Data 102: Trends in U.S. Transportation Surrounding the COVID-19 Pandemic

Cory McCubbrey, Alexis Llamas, Aryan Kanoi

1. Introduction

For our project, we wanted to explore trends in transportation, using a dataset provided by the United States Bureau of Transportation Statistics. We decided to focus specifically on automobile sales, given that it is an important sector of the economy. We formulated two main research questions.

1. What is the overall trend in automobile sales in the United States, and how will this trend continue past the start of the COVID-19 pandemic?
2. How did the COVID-19 pandemic affect trends in automobile sales, and can we determine a causal relationship?

To answer the first research question, we decided to use a Bayesian regression model to capture the trends in automobile sales after the pandemic began. To answer the second research question, we wanted to compare the trend in automobile sales from before the beginning of the pandemic to after it began, and attempt to determine a causal relationship using Google's Causal Impact model.

1.1 Overview of Research Question 1

Our first research question involves simply examining the trend of auto sales in the U.S. since the COVID-19 pandemic began. We used a Bayesian linear model with a Gaussian likelihood, which equates to ordinary least squares regression. After exploring various trends in the transportation dataset, we decided to predict auto sales. The trend of post-pandemic auto sales wasn't exactly linear, but we made it work as best we could.

1.2 Overview of Research Question 2

<u>Motivation for Using Google's Causal Impact</u>

Google's Causal Impact model is not something that has been introduced in this course, and thus a more in depth explanation is warranted. Our motivation for using this package originated from the fact that we are analyzing a dataset that is organized by a time-series. We understand that given the nature of this data, the Stable Unit Treatment Value Assumption (SUTVA) does not hold. This is due to the fact that the units of time in our auto sales dataset are not independent of each other. After conducting EDA and understanding the transportation data, we felt committed to sticking with it, so we conducted research on what other methods were available to conduct such analysis. It was at this point that we discovered Google's Causal Impact Model.

<u>Differences Between Google's Causal Impact and Traditional Causal Inference</u>

Note that our auto sales data would be considered an observational study. To estimate the Average Treatment Effect (ATE), we would need to at a minimum calculate the difference in counterfactual means from before the treatment (COVID-19) to after treatment. We do not have this option, considering that we have aggregate data, meaning that there would be only a single data point each for pre and post treatment. The unconfoundedness assumption does not hold,

given that auto sales is part of many ever-changing variables in the greater U.S. economy which all affect each other.
Google's Causal Impact model utilizes a Bayesian Structural Time-Series model, which will be elaborated on below. Given a time period for pre-treatment and post-treatment, the model will forecast the trend in auto sales for post-treatment using only the pre-treatment data. This projection, or synthetic control, will be used as the counterfactual, which represents how the trend in auto sales might have continued had COVID-19 not occurred (**Fig. 1**). The model can also incorporate covariate trends to help with its projection. In this case, it would be desirable to use another time-series that is correlated with auto sales but is unaffected by COVID-19.

2. Data Overview

Note: The Google Community Mobility dataset was used for our EDA but ultimately was not included in either of our research questions. In addition, the dataset ran in Google Collab but crashed the kernel when it was loaded into JupyterHub, so I elected to remove the code related to the mobility dataset from our Jupyter notebook submission.

The data has been generated using the datasets published by the Bureau of Transportation statistics and contains monthly data transportation utilization spending. The dataset includes information on air- line traffic, transit ridership, transportation employment, construction spending, and trans-border movement. We filtered the data based on auto sales as we intend to understand the state of the automobile market due to the COVID 19 pandemic. We also filtered according to the dates when the pandemic hit to determine the changes. We also used the mobility dataset by Google (daily Community Mobility data) We filtered the mobility dataset to get data pertaining to the United States only. Moreover we also filtered according to date (04-01-2020 and higher) in order to understand how the data changed with the pandemic hitting the market.

We used publicly available data from the US Department of Transportation. The data contained certain columns which were unnecessary for our purpose and hence we chose the columns which were necessary according to the need of our questions. The rows of our data represent monthly statistics pertaining to columns such as auto sales, Highway Vehicle Miles Travelled, Unemployment rate, Labor Force participation. The rows help us in determining how the auto sales changed over the course of a month and how the pandemic months were different from the normal course.

There may be selection bias as we chose certain columns and may have neglected some useful columns which could be confounding factors. Using just the data for the US could potentially be convenience sampling. Having data about the number of people per state could help market size and understand the trend related to auto sales. Moreover Tax rates in different states could also be useful in determining why auto sales are higher or lower in certain states as opposed to others.

Links to the dataset:
Bureau of Transportation Statistics: Monthly Transportation Statistics:
https://data.bts.gov/Research-and-Statistics/Monthly-Transportation-Statistics/crem-w557

Google: Daily Community Mobility Data : https://www.google.com/covid19/mobility

3. Exploratory Data Analysis

3.1 EDA for GLM Prediction

Our first round of EDA involved the exploration of various trends in the transportation dataset to discover one that was interesting. We viewed post-COVID-19 trends for several metrics in the transportation dataset, including total highway miles travelled, the unemployment rate, and government spending on roads. Highway travellers began to gradually increase in number following their initial decline after March 2020, and continued to increase until late 2020 when they began to fall once more. The unemployment rate, as one might expect, spiked around March 2020 but has been slowly declining since then and continues to decline into 2021. We finally decided on auto sales, as we thought it would be an interesting metric to attempt to model.

3.2 EDA for Causal Impact Analysis

The analysis for this section had two primary motivations: First, to prepare our auto sales data for use with our model. Second, to explore trends in other features of our transportation data in order to identify a trend that was correlated with auto sales but unaffected by COVID-19.

In order to run the model, we needed to establish time periods for pre and post COVID-19. Deciding on the post-period was more straightforward; we chose the beginning of this period to be March of 2020, the same month in which President Trump declared a national emergency (whitehouse.gov). For the pre-period, we chose the start date to be January 1st 2016. The duration of the pre-period might not have mattered due to the fact that the duration of the post-period was only a single year.

In order to find a trend correlated with auto sales, we paired up each feature with auto sales and examined their correlation over the entire time period. We decided on a feature representing government spending on land transport terminals, as it was highly correlated with auto sales during this period. Although this correlation is probably spurious, it could be used in this hypothetical case as a control for how auto sales might have continued through 2020 had COVID not occurred. Considering that their values are on different scales, we decided to normalize each trend in order to properly visualize them and later feed them to our model (**Fig. 2).**

4. Research Questions, Methods, and Results

4.1 GLM Prediction

GLM are essential as they Can make use of subject matter knowledge to increase sample efficiency. Models created from GLMs are naturally interpretable and have easier uncertainty quantification.

Our GLM model selection and specification largely come from lab 03, the turbine problem. We noticed that the distribution of our outcome variable, auto sales , was apparently normal-appearing. As were quite a few of our explanatory variables, which led us to believe that a Gaussian GLM would be an optimal choice. The assumption behind this model is that Y | X = x is an independent normal distribution N(βT X, σ2). We have trained our model with Gaussian likelihood in order to study monthly rates of auto sales. Using lectures on GLM from class, we also imported the posterior plotting function.

Results and Analysis

We ultimately ran our GLM model on two plots. The first was the cumulative total of auto sales since the start of the pandemic. Reshaping the data in this way caused the resulting plot to be almost perfectly linear, which makes sense given that the rate of auto sales increased rather sharply since its initial decline and then remained relatively constant (**Fig. 5**). Reading the model's results, the coefficient on month was around 294, meaning that for each month auto sales increased at an average rate of 294 thousand, given that the metric we used was in thousands. The bulk of the distribution of this coefficient was between 285 and 300 (**Fig. 6**). Using an edited function to plot the posterior predictive distribution, we can conclude that the model is very confident in its prediction (**Fig. 7**). This is to be expected, given that the data was extremely linear.

The second plot we trained our model on had a log transformed x axis. We did this because the original shape of the data, the rate of auto sales by month, appeared to be growing logarithmically. Transforming the data made it appear slightly more linear but was ultimately not very effective (**Fig. 8**). The model's prediction had a great degree of uncertainty, with the edge of the distribution of the coefficient on month actually passing past 0 into the negative numbers (**Fig. 9**). Given that our transformed data was still not linear, this is not a surprise. Plotting the posterior predictive distribution demonstrates the model's uncertainty (**Fig. 10**).

4.2 Causal Impact Analysis

Model Structure and Components of a Time Series

A time series model can be represented by 3 major components: A trend component, a seasonal component, and random noise (Chris Price). Each component in Google's model is represented by the following equation: $y_t = \mu_t + \gamma_t + \beta X_t + \epsilon_t$ where y_t is the response variable, μ_t is the trend component, γ_t is the seasonal component, and ϵ_t is the random noise.

Analysis

Running the model itself is actually quite simple as it fits all of its parameters automatically. We fed into the model the dates for our pre-period and post-period, as well as the data for auto sales and government spending.

Figure 3 displays the model's 3 major numerical results: The model's predicted counterfactual post-covid trend, the absolute effect, and relative effect of treatment. It also quantifies the uncertainty in these above values.

The model showed that our response variable, auto sales, experienced a decrease of -14.07% at the beginning of the post-period. However, the 95% confidence interval of this percentage is [-23.72%, -4.6%]. The model is able to print out verbal results, and a truncated version is as follows:

"During the post-intervention period, the response variable had an average value of approx. 272936.36. By contrast, in the absence of an intervention, we would have expected an average response of 317638.4. The 95% interval of this counterfactual prediction is [287541.18, 348278.43]. Subtracting this prediction from the observed response yields an estimate of the causal effect the intervention had on the response variable."

Figure 4 shows a visualization of the model's predictions and results. The first plot is the most revealing for our analysis. The dip in the trend in auto sales around March 2020 is what results in the predicted trend to be greater than the actual trend. As expected, our covariate trend was not effective in increasing the model's predictive accuracy, and thus the confidence interval remains quite wide.

Results and Discussion

There are flaws in our modeling decisions, which means our results are not scientifically viable. Even so, we were able to identify the causal impact that covid had on auto sales, albeit with a vast amount of uncertainty and inaccuracy. Due to this, the majority of this discussion will be about our model's limitations, flaws, and how they can be improved.

Our selection of our response variable may not have been the best choice due to how quickly auto sales recovered. Generally, Google's Causal Impact performs best on data that experiences a change that lasts for a longer period of time. If we had conducted our analysis in the future, we could have accessed more time series data post-COVID-19 and would have had more accurate results to quantify the impact on auto sales.

The selection of a covariate trend was also flawed. As stated above, government spending on land transportation terminals is quite likely spuriously correlated with auto sales and thus was not the most effective trend to include in the model. In fact, there are many confounding variables affecting all kinds of economic trends in the U.S. that cause more inaccuracy in our model. As

mentioned previously, having auto sales data from another country less impacted by covid would have allowed the model to create a much better counterfactual projection.

Determining the impact of COVID-19 on auto sales is difficult simply because COVID-19 impacted auto sales across the globe. The point of a counterfactual is to estimate the state of the world had the treatment not occurred, and in the case of our treatment, it was indeed global.

5. Conclusion

GLM Prediction

Our prediction with GLMs was overly simple but yielded some positive results in the case where we modelled aggregate linear data. An automobile company that conducted a more robust analysis could predict how many autos will be sold in the next few months and adjust production to accommodate that prediction. To conduct a more accurate prediction of sales data that does not appear linear, one could use a machine learning method that works with time series data, such as a Recurrent Neural Network.

Causal Impact Analysis

Given the flaws in our method described above, we would not use it to inform any real world decisions. However, such an analysis done right would be very beneficial for informing real-world business decisions. It is very important for an automobile company to know how their revenue was impacted by the pandemic, as this could inform how they cut costs within their company and recover from the financial hit. A potential future study could be conducted on how the recent hack and resulting shutdown of the Colonial Pipeline affected gas prices, which in turn could also have affected auto sales, especially in the eastern U.S which was hit the worst (Cnet).

Flaws aside, our model still may have produced results in the ballpark of reality. An article posted at the end of 2020 on CNBC detailed that auto sales declined by at least 15% compared to 2019 (Wayland). Our model's prediction was less than 1% away from this statistic, proving the power, or possibly luck, of Google's Causal Impact model in our analysis.

References:

https://www.cnet.com/news/gas-shortage-2021-when-service-will-restart-on-the-hacked-pipeline/

https://www.cnbc.com/2020/12/23/covid-19-crippled-us-auto-sales-in-2020-but-it-could-have-been-worse.html

Appendix

Figure 1: Explanation of causal impact analysis, using revenue before and after a marketing campaign as an example.

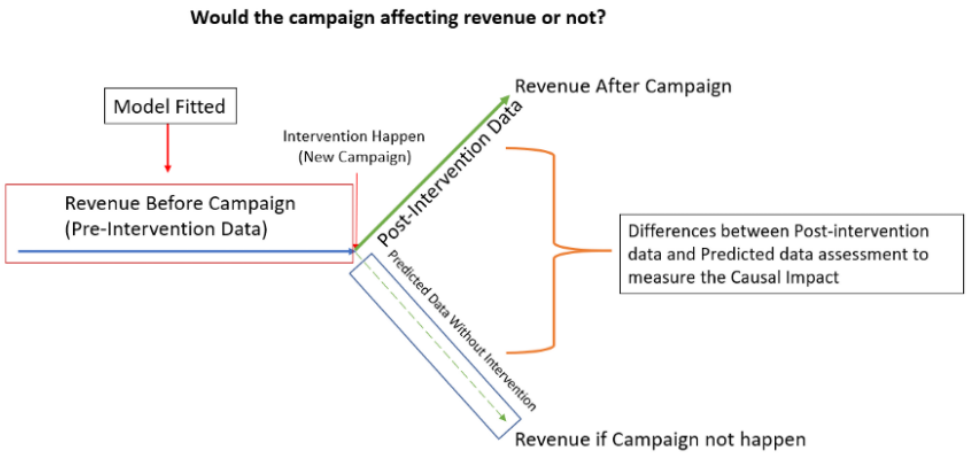


Figure 2:

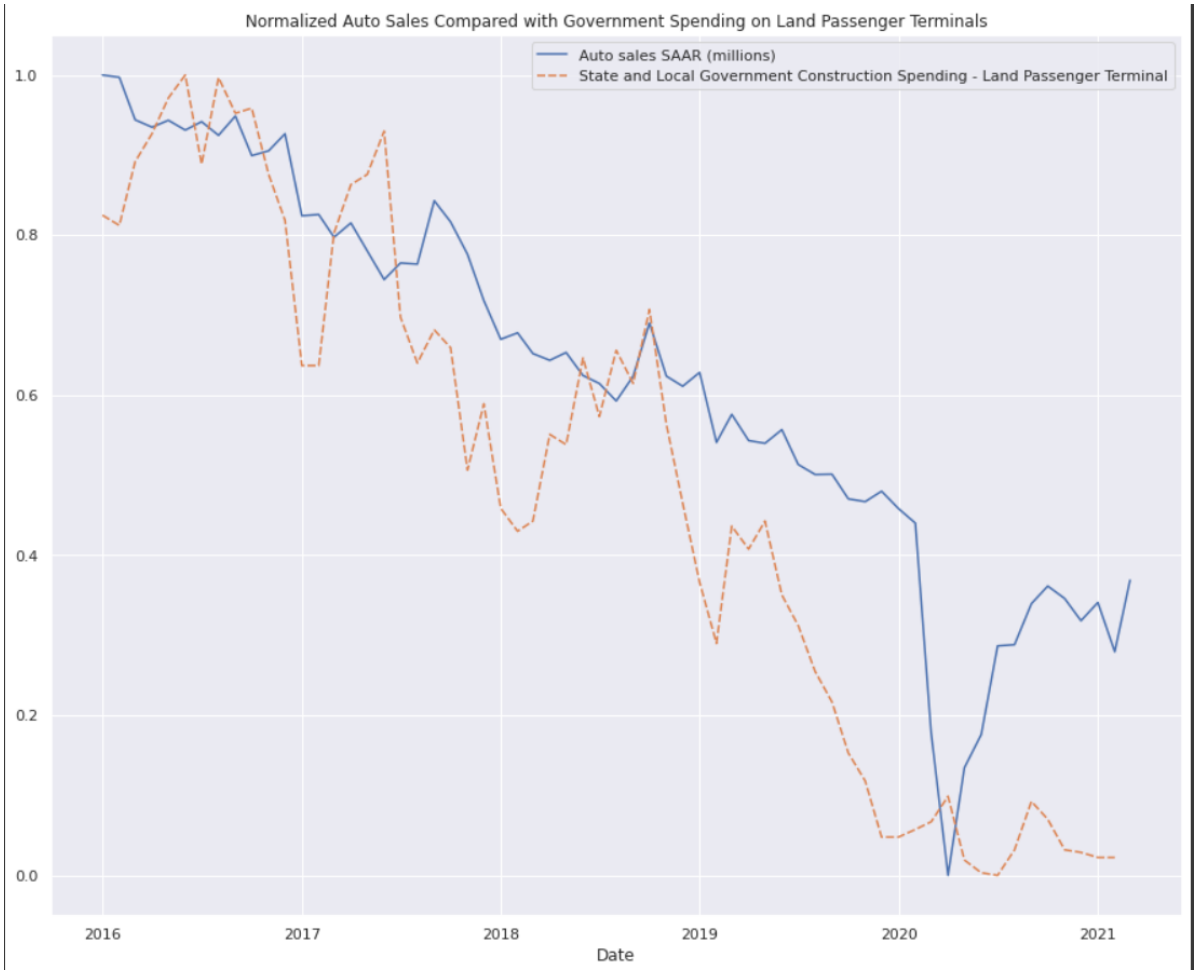


Figure 3:

```
Posterior Inference {Causal Impact}
                          Average            Cumulative
Actual                    272936.36          3002300.0
Prediction (s.d.)         317638.4 (15494.48)3494022.43 (170439.3)
95% CI                    [287541.18, 348278.43][3162952.95, 3831062.73]

Absolute effect (s.d.)    -44702.04 (15494.48)-491722.43 (170439.3)
95% CI                    [-75342.07, -14604.81][-828762.73, -160652.95]

Relative effect (s.d.)    -14.07% (4.88%)    -14.07% (4.88%)
95% CI                    [-23.72%, -4.6%]   [-23.72%, -4.6%]

Posterior tail-area probability p: 0.0
Posterior prob. of a causal effect: 99.8%
```

Figure 4:

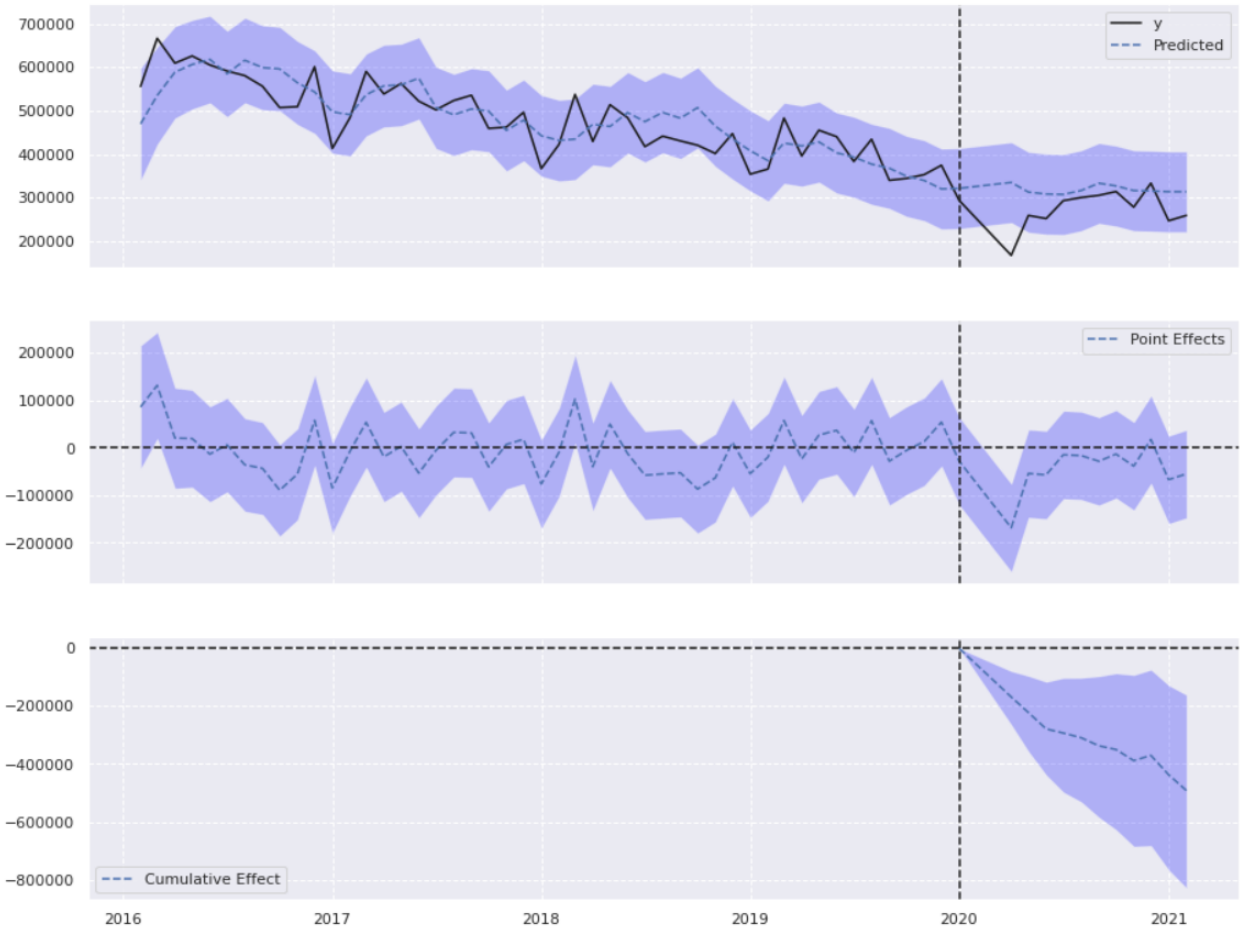

y
Predicted
700000
600000
500000
400000
300000
200000
Point Effects
200000
100000
0
−100000
−200000
0
−200000
−400000
−600000
−800000
Cumulative Effect
2016
2017
2018
2019
2020
2021

Figure 5:

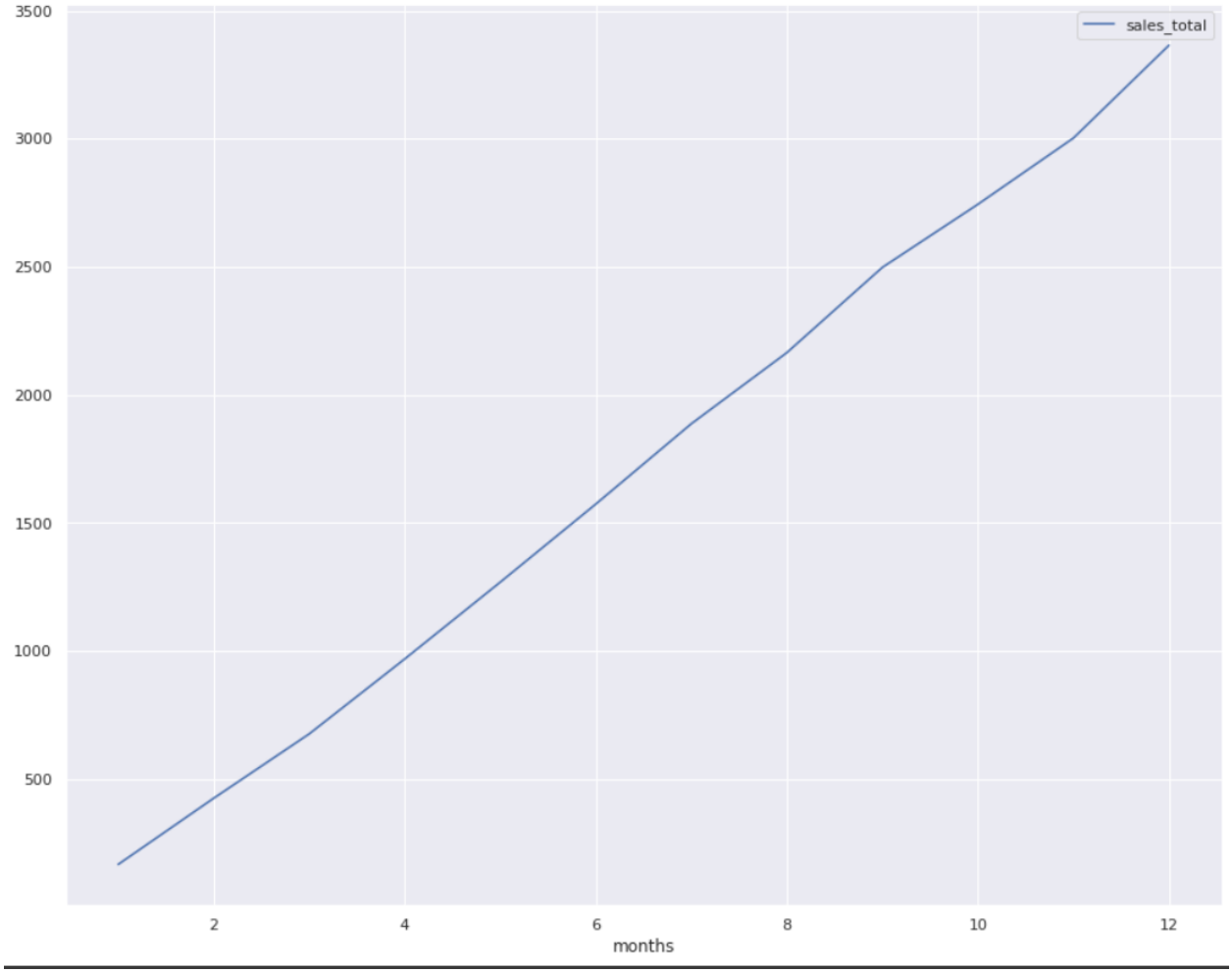

sales_total
3500
3000
2500
2000
1500
1000
500
2
4
6
8
10
12
months

Figure 6:

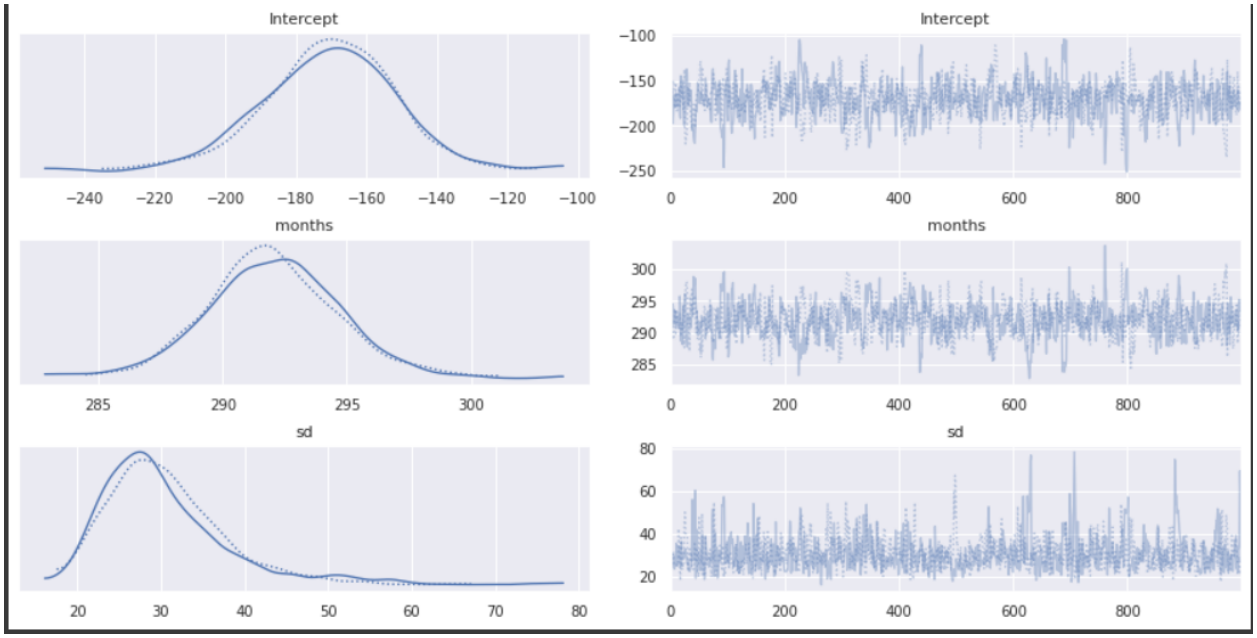


Figure 7:

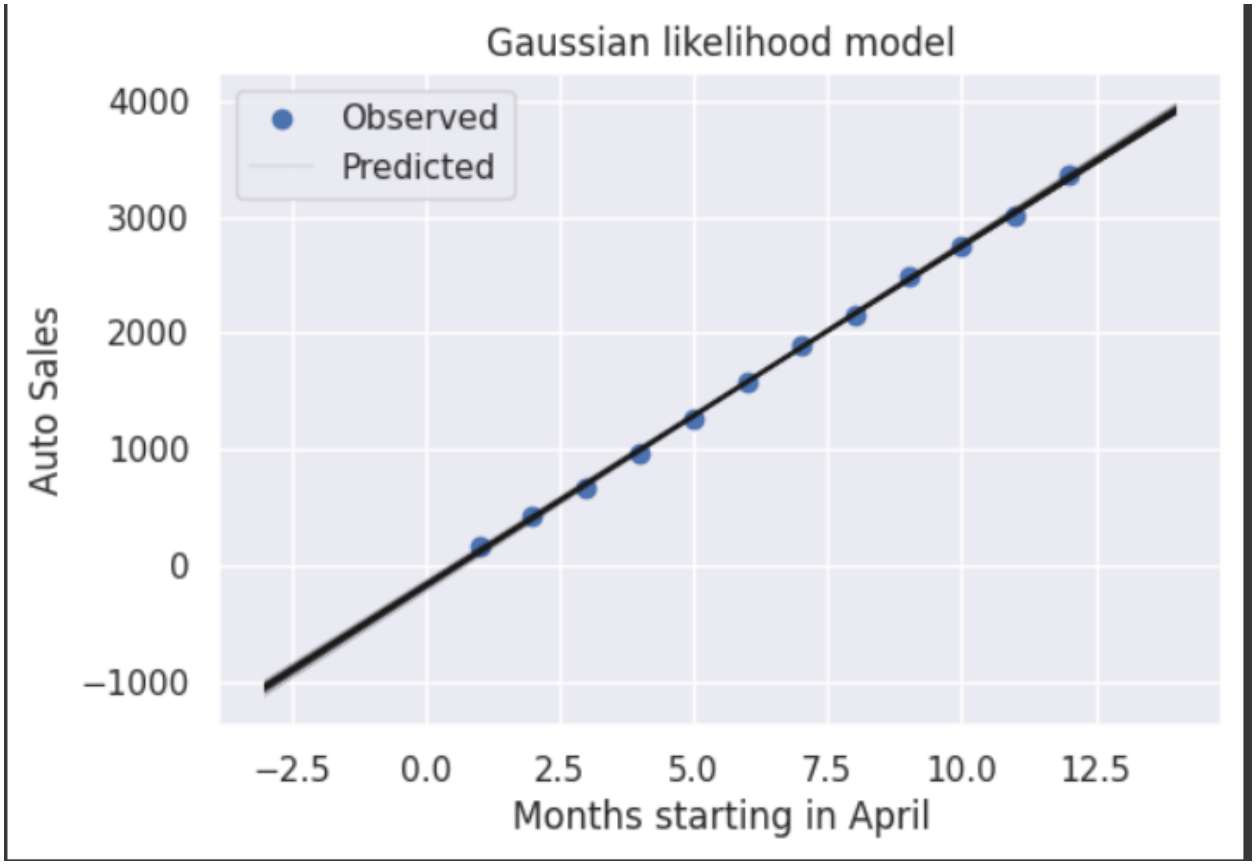


Figure 8:

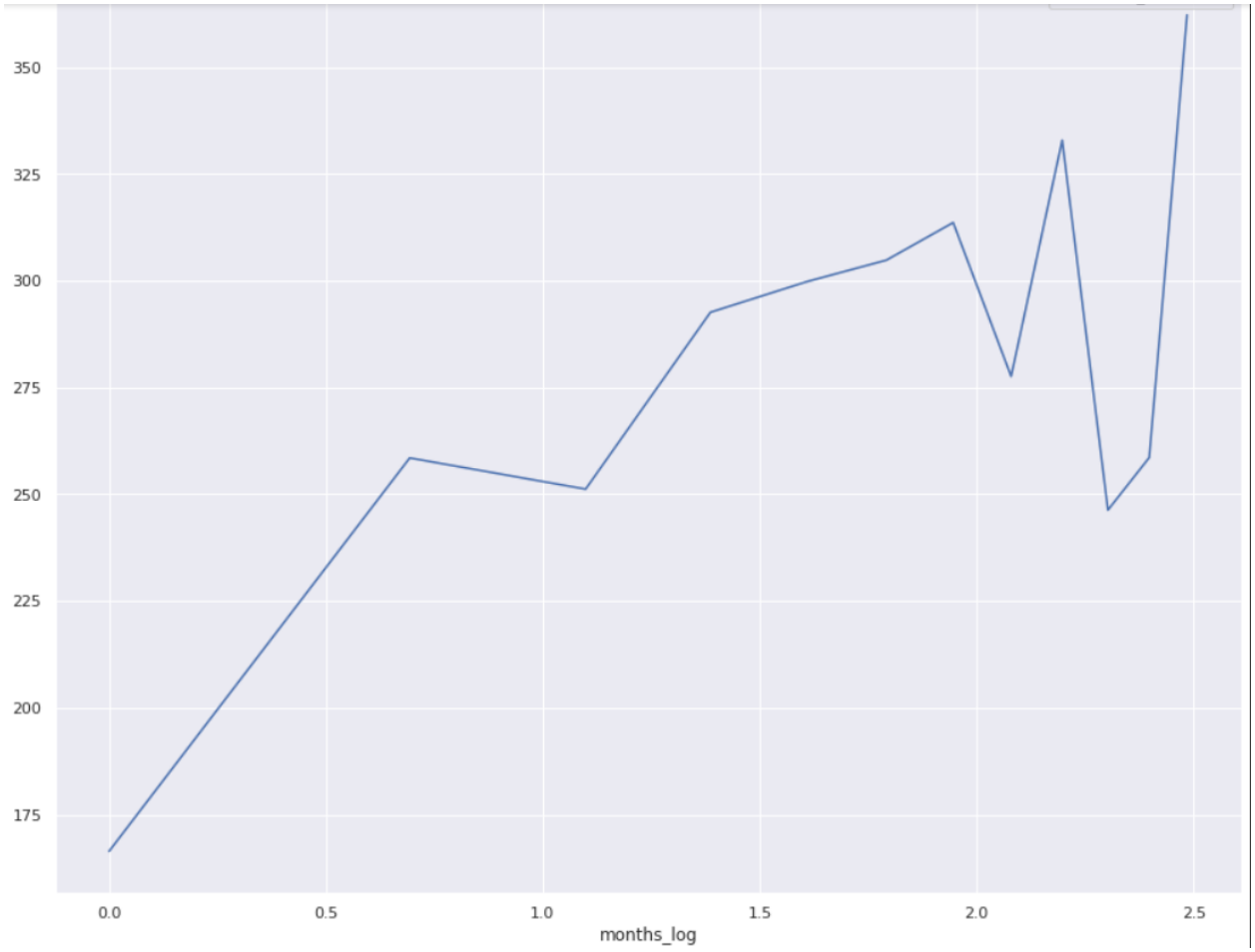


Figure 9:

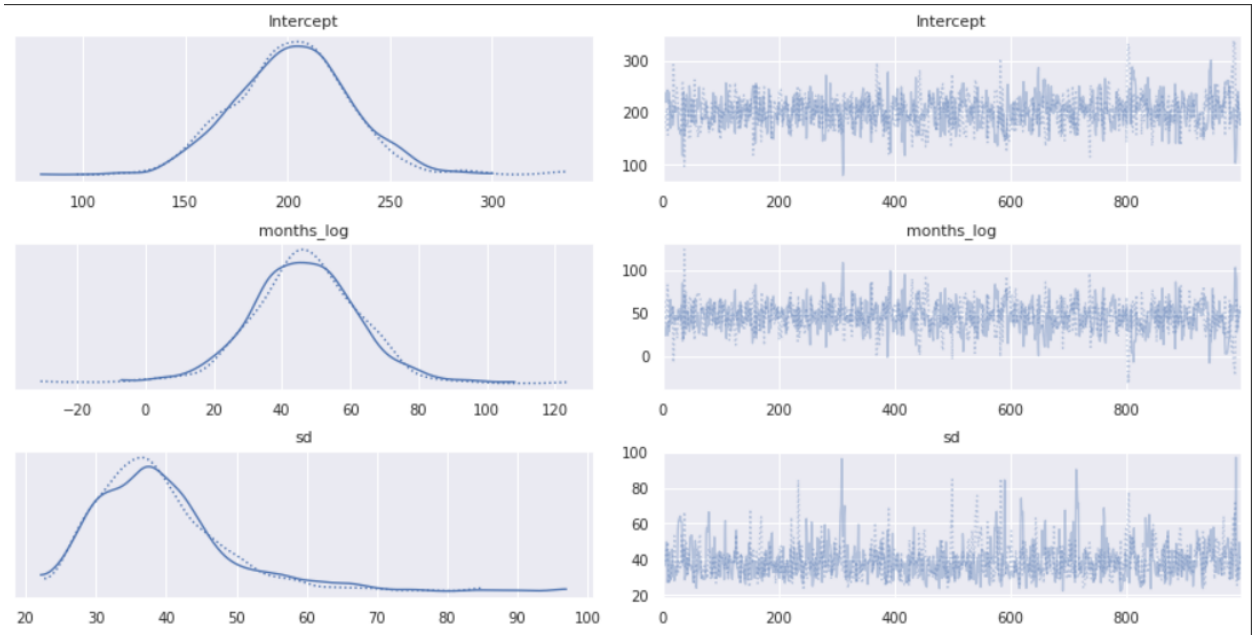
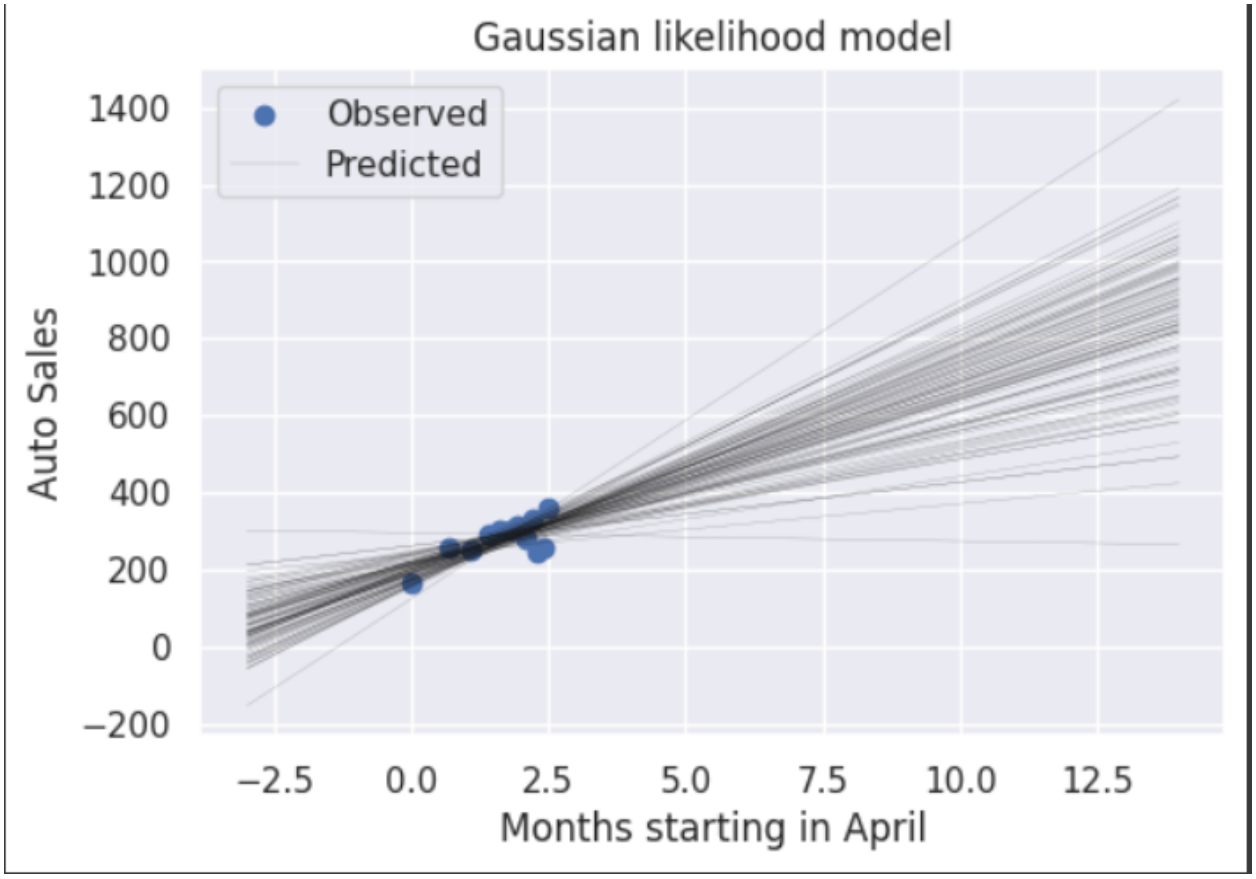


Figure 10:


Gaussian likelihood model
Observed
Predicted
1400
1200
1000
800
600
400
200
0
−200
Auto Sales
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
Months starting in April