R Teams Assignment Report 3

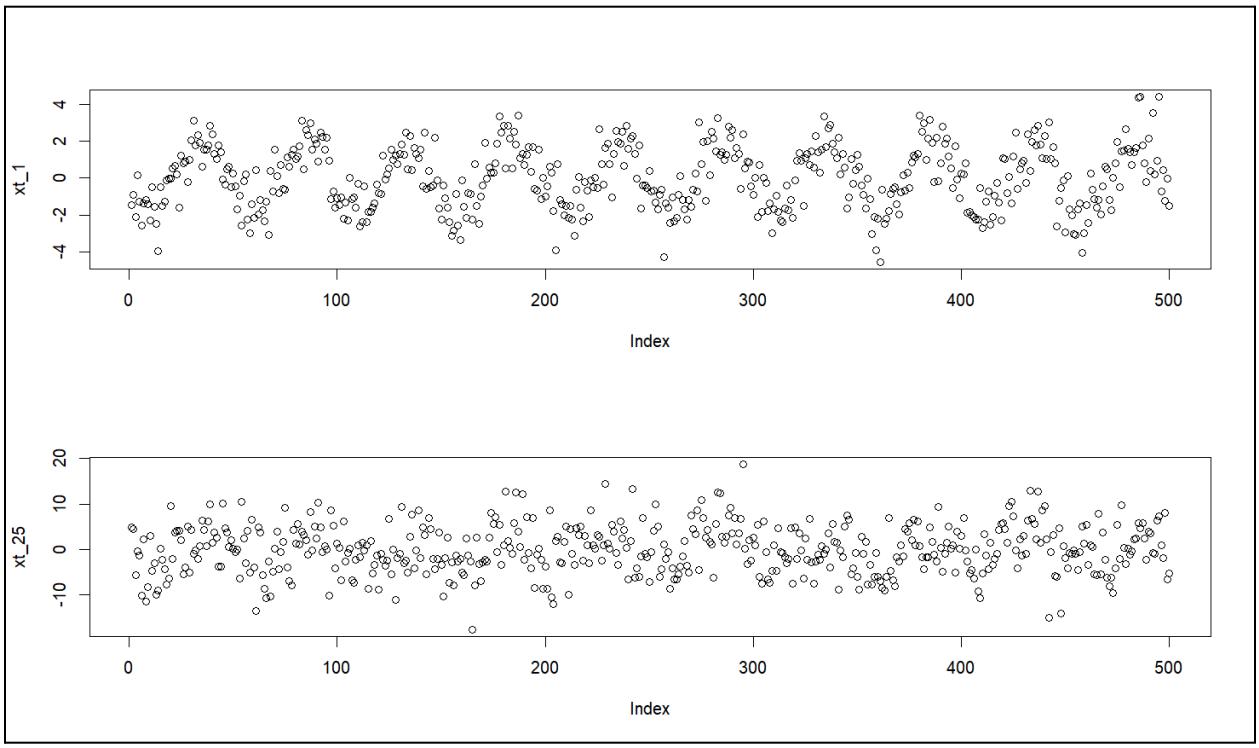
Roman Kolosok, Brian Myers, Nick Treadwell

Lebanon Valley College

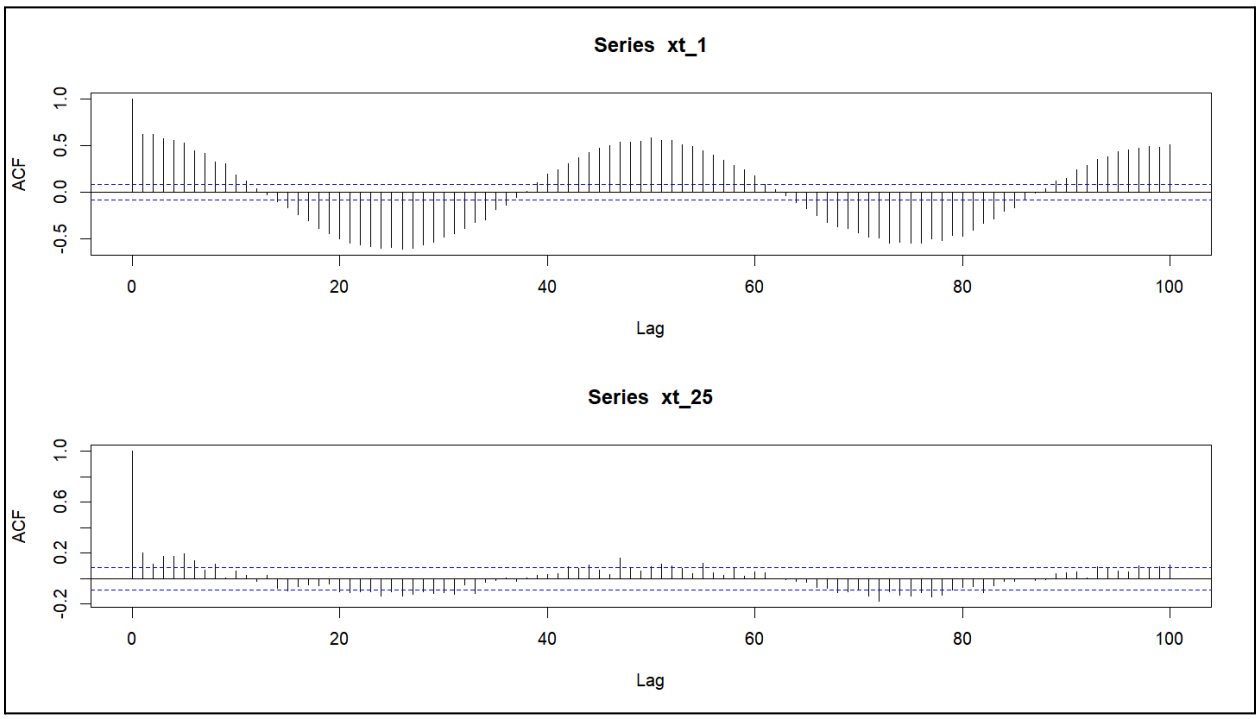
MAS-372: Statistical Modeling

Problem 23

In this problem, we simulate two different signal-plus-noise models to determine the effect that the size of the variance of the white noise process has on the model. The first model is simulated while setting the variance of the white noise process to 1. The following chart displays both series (variance 1 on top, variance 25 on bottom):



We note that the signal is more clearly visible in the plot with less variance, and the signal is less clear in the plot with greater variance. We now plot the sample autocorrelation functions for both series (plotted in the same order):



We observe that the sample ACF for the variance-1 series closely follows the typical sinusoidal pattern displayed by the theoretical ACF of a signal series. Though the variance-25 series generally follows a similar pattern, the autocorrelations are much smaller, and thus we could not use the sample ACF as significant justification for the claim that the data is modeled well by a signal-with-noise series.

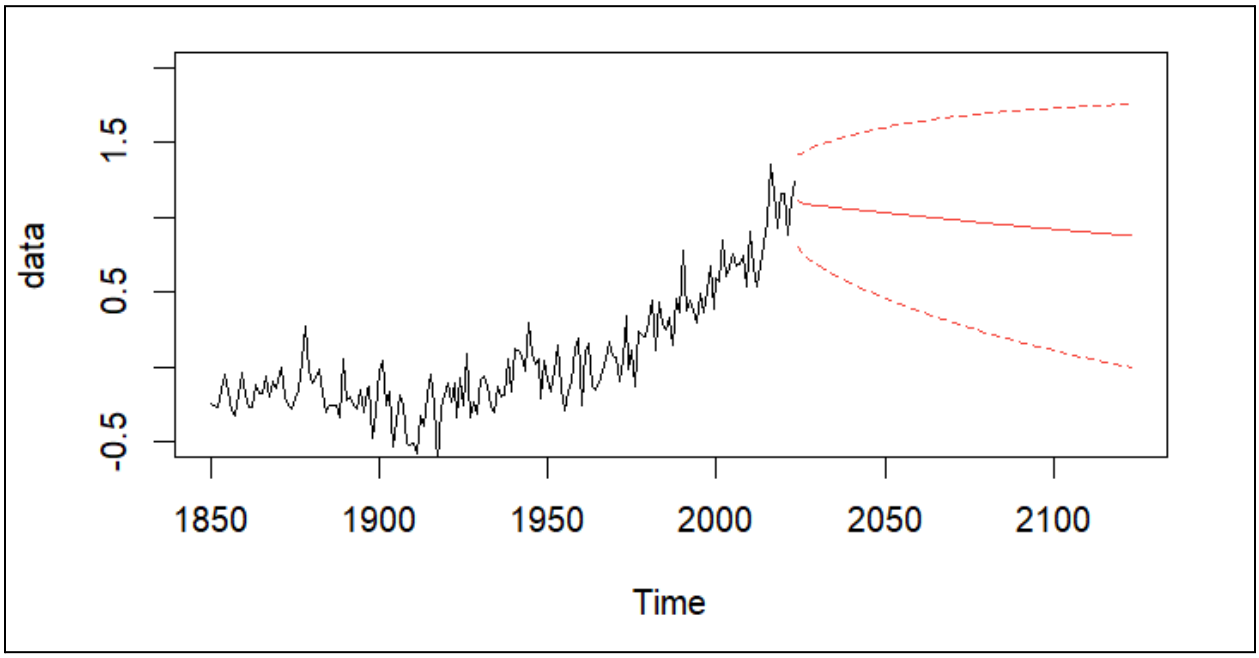
Supplemental Problem #1

Next, we examine global temperature data and analyze for autoregressive trends. First, we fit an ARMA(2,1) model to the data. The summary output for the model is shown below:

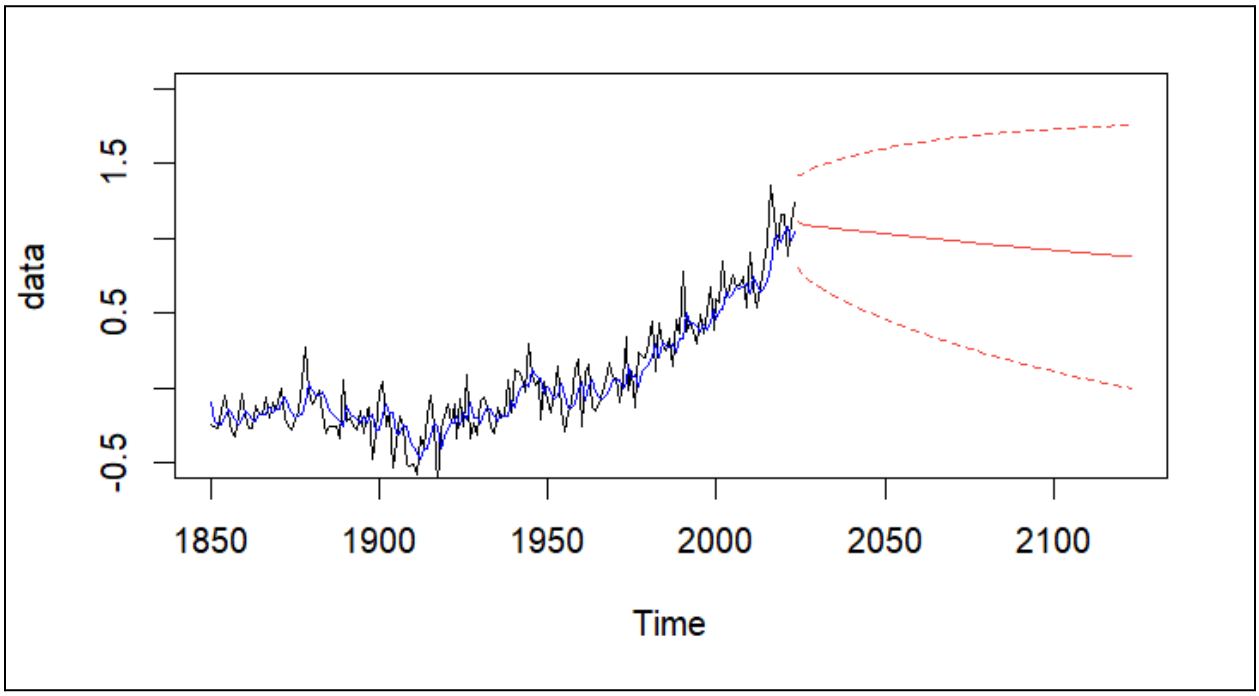
```
Coefficients:
          ar1      ar2      ma1  intercept
      1.1045  -0.1076  -0.7196     0.3420
s.e.  0.1040   0.1033   0.0677     0.5314

sigma^2 estimated as 0.02504:  log likelihood = 72.29,  aic = -134.57
```

The order-1 autoregressive and moving-average coefficients are all large in magnitude relative to their standard errors, so these parameters are very significant in determining the underlying relationships in the data. The ar2 coefficient is noticeably less significant.

We now display a prediction line along with a 95% prediction interval for the following 100 years. The plot is shown below:



For model evaluation purposes, we now overlay the model's predictions for the years where data exists.



We observe that the model's predictions tend to fit the overall trend of the data very well, but the long-term forecast does not seem to accurately gauge the underlying increasing trend in temperature. This is due to the fact that an ARMA model assumes that the data are stationary; the growth-over-time trend of global temperatures is not reflected in the predictions of this model. As a result, we determine that the efficacy of this model depends on the goal of using it; if short-term predictions are desired, the model will likely work well. However, the model fails to accurately forecast ahead several years in the future.

Supplemental Problem #2

We now consider the autoregressive time series defined by the recursive equation:

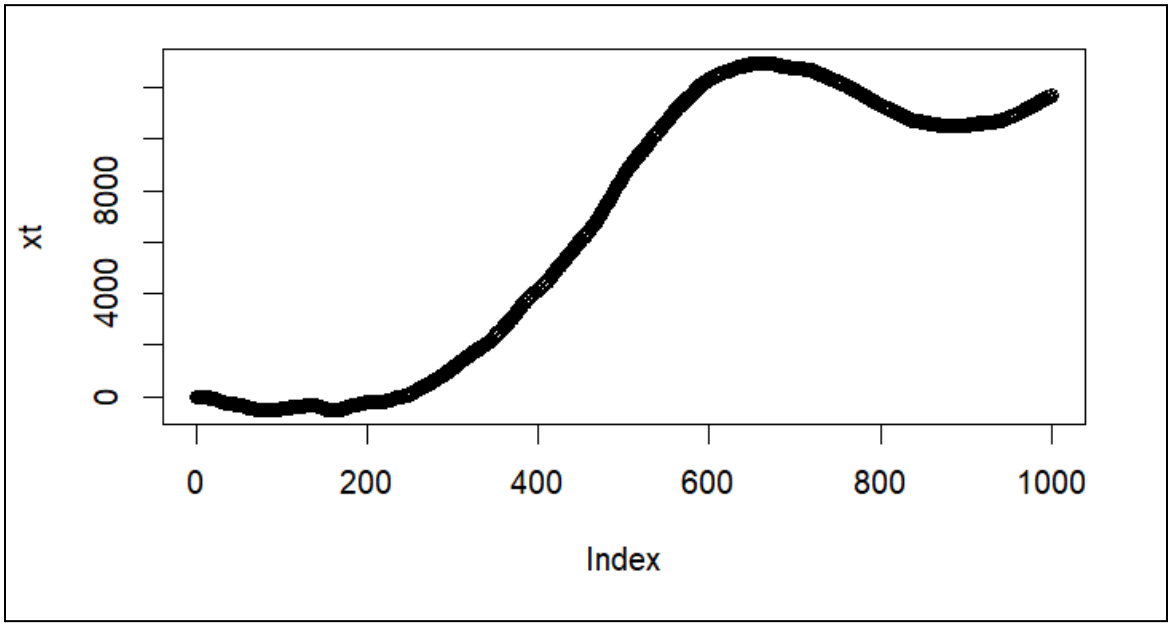
$$x_t = \frac{5}{2}x_{t-1} - 2x_{t-2} + \frac{1}{2}x_{t-3} + w_t$$

We first demonstrate that the series is not causal by showing that there exists a root of φ(B) which lies on/outside of the unit circle.

$$x_t - \frac{5}{2}x_{t-1} + 2x_{t-2} - \frac{1}{2}x_{t-3} = w_t \implies \varphi(B) = 1 - \frac{5}{2}B + 2B^2 - \frac{1}{2}B^3 = (1-B)^2 \cdot \left(1 - \frac{1}{2}B\right)$$

The factor $(1-B)^2$ implies that the function has a root of B=1, which is on the unit circle. As a result, the series is not causal.

We now simulate 1000 consecutive values of the series. Below is a plot of the series:



We now fit an AR(3) model to all observations. A summary for the model is displayed below:

```
Coefficients:
         ar1  ar2      ar3  intercept
      0.9986    1  -0.9986   6491.553
s.e.  0.0011  NaN   0.0011  31711.827

sigma^2 estimated as 3.89:  log likelihood = -2102.14,  aic = 4214.28
```

Note that the standard error for the ar2 coefficient is listed as NaN. This is because the coefficient for ar2 is 1, implying that $\varphi(B)$ has a unit root and therefore the series is nonstationary. The theory and formulas used to calculate the standard error of the AR model rely on the stationarity assumption; since this rule is violated, the standard error cannot be computed using standard methods.

Next, we fit an AR(3) model to only the first 100 observations in the series. The following summary is displayed by R for this model:

```
Coefficients:
         ar1      ar2     ar3  intercept
      2.4351  -1.9117  0.4756  -468.1198
s.e.  0.0857   0.1689  0.0840   113.6925

sigma^2 estimated as 0.6677:  log likelihood = -123.96,  aic = 257.92
```

The error caused by using all data has now been removed. Using this model causes a decrease in log-likelihood and AIC when compared with the original model including all data.

We now consider the differenced series $y_t = x_t - x_{t-1}$. First, we find a formula to express y_t in terms of y_{t-1}, y_{t-2}, and w_t.

$$y_t = x_t - x_{t-1} = \frac{3}{2}x_{t-1} - 2x_{t-2} + \frac{1}{2}x_{t-3} + w_t$$

$$y_t - \frac{3}{2}y_{t-1} = -\frac{1}{2}x_{t-2} + \frac{1}{2}x_{t-3} + w_t$$

$$y_t - \frac{3}{2}y_{t-1} + \frac{1}{2}y_{t-2} = w_t \implies \mathbf{y_t = 1.5 \cdot y_{t-1} - 0.5 \cdot y_{t-2} + w_t}$$

We conclude that $y_t = 1.5y_{t-1} - 0.5y_{t-2} + w_t$. Note that $\varphi(B) = 1 - 1.5B + 0.5B^2 = (1 - B)(1 - \frac{1}{2}B)$. This function has a root of 1, indicating that the series is nonstationary.

We now fit an AR model to the empirically-determined values of y_t. Since y_t is theoretically dependent only on the lag-1 and lag-2 values, we choose to fit an AR(2) model. A summary for this model is displayed below:

```
Coefficients:
         ar1      ar2  intercept
      1.4870  -0.4896    11.8556
s.e.  0.0276   0.0276    10.0480

sigma^2 estimated as 0.9929:  log likelihood = -1417.06,  aic = 2842.12
```

Note that the estimated coefficient values are within one standard error of the theoretical parameter values found above. An error still results from the R code because the data series is nonstationary, meaning that the optimization algorithm used for a stationary series might not function correctly as its assumptions are not met.

Finally, we consider z_t, the differenced series of y_t. We first derive a formula for z_t in terms of z_{t-1} and w_t.

$$z_t = y_t - y_{t-1} = 0.5y_{t-1} - 0.5y_{t-2} + w_t$$

$$z_t - 0.5z_{t-1} = w_t \quad \Longrightarrow \quad \mathbf{z_t = 0.5z_{t-1} + w_t}$$

It is worth noting that in this case, φ(B) is now equal to (1-½B), so all roots of φ(B) lie outside the unit circle. We finally have arrived at a series which is stationary.

We now fit an AR model to z□. Since z□ only theoretically depends on the lag-1 term, we fit an AR(1) model. A summary for this model is displayed below:

```
Coefficients:
          ar1  intercept
      0.4886     0.0144
s.e.  0.0276     0.0617

sigma^2 estimated as 0.9959:  log likelihood = -1414.2,  aic = 2834.39
```

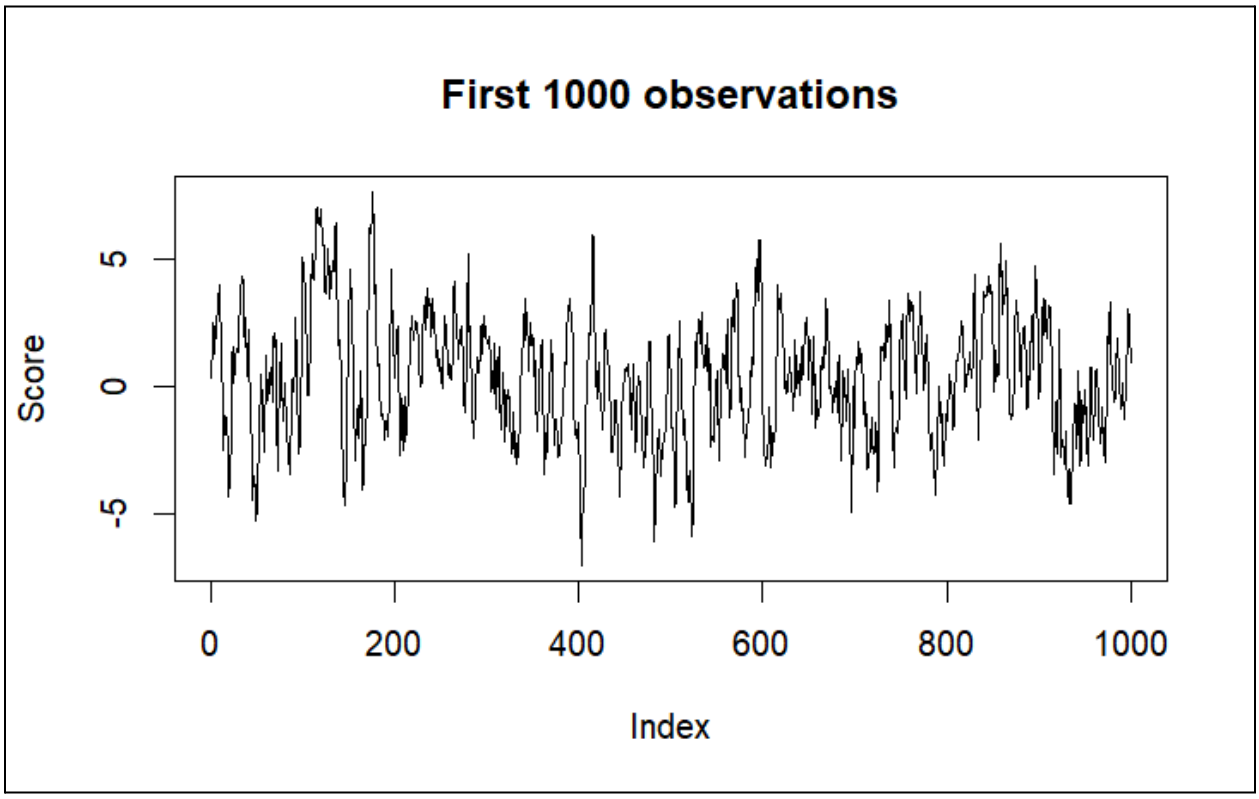
Finally, we calculate a 95% confidence interval for the true value of the lag-1 autocorrelation coefficient in the series z□.

$$0.4886 \pm z_{0.975} \cdot 0.0276 \equiv [0.4345,\ 0.5427]$$

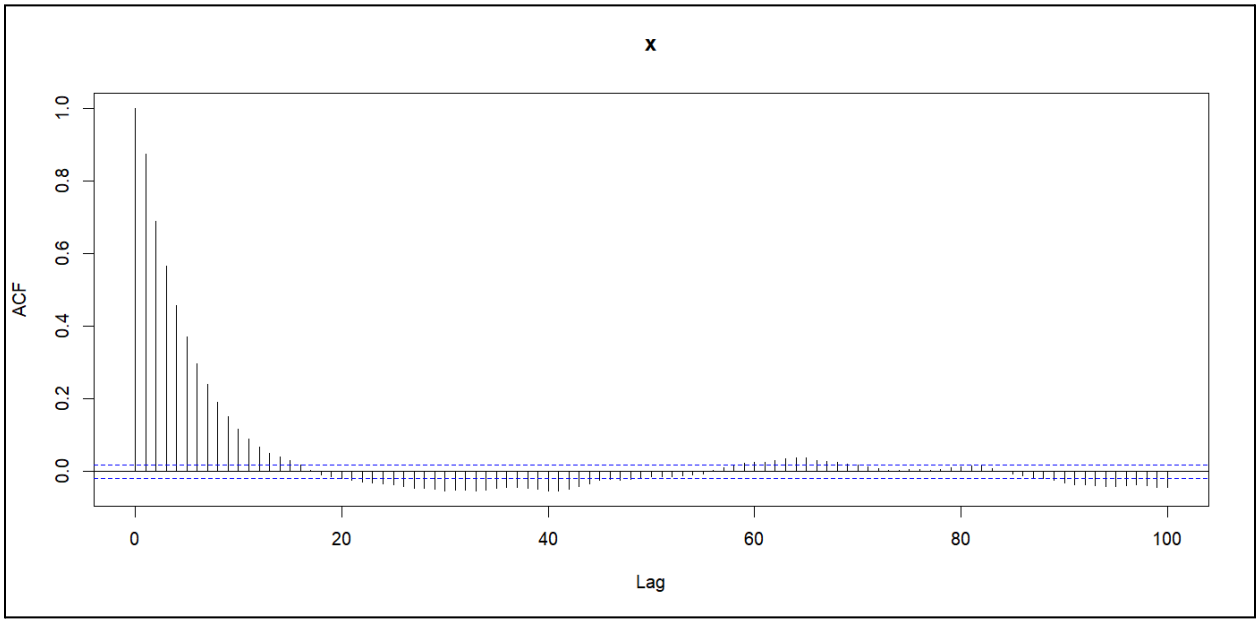
This interval contains the true value of the lag-1 autocorrelation coefficient, 0.5.

Problem 3

We now consider a dataset of historical air quality values in classrooms. We plot the first 10% of the data below:

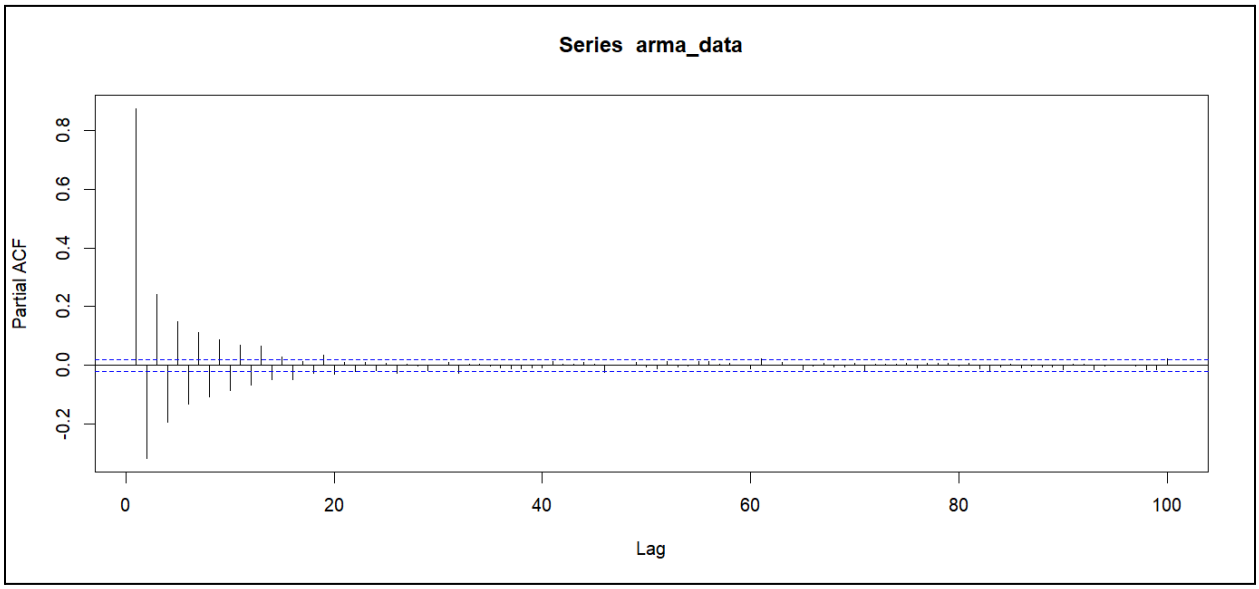


The repetitive jagged peaks appear to indicate autocorrelation. To confirm our suspicions, we will analyze the sample ACF and PACF of the series. Below is the plot of the sample ACF for the first 100 lags:



We observe that the first 15 lags are all significantly correlated. Additionally, the narrow sinusoidal pattern in later lags is reminiscent of a signal-with-noise series.

We now consider the PACF, which is displayed below:



Almost every lag up to lag 20 has a significant PACF value – however, after lag 2, most autocorrelations are less than 0.2, indicating weak autocorrelation. This cutoff in the PACF is indicative of the autoregressive part of the ARMA model, and as a result, we will use a value of p in the ARMA(p,q) model that is large relative to q.

We consider all models with $1 \leq p \leq 6$ and $0 \leq q \leq 3$, and we use AIC as an evaluation statistic. Below is a table of the AIC computed for the ARMA(p,q) model with tabulated values of p and q.

p / q	0	1	2	3
1	31308.70	28780.63	28335.07	28288.10
2	30258.60	28284.39	28285.68	28287.43
3	29667.10	28285.65	28284.28	28288.23
4	29293.25	28287.38	28286.77	28288.62
5	29068.85	28289.28	28288.14	28290.18
6	28895.70	28289.39	28287.69	28290.68

For reference, we calculate the AIC of a linear regression:

45678.34

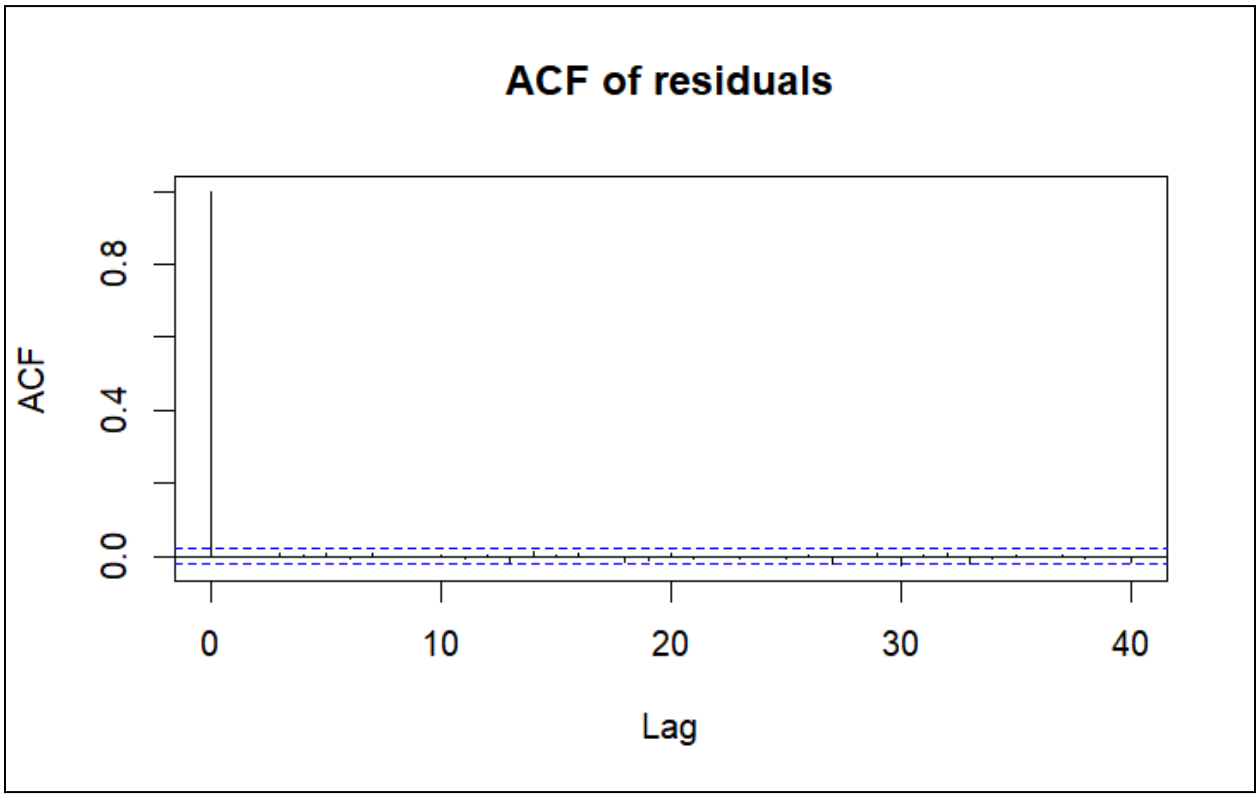
The smallest AIC values are 28284.39 ($p = 2, q = 1$) and 28284.28 ($p = 3, q = 2$). For the sake of simplicity, we opt for the ARMA(2,1) model.

A summary output for this model is displayed below:

```
Coefficients:
          ar1     ar2     ma1  intercept
      0.4855  0.2626  0.9052    -0.0318
s.e.  0.0114  0.0113  0.0050     0.0752

sigma^2 estimated as 0.9893:  log likelihood = -14137.19,  aic = 28284.39
```

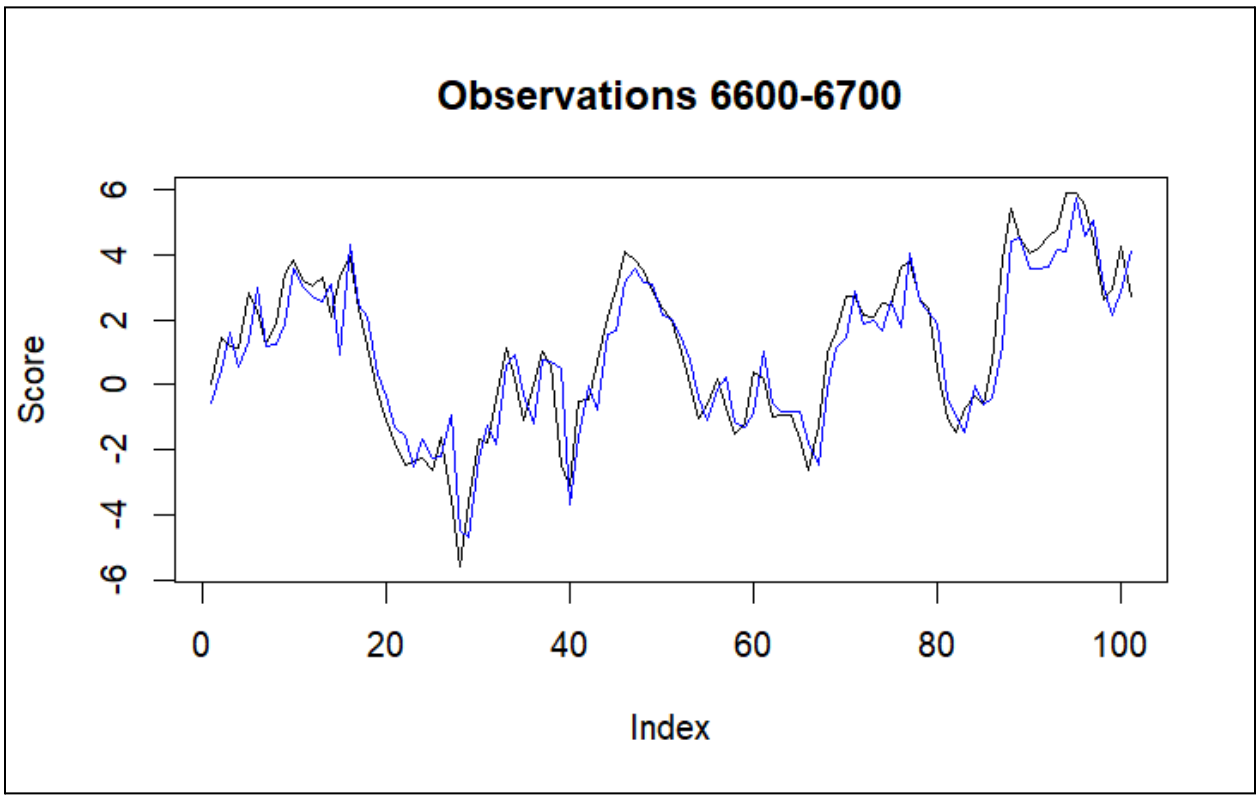
We observe that each coefficient is several times larger than its standard error, indicating that these predictors are extremely significant.

Next, to verify the goodness-of-fit of our model, we plot the ACF of the residuals to ensure there is no autocorrelation among them.



The above ACF plot demonstrates that the model is a good fit and leaves behind little autocorrelation, if any, among the residuals.

Finally, we plot a small portion of the data alongside the model's predictions to obtain a visual metric to gauge the goodness-of-fit of the model.



By both numerical and observational methods, we have determined that the ARMA(2,1) model is a highly suitable model for this data.

Non-technical Summary:

Imagine our classroom's air quality as a dynamic variable determined by two properties: recent air quality history and unexpected changes (errors).

We tested 24 models that predict air quality and carefully compared them using a scoring system called AIC - think of it as a rating system where lower scores are better. After comparing all these scores, we found two best models: a model that looks back at 2 time periods with one adjustment factor (ARMA(2,1)) and another that looks back at 3 time periods with two adjustment factors (ARMA(3,2)). Since both performed almost equally well (with AIC scores of

about 28284), we chose the simpler one - ARMA(2,1) - because it's easier to use while being just as accurate. Our ARMA(2,1) model looks back two time steps to understand how recent air quality levels interact: if the previous two measurements showed high air quality, the model predicts the subsequent measurement will likely continue that trend. Simultaneously, the moving average component helps smooth out sudden, random fluctuations, ensuring our predictions aren't overly reactive to instantaneous changes. By combining these two elements, the model creates an intuitive prediction that accounts for trends over time in classroom air quality.

Technical Summary:

To select the optimal ARMA(p,q) model, we systematically explored the model space by computing AIC across various parameter combinations through a grid search of p and q values. We examined models with p ranging from 1 to 6 and q from 0 to 3, leveraging the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) as initial guidance. The ACF revealed significant correlations in the first 15 lags, while the PACF suggested potential autoregressive complexity, particularly with weak autocorrelations after lag 2. Comprehensively comparing AIC values for ARMA(p, q) we identified two competitive models: ARMA(2,1) and ARMA(3,2) with very similar AIC values, ultimately selecting ARMA(2,1) for its simplicity and strong statistical significance. The final model does not leave much autocorrelation in its residuals and has extremely significant predictors that contribute to the goodness of fit.