AAEC 5484: Applied Economic Forecasting

Your Name Here

Homework #4 - Spring 2024

Instructions: In all cases, please ensure that your graphs and visuals have proper titles and axis labels, where necessary. Refer to the output, whenever appropriate, when discussing the results. **Lastly, remember that creativity (coupled with relevance) will be rewarded.**

Question 1: Regressions with Segmented Trends

In this module, we covered the idea that the `trend` variable in a regression model can capture a number of real-world phenomena. In particular, we discussed how the `trend` variable can capture improvements in technology, changes in consumer preferences, and other structural changes in the economy.

In this question, we will explore the idea of segmented trends. In particular, we will model the `trend` to capture a structural break in US corn yields (**in bushels per acre**).

I have pulled the data from the USDA NASS Quickstat database https://quickstats.nass.usda.gov/results/39F2851C-874B-310A-8A78-DF4D529A9E93 and stored it in the `CORN_yield.csv` file in the `HW4` folder on GitHub. We will use the codes below to retrieve the data and then proceed to the questions.

```r
corn <- read.csv("https://raw.githubusercontent.com/Shamar-Stewart/Forecasting/main/Homework/HW4/CORN_yiel
```

1. If you view the first few rows of the `corn` data (**in your console or on GitHub**), you will notice that there is a lot of information on actual and forecasted yields at different points in the calendar and market years. Let us work on cleaning the dataset up a bit.

a. Since we are interested in annual yields, filter the `corn` data to include only `Period` equal to `YEAR`. Next, drop all columns except the `Year` and `Value` columns.

b. Declare the results above as a `tsibble` object with the appropriate `index`.

c. Store the result into an object called `yield`.

I anticipate that you will achieve this using a single line of code.

2. We are ready to proceed to the next step. Using the `autoplot()` function, present a plot of the `yield` data. Comment on any discernible patterns in the data. Do not limit your discussion to just the trend, but be sure to discuss any other dynamics in the data.

3. Estimate a trend regression model of the form:

$$yield_t = \beta_0 + \beta_1 \cdot t + \epsilon_t.$$

Report all relevant regression results and interpret the coefficient on `t`, the trend term.

4. Obtain and report a plot of the actual (presented as points) and fitted values (presented as a line) over time. Does it appear that this model fits the data very well? Why or why not? Discuss.

5. Segmented trend. Commercial hybrids were introduced to farmers on a large scale in the late 1930s, a significant event in the history of US corn yields. We will model the trend to capture this structural break.

Create a dummy variable that is equal to 1 beginning in 1938 and all subsequent periods and is zero otherwise. Recall that you can use an `ifelse` statement to accomplish this. Call the dummy `d38`. For completeness, the dummy is defined as:

$$d38 = \begin{cases} 1 & \text{if year } \geq 1938 \\ 0 & \text{if year } < 1938 \end{cases}$$

Run the following regression for corn yields that allows for structural change before and after 1938:

$$yield_t = \beta_0 + \beta_1 \cdot t + \delta_0 \cdot d38 + \delta_1 \cdot t \cdot d38 + \epsilon_t.$$

Note that we have also included an interaction term between `t` (our `trend()`) and `d38`. This interaction allows for (captures) a change in the slope of the yield data. *In this step, I only want you to store the results of the regression.*

6. Report all relevant regression results. Obtain and report a plot of the actual (presented as points) and fitted values over time. Does it appear (at least visually) that this model fits the data better than the simple linear trend model? How about looking at the adjusted R^2? Discuss.
7. Your friend, Roman, is not convinced that the segmented trend model is accurate. He argues that 1938 is an arbitrary year and that you can improve the model by allowing for a more flexible trend.

To address this, he suggests estimating a model with a quadratic trend. In particular, you will estimate a model of the form:

$$yield_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \epsilon_t.$$

Hint: `R` can automatically create the squared trend in your regression by simply including the syntax `I(trend()^2)` as an additional variable in the `TSLM` function.

8. In a last-ditch effort to convince you, Roman called on Ryan as an "expert witness" (as if this were a trial, sigh). Ryan suggests that you use sine and cosine functions to capture the nonlinear trend and structural breaks in the data. He argues that this would allow for an even more flexible trend and eliminate the need to know the structural breaks (and whether they were smooth or sharp).

You decided to indulge them and estimate a model with Fourier terms. In particular, you will estimate a model of the form:

$$yield_t = \beta_0 + \beta_1 \cdot t + \gamma_1 \cdot \sin\left(\frac{2\pi t}{T}\right) + \gamma_2 \cdot \cos\left(\frac{2\pi t}{T}\right) + \gamma_3 \cdot \sin\left(\frac{2\pi 2t}{T}\right) + \gamma_4 \cdot \cos\left(\frac{2\pi 2t}{T}\right) + \epsilon_t.$$

where T is the number of observations in the data.

Hint: Notice that this is slightly different from the seasonal version in the notes but can be created using similar logic. When creating your variables, you can use `n()` function to get the total number of observations, T

Report your model fit and present the plot of the actual (presented as points) and fitted values over time. Does it appear that this model fits the data better than the simple linear trend model? How about compared to the segmented trend model? Discuss both from a visual and an adjusted R^2 perspective.

9. In a single step, use the `model` command to reestimate all four (4) models. Call your models `Trend`, `Segmented`, `Flexible`, and `Fourier`, respectively. Next, use the `glance` function to compare the models based on (i) adjusted R^2, (ii) `AIC`, and (iii) `BIC`. Which model is preferred based on each criterion? Be sure to explain your answer.

Present the model selection results in a table using the `knitr::kable()` function. Round your values to 3 digits and express in `,000s` using the `big.mark` argument.

10. Using the `forecast()` function, produce the predictions for the next ten (10) years using the model preferred by the `AIC` criterion. Produce a plot of these results. Set your `level = 95`. You will need to use the `new_data` argument to produce the forecasts since you must provide future values for your `d38` dummy.

Be sure to include the actual data in the plot.

Question 2: Revisiting the US Finished motor gasoline product supplied

In the last homework, we explored using the four (4) basic models to forecast the US finished motor gasoline product supplied. Additionally, we observed that the `mean` model was deemed to be the best at forecasting the data.

Unfortunately, none of the models were able to capture the trend and seasonality in the data simultaneously. Here enters our time series regressions.

1. Using the codes from the last homework, recover the `gas.ts` variable up to Dec 31, 2023. In effect, you are repeating steps (a) and (b) from Q#2.

2. Next, produce a plot of the data along with the `ACF` (with a maximum lag of 3 years) to get reacquainted with the data.

3. In a single step, use the model() and TSLM() functions to fit the following models to the `gas.ts` data:

i. a model with a trend and seasons.
ii. a model with a trend, squared trend, and seasons.
iii. a seasonal naive model.
iv. a RW model.
v. a RW model with drift.
vi. a mean forecast model.

Store the model fits as `mod.fit`. Be sure to give your models appropriate names here.

4. Using the `report()` function, print the model summary for the model (i) in part 3 above.

Hint: Remember that you will need to employ the `select` function at some point here.

5. Interpret the coefficient on the `intercept`, `trend` and `season()year24` variables, respectively. **In your explanations, remember to pay attention to the exact units in which our dependent variable is measured.**

6. Based on your results in part 4, what would the model with trend and seasonality predict as the average value for the 36th week of the year, holding all other factors constant?

7. Visualize the model fits in `mod.fit` against the actual data. **Be sure that each series is appropriately labeled in your legend.**

8. From the visuals above, which of your TSLM (regression) models appears to do a better job of predicting the data? Also, use the `glance()` function and any three statistics you deem necessary to bolster your conclusion from your "eyeball test". Be sure to present your results in a table using the `kable` function and express your values in 3 decimal places and in '000s.

9. Conduct a diagnostic test of the residuals from the "preferred model above" and comment on your observations from each graph of the `gg_tsresiduals()` function.