ECON144 Lab1

Andrew Grove and Thomas Santosa

304785991 and 504797813

Part I.

The data we have chosen is the amount of pork produced in the United States, sourced from the Industrial Production Index. The units of these values are in percentages, where we use 2012 as the base year. Thus, for example, a value of 62.1 means that that month's production was 62.1% of the average amount of pork produced in a month in the year 2012. These values are input every month without seasonal adjustment. For the purposes of this assignment, and to better see seasonality, we have decided to look at only the data from 2000 onwards.

The reason for this seasonality is due to the weather. Note that the data peaks in the autumn and hits its lowest point in the summer months. The usual cycle for pork production is that the pigs breed better in cooler temperatures. This causes them to breed more in said months, and them being born in the spring / summer. They then are given food, and will grow up in time for a harvest in the next fall / winter. In contrast, they breed less it higher temperatures, and thus a smaller amount are bred in spring and summer. Naturally, for the opposite reason as above, less pigs are then able to be harvested in the next spring and summer.

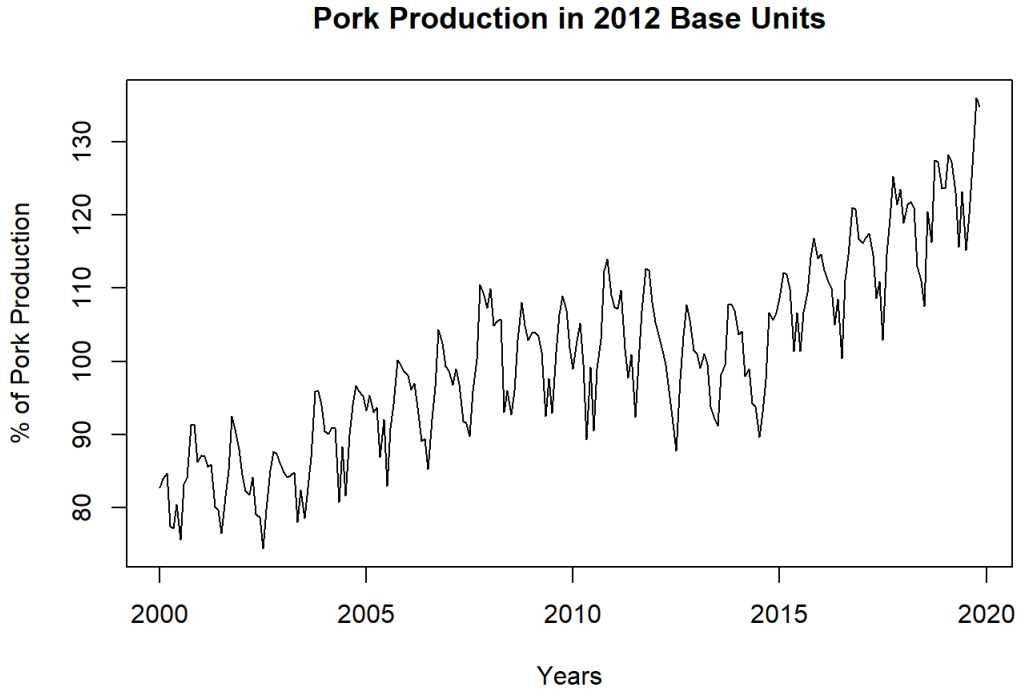
Additionally, pigs end up eating less, and thus growing slower, in very hot temperatures. This causes pigs to be undersized during the summer, and many ranch owners will postpone harvesting until the autumn season so that their pigs grow fat enough to produce a larger amount of meat. These factors combined causes the graph shown below.

Part II.

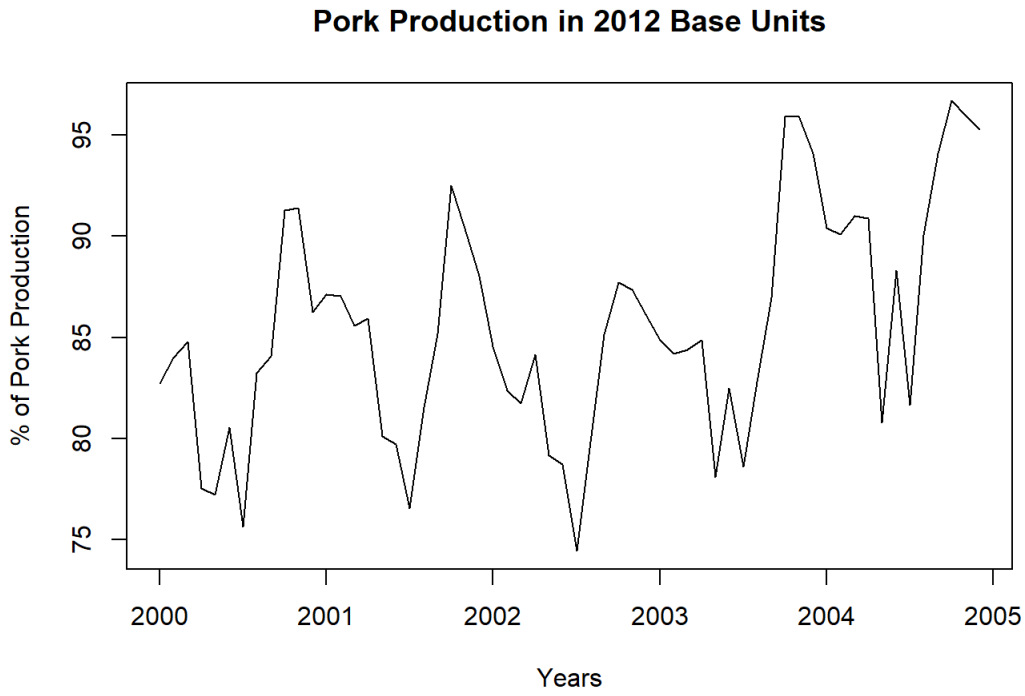
Section 1.

a.

The desired time series is shown below:



To better see the seasonality, we can zoom in on a specific period of time, e.g. 2000 to 2005:

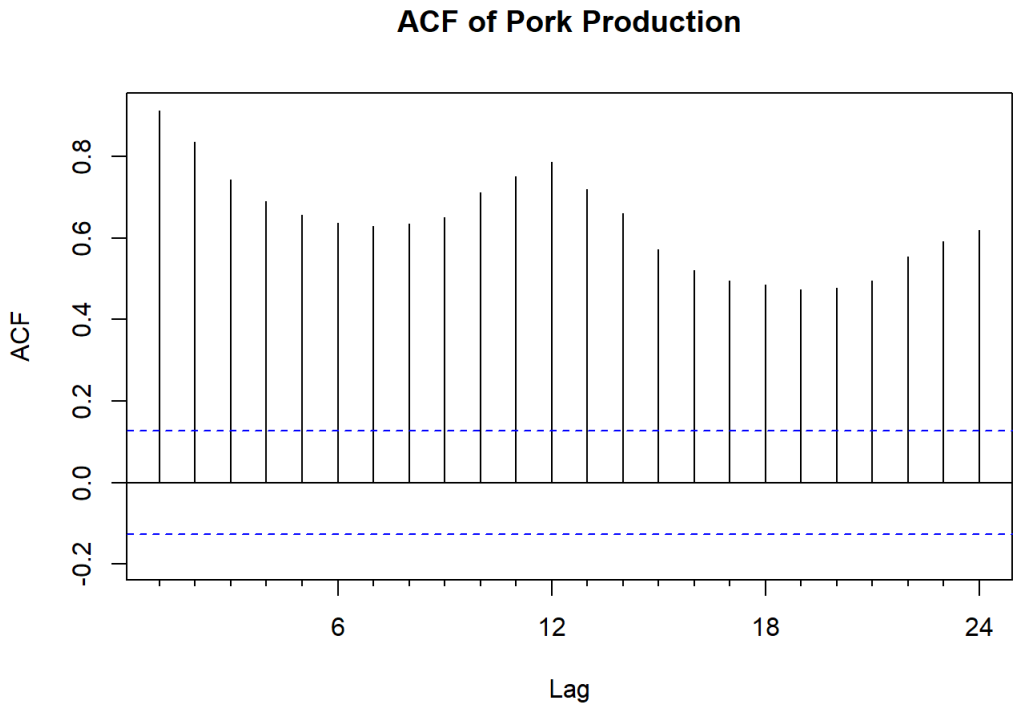


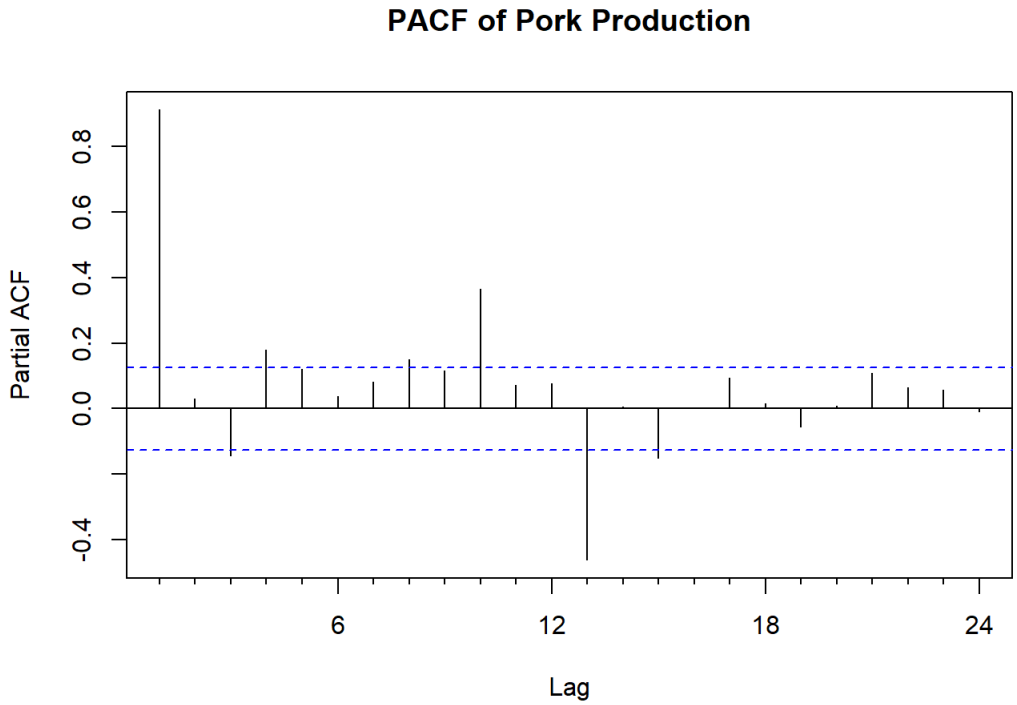
b.

No, the plot suggests that the data is not stationary, in any order. We can see extremely clear seasonality and a clear slight positive trend over the years. By the definition of stationary, any order stationary data must have the same mean, which this plot clearly doesn't. The covariance is also clearly dependent on time and thus $Cov[X_n, X_{n-j}] = r_j$ is dependent on n.

c.

The ACF and PACF plots are shown below:





As we can see from the ACF plot, there is clearly statistically significant time dependence in the data. For every lag from 1 to 24 lags, there is an extreme amount of autocorrelation, meaning that, on average, there is a significant amount of correlation between x_n and x_{n-j} where x_n is the nth data point, and j is the lag difference. The PACF plot does not have the same extremes amount of correlation as ACF does, while it does have a few spikes. This can be interpreted such that when the intermediary time steps are removed, most time gaps are not too autocorrelated with each other, especially relative to the very significant correlation found in the ACF.

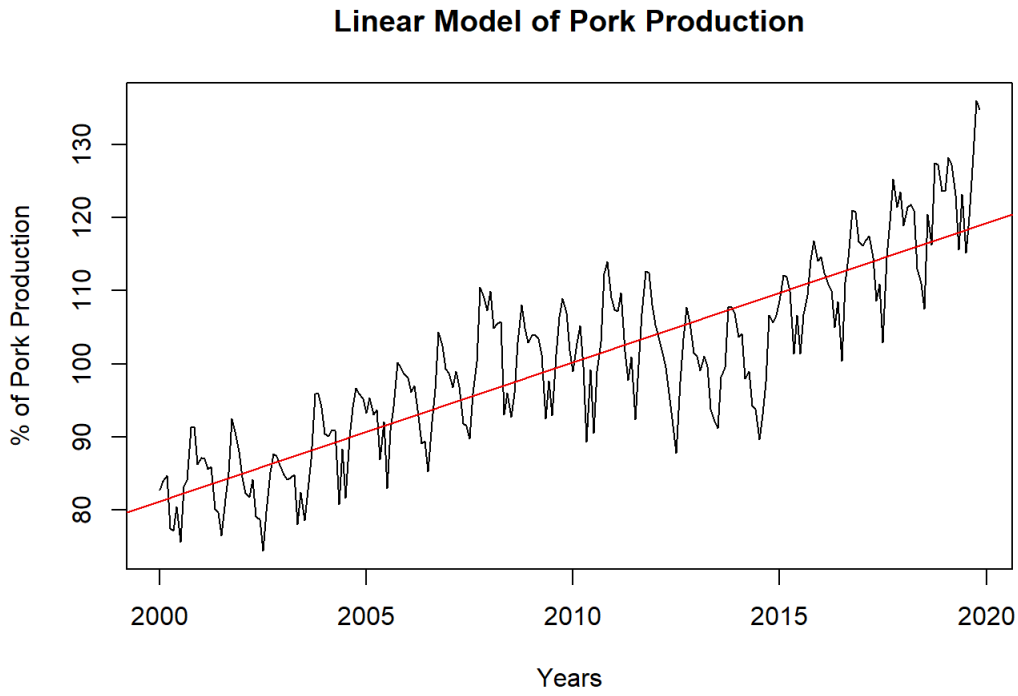
d.

Linear model:

As we can see here our linear model only represents the trend of the data, excluding any of the seasonality that occurs within the data set. It is also very clear that our dataset has an upward trend. We use this equation to represent the linear fit:

$$Data_t = \beta_0 + \beta_1 Years_t$$

```
##
## Call:
## lm(formula = data[, 3] ~ years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2248  -4.0988   0.4624   4.6218  17.0459
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.717e+03  1.491e+02  -24.92   <2e-16 ***
## years        1.899e+00  7.419e-02   25.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.595 on 237 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7332
## F-statistic: 655.1 on 1 and 237 DF,  p-value: < 2.2e-16
```



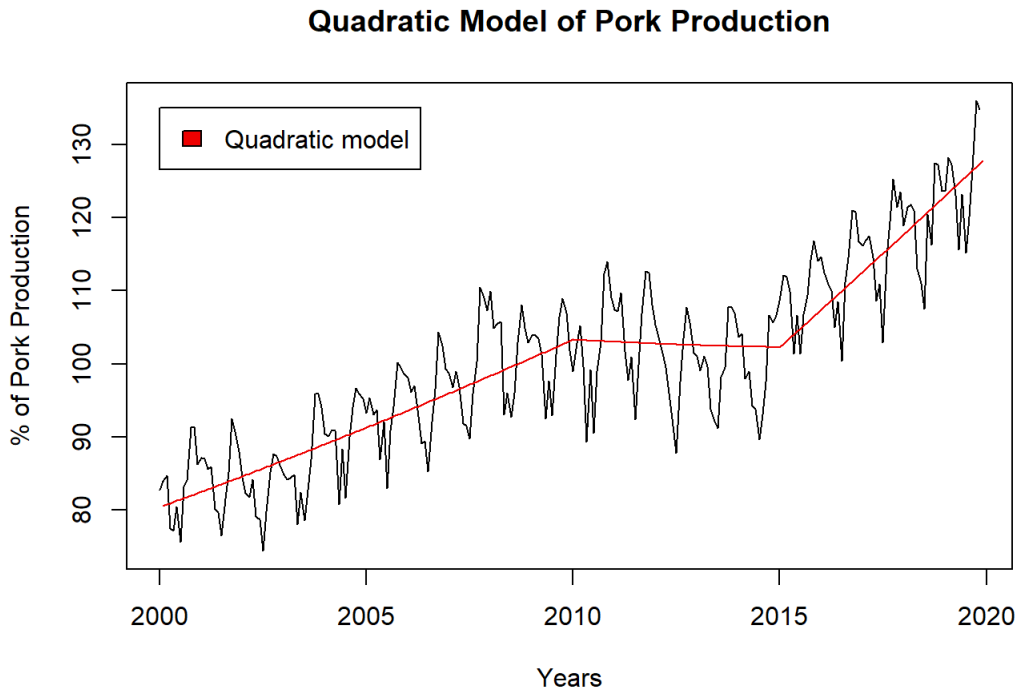
Nonlinear model:

a.) Quadratic:

This quadratic model, with breaks included in 2010 and 2015, certainly represents the data better than the linear model. We chose to add the breaks in both 2010 and 2015 because we can see that after around 2010 because there is a constant (or slightly downward) trend in the data. This might be caused by the economy crisis in 2008 which decreases pork production. This slightly downward trend ends at around 2015 in which the data shows an increasing trend again. In this model, we represent the data using this equation, in which tb_1 and tb_2 are just break variables.

$$Data_t = \beta_0 + \beta_1 Years_t^2 + \beta_2 Years_t + \beta_3 tb_1 + \beta_4 tb_2$$

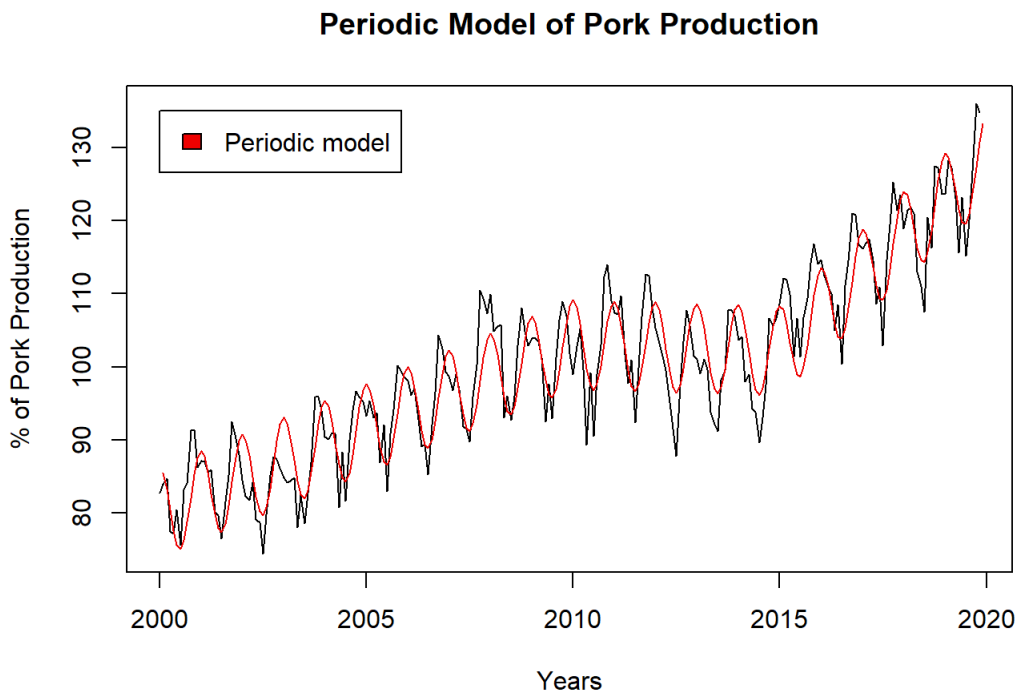
```
##
## Call:
## lm(formula = data[, 3] ~ time2 + years + tb1 + tb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8708  -3.9228   0.7526   4.1529  12.5342
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.750e+04  2.516e+05   0.348   0.7283
## time2        2.289e-02  6.257e-02   0.366   0.7149
## years       -8.948e+01  2.509e+02  -0.357   0.7217
## tb1         -2.838e+00  1.099e+00  -2.583   0.0104 *
## tb2          5.168e+00  7.770e-01   6.652 2.03e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.801 on 234 degrees of freedom
## Multiple R-squared:  0.797,  Adjusted R-squared:  0.7935
## F-statistic: 229.7 on 4 and 234 DF,  p-value: < 2.2e-16
```



b.) Periodic:

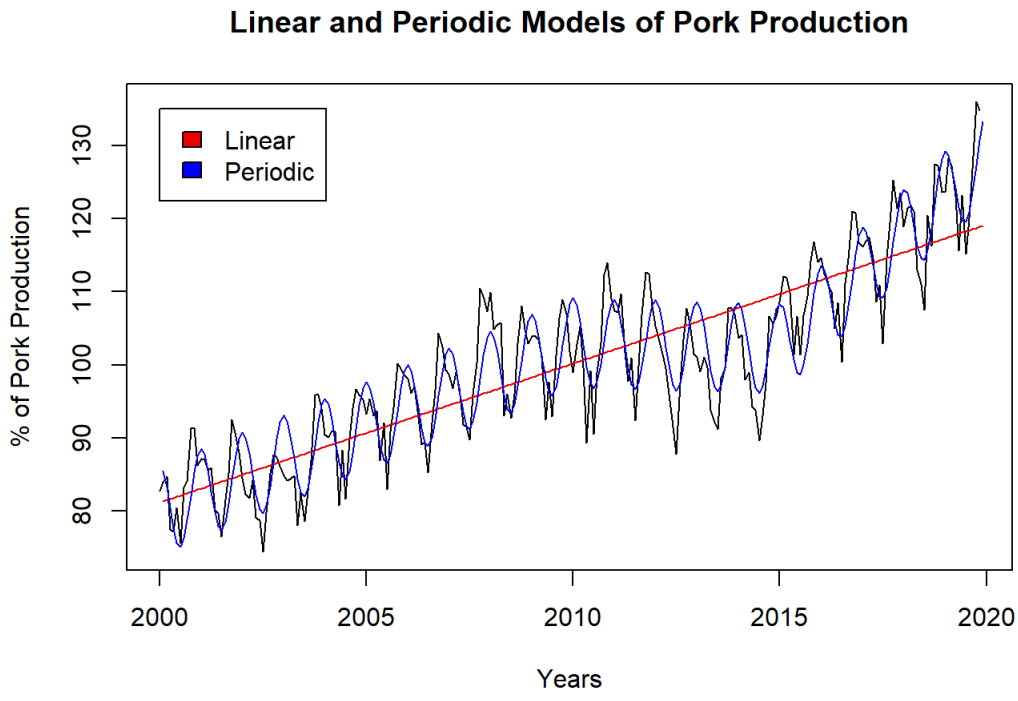
In this part, we added a periodic function to our previous model because it seems that there is some periodicity and seasonality in the actual dataset. We chose to combine both linear with breaks and periodic, not the quadratic, because it seems that the quadratic model with breaks does not seem to be statistically significant with marginally different AIC and BIC, and we want to keep the model as simple as possible. By adding the periodicity to the linear model, we can see from the plot below that our fit is significantly better in representing the dataset. we use this equation to represent the dataset:

$$Data_t = \beta_0 + \beta_1 Years_t + \alpha_1 sin(2\pi Years_t) + \alpha_2 cos(2\pi Years_t) + \beta_3 tb_1 + \beta_4 tb_2$$



```
##
## Call:
## tslm(formula = data.ts ~ years + I(sin(2 * pi * years)) + I(cos(2 *
##     pi * years)) + I(ts(pmax(0, years - tbreak1), start = 2000)) +
##     I(ts(pmax(0, years - tbreak2), start = 2000)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6558 -2.5174  0.1411  2.5240  9.8673
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                -4501.5089   208.6702 -21.572
## years                                          2.2908     0.1040  22.018
## I(sin(2 * pi * years))                        -0.2445     0.3505  -0.698
## I(cos(2 * pi * years))                         6.1143     0.3515  17.395
## I(ts(pmax(0, years - tbreak1), start = 2000))    -2.4515     0.2696  -9.094
## I(ts(pmax(0, years - tbreak2), start = 2000))     5.3702     0.4169  12.882
##                                            Pr(>|t|)
## (Intercept)                                  <2e-16 ***
## years                                        <2e-16 ***
## I(sin(2 * pi * years))                        0.486
## I(cos(2 * pi * years))                       <2e-16 ***
## I(ts(pmax(0, years - tbreak1), start = 2000))   <2e-16 ***
## I(ts(pmax(0, years - tbreak2), start = 2000))   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.834 on 233 degrees of freedom
## Multiple R-squared:  0.9117, Adjusted R-squared:  0.9098
## F-statistic: 481.3 on 5 and 233 DF,  p-value: < 2.2e-16
```

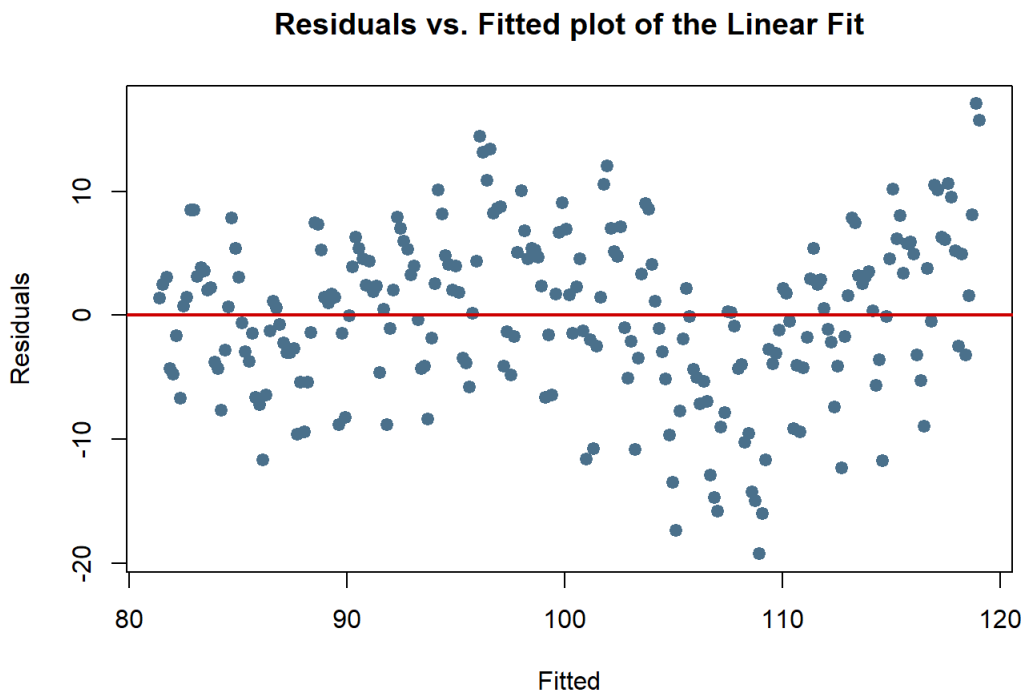
The plot with the linear and chosen nonlinear (periodic) model is shown below:



e.

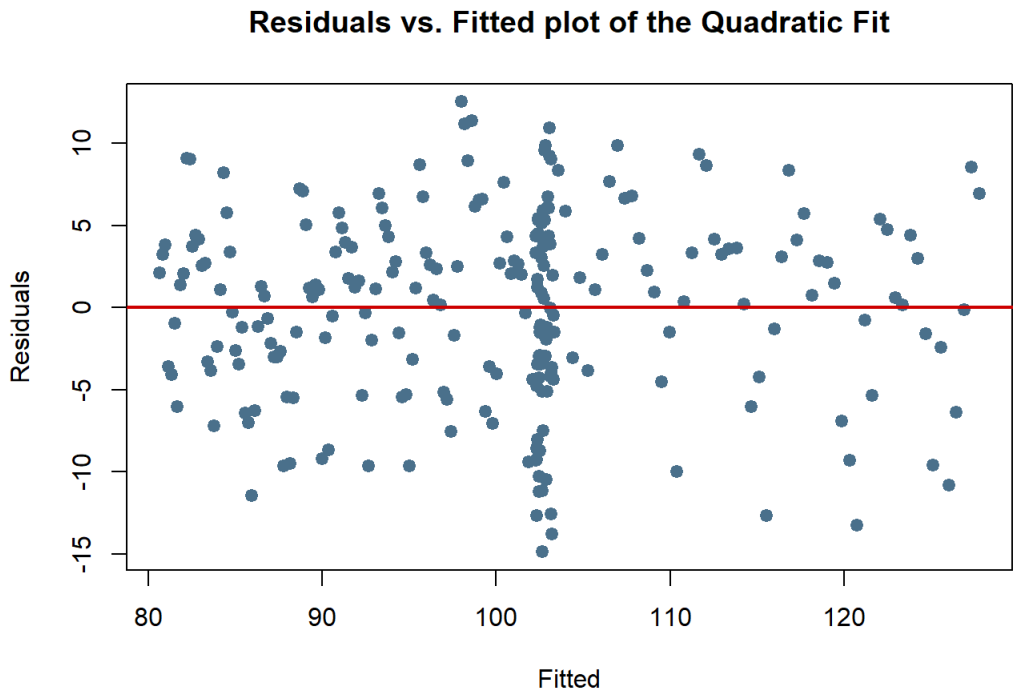
Linear fit residuals vs. fitted plot:

It is very clear that the residuals of this model have a certain pattern and not randomly distributed around 0. This conclude that this model is a not good enough model in representing out dataset.



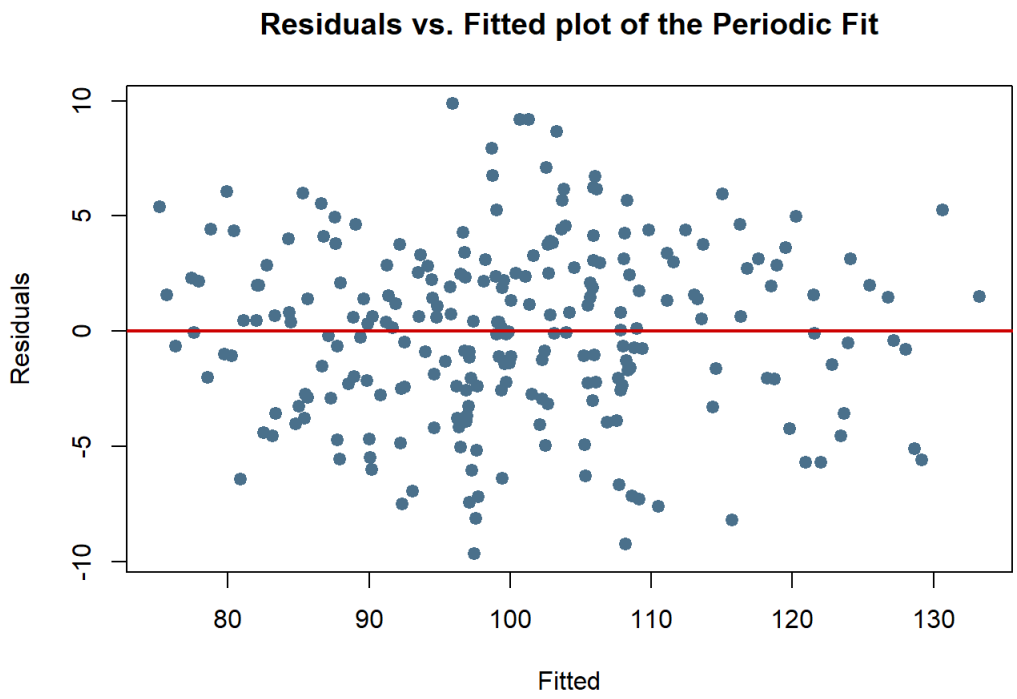
Quadratic fit residuals vs. fitted plot:

Similar to that of Linear fit residuals vs. fitted plot, this model has a certain pattern, which is there seem to be a cluster of points centered in around the 100 fitted value. This shows that this model is not a good enough model to represent the dataset.



Periodic fit residuals vs. fitted plot:

Certainly we can see that this model is significantly better because the plot seems to show a random distribution around the 0 line, which shows that this model is a good enough representation of our dataset.

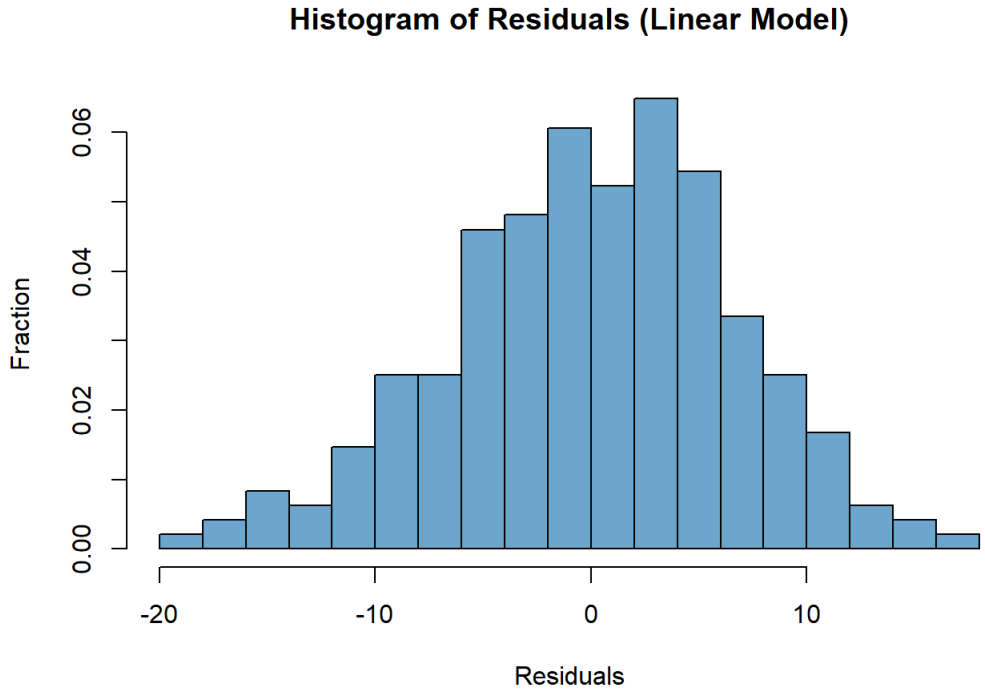


It seems that the best model to represent the data is the Periodic model. We can deduce this by looking at the residuals vs. fitted value. The first two plots seems to have a certain pattern that resembles the data, meaning that they are not a good model to represent the actual data. The last model, however, seems to be evenly distributed around 0, which shows that this is a good enough model in representing the actual data.

f.

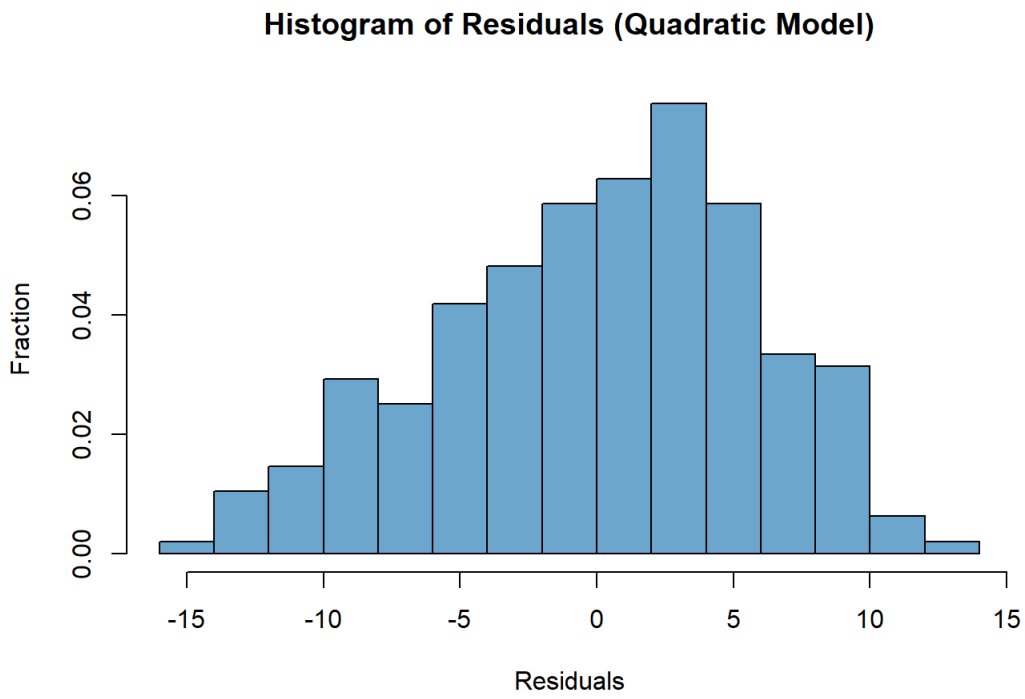
Linear fit residuals histogram:

Though the histogram of residuals of this model fairly resembles a normal distribution, we can see that the median of the residuals is not located in the 0 point. This show that a linear fit can be improved to a better model.



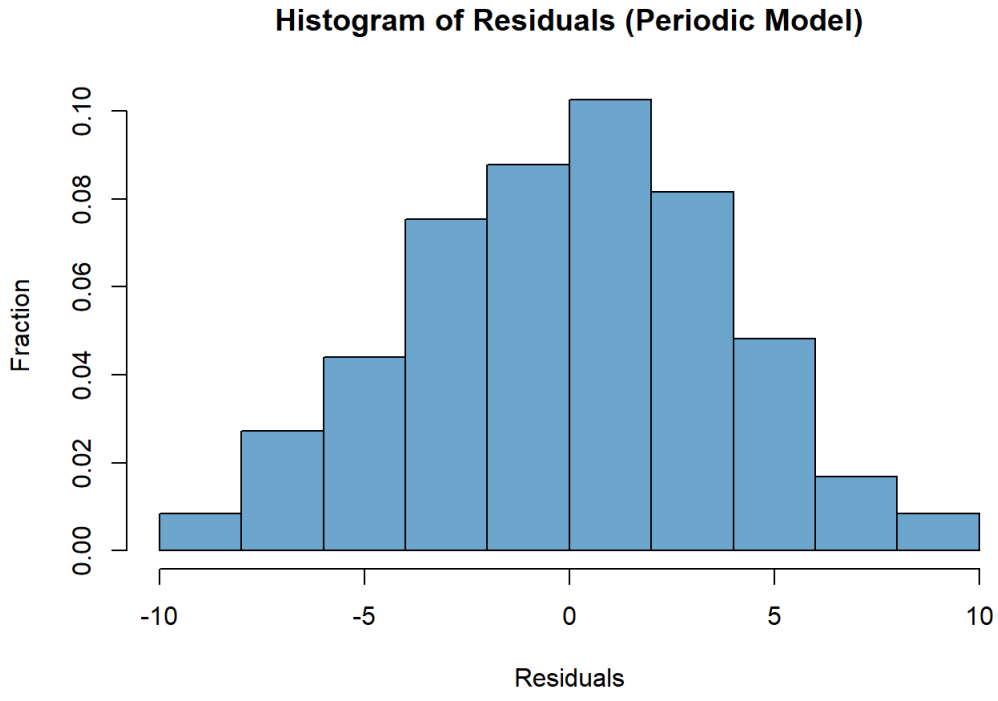
Quadratic fit residuals histogram:

Consistent with our argument of the residuals vs. fitted value plot, the quadratic model shows a histogram of residuals that although resembling a normal distribution, it is not a good enough fit to model out dataset. This came from the fact that the histogram is slightly positively skewed.



Linear + Periodic fit residuals histogram:

Compared to the first two models, the linear and the quadratic, the linear + periodic model has a very good histogram of the residuals. we can see that this histogram resembles a normal distribution the most, which makes us conclude that this model is the best among the three models.



We can see that all of the histogram of the residuals resembles a normal distribution, which means that the models represent the data good enough. But among these three models, the third one, which is the Linear + Periodic fit model, is the best model to represent the data, judging from the histogram of the residuals, because the third histogram resembles normal distribution the most. The first and the second histogram, which are those of Linear and Quadratic fit residuals, also resembles a normal distribution, but they do have some right skewness, respectively. The third one, however, also have a slight left skew, but it is very slight compared to those of the first two histrograms.

g.

Linear fit regression summary:

```
##
## Call:
## lm(formula = data[, 3] ~ years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2248  -4.0988   0.4624   4.6218  17.0459
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.717e+03  1.491e+02  -24.92   <2e-16 ***
## years        1.899e+00  7.419e-02   25.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.595 on 237 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7332
## F-statistic: 655.1 on 1 and 237 DF,  p-value: < 2.2e-16
```

We first note that the R-squared value and adjusted R-squared value is relatively low, with a 73%. This means that 73% of the variation found in pork production from 2000 to 2020 can be explained by the model. We note however though that according to the t-tests, the coefficient associated with years is statistically significant, i.e. that we reject the null hypothesis that number of years has no effect on pork production. Lastly, The F-statistic is the same as the t-test in this case, since it only has one variable, and once again it shows that we reject the null hypothesis that number of years has no effect on pork production.

Quadratic fit regression summary:

```
##
## Call:
## lm(formula = data[, 3] ~ time2 + years + tb1 + tb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8708  -3.9228   0.7526   4.1529  12.5342
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.750e+04  2.516e+05   0.348   0.7283
## time2        2.289e-02  6.257e-02   0.366   0.7149
## years       -8.948e+01  2.509e+02  -0.357   0.7217
## tb1         -2.838e+00  1.099e+00  -2.583   0.0104 *
## tb2          5.168e+00  7.770e-01   6.652 2.03e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.801 on 234 degrees of freedom
## Multiple R-squared:  0.797,  Adjusted R-squared:  0.7935
## F-statistic: 229.7 on 4 and 234 DF,  p-value: < 2.2e-16
```

We first note that the R-squared value and adjusted R-squared value is relatively low, with a 80%. This means that 80% of the variation found in pork production from 2000 to 2020 can be explained by the model. Due to the addition of break variables (tb1 and tb2) and the addition of time2 = years^2, each respective component becomes less significant. As such, the t-test results are inconclusive, where it seems that all variables except for the break variables are not statistically significant. The F-statistic however shows that we have statistically significant evidence that at least one of the coefficents is not zero, and thus at least one of these has an effect on pork production.

Periodic fit regression summary:

```
##
## Call:
## tslm(formula = data.ts ~ years + I(sin(2 * pi * years)) + I(cos(2 *
##     pi * years)) + I(ts(pmax(0, years - tbreak1), start = 2000)) +
##     I(ts(pmax(0, years - tbreak2), start = 2000)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6558 -2.5174  0.1411  2.5240  9.8673
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                -4501.5089   208.6702 -21.572
## years                                          2.2908     0.1040  22.018
## I(sin(2 * pi * years))                        -0.2445     0.3505  -0.698
## I(cos(2 * pi * years))                         6.1143     0.3515  17.395
## I(ts(pmax(0, years - tbreak1), start = 2000))    -2.4515     0.2696  -9.094
## I(ts(pmax(0, years - tbreak2), start = 2000))     5.3702     0.4169  12.882
##                                            Pr(>|t|)
## (Intercept)                                  <2e-16 ***
## years                                        <2e-16 ***
## I(sin(2 * pi * years))                        0.486
## I(cos(2 * pi * years))                       <2e-16 ***
## I(ts(pmax(0, years - tbreak1), start = 2000))   <2e-16 ***
## I(ts(pmax(0, years - tbreak2), start = 2000))   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.834 on 233 degrees of freedom
## Multiple R-squared:  0.9117, Adjusted R-squared:  0.9098
## F-statistic: 481.3 on 5 and 233 DF,  p-value: < 2.2e-16
```

We can see that unlike the previous models, this has a much higher R-squared value at around 91%. Once again, this means that 91% of the variation found in pork production is explained by the model. Additionally, we can see that most of our variables are statistically significant according to t-tests, with the only one not being statistically signicant being the sin() period. This means that we are able to reject the null hypothesis that these variables are equal to 0, and thus we can have some evidence to say that each of the four statistically significant variables are correlated with pork production. Lastly, the F-statistic naturally only indicates a similar sentiment, where we have statistically significant evidence that at least one of the coefficients of these variables is not 0.

h.

We can see that from the AIC and BIC oberved below, the best model is the third model. We selected this model by choosing the one with the lowest AIC and BIC, and the third model, which is the y1 model with the Linear and Periodic fit, has the lowest both AIC and BIC values.This is consistent with our aforementioned arguments regarding which model is the best in representing the data.

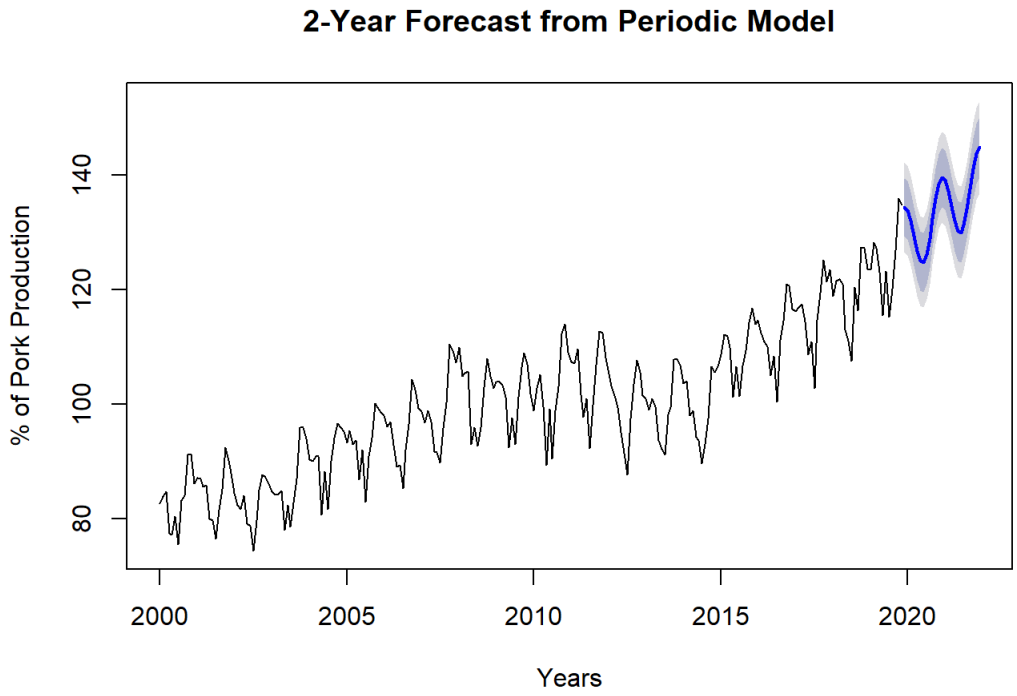
y
y_quad
y1

3 rows | 1-1 of 3 columns

y
y_quad
y1

3 rows | 1-1 of 3 columns

g.



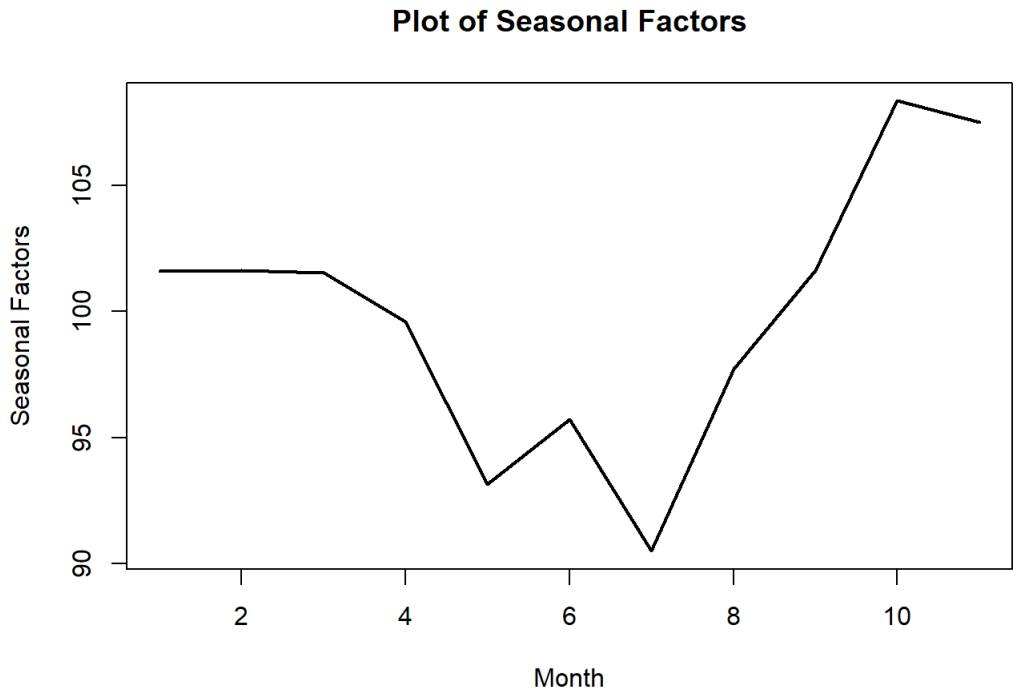
Section 2.

a.

The current model (linear with breaks) with a full set of seasonal dummies is as follow:

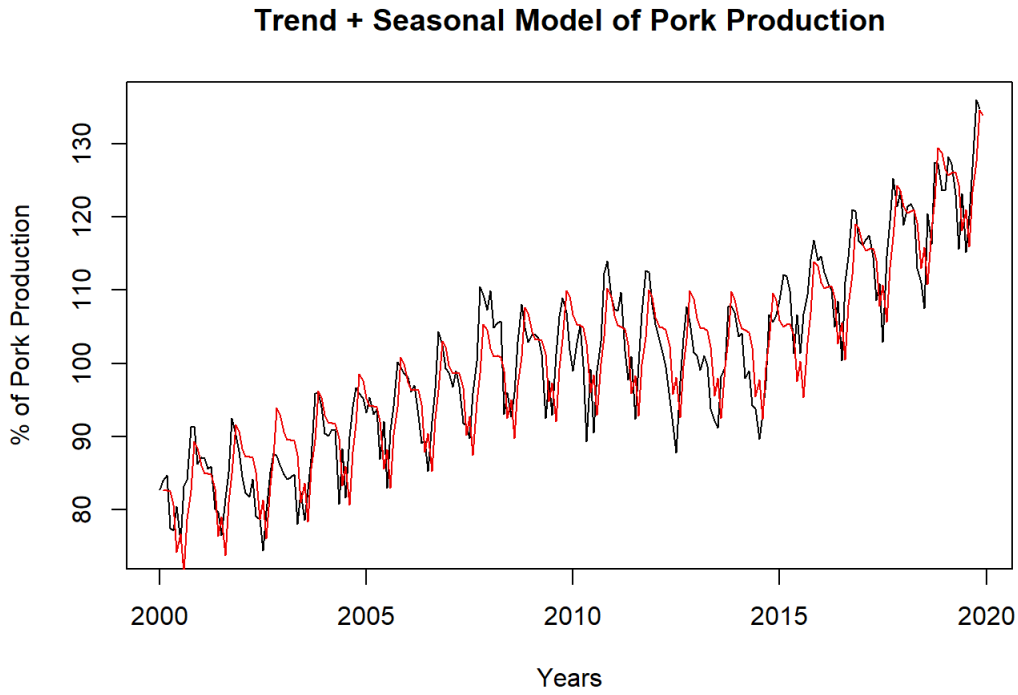
```
##
## Call:
## tslm(formula = data.ts ~ dummies + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.060  -7.442   0.320  10.906 123.579
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## dummiesJan  101.598      7.207   14.10   <2e-16 ***
## dummiesFeb  101.638      7.207   14.10   <2e-16 ***
## dummiesMar  101.525      7.207   14.09   <2e-16 ***
## dummiesApr   99.597      7.207   13.82   <2e-16 ***
## dummiesMay   93.149      7.207   12.93   <2e-16 ***
## dummiesJun   95.730      7.207   13.28   <2e-16 ***
## dummiesJul   90.501      7.207   12.56   <2e-16 ***
## dummiesAug   97.721      7.207   13.56   <2e-16 ***
## dummiesSep  101.644      7.207   14.10   <2e-16 ***
## dummiesOct  108.362      7.207   15.04   <2e-16 ***
## dummiesNov  107.532      7.207   14.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.23 on 228 degrees of freedom
## Multiple R-squared:  0.9029, Adjusted R-squared:  0.8982
## F-statistic: 192.7 on 11 and 228 DF,  p-value: < 2.2e-16
```

b.

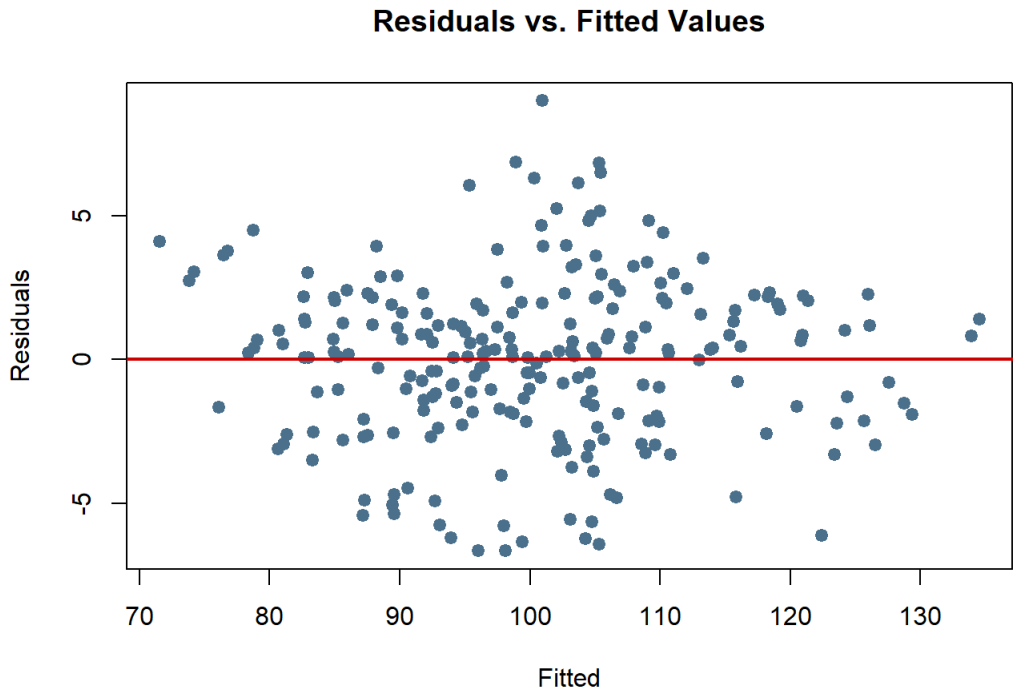


We can see that pork production seems to spike in the autumn, particularly in the month of October. It remains very quite high allthrought winter, before declining, reaching its trough in July.

c.



For this section, we choose to instead of using the periodic model, to instead use the linear model with breaks by itself, and add the seasonality to that. This is because our seasonality accounts for the periodic movement, and thus the periodic portion becomes redundant. This is further shown when running the regression with the periodic parts put in. If we do this, the coefficients for the sin and cos becomes very close to zero (NA), and has very little effect on the overall regression.



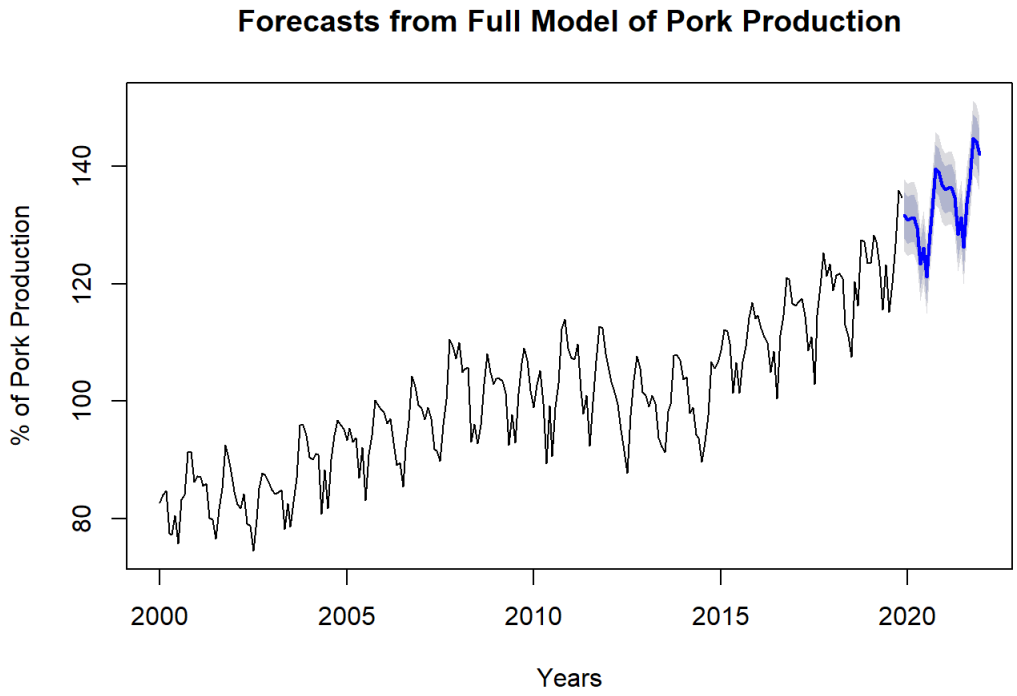
We can see in the plot itself that outside of a few anomalies in 2003, 2008, and a few other years, this model does a pretty good job of capturing the data. When plotting the residuals, we can see that most points are evenly spaced out around 0, and there is no clear trend or pattern that can be found. As such, we can say that the points are roughly normally distributed, and the residuals show no cause for alarm or reason to doubt our model.

d.

```
##
## Call:
## tslm(formula = data.ts ~ years + +I(ts(pmax(0, years - tbreak1),
##     start = 2000)) + I(ts(pmax(0, years - tbreak2), start = 2000)) +
##     season)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.649 -1.894  0.246  1.952  8.984
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                -4.490e+03  1.614e+02 -27.825
## years                                       2.286e+00  8.047e-02  28.414
## I(ts(pmax(0, years - tbreak1), start = 2000)) -2.433e+00  2.085e-01 -11.673
## I(ts(pmax(0, years - tbreak2), start = 2000))  5.309e+00  3.225e-01  16.463
## season2                                    -1.590e-01  9.375e-01  -0.170
## season3                                    -4.722e-01  9.375e-01  -0.504
## season4                                    -2.600e+00  9.376e-01  -2.773
## season5                                    -9.248e+00  9.376e-01  -9.863
## season6                                    -6.866e+00  9.377e-01  -7.322
## season7                                    -1.230e+01  9.378e-01 -13.111
## season8                                    -5.275e+00  9.379e-01  -5.625
## season9                                    -1.552e+00  9.380e-01  -1.654
## season10                                    4.967e+00  9.381e-01   5.294
## season11                                    3.937e+00  9.383e-01   4.196
## season12                                    1.284e+00  9.500e-01   1.352
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## years                                       < 2e-16 ***
## I(ts(pmax(0, years - tbreak1), start = 2000))  < 2e-16 ***
## I(ts(pmax(0, years - tbreak2), start = 2000))  < 2e-16 ***
## season2                                     0.86547
## season3                                     0.61500
## season4                                     0.00603 **
## season5                                     < 2e-16 ***
## season6                                    4.34e-12 ***
## season7                                     < 2e-16 ***
## season8                                    5.52e-08 ***
## season9                                     0.09952 .
## season10                                   2.85e-07 ***
## season11                                   3.92e-05 ***
## season12                                    0.17776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.965 on 224 degrees of freedom
## Multiple R-squared:  0.9493, Adjusted R-squared:  0.9461
## F-statistic: 299.3 on 14 and 224 DF,  p-value: < 2.2e-16
```

The R-squared seems to agree with the sentiment that it is a good model, since this model accounts for 94% of the variation found in the data. We also note that most of the seasonal dummies are statistically significant according to the p-values. The F-statistic found signifies that there is statistically significant evidence that at least one of the coefficients is not zero. We note that the residuals have a median of close to 0, thus backing up our claim that the distribution of the residuals is relatively normal. The standard error of the residual is not very high, signfiying that 95% of the variation in the data can be found six points away from our modeled line.

e.



Part III

We can conclude that our model works relatively well, especially considering we have no cycle parameter that we can use and were solely limited to trend and seasonality. We can see that we have an R-squared of 94% and are explaining a large proportion of the variability in the percentage of pork produced. We also note that the residual plot seems to be normally distributed, and as such, there does not seem to be any major issues that will massive throw off our forecast. We do note however there are some years whose peaks either fall short or surpass that of our prediction. These shocks seem to be random, and it would be difficult for our model to account for them and have them remain to be useful in our forecast.

We could improve this model in a variety of ways. Firstly, the trend line we use for the full model, i.e. linear with breaks, is difficult to extend into the future accurately. We would need further information on why said break occurs in order to better predict it happening again in the future. On a similar note, we could also improve this model by having information that explains why a certain year has higher peaks that others, or lower peaks than the others, e.g. 2008 and 2003 respectively. We also could use a more sophisticated seasonality component, making use of S-ARMA functions or similar to attempt to match the seasonality peaks and troughs better. We could also attempt to implement a cycle portion into the model, and see if that improves the model's accuracy.

Part IV

http://www.economagic.com/em-cgi/data.exe/frbg17/n311611t3p_ipnsa (Dataset)

https://cran.r-project.org/web/packages/forecast/forecast.pdf

https://www.rdocumentation.org/packages/forecast/versions/8.10/topics/tslm

https://www.rdocumentation.org/packages/forecast/versions/8.10/topics/seasonaldummy

https://stat.ethz.ch/R-manual/R-devel/library/base/html/AsIs.html

Part V