Forecast of Weekly Median Household Earnings in the US

Daksh Prashar Dhruv Tantia Laura Edward Sofya Malashchenko

2024-12-06

Problem

The idea of the project is to analyse the earnings dataset and fit a model to it to explore trends in earnings as well as predict the median earnings for the next year. This is particularly important now as a lot of us are graduating within the next year and are searching for full-time jobs.

Plan

The goal of this project is to create a model that can be used to predict the earnings in the upcoming year. For that, we will need to go through the following steps

1. Identify any sources of non-stationarity in the dataset.
2. If the data doesn't have constant variance, use Box-Cox transformation to address the issue.
3. If the trend and/or seasonality are present, test out appropriate models that.
4. Using APSE, select the model with the highest prediction power.
5. Fit the model from step 4 on the whole dataset and predict the upcoming year.

Data

As concluded from the description of the dataset, this data contains quarterly, seasonally adjusted data on the median weekly earnings.

The data was collected by surveying the participants. Note that self-employed individuals were not considered for this survey. As indicated in the data description, there was a change in a data collection method in 1994. Prior to 1994, the participants were asked to provide their weekly income while after January 1994 they were asked to provide this information in a way that is easiest for them and that was later converted to weekly earnings. In both cases, the values in the dataset are weekly median earnings ordered by quarters.

Some of the things we needed to keep in mind when working with this dataset:

1. The data is quarterly rather than monthly
2. As provided in the data description, there was a change in the data collection process around 1994 which is the definition of a change point. this might make the patterns in the data more complex.
3. Since the data was collected through surveying, it likely contains biased and should be taken with a grain of salt.
4. There might be a change point around 2020 caused by covid that is not included in the data description

As a note: the data does not contain any missing values so we did not have to address this issue

Exploratory data analysis

The first step is to plot the whole dataset.

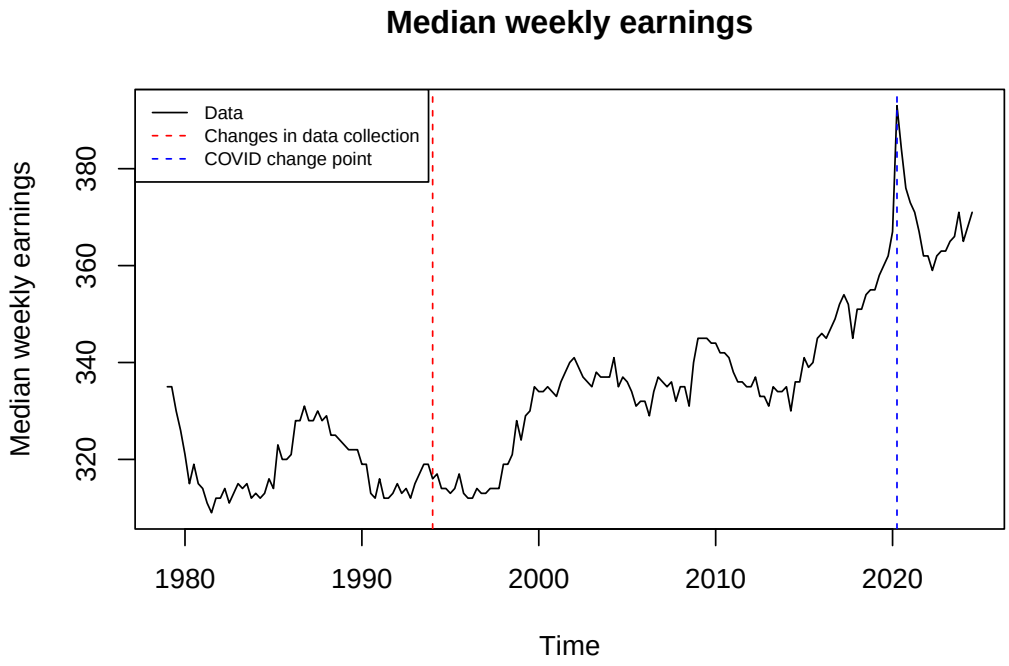

Figure 1: Plot of the full earnings dataset

Looking at Figure 1, the first thing we notice is that Covid indeed had an impact on the data. There seem to be a sharp increase in the median weekly earnings right around that period which might affect our future analysis. Hence, it was decided to remove this period from the dataset. That is, we removed all of 2020 from our data (see Figure 2 for a visual). For the rest of the report we will be working with this dataset.

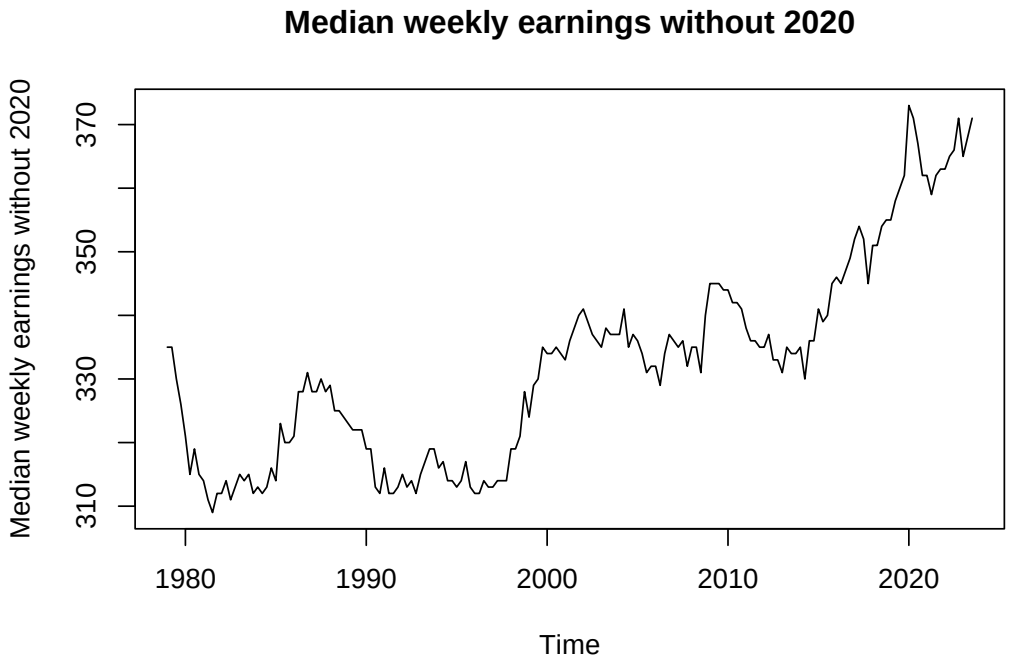

Figure 2: Plot of earnings data with 2020 removed

At a first glance, the data does not contain a seasonal component but it does seem to have an upward trend. There also does seem to be some changes in variance. We verify this using Fligner-Killeen test for constant

variance. After running the test, a p-value of `6.138e-06` was observed which indicates that there is strong evidence against the null hypothesis of normal variance. This will be addressed later.

The next step is to take a look at the ACF plot of the whole dataset to see if there are trend and/or seasonality components in the data. Investigating the ACF plot in Figure 3, we see further evidence that this data does not have a seasonal component. This is expected as the data has been seasonally adjusted. However, we do see a slow (not exponential) drop in the ACF spikes. This provides more reason to believe that this data has a trend and is not stationary. As it can be seen from the ACF plot in Figure 3, there seems to be a slow linear decay in the spikes, thus, we can conclude that the process is not stationary as a trend exists.

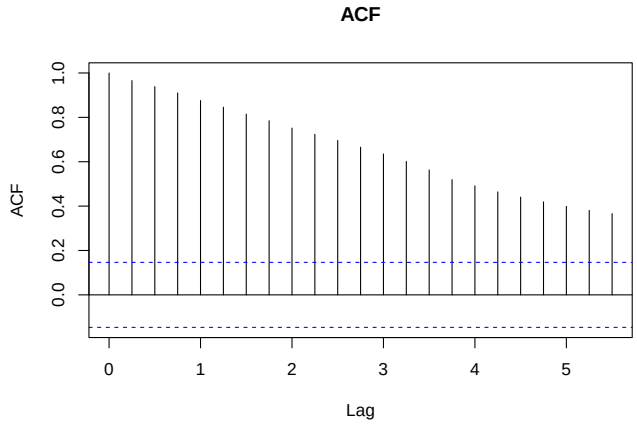

Figure 3: ACF plot of the whole dataset

For more context on the underlying process in this data, we take a look at the PACF plot. On the PACF plot in Figure 4, we can see that there is a spike at lag 1, which leads us to believe that this might be an AR(1) process. We will investigate and confirm this in the following sections.

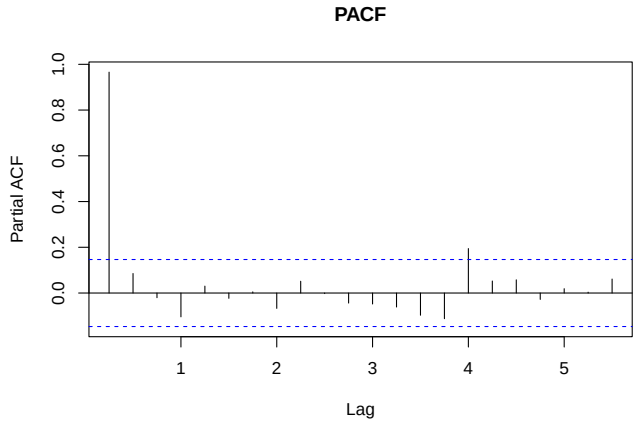

Figure 4: PACF plot of the whole dataset

Note that since there is a decreasing trend in the ACF plot, we likely have a trend in the data. Hence, we will see if first-order differencing can remove this trend. Taking a look at the differenced data, we get the following ACF plot.

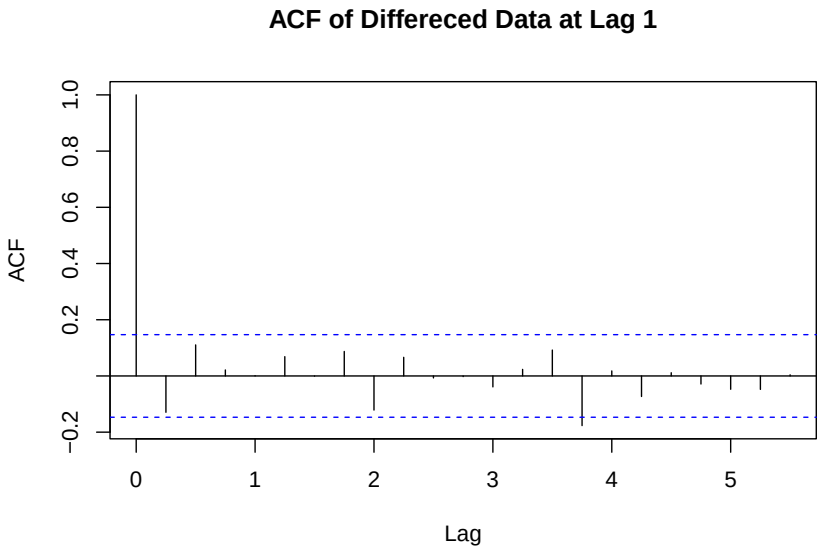

Figure 5: ACF plot of differenced data

Generally there does not appear to be correlation in the differenced data with a few false positives around lag 8 and lag 15. This plot in Figure 5 suggests that all of the information about the data is provided in the trend.

Variance stabilization

Earlier we noticed that the variance in the dataset is not constant. To remove this source of non-stationarity we will attempt to stabilize the variance using a Box-Cox transformation. Running the `boxcox` model suggest an optimal lambda value of -6.111111. However, performing the Fligner Killeen test on the transformed data, we get a p-value of `1.27e-06`. This indicates that the variance cannot be made constant using this type of transformation. Thus, for the rest of the report we will proceed with the untransformed data.

Trend estimation

In this step we will consider multiple models like simple linear regression, exponential smoothing, double-exponential smoothing, elastic net regression (with multiple orthogonal polynomial degrees), and Box-Jenkins models and compare them based on their prediction power or APSE. We will also asses whether the residuals are stationary or not and combine models if necessary.

With this dataset, approximately the last 10% of the observations were part of the test set and everything prior is in the training set.

1. Simple linear regression

Here, we fit a simple linear regression model with polynomial degree $p = 1$.

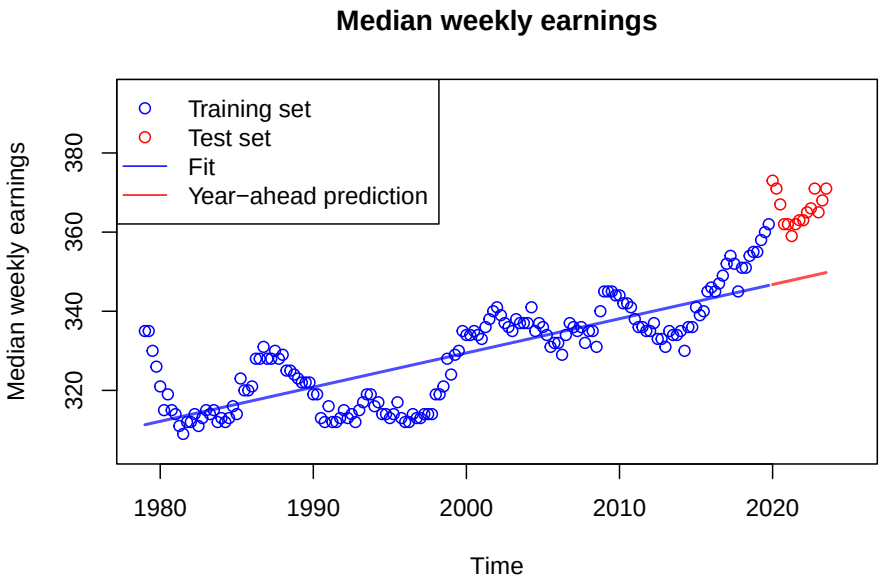

Figure 6: Plot of the data with linear regression model

The APSE value for the linear model in figure 6 is `326.5539`. Looking at the plot we notice that the model doesn't seem to capture the full trend. In particular, from around 2013 to 2020 the trend is very underestimated. Hence, this will likely not be the final model as it doesn't perform well with the given data.

2. Exponential smoothing

Now we will see if regular exponential smoothing is a good fit for this dataset.

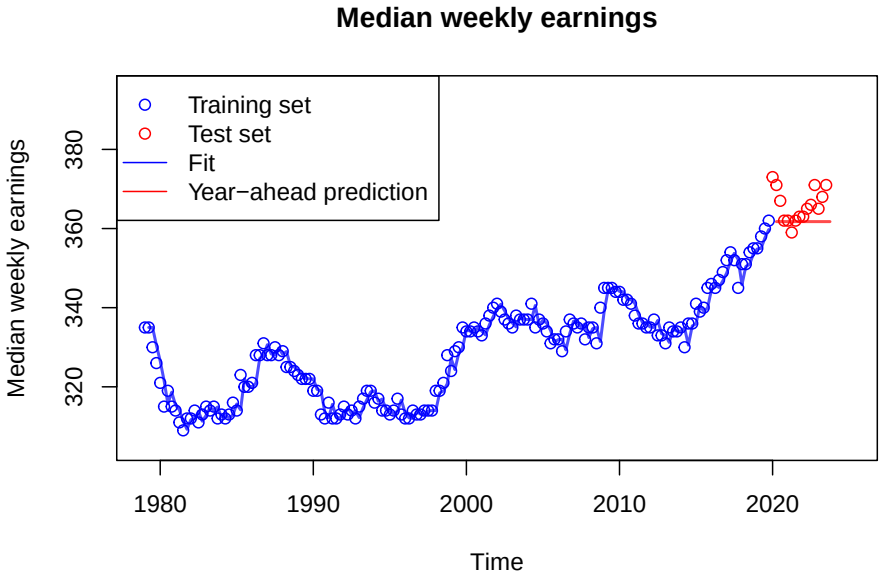

Figure 7: Plot of the data with exponential smoothing model

As we can see the exponential smoothing model seems to fit the training data well, however, the prediction based on the model is not performing as we hoped as it is simply a straight horizontal line.

3. Double exponential smoothing Next, we apply the double exponential smoothing. This type of a model predicts in a straight line, strictly based on the last observation.

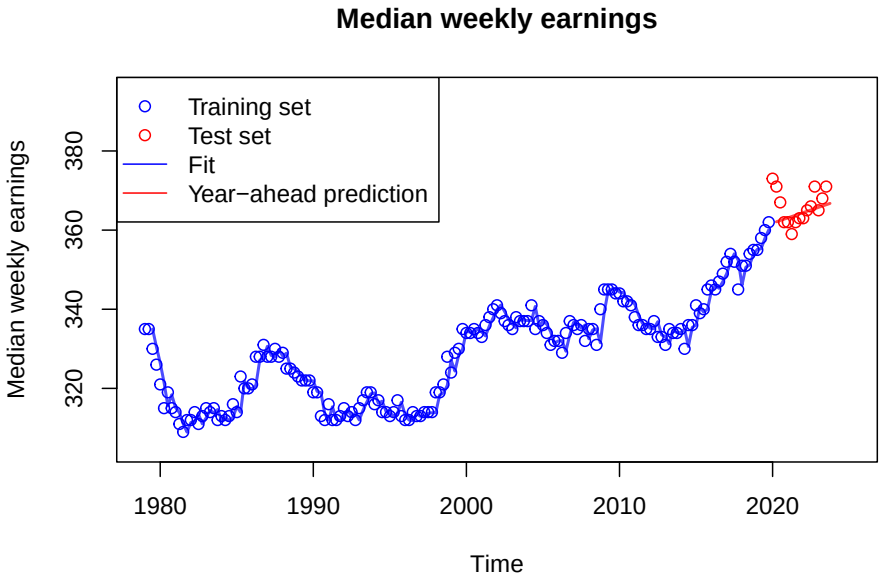

Figure 8: Plot of the data with double exponential smoothing model

Similiar to the exponential smoothing, the double-exponential smoothing model also seems to fit the training data well, however, the prediction based on the double exponential model seems to be able to capture the trend of our test set.

4. Elastic net regression with various orthogonal polynomial degrees and various α values

We aim to fit a model using orthogonal polynomials of degree $p \in 2, 3, ..., 15$ with $\alpha \in 0, 0.5, 1$ to emulate Ridge, Elastic Net, and LASSO models.

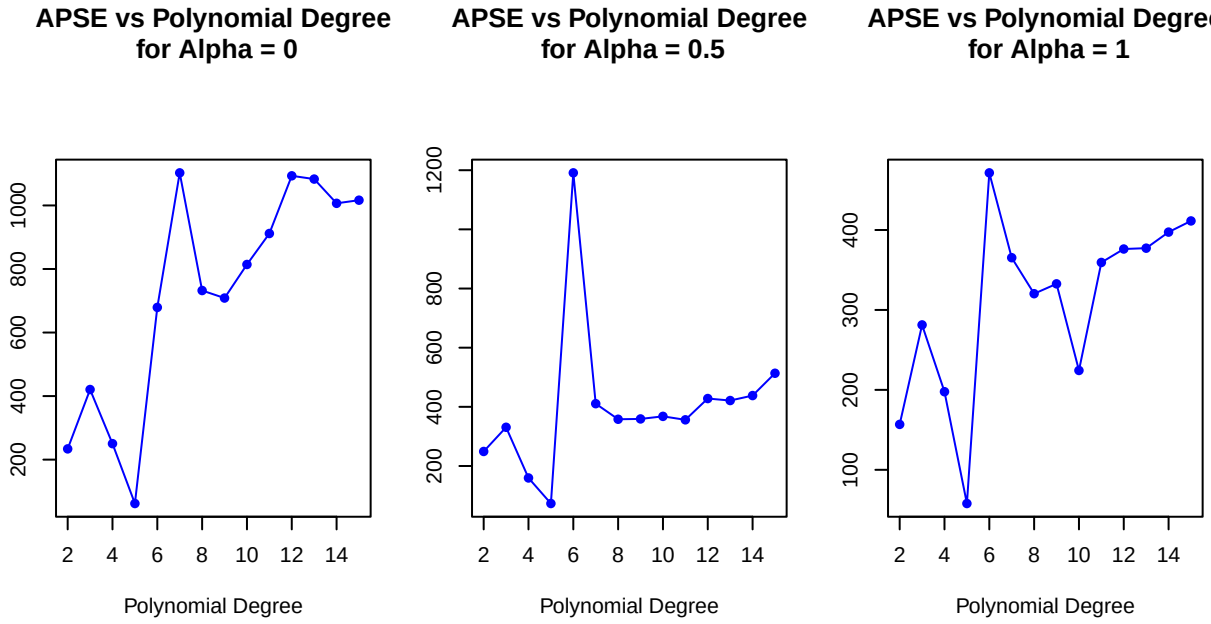

Figure 9: APSE plots for various alpha values and polynomial degrees

We notice from the results of Figure 9, for all α values, the lowest APSE value occurs when the polynomial degree $p = 5$.

Table 1: Alpha Values and APSE Results

Alpha	APSE
0.0	61.8476
0.5	73.1021
1.0	57.8334

Thus, from the results of Table 1, we can conclude that our best model is when alpha = 1 or the LASSO regression with polynomial degree 5.

APSE Analysis

Table 2: Intermediate APSE Results

Model	APSE
Simple Linear Regression	326.5539
Exponential	33.3468
Double Exponential	19.6808
Elastic Net w/ alpha = 1	57.8334

After summarizing all the APSE values we notice that the double exponential model has the best predicting power out of all the other models. Thus, we will select this model and make our predictions from here onward.

Residual Analysis

In this section, we will perform the residual analysis for exponential smoothing, double exponential smoothing, and elastic net model with $\alpha = 1$ and polynomial degree $p = 5$.

Notice that in this section, out of elastic net and simple linear regression, we only perform residual analysis only on elastic linear regression. First, the two models are very similar conceptually and second, the APSE values for simple linear regression are really high, indicating poor performance even in comparison to other elastic net models.

We will check the stationarity of the residuals and try to fit a Box-Jenkins model on them.

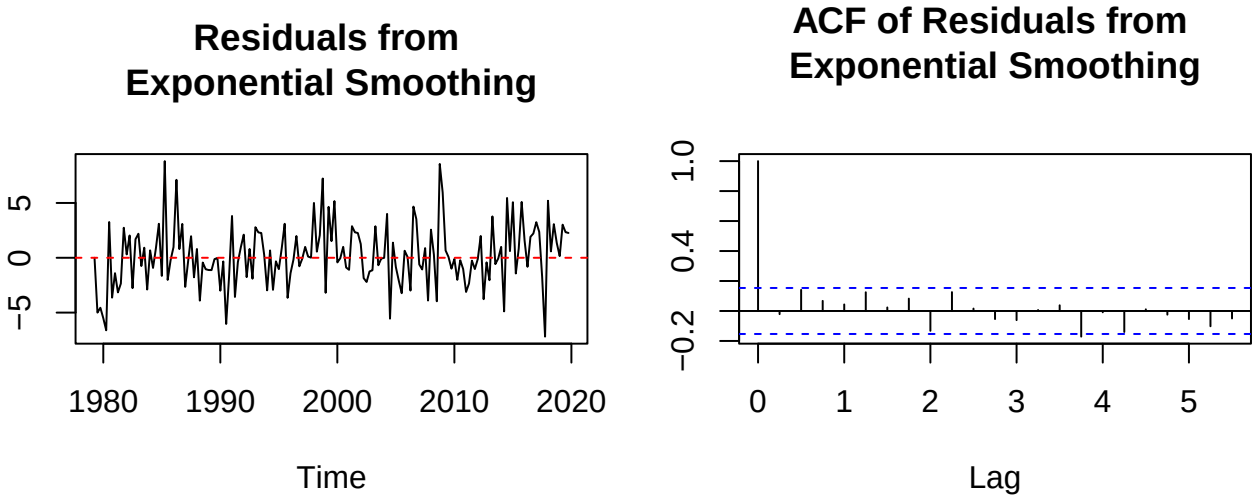

Figure 10: Residual analysis for exponential smoothing

There does not seem to be any correlation in the acf plot. Additionally, a trend does not seem to exist in the variance plot. Hence, we can conclude stationarity for the residuals of exponential smoothing model.

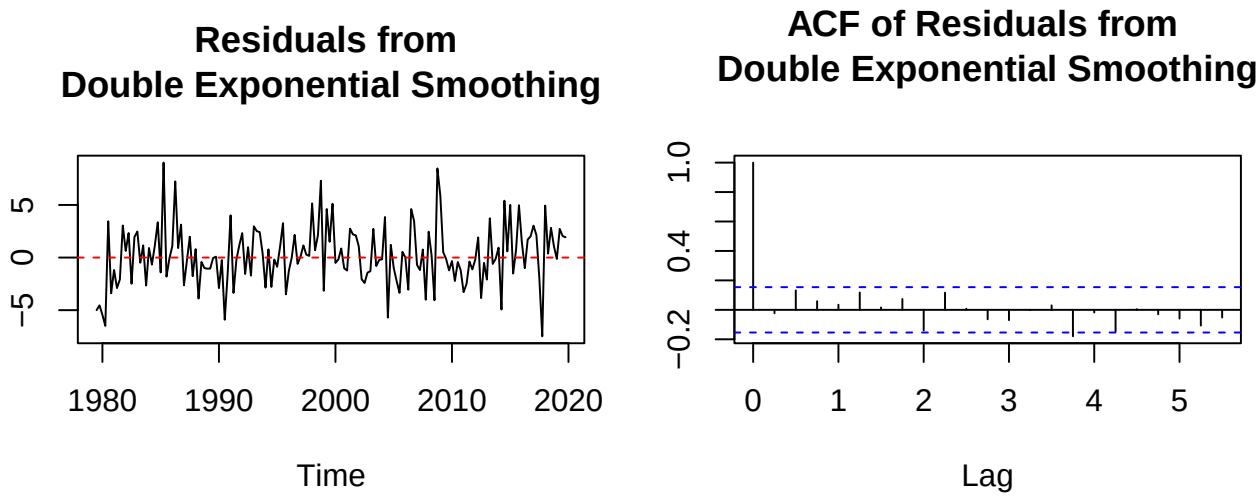

Figure 11: Residual analysis for double exponential smoothing

There does not seem to be any correlation in the acf plot. Additionally, a trend does not seem to exist in the variance plot. Hence, we can conclude stationarity for the residuals of double exponential smoothing model.

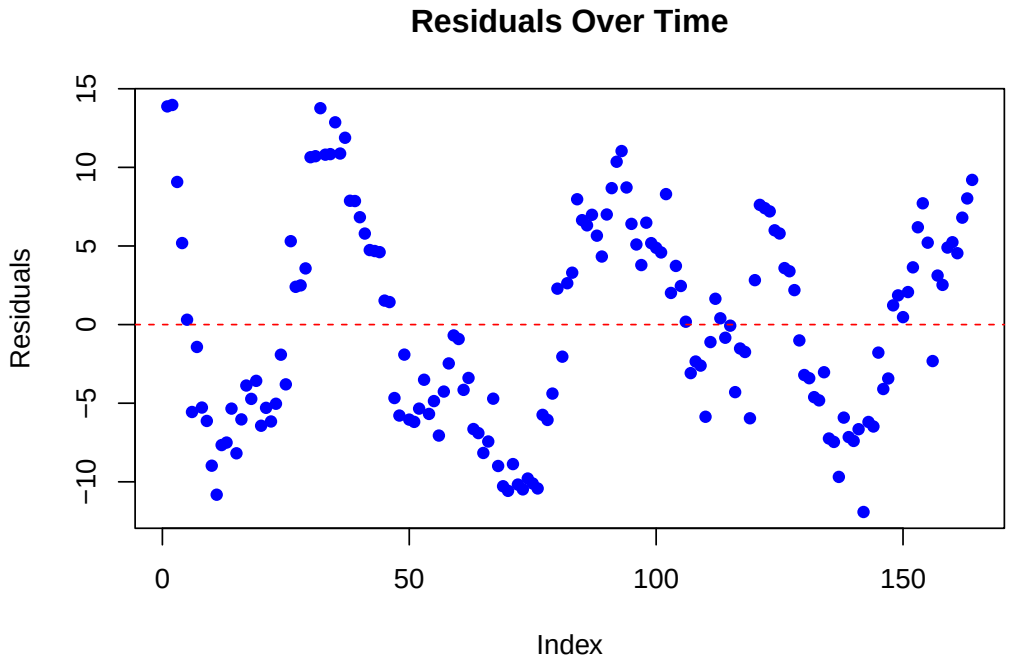

Figure 12: Residuals of the LASSO Model

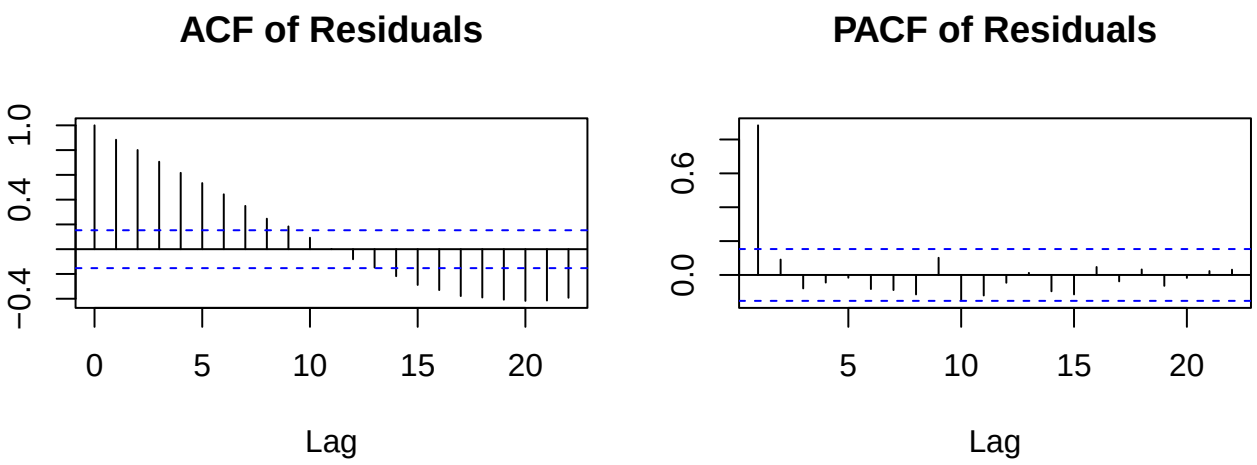

Figure 13: ACF of LASSO model residuals

The residual plot in Figure 12, seems to have a cyclic pattern. Additionally, we see a linear trend in the acf plot in Figure 13. Thus, we can conclude that these residuals are not stationary.

The ACF plot shows a clear linear trend in the lag spikes. This suggests the presence of autocorrelation in the residuals, The PACF plot in Figure 14 shows a strong spike at lag 1 and no significant spikes afterward. This suggests a first-order autoregressive process (AR(1)).

We will try ARIMA models with various p and q values and select the one with the best AIC value. Some of the ARIMA(p,d,q) we will try are: (1,0,1) (1,0,0) (0,0,1) (0,0,2) (1,0,2) (1, 0, 3) (1,0,9). Note again, since our is seasonally adjusted, we will be doing ARIMA and not SARIMA.

Table 3: AIC of Different ARIMA Models

Model	AIC
ARIMA(1,0,1)	806.2966
ARIMA(1,0,0)	806.2906
ARIMA(0,0,1)	955.9536
ARIMA(0,0,2)	894.8664
ARIMA(1,0,2)	804.3668
ARIMA(1,0,3)	806.0863
ARIMA(1,0,9)	805.2102

As we see the lowest AIC value occurs for ARIMA(1,0,2). Now, we will fit this ARIMA model on the residuals of the Elastic Net Model.

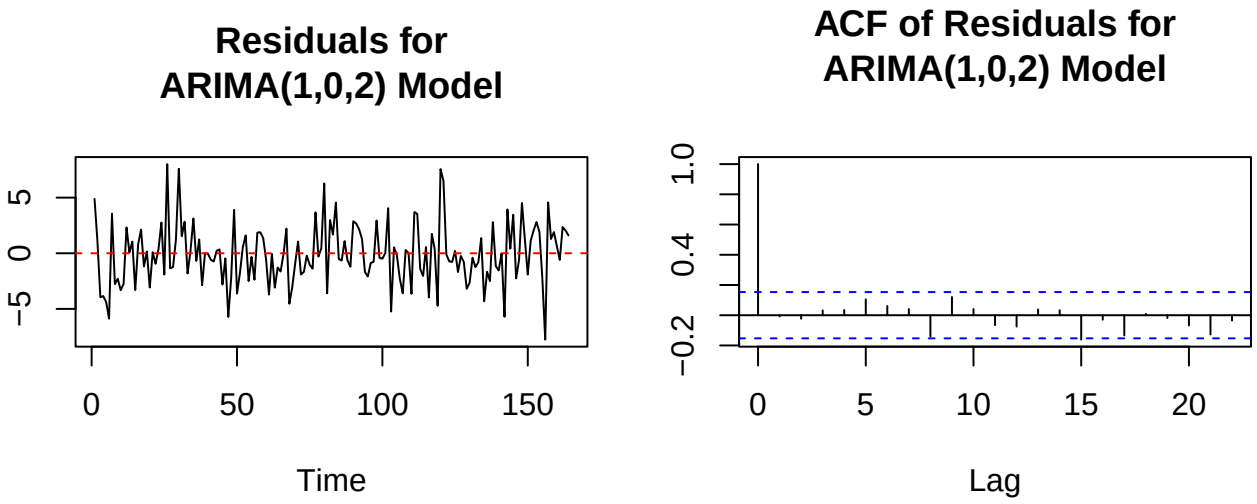

Figure 14: Residual analysis for ARIMA(1,0,2) Model

Now, after the applying the ARIMA model, we see that the residuals are indeed stationary as they seem to be randomly scattered and the spikes in the acf plot do not seem to be correlated.

Next, we generate predictions for residuals of the test set using the ARIMA process and add the trend component back to get forecasted values for the test set indices. Then we will extract the new APSE for the Elastic Net model.

Table 4: Final APSE Results

Model	APSE
Simple Linear Regression	326.5539
Exponential	33.3468
Double Exponential	19.6808
Elastic Net w/ alpha = 1 and ARIMA	28.2608

The new APSE value for the Elastic Net Regression after applying ARIMA(1,0,2) is approximately 28.261. This is a significant improvement from our previous APSE, however, the APSE value of our double-exponential model is still better.

Prediction

Recall that in one of the earlier sections we removed a portion of the data that contained the change point. Since we are interested in predicting the the median earnings for year 2025, we will predict it using the data with the change point removed which will lead to a gap of 1 year on the plot.

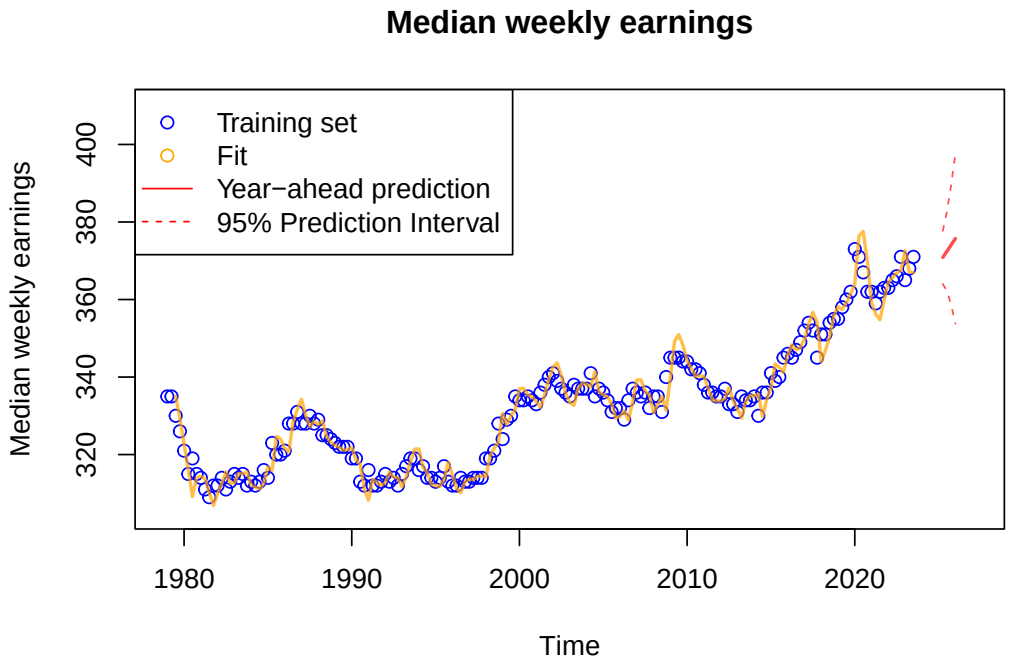

Figure 15: Prediction for the Next Year with Prediction Intervals

Table 5: Prediction for Next Year

Time	Prediction Values
2024 Q4	370.8584
2025 Q1	372.4842
2025 Q2	374.1100
2025 Q3	375.7359

Overall, although the prediction for next year seems to follow the increasing trend of the data, however, our prediction bands are very wide. The reason for this wide interval could be due to the changing variance throughout out dataset which we were not able to address through the Box-Cox transformation. Additionally, the change of data collection process might have had a more significant impact than we had anticipated. As a result of this disturbance in our data, our model is not very confident about its predictions.

Conclusions

In this report we investigated the earnings dataset. There were a few issues that were identified with the original data, but only some of them could be addressed

- We removed the change point due to Covid-19
- We did not think that the change point due to changes in data collection affected the data so we did not address that
- The change in variance could not be addressed

After addressing all issues with the data, we went to investigate several models that we found to be appropriate for this dataset. Specifically, we focused on

- Simple linear regression
- Exponential smoothing
- Double Exponential Smoothing

- Elastic Net Regression

which were compared based on their prediction power.

However, after performing residual analysis we noticed that Elastic Net Regression leaves out correlated reiduals. Hence, it was decided to add an ARIMA model to the residuals of the chosen elastic net model. After that, the residuals for all models were uncorrelated so we could focus on making predictions for the whole dataset.

From the APSE analysis of all 4 models, we found that double exponential smoothing performed the best in terms of prediction. We used that model to perform prediction on the whole dataset (ommitting the Covid-19 change point).

Lastly, after predicting the median earnings for the year 2025 we noticed that the prediction intervals were very wide. This means that one of the issues that we didn't consider at the start. For example, the change point due to changes in the data collection process might have had more impact on the final result than anticipated.

Overall, while we were able to achieve stationary residuals for all models, we were not able to create a model that has high confidence in its predictions.