AAEC4484/AAEC(STAT)5484: Applied Economic Forecasting

Your Name Here

Homework #3 - Spring 2025

Instructions: Where necessary ensure that your graphs and visuals have proper titles and axis labels. Refer to the output, whenever appropriate, when discussing the results. **Creativity (coupled with relevance) will be rewarded.**

This week, our emphasis will be on time series regressions. We will explore the idea of segmented trends and how they can be used to capture structural breaks in the data. We will also revisit the US finished motor gasoline product supplied and use time series regressions to forecast the data.

Question 1: Regressions with Segmented Trends

> Historical trends of grain yield improvement offer us a glimpse of yields yet to come, although, like the stock markets, past performance is no guarantee of the future. The historical yield data for corn in the U.S. illustrate the positive impact of improved crop genetics and improved crop/soil management practices. American farmers grew open-pollinated corn varieties until the rapid adoption of hybrid corn began in the late 1930's. From 1866, the first year USDA began to publish corn yield estimates, through about 1936, yields of open-pollinated corn varieties in the U.S. were fairly stagnant and only averaged about 26 bu/ac (1.6 MT/ha) throughout that 70-year period. (Source: Corn Yield Trends)

The excerpt above provides a precursor to this week's assignment. Our regression module highlighted the use of a trend in forecasting historical data. We noted that the time trend can capture improvements in technology, productivity, changes in consumer preferences, and other structural changes in the economy. However, the trend is often assumed to be linear, which may not always be the case. In this question we will explore various functional forms for the trend, including segmented trends, to capture structural breaks in the data.

To begin, I have pulled US corn yields (**in bushels per acre**) data from the USDA NASS Quickstat database and stored it in the `corn_yield.csv` file in the `Data` folder on our GitHub page.

1. Pull the data into `R` **directly from GitHub**, keeping only the **national** corn yield (in `BU/ACRE`) for the `Period = YEAR`. Drop all columns, save for the `Year` and `Value` columns, and declare the data as a `tsibble` object with the appropriate `index`. Store the results in an object called `yield`.
2. Using the `ggplot()` and `geom_point()` functions, present a plot of the `yield` data over time. 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. Produce a gridded plot with two (2) panels:

(i) a plot of the actual (presented as points) and fitted values (presented as a line) over time. **Ensure that both series are appropriately labeled in the legend.**
(ii) a plot of the residuals (as points) over time.

Does it appear that this model fits the data very well? Why or why not? Be sure to link your discussions of both graphs back to period(s) where the model overestimated and/or underestimated the yield.

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. The article highlights further that "...the drought of 1936 sped the process of adoption after it revealed the drought resistance of hybrid corn. [...] After 1937, a new dynamic was set in motion. The explosion of demand for hybrid corn generated large profits for the major hybrid seed companies: Pioneer, Funk, and DeKalb."

To address this new development, we run the following regression for corn yields that allows for structural change before and after 1936:

$$yield_t = \beta_0 + \beta_1 \cdot t + \delta_0 \cdot d37 + \delta_1 \cdot t \cdot d37 + \epsilon_t.$$

Here $d37$ is a dummy variable that is equal to 1 beginning in 1937 and all subsequent periods and is zero otherwise. Including the $d37$ dummy allows for a change in the level (intercept) of the yield data while the interaction term between `t` and `d37` allows for a change in the slope.

For completeness, the dummy is defined as:

$$d37 = \begin{cases} 1; & t > 1936 \\ 0; & t \leq 1936 \end{cases}$$

Run and store the results of the regression. Recall that you will need to create the dummy variable `d37` before running the regression. **You are not required to print any results at this stage.**

6. Report all relevant regression results. Next repeat the tasks from part 4 above for this segmented trend model. Center your discussions on:

(i) whether this model appears (at least visually) to fit the data better than the simple linear trend model?
(ii) Which model is preferred based on the adjusted R^2?

7. **Segmented trend2:** The article argues that three distinct regimes exists, instead of the two we have modeled so far. The first regime is the period up to and including 1936, the second regime is the period from 1937 to 1955, and the third regime is the period from 1956 to the present.

To account for these regimes, we will estimate a segmented trend model with two (2) dummies, `ds37` and `ds56`, that allow for changes in the level and slope of the trend in 1938 and 1956, respectively. The model is of the form:

$$yield_t = \beta_0 + \beta_1 \cdot t + \delta_0 \cdot d37 + \delta_1 \cdot t \cdot d37 + \delta_2 \cdot ds56 + \delta_3 \cdot t \cdot ds56 + \epsilon_t.$$

For clarity, the $ds37$ dummy is defined as:

$$ds37 = \begin{cases} 1; & 1937 \leq t \leq 1955 \\ 0; & \text{otherwise} \end{cases}$$

and the $ds56$ dummy is defined as:

$$d56 = \begin{cases} 1; & t > 1955 \\ 0; & t \leq 1955 \end{cases}$$

Conduct this regression and present the graph of the fitted values (as a line) and actual values (as points) over time. Also, present the residuals over time. Does this model appear to offer a vast improvement over the segmented trend model?

8. In class, we mentioned that Fourier terms can be used to capture nonlinear trends and potential structural breaks in the data. The use of these sine and cosine functions can allow for an even more flexible trend and eliminate the need to know the dates of the structural breaks (and whether they were smooth or sharp).

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 the `n()` function to get the total number of observations, T

Along with your regression model results, present:

(i) the plot of the actual (presented as points) and fitted values over time.
(ii) the plot of the residuals from this model. Does it appear that this model fits the data better than the simple linear trend model? How about compared to the two segmented trend models? 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`, `Segmented1`, `Segmented2`, 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 under 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 `,000`s 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`. Recall that you will need to use the `new_data` argument to produce the forecasts since you must provide future values for your dummy variable(s).

Be sure to include the actual data in the plot.

Question 2: US Finished motor gasoline product supplied (Revisted)

In the last homework, we explored using the four (4) basic models to forecast the US finished motor gasoline product supplied. Although, the seasonal naive model emerged as the "best" model based on the `AIC` and `BIC` criteria, it was evident that the model was not capturing the trend in the data. In fact, none of the models could simultaneously capture the trend and seasonality in the data.

1. Using the codes from the last homework, recover the `gas` variable up to Dec 31, 2024. In effect, you are repeating steps (a) and (b) from Q2.
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` 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.

Hint: A squared trend can be created using the `I()` function in `R`. For example, `I(trend()^2)` will create a squared trend. 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 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()year9` 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 November each 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 **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 on the residuals from the "preferred model above" and comment on your observations from each graph of the `gg_tsresiduals()` function. Use a "Ljung-Box" test to support your observations. Be sure to discuss the hypothesis and your conclusion.
10. Using the `forecast()` function, produce the predictions for the next three (two) years using the model preferred by the `AIC` criterion. Produce a plot of these results along with the original data. Set your `level = 95`.

Hint: Since all your variables were created using the built-in `trend()` and `season()` functions, you do not need to create any new variables when you forecast into the future. Isn't that nice?!?