STAT 153 & 248 - Time Series
Homework One

Spring 2025, UC Berkeley

Due by 11:59 pm on 10 February 2025

Total Points = 70 (STAT 153) and 85 (STAT 248)

For problems involving data analysis, you are free to use inbuilt functions from libraries in R or Python. Attach code snippets in your solutions corresponding to each problem.

1. Download the dataset on the annual size of the Resident Population of California from `https://fred.stlouisfed.org/series/CAPOP`. This dataset gives the annual population of California from 1900 to 2024 (units are in thousands of persons). Drop the data point for 2024 and only consider the data 1900 to 2023.

   a) To the observed data (from 1900 to 2023), fit the model:

   $$y_t = \beta_0 + \beta_1 t + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{1}$$

   Provide point estimates for β_0, β_1 along with appropriate uncertainty intervals. Interpret your point estimates and explain why they make sense. (**4 points**)

   b) Along with a plot of the observed dataset, plot lines corresponding to 100 samples from the posterior distribution of (β_0, β_1) (under the prior $\beta_0, \beta_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$ for a very large C). Comment on the range of uncertainty revealed in this plot. (**4 points**)

   c) Based on model (1), provide a point estimate along with appropriate uncertainty quantification for the Resident Annual Population of California for the year 2024. (**2 points**)

   d) To the observed data (from 1900 to 2023), fit the model:

   $$\log y_t = \beta_0 + \beta_1 t + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{2}$$

   Provide point estimates for β_0, β_1 along with appropriate uncertainty intervals. Interpret your point estimates and explain why they make sense. (**4 points**)

   e) For the model (2) (and under the prior $\beta_0, \beta_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$ for a very large C), calculate the posterior probability that β_1 is larger than 0.03. (**2 points**)

   f) Based on model (2), provide a point estimate along with appropriate uncertainty quantification for the Resident Annual Population of California for the year 2024. (**2 points**)

   g) To the observed data (from 1900 to 2023), fit the model:

   $$\log y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{3}$$

Based on the fitted model, provide a point estimate along with appropriate uncertainty quantification for the Resident Annual Population of California for the year 2024. (**2 points**)

h) Which of the three models (1), (2), (3) would you recommend for this dataset and why? Also report which of your predictions for the 2024 population is closest to the actual observed value. (**3 points**)

2. Download the FRED dataset from `https://fred.stlouisfed.org/series/MRTSSM4453USN` which gives monthly data on beer, wine and liquor sales in the United States. Let y_t denote the observed data value for time t.

a) Is the linear model:

$$\log y_t = \beta_0 + \beta_1 t + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{4}$$

appropriate for this dataset? Why or why not? (**2 points**)

b) Let t_0 be the time point which corresponds to April 2020. Fit the model:

$$\log y_t = \beta_0 + \beta_1 t + \beta_2 I\{t \geq t_0\} + \beta_3 t I\{t \geq t_0\} + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{5}$$

to this dataset. Here $I\{t \geq t_0\}$ denotes the variable which takes the value 0 for $t < t_0$ and 1 for $t \geq t_0$; and $tI\{t \geq t_0\}$ denotes the variable which takes the value 0 for $t < t_0$ and t for $t \geq t_0$. On a single figure, plot the fitted values of models (4) and (5) along with the original data. Based on this plot, is (5) a better model for this dataset compared to (4)? Why or why not? (**4 points**)

c) In model (5) (along with the prior $\beta_0, \beta_1, \beta_2, \beta_3, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$ for a very large C) , calculate the posterior probability that $\beta_3 < 0$. Also calculate the posterior probability that $\beta_2 + \beta_3 t_0 > 0$. (**6 points**)

d) For an appropriate $k \leq 4$, fit the model:

$$\begin{aligned} y_t = {} & \beta_0 + \beta_1 t + \beta_2 I\{t \geq t_0\} + \beta_3 t I\{t \geq t_0\} \\ & + \sum_{i=1}^{k} \left(a_i \cos(2\pi \frac{it}{12}) + b_i \sin(2\pi \frac{it}{12}) \right) + \epsilon_t \end{aligned} \tag{6}$$

to the data. Give reasons for your choice of k. (**4 points**)

e) Using model (6) (with your chosen k), predict the sales data for the next 36 future months. Plot the predictions along with the original data and explain why your forecast is reasonable. (**4 points**).

3. 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 January 2025. The goal of this exercise is to fit the polynomial trend model

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_k t^k + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

to this dataset for an appropriate value of $k \leq 5$.

a) Visually evaluate the fit of the least squares estimate for this model to the observed data to pick an appropriate value of $k \leq 5$. Explain the reason for your choice of k. (**5 points**).

b) On a plot of the observed dataset, plot the polynomial corresponding to the least squares estimate for the model with your chosen value of k. On the same figure, plot polynomials corresponding to 300 samples from the posterior distribution of the coefficients. Comment on the range of uncertainty revealed in this plot. (**5 points**)

4. Consider the multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon_i \qquad \text{with } \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{7}$$

We observe data $(x_{i1}, \ldots, x_{ip}, y_i), i = 1, \ldots, n$. In the model above, the covariates x_{ij} are treated as fixed constants. For Bayesian analysis with the prior,

$$\beta_0, \beta_1, \ldots, \beta_m, \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C).$$

we showed (in Lecture 4) that when $C \to \infty$,

$$\beta \mid \text{data} \sim t_{n-m-1,m+1}\left(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}\right) \tag{8}$$

where $X, \beta, \hat{\beta}, \hat{\sigma}$ are all as defined in Lecture 4.

This exercise provides another slightly different proof of this same result. Along the way, it also discusses inference for the parameter σ. Throughout, assume that $C \to \infty$ (i.e., C is very large).

a) Prove that

$$\beta \mid \text{data}, \sigma \sim N_{m+1}\left(\hat{\beta}, \sigma^2 \left(X^T X\right)^{-1}\right). \tag{9}$$

The left hand side above is referring to the conditional density of β given the data as well as the parameter σ. (**4 points**)

b) Prove that the posterior density of σ (i.e., the density of $\sigma \mid$ data) is proportional to

$$\sigma^{-n+m} \exp\left(-\frac{S(\hat{\beta})}{2\sigma^2}\right) I\{\sigma > 0\}, \tag{10}$$

where $S(\hat{\beta})$ is the Residual Sum of Squares. (**4 points**)

c) Use the results of the previous two parts along with the formula

$$f_{\beta|\text{data}}(\beta) = \int_0^\infty f_{\beta|\text{data},\sigma}(\beta) f_{\sigma|\text{data}}(\sigma) d\sigma$$

to deduce (8). (**4 points**)

d) Use the result in part (b) to show that (**3 points**)

$$\frac{S(\hat{\beta})}{\sigma^2} \mid \text{data} \sim \chi^2_{n-m-1}.$$

e) Use the result in the previous part to argue that the usual estimator:

$$\hat{\sigma} := \sqrt{\frac{S(\hat{\beta})}{n-m-1}}$$

is a reasonable point estimate of σ. (**2 points**).

5. [**This question is only for students taking STAT 248**] This problem discusses predictions. Consider the same setting as the previous problem.

   a) Suppose we are interested in the parameter $a^T\beta$ for some fixed vector a. Using (9), prove that

   $$a^T\beta \mid \text{data}, \sigma \sim N\left(a^T\hat{\beta}, \sigma^2 a^T \left(X^T X\right)^{-1} a\right)$$

   Note that we are conditioning on the data as well as the parameter σ in the left hand side above. (**3 points**)

   b) Using (8), prove that (**3 points**)

   $$a^T\beta \mid \text{data} \sim t_{n-m-1}\left(a^T\hat{\beta}, \hat{\sigma}^2 a^T \left(X^T X\right)^{-1} a\right)$$

   c) Suppose we are interested in predicting the value Y_{n+1} of the response variable at a new set of covariate observations $x_{n+1,1}, \ldots, x_{n+1,m}$. Assuming that

   $$Y_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \cdots + \beta_m x_{n+1,m} + \epsilon_{n+1}$$

   for an independent error $\epsilon_{n+1} \sim N(0, \sigma^2)$, prove that (**4 points**)

   $$Y_{n+1} \mid \text{data}, \sigma \sim N\left(\tilde{x}_{n+1}^T\hat{\beta}, \sigma^2 + \sigma^2 \tilde{x}_{n+1}^T \left(X^T X\right)^{-1} \tilde{x}_{n+1}\right)$$

   where $\tilde{x}_{n+1}$ is the vector with entries $1, x_{n+1,1}, \ldots, x_{n+1,m}$.

   d) Using the result from the previous part, the formula

   $$f_{Y_{n+1}|\text{data}}(y) = \int_0^\infty f_{Y_{n+1}|\text{data},\sigma}(y) f_{\sigma|\text{data}}(\sigma) d\sigma,$$

   and the density $\sigma \mid$ data from (10), prove that

   $$Y_{n+1} \mid \text{data} \sim t_{n-m-1}\left(\tilde{x}_{n+1}^T\hat{\beta}, \hat{\sigma}^2 + \hat{\sigma}^2 \tilde{x}_{n+1}^T \left(X^T X\right)^{-1} \tilde{x}_{n+1}\right).$$

   This result can be used for obtaining predictions of Y_{n+1} along with uncertainty intervals. (**5 points**).