MATH 2210: Project 2

Modeling US Ice Cream Production by Month

Sample Student 1

Introduction

In this project, I'll use least-squares approximation to calculate two different functions to estimate the level of ice cream production in United States over time. I obtained the following production data from the United States Federal Reserve.

Month	Jan '18	May '18	Aug '18	Nov '18	Feb '19	June '19	Oct '19	Mar '20	July '20
Ice Cream Production	86.4	113.2	117.5	84.8	104.3	120.8	92.0	108.0	117.0

Here, the production value is the percent of the average monthly production in the year 2017 (so the 86.4 in January 2018 means that in that month, the ice cream production was 13.6 percent less than in the average month in 2017). In my models, I will take the independent variable x to represent the number of months after January 2018, and I'll take the dependent variable y to be the ice cream production level.

Best-Fit Line

First, we wish to compute the best fit line for the data using least-squares approximation. That is, we wish to find coefficients r_0, r_1 for which the line $y = r_0 + r_1 x$ best fits our ice cream data. To compute these coefficients, we use the normal equations $\mathbf{A}^\intercal \mathbf{A}\mathbf{r} = \mathbf{A}^\intercal \mathbf{y}$, with

$$\mathbf{A}^\intercal = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 4 & 7 & 10 & 13 & 17 & 21 & 26 & 30 \end{bmatrix},$$

and $\mathbf{y}$ as the column vector of populations. The normal equations simplify to

$$\begin{bmatrix} 9 & 128 \\ 128 & 2640 \end{bmatrix} \begin{bmatrix} r_0 \\ r_1 \end{bmatrix} = \begin{bmatrix} 944 \\ 13782.8 \end{bmatrix},$$

which have solution $r_0 \approx 98.693$ and $r_1 \approx 0.436$. Therefore, the equation of the best-fit line is

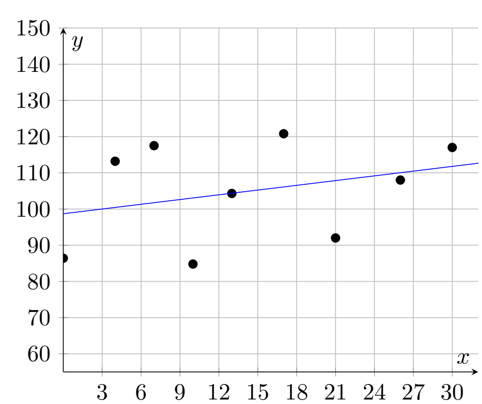
$$y = 98.693 + 0.436x.$$

Visualizing the Data

A graph of the ice cream production data is shown to the right. The best fit line has been superimposed onto the plot in blue.

The production data appears to oscillate with period around 12 months. This makes sense intuitively, as I would suspect that more ice cream is sold during the summer months than the winter months.

In order to capture this oscillating production behavior, I will consider a basis which includes sine and cosine functions with a period lasting 1 year.



A Model that Captures a Yearly Oscillation

For the second model, I'll consider the basis functions $f_0(x) = 1$, $f_1(x) = \sin(\frac{\pi x}{6})$, and $f_2(x) = \cos(\frac{\pi x}{6})$. The trigonometric functions were chosen to repeat every 12 months. The revised coefficient matrix is,

$$\mathbf{A}^\intercal = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & \frac{\sqrt{3}}{2} & \frac{-1}{2} & \frac{-\sqrt{3}}{2} & \frac{1}{2} & \frac{1}{2} & -1 & \frac{\sqrt{3}}{2} & 0 \\ 1 & \frac{-1}{2} & \frac{-\sqrt{3}}{2} & \frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} & 0 & \frac{1}{2} & -1 \end{bmatrix}.$$

Plugging this into the normal equations, we obtain the linear system,

$$\begin{bmatrix} 9 & \frac{-1+\sqrt{3}}{2} & \frac{1-\sqrt{3}}{2} \\ \frac{-1+\sqrt{3}}{2} & 4 & 0 \\ \frac{1-\sqrt{3}}{2} & 0 & 5 \end{bmatrix} \begin{bmatrix} r_0 \\ r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} 944 \\ 79.926 \\ -106.847 \end{bmatrix}.$$

Solving this linear system, we find that $r_0 \approx 103.932$, $r_1 \approx 11.251$, and $r_2 \approx -14.385$, giving us the approximation

$$y = 103.932 + 11.251 \sin(\frac{\pi x}{6}) - 14.385 \cos(\frac{\pi x}{6}).$$

Analysis of Models

To quantitatively analyze our models, we compute their coefficients of determination. The average production level from our samples is $\overline{\mathbf{y}} \approx 104.9$. Plugging in to the coefficient of determination formula, we find that

$$R^2 = 0.100$$

for the best-fit line, and

$$R^2 = 0.886$$

for the trigonometric approximation. The coefficient of determination for the best-fit line is very close to 0, meaning that the line does not represent the relationship in the data very well. This makes sense, since the data has an apparent oscillatory behavior which cannot be captured by a line. The coefficient of determination for the trigonometric approximation is significantly closer to 1, meaning that this model is a fairly accurate representation of the data. This gives us some quantitative confirmation that our intuition about higher ice cream production in the summer months is correct.

Citations

I utilized some of the data from the St. Louis Federal Reserve's FRED database:

- Board of Governors of the Federal Reserve System (US), Industrial Production: Manufacturing: Non-Durable Goods: Ice Cream and Frozen Dessert (NAICS = 31152) [IPN31152N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/IPN31152N, July 22, 2021.