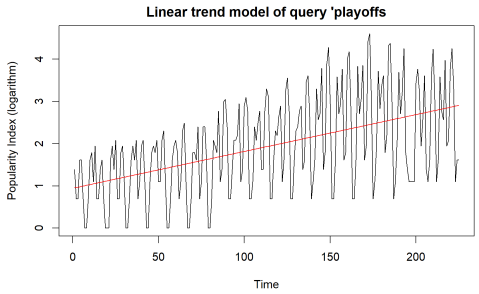
(a) Original Data



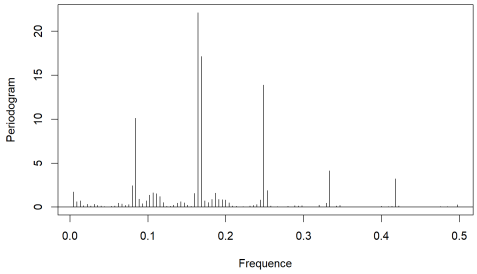
(b) Logarithm of the data

6.2.2 (b)

The linear trend model captures the long-run increasing trend of the data. Using the residuals of linear model, we could filter out the linear trend and observe pure oscillation of the data. The periodogram shows there are three main spikes, which lie around frequencies 0.08, 0.16, 0.25. These frequencies represent the periods of 12 months, 6 months, and 4 months. We could guess that they are caused by playoffs of NBA and other popular sports.



(a) Linear Trend Model



(b) Periodogram of the residuals

frequency	Amplitude
0.164444444	2.212248e+01
0.168888889	1.714009e+01
0.248888889	1.388801e+01
0.084444444	1.013098e+01
0.333333333	4.143273e+00

Table 5: Periodogram Table

6.2.3 (c)

From the periodogram, we could find there are three main spikes, which lie around frequencies 0.16, 0.25, and 0.08. So here we consider the f_1, f_2, and f_3 are around 0.08, 0.16, and 0.25. Using the grid search to construct samples of $\mathbf{f}$, we can compute the posterior of $\mathbf{f}$ by the similar method in Q5. After that, we sample these parameters and find the point estimates of each parameters. In the figure 12, the black line is the actual data, the red line is the point estimate of g, and the blue area is the uncertainty interval generated from posterior distributions of parameters.

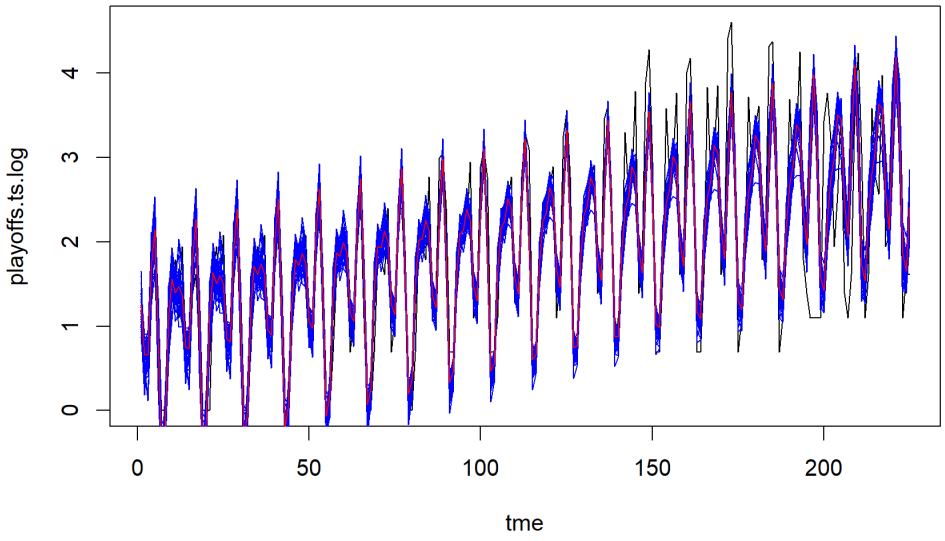


Figure 18: Estimate of g and the actual data

	Point estimate	95% Uncertainty Interval
f_1	0.083	0.083
f_2	0.166	0.166
f_3	0.250	0.250
β_0	0.935	[0.772, 1.093]
β_1	0.0088	[0.0076, 0.0101]
β_2	0.3201	[0.054, 0.460]
β_3	0.3204	[0.202, 0.494]
β_4	0.1077	[-0.083, 0.63]
β_5	-0.8096	[-0.953, -0.558]
β_6	-0.0246	[-0.141, 0.121]
β_7	0.5115	[0.379, 0.626]

Table 6: Uncertainty quantification