Chapter 6
Model Evaluation and the Forecasting Process

1 Introduction

We have learned about a few ways to make quantitative forecasts. In future chapters, we will learn about even more forecasting methods. However, it is not always clear which method will perform best for a given data set. For example, maybe you have 50 data points, and it looks like there might be a slight upward trend, but it's hard to tell from the naked eye. Although we have measures that we can use to judge forecast accuracy, we need to be able to distinguish between models *before* we start forecasting. In this chapter, you will learn about how you can evaluate models "in-sample", in order to select the ones that will be good candidates for use in forecasting.

2 Model Evaluation

In the past, we covered measures of forecast accuracy. These measures of forecast accuracy correspond with our loss functions, and inform us which models forecast most accurately over a given time period. However, when you are first starting out, this probably seems backwards — isn't it more useful to know if a forecasting method will do a good job *before* we start using it?

There are several ways to assess the fitness of models before you start actually using the models to forecast. In this chapter, we will focus on in-sample performance.[1] All of our statistical models place assumptions on the error terms (such as having mean zero and being i.i.d.). One common approach is to ensure that the estimated error terms actually have the properties that we assumed they would. Another common approach is to strike a balance between measures of goodness-of-fit and model complexity. In other words, models that do a better job of fitting the in-sample data are rewarded, but those that have more parameters are

[1] A common but more complicated alternative to using in-sample-performance is to use "pseudo-out-of-sample forecasts." We will not cover pseudo-out-of-sample forecasts in this course.

punished. There are statistics that can summarize this tradeoff using only one number, making it easy to quickly compare different models.

2.1 Evaluating Error Terms

After estimating a statistical model, we are left with the in-sample fit, or the in-sample "predictions" and the in-sample error terms. For example, recall the intercept model:

$$y_t = b + \varepsilon_t$$
$$\varepsilon_t \sim \mathcal{N}(0, \sigma)$$

Under squared error loss, the optimal estimate of b, given the data $Y = [y_1, y_2, \cdots, y_T]$, is simply the mean of the data:

$$\hat{b} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

This optimal in-sample estimate will not perfectly match the data in every time period. For example, consider a sports example — suppose your favorite basketball player averages 20 points per game. Clearly, he will not score exactly 20 points per game every night, rather, that's how many points he will score in an "average" game.

The in-sample error terms are the difference between the model fit and what actually happened:

$$\varepsilon_t = y_t - \hat{y}_t$$

where $\hat{y}_t$ is the model fit. In the intercept model under squared error loss:

$$y_t = b + \varepsilon_t$$
$$\hat{y}_t = \hat{b}$$

Let's consider a more concrete example — points per game for Joel Embiid in the last 10

regular season NBA games he played last season:

$$Y = \begin{bmatrix} 33 \\ 17 \\ 21 \\ 40 \\ 37 \\ 27 \\ 20 \\ 39 \\ 34 \\ 20 \end{bmatrix}$$

We will continue to use the intercept model under squared error loss for this example. His average over this span was:

$$\begin{aligned} \hat{b} &= \frac{1}{T}\sum_{t=1}^{T} y_t \\ \hat{b} &= \frac{1}{10}(33 + 17 + 21 + 40 + 37 + 27 + 20 + 39 + 34 + 20) \\ \hat{b} &= \frac{1}{10}(288) \\ \hat{b} &= 28.8 \end{aligned}$$

So under this model, our "best guess" for every game is the same, 28.8. Using this in-sample fit, we can calculate the error terms of our model. In time period one (i.e. game one):

$$\begin{aligned} \varepsilon_1 &= y_1 - \hat{y}_1 \\ \varepsilon_1 &= 33 - 28.8 \\ \varepsilon_1 &= 4.2 \end{aligned}$$

At time period two:

$$\varepsilon_2 = y_2 - \hat{y}_2$$

$$\varepsilon_2 = 17 - 28.8$$

$$\varepsilon_2 = -11.8$$

We can keep doing this for each time period. In vector notation, we would have:

$$\varepsilon = Y - \hat{Y}$$

$$\varepsilon = \begin{bmatrix} 23 \\ 25 \\ 23 \\ 20 \\ 21 \\ 24 \\ 22 \\ 26 \\ 18 \\ 32 \end{bmatrix} - \begin{bmatrix} 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \\ 28.8 \end{bmatrix} = \begin{bmatrix} 4.2 \\ -11.8 \\ -7.8 \\ 11.2 \\ 8.2 \\ -1.8 \\ -8.8 \\ 10.2 \\ 5.2 \\ -8.8 \end{bmatrix}$$

Now, since we know that the errors should be independent and identically distributed with mean zero, we can analyze the properties of the actual observed error terms to see if they have these properties. **With the estimates formed under squared error loss, the error terms will always have mean zero, by construction.** Therefore, for the vast majority of our models, we do not need to check to see if the error terms have mean zero — they always will. Instead, we will check to see whether the error terms are independent and identically distributed.

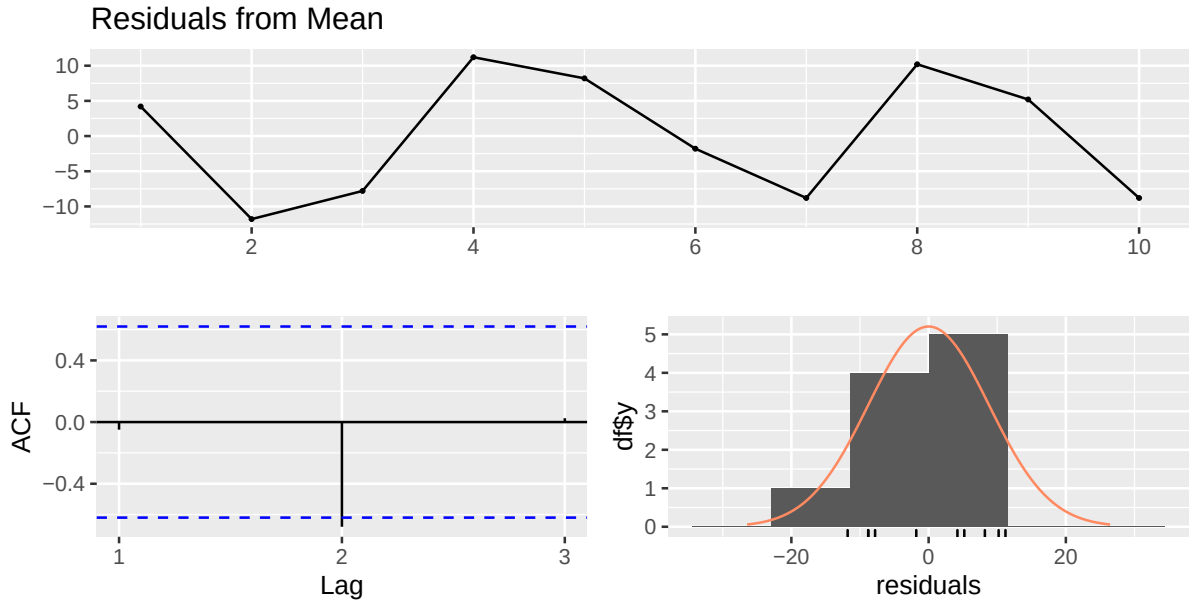
Of the two properties we can actually test, it is easiest to test independence. Recall that time-series independence means that the elements in a time series have no autocorrelation. In other words, if the error terms are truly independent, old values of the error term should not tell us anything about the value of the error term today.

We can check independence of the error terms both graphically by looking at an ACF plot of

the residuals, as well as statistically with an Ljung-Box test. With an ACF plot, we do not want to see any large spikes in the ACF. In R, the ACF plot automatically includes the associated 95% confidence interval. If the autocorrelation at any lag is outside of this interval, it is statistically significantly different from zero — this is a sign that the error terms are autocorrelated and therefore do not satisfy the assumptions of our statistical model.

The null hypothesis of the Ljung-Box test is that the series being tested is independent. Therefore when we apply this test to the residuals of our series, the null hypothesis is that our independence assumption is satisfied. We typically use a critical value of 0.05 (5%). If the p-value is less than the critical value, then we reject the null hypothesis and are forced to conclude that the residuals are NOT independent, and therefore violate our model assumptions.

In R, it is relatively straightforward to apply both of these tests with the checkresiduals() function:




```
##
##  Ljung-Box test
##
## data:  Residuals from Mean
## Q* = 10.566, df = 3, p-value = 0.01432
##
## Model df: 1.   Total lags used: 4
```

The checkresiduals() function produces three plots and the results of the Ljung-Box (LB)

test. The top graph is simply a time series plot of the residuals. The bottom left is the plot we are most interested in — the ACF plot for the residuals. In this example, we can see that only three lags are plotted (since the time series has only 10 observations). The autocorrelation at the first and third lags is roughly zero, but the autocorrelation at the second lag is highly negative. This spike at the second lag suggests that the error terms may NOT be independent. Finally, on the bottom right we see a histogram of the residuals (the dark bars) along with a Normal distribution overlaid (the orange line). We typically assume that the error terms are Normally distributed, so it can be useful to check to see if the histogram bars roughly line up with the orange line.

The checkresiduals() function also produces the results from the LB test. When looking at this test, the most important value to look at is the p-value. If the p-value ≥ 0.05 then we fail to reject the null (i.e. we conclude that there is not enough evidence to overturn the hypothesis that the residuals are independent). If the p-value < 0.05, then we reject the null and conclude that the residuals are autocorrelated. In this specific example, we can see that the p-value is 0.014, so we reject the null. This formal statistical test suggests that our model assumption of no autocorrelation is NOT satisfied. This suggests that it is likely possible to improve on the intercept model (i.e. the intercept model is probably not the "best" forecasting model to use for this data).[2]

2.2 What To Do About Autocorrelation in Residuals?

Suppose that we encountered autocorrelation in the error terms. This is a bad sign for the model we are using, as it means that the model we are using does not fully take advantage of the data. To see this, note that if there is autocorrelation in the residuals, it means that the old error terms contain information about the error term today. Therefore, the errors are predictable, so the fitted values can be improved.

It is not always clear how best to eliminate the autocorrelation in error terms. Sometimes, if the errors from the intercept model are autocorrelated, it is because there is a trend or seasonality that are unaccounted for. Other times, the source of the autocorrelation is not as obvious, and we need to turn to more sophisticated methods or transform the data before fitting a model to it. Finally, it may be the case that you try a wide variety of models, but you

[2] In this case, since we only have ten observations, it may be wise to collect more historical data before making a final determination for or against the intercept model.

can never get all of the autocorrelation to go away. In that case, you should NOT simply give up and refuse to forecast. Instead, you will need to use a different metric to choose between models.[3]

2.3 Model Fit

Another useful property to investigate is how well the model fits the in-sample data. As you add more and more parameters to a model, your model will automatically fit the data better. However, it is not always desirable to have a perfect or even a near perfect in-sample fit. For example, adding seasonality to an intercept model will always produce a better in-sample fit, even if your data does not have seasonality. Therefore, it is important to balance both model fit and model complexity — continuing with the previous example, we should only use seasonality if it improves the model fit a great deal.

As with most other concepts, there are many different ways that forecasters can try to determine model fit. For now, we will focus on two: (1) the standard deviation of the error terms and (2) the likelihood of the data given the model.

2.3.1 Standard Deviation of Error Terms

One way to see how well a model fits is to measure the standard deviation of the error terms. The smaller the standard deviation, the closer the model fit is to the actual data values. Mathematically, recall that the error terms are given by:

$$\varepsilon_t = y_t - \hat{y}_t$$

where $\hat{y}_t$ is the model fit. The standard deviation of the error terms is given as:

$$\hat{s} = \sqrt{\frac{1}{T_\varepsilon - k} \sum_{t=1}^{T} \varepsilon_t^2}$$

where k is the number of parameters estimated in the model, T_ε is the total number of residuals, and $\hat{s}$ is the estimated standard deviation of the error terms. Therefore the denominator of the fraction above is the number of residuals minus the number of parameters. For example, if we were using the intercept model, $k = 1$ since we need to estimate the intercept, but that is the

[3] You will also need to do this if several models have independent errors — if they all satisfy the assumptions, you will need to use something else in order to choose the "best" candidate model for forecasting.

only parameter that needs to be estimated. In the intercept model, we can compute a residual starting at the first time period, so $T_\varepsilon = T$. Note that since we assume the mean of the error term is 0 (i.e. $\varepsilon_t \sim N(0, \sigma)$) we do NOT need to subtract the mean of the residuals from the residuals before squaring. We simply square each residual.

Since a smaller standard deviation means a more accurate model, one way to choose the "best" model is to estimate a number of models, and then choose the model that has the smallest standard deviation of error terms. However, you need to be very careful when doing this — in general, models with more parameters will have smaller standard deviations. This means that including irrelevant parameters (such as seasonality or trend when they are not really present) can decrease the standard deviation and lead you to choose a model that fits well in-sample, but will probably forecast poorly.

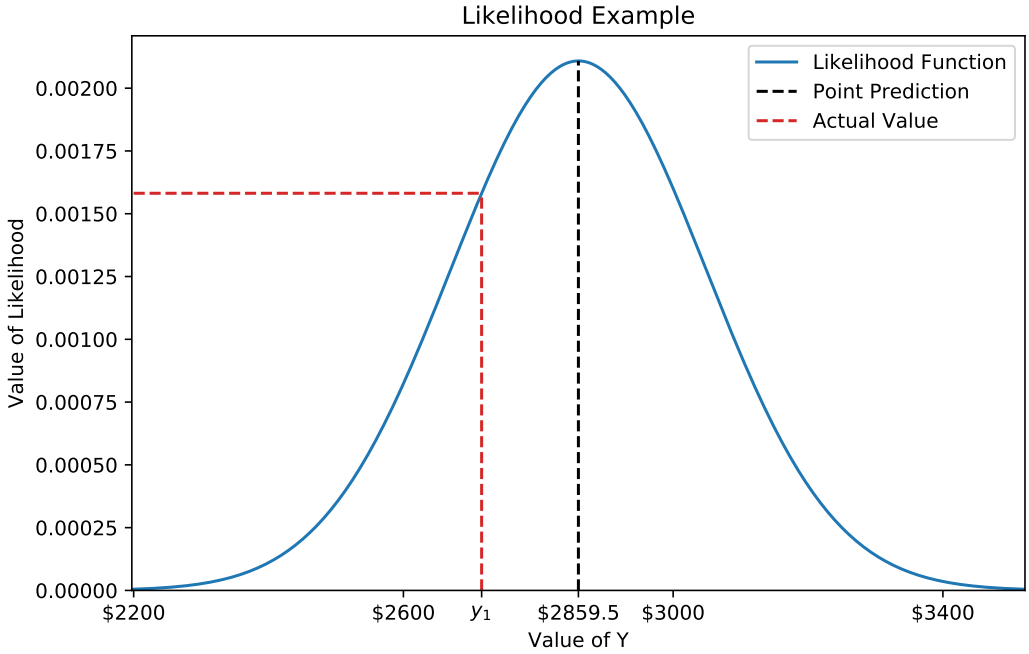
2.3.2 Information Criteria

Choosing a model based only on size of the standard deviation may lead to selecting models that have too many parameters. Therefore, it is better to use something called an **information criterion**, which rewards models with small standard deviations but punishes models for including additional parameters. A general philosophy that is smart to follow when forecasting is *Occam's Razor* — all else equal, a less complicated model is preferred to a more complicated model.

Information criteria consist of two pieces: (1) a reward for better fit and (2) a punishment for model complexity. The reward is based on the total likelihood, calculated from the likelihood function. A **likelihood function** summarizes how likely the observed data is, conditional on being generated by the model under consideration. For example, if the data has a strong upward trend and you use an intercept model to fit the data, the total likelihood will be very low, since it is very unlikely that we would observe data with a strong trend if the true model was the intercept model.

The total likelihood is the product of the likelihood at each point, evaluated in-sample. For example, suppose we were using the intercept method to forecast sales, so our point prediction was the same at every point in-sample. In this case, let's assume the point forecast was \$2,859.5. If we assume that the error terms are Normally distributed (and we estimate the standard deviation), we would have all the information needed for the likelihood function. Then at each point, $y_1, y_2, \cdots, y_T$, we could evaluate the likelihood function at the value of the data that

actually occurred to compute the value of the likelihood at that point. For example, suppose $y_1 = \$2,700$ Then we would have:



where the value of the likelihood function is given by the valule on the y-axis (in this case, roughly 0.0016). We would compute this number for each observation, and then multiply them all together to find the total likelihood for this model. The total likelihood measures how well our model fits the in-sample data.

The precise formula for the "punishment" term differs across different information criteria. However, in all information criteria, models that include more parameters are punished more. This punishment is included because, similar to standard deviation, the likelihood will only increase as we add more parameters, but we don't always want to choose the most complicated model. Instead, we want to choose the model that both explains the data well and is relatively simple.

The two information criteria we will learn about are the **corrected Akaike Information Criterion (AICc)** and the **Bayesian Information Criterion (BIC)** (BIC is sometimes referred to as the Schwartz Information Criterion (SIC)). The two information criterion have the same reward — they both use the likelihood in the same way. However, they differ in how harshly they punish model complexity. BIC punishes models with more parameters more harshly, so typically chooses smaller models then AICc.

Mathematically, BIC is calculated as:

$$BIC = (k+1)\ln(T_\varepsilon^*) - 2\ln(\hat{L})$$

where k is the number of parameters of the model (including the intercept...if we only have an intercept and no other parameters, $k = 1$), T_ε^* is the number of residuals that can be computed for all models that we are comparing, $\hat{L}$ is the estimated likelihood of the model, and ln denotes the natural log. **Since models with larger likelihoods are preferred, and the likelihood is subtracted, models with smaller BIC are preferred.** In other words, if you calculate the BIC for several models, you would choose the model with the smallest value of BIC, where "smallest" means the closest to $-\infty$

Mathematically, AICc is calculated as:

$$AICc = 2(k+1) + \frac{2(k)(k+1)}{T_\varepsilon^* - k - 2} - 2\ln(\hat{L})$$

Like BIC, **models with smaller AICc are preferred.**

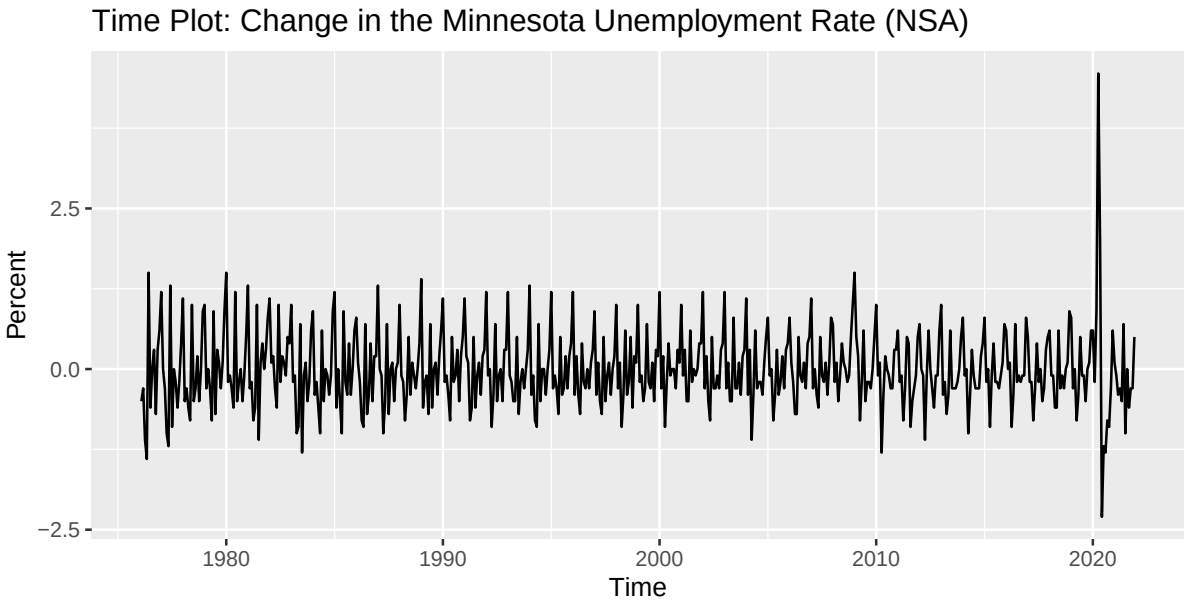
While both AICc and BIC are well-grounded and have (different) desirable properties, research suggests that if AICc and BIC choose different models, the model chosen by AICc typically forecasts better. In other words, if our goal is forecasting, we should probably use AICc when choosing a model.

3 Application: MN Unemployment Rate

We will continue with the data series from Chapter 5, the unemployment rate in Minnesota since 1976. Recall that we determined there was a unit root, so fit the three benchmark models on the differenced data (i.e., the monthly change in the unmployment rate).

```
library(fredr)
data_set <- fredr(series_id = "MNURN")
Y <- data_set$value
Y <- ts(Y,start=c(1976,1),frequency=12)
DY <- diff(Y)
# Time Plot
```

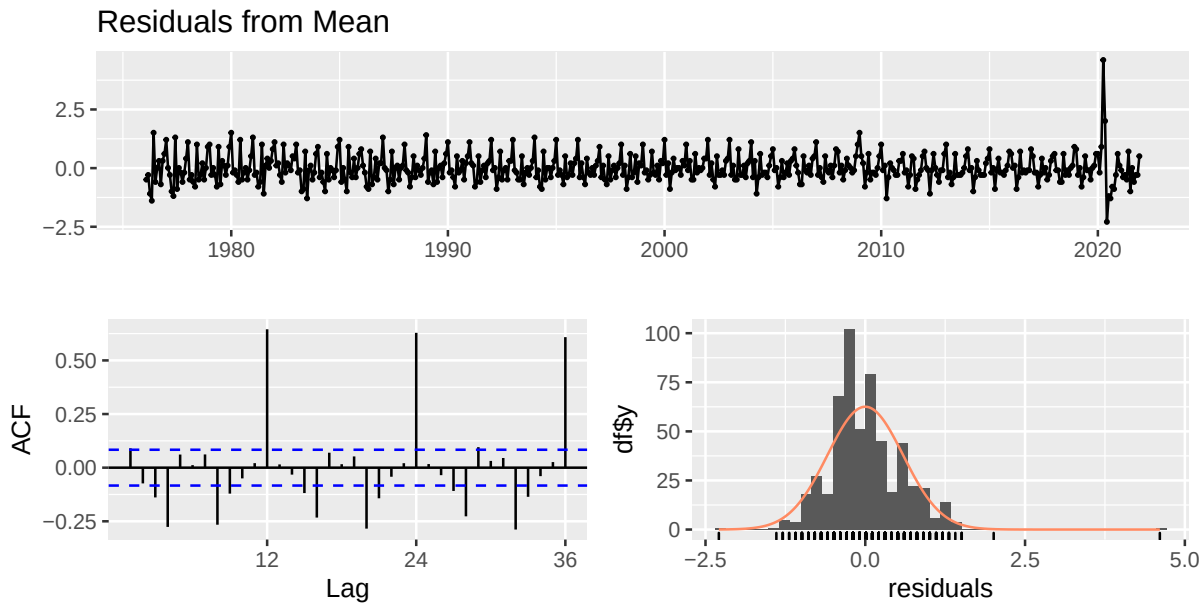
```
autoplot(DY) +
  ggtitle("Time Plot: Change in the Minnesota Unemployment Rate (NSA)") +
  ylab("Percent")
```



3.1 Intercept Model

First, let's inspect the in-sample fit from the intercept model.

```
# Fit the intercept model
fcst_mean <- meanf(DY,h=12)
# Check the in-sample residuals from the intercept model:
checkresiduals(fcst_mean)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Mean
## Q* = 681.82, df = 23, p-value < 2.2e-16
##
## Model df: 1.   Total lags used: 24
```

We can see that this model is (unsurprisingly) not able to capture the predictable seasonal variation in the data. Instead, it treats this part of the data, which is predictable, as unpredictable, pushing it into the error term. Therefore, the errors are highly predictable, and the forecast errors will likely be highly predictable as well. For example, if we forecasted that all future time periods would be equal to the mean of the data, then each January we would neglect to forecast the highly predictable increase in the unemployment rate that has occurred every January since data collection began in 1976.

3.2 Random Walk Model

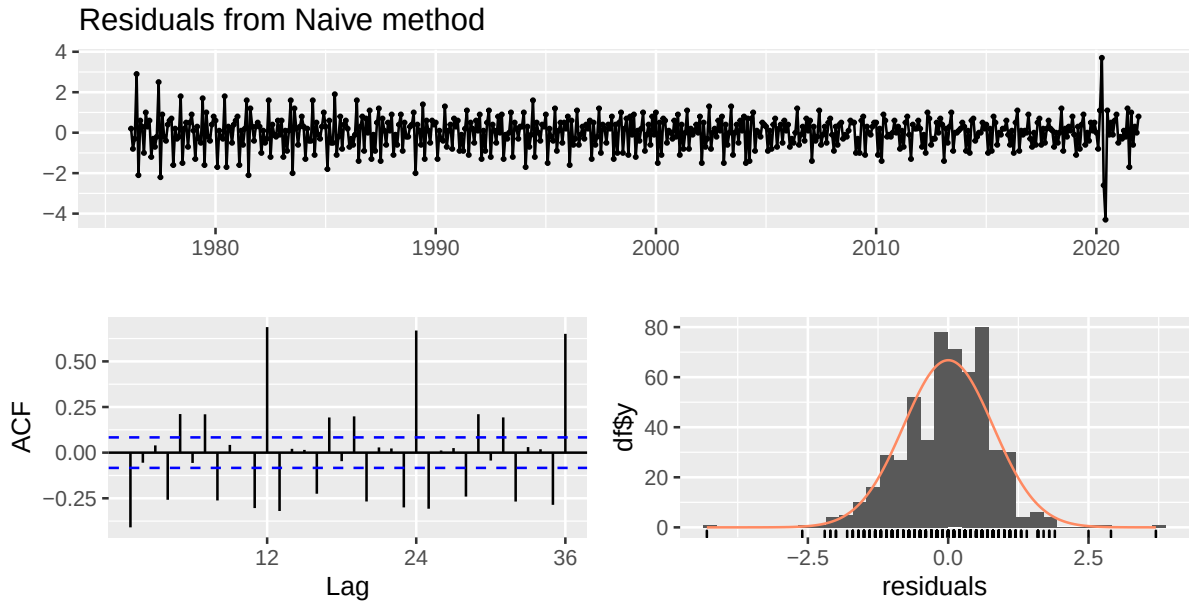
Next, let's inspect the in-sample fit from the random walk model.

```r
# Fit the random walk model

fcst_rw <- naive(DY,h=12)

# Check the in-sample residuals from the random walk model:

checkresiduals(fcst_rw)
```



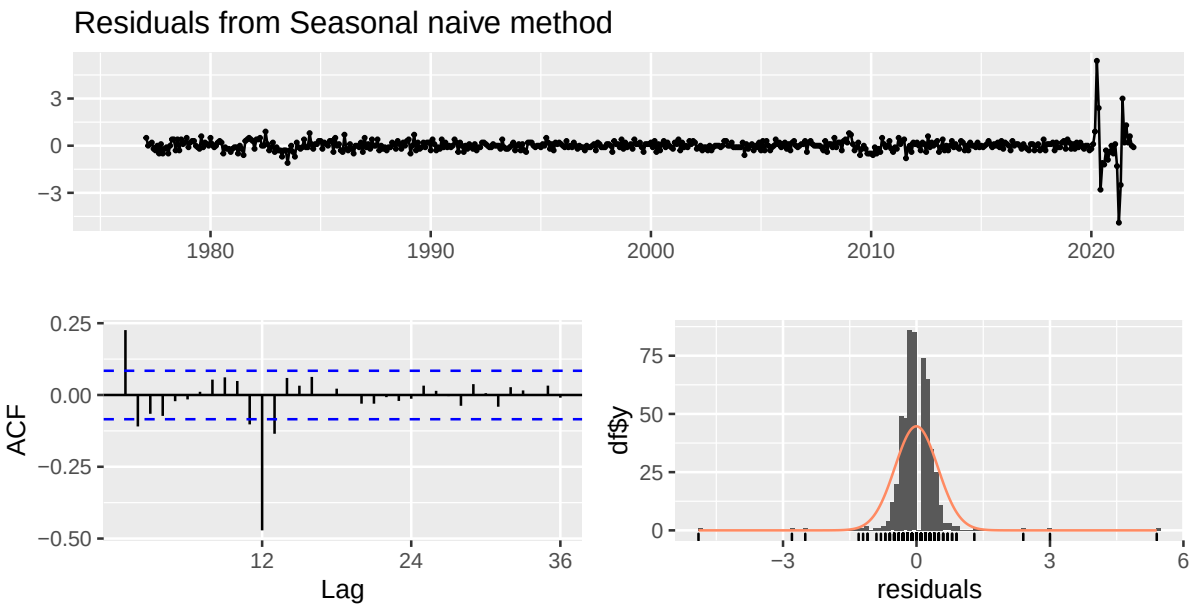

```
##
##  Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 1025.7, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24
```

Similar to the intercept model, the in-sample fit is poor, since this model cannot account for the predictable seasonal patterns, instead pushing them into the error term. This means that the errors are predictable, and the forecast errors likely will be too.

3.3 Seasonal Random Walk Model

Next, let's inspect the in-sample fit from the seasonal random walk model.

```r
# Fit the seasonal random walk model
fcst_srw <- snaive(DY,h=12)
# Check the in-sample residuals from the seasonal random walk model:
checkresiduals(fcst_srw)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Seasonal naive method
## Q* = 190.15, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24
```

This model does a much better job of capturing the seasonality. However, there is now negative autocorrelation in the errors at the 12th lag. This suggests that the forecast errors are still predictable, in the following sense — if the model under-predicted what actually happened a year ago (i.e., the actual value was greater than the predicted value), it will likely over-predict this year (i.e., the predicted value will likely be greater than the actual value). Therefore, while better, this model is still not an ideal fit for the data.

3.4 In-Sample Accuracy: AICc

Aside from checking the assumptions on the residuals, we can calculate model accuracy statistics. Unfortunately, for these benchmark models, RStudio does not automatically produce AICc values, and they need to be calculated by hand. Doing so, we get the following results:

Model	AICc
Intercept	988
Random Walk	1314
Seasonal Random Walk	745

Since smaller AICc values indicate a better in-sample fit, this suggests that of the models considered, the Seasonal Random Walk model is the most appropriate.

4 Forecasting Process

Finally, now that we have a few ways of choosing between models, let's review the five-step forecasting process one more time. The process is as follows:

1. Define the forecasting problem.
    - What is your main goal? Is it to help boost sales at the company? Is it to better predict sales so you know how much inventory to carry?
    - What data will you need to help you meet this goal?
    - What is the forecasting horizon?
2. Gather information.
    - Does the data already exist? Has your company been collecting it already, is it publicly available, etc.?
    - Or does the data not exist yet? Was it collected at irregular frequency? How much will it cost to collect this data?
    - Is this a feasible project?
3. Analyze the data.
    - Compute summary statistics and summary plots (time, seasonal, ACF).
    - What features does the data have and what models are likely to fit the best?

4. Fit models & analyze properties.

    - Based on results from step 3, fit models you think are good candidates.
    - Are residuals of these models independent?
    - Which model has the best AICc? Is the 2nd best model close?

5. Use and evaluate the forecasts.

    - Start to make forecasts and track your results.
    - Is your RMSFE low? Is your model doing well or are the forecasts getting more inaccurate over time?
    - Is there any extra information (or qualitative information) that exists that is not embedded in the forecast model (policy change, new competition, etc.)? Use this to improve forecasts! Especially if long-term.

For the rest of the course, we will focus on learning methods to apply during step 4. The three main methods we will cover are: (1) exponential smoothing, (2) ARIMA models, (3) linear regression models.

5 Exercises

1. Suppose consider the following two models:

	Model A	Model B
Likelihood ($\hat{L}$)	100	120
Parameters (k)	5	10
Observations (T)	150	150

    (a) Compute AICc and BIC for each model.

    (b) Which model is favored by AICc? Which model is favored by BIC?

    (c) Why do we use AICc and BIC?

    (d) In practice, which measure (AICc or BIC) tends to prefer smaller models (i.e. models with fewer parameters)?

2. Consider the following in-sample error terms from two different models:

Time (t)	Error under Model A	Error under Model B
1	0.0	0.28
2	0.12	-2.22
3	1.70	1.65
4	1.72	0.73
5	0.87	-0.63
6	-1.01	-0.33
7	-0.83	0.76
8	0.12	-3.23
9	0.09	-0.72
10	0.17	0.57

(a) Suppose that Model A had 10 parameters and Model B had 7 parameters. Compute the standard deviation of the error terms from Model A and Model B.

(b) If you were selecting a model based on the size of the standard deviation, which model would you choose?

3. Download the series "All Employees: Total Nonfarm in Eugene, OR (MSA)" NOT seasonally adjusted from FRED. Use R to read in the data and declare it as a time series.

(a) Based on your preliminary analysis, which of the benchmark forecast methods do you think will forecast best?

(b) Use R to fit the Intercept Method (possibly to transformed data). Use the checkresiduals() to analyze the model fit. Does the model appear appropriate? Explain.

(c) Use R to fit the Seasonal Random Walk Method (possibly to the transformed data). Use the checkresiduals() to analyze the model fit. Does the model appear appropriate? Explain.

(d) If you were forced to choose between the intercept and the seasonal random walk model, which would you choose? Explain.