https://workshop.f4sg.org/africast/

Africast-Time Series Analysis & Forecasting Using R

9. Basic training and test accuracy

tsibble

feasts

Fable

Outline

1. Evaluating forecast accuracy
2. Evaluating point forecast accuracy
3. Evaluating distributional forecast accuracy

Outline

1. Evaluating forecast accuracy
2. Evaluating point forecast accuracy
3. Evaluating distributional forecast accuracy

Evaluate forecast accuracy

- In order to evaluate the performance of a forecasting model, we compute its forecast accuracy.
- Forecast accuracy is compared by measuring errors based on the test set.
- Ideally it should allow comparing benefits from improved accuracy with the cost of obtaining the improvement.

Evaluate forecast accuracy- Business impact

- We should be choosing forecast models that lead to better business decisions
  - least staffing costs, least emission, highest service level, least stock-out, least inventory, fastest response, least change in planing, for example.
- However, this is not always easy to obtain, therefore we might simply use methods that provide the most accurate forecast.

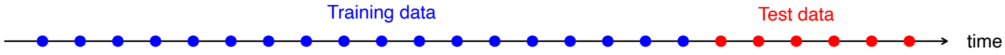
In-sample (training) vs. out-of-sample (test)

- Fitting and its residual are not a reliable indication of forecast accuracy
- A model which fits the training data well will not necessarily forecast well
- A perfect fit can always be obtained by using a model with enough parameters
- Over-fitting a model to data is just as bad as failing to identify a systematic pattern in the data

Forecast accuracy evaluation using test sets

- We mimic the real life situation
- We pretend we don't know some part of data (new data)
- It must not be used for *any* aspect of model training
- Forecast accuracy is computed only based on the test set

Training and test sets



Outline

1 Evaluating forecast accuracy

2 Evaluating point forecast accuracy

3 Evaluating distributional forecast accuracy

Forecast errors

Forecast “error”: the difference between an observed value and its forecast.

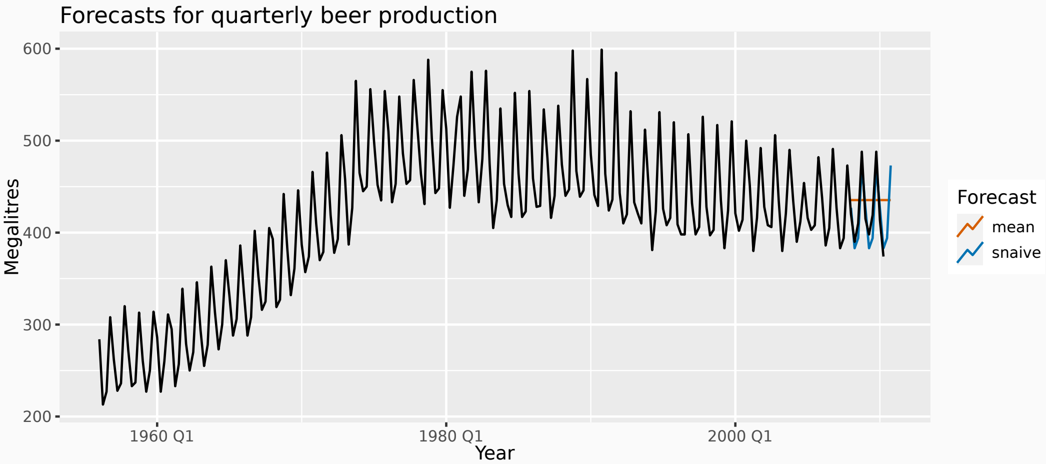
$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T},$$

where the training data is given by $\{y_1, \dots, y_T\}$

Measures of forecast accuracy

```r
beer_fit <- aus_production |>
  filter(between(year(Quarter), 1992, 2007)) |>
  model(
    snaive = SNAIVE(Beer),
    mean = MEAN(Beer)
  )
beer_fit |>
  forecast(h = "3 years") |>
  autoplot(aus_production, level = NULL) +
  labs(title ="Forecasts for quarterly beer production",
       x ="Year", y ="Megalitres") +
  guides(colour = guide_legend(title = "Forecast"))
```

Measures of forecast accuracy



Measures of forecast accuracy

$$
\begin{aligned}
y_{T+h} &= (T+h)\text{th observation, } h = 1,\dots,H \\
\hat{y}_{T+h|T} &= \text{its forecast based on data up to time } T. \\
e_{T+h} &= y_{T+h} - \hat{y}_{T+h|T}
\end{aligned}
$$

$$
\begin{aligned}
\text{MAE} &= \text{mean}(|e_{T+h}|) \\
\text{MSE} &= \text{mean}(e^2_{T+h}) \qquad \text{RMSE} = \sqrt{\text{mean}(e^2_{T+h})} \\
\text{MAPE} &= 100\text{mean}(|e_{T+h}|/|y_{T+h}|)
\end{aligned}
$$

Measures of forecast accuracy

$$\begin{aligned} y_{T+h} &= (T+h)\text{th observation}, h = 1, \dots, H \\ \hat{y}_{T+h|T} &= \text{its forecast based on data up to time } T. \\ e_{T+h} &= y_{T+h} - \hat{y}_{T+h|T} \end{aligned}$$

$$\begin{aligned} \text{MAE} &= \text{mean}(|e_{T+h}|) \\ \text{MSE} &= \text{mean}(e^2_{T+h}) \qquad \text{RMSE} = \sqrt{\text{mean}(e^2_{T+h})} \\ \text{MAPE} &= 100\text{mean}(|e_{T+h}|/|y_{T+h}|) \end{aligned}$$

- MAE, MSE, RMSE are all scale dependent.
- MAPE is scale independent but is only sensible if $y_t \gg 0$ for all t, and y has a natural zero.

Measures of forecast accuracy

Mean Absolute Scaled Error

$$\text{MASE} = \text{mean}(|e_{T+h}|/Q)$$

Measures of forecast accuracy

Mean Absolute Scaled Error

$$\text{MASE} = \text{mean}(|e_{T+h}|/Q)$$

- For non-seasonal series, scale uses naïve forecasts:

$$Q = \frac{1}{T-1} \sum_{t=2}^{T} |y_t - y_{t-1}|$$

- For seasonal series, scale uses seasonal naïve forecasts:

$$Q = \frac{1}{T-m} \sum_{t=m+1}^{T} |y_t - y_{t-m}|$$

where m is the seasonal frequency

Measures of forecast accuracy

Mean Absolute Scaled Error

$$\text{MASE} = \text{mean}(|e_{T+h}|/Q)$$

- For non-seasonal series, scale uses naïve forecasts:

$$Q = \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|$$

- For seasonal series, scale uses seasonal naïve forecasts:

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|$$

where m is the seasonal frequency

Proposed by Hyndman and Koehler (IJF, 2006).

Measures of forecast accuracy

Root Mean Squared Scaled Error

$$\text{RMSSE} = \sqrt{\text{mean}(e_{T+h}^2/Q)}$$

- For non-seasonal series, scale uses naïve forecasts:

$$Q = \frac{1}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1})^2$$

- For seasonal series, scale uses seasonal naïve forecasts:

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T}(y_t - y_{t-m})^2$$

where m is the seasonal frequencyq

Proposed by Hyndman and Koehler (IJF, 2006).

Measures of forecast accuracy

```r
beer_fc <- forecast(beer_fit, h = "3 years")
accuracy(beer_fc, aus_production)
```

```
# A tibble: 2 x 10
  .model .type     ME  RMSE   MAE   MPE  MAPE  MASE RMSSE    ACF1
  <chr>  <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
1 mean   Test  -13.8   38.4  34.8 -3.97  8.28 2.20  1.96  -0.0691
2 snaive Test    5.2   14.3  13.4  1.15  3.17 0.847 0.729  0.132
```

Outline

1 Evaluating forecast accuracy

2 Evaluating point forecast accuracy

3 Evaluating distributional forecast accuracy

Prediction interval accuracy using winkler score

Winkler proposed a scoring method to enable comparisons between prediction intervals:

- it takes account of both coverage and width of the intervals.

Winkler score

$$W(l_t, u_t, y_t) = \begin{cases} u_t - l_t & \text{if } l_t < y_t < u_t \\ (u_t - l_t) + \frac{2}{\alpha}(l_t - y_t) & \text{if } y_t < l_t \\ (u_t - l_t) + \frac{2}{\alpha}(y_t - u_t) & \text{if } y_t > u_t \end{cases}$$

Prediction interval accuracy

```
  # Compute interval accuracy
beer_fc |> accuracy(aus_production,
   measures = interval_accuracy_measures)
```

```
# A tibble: 2 x 3
  .model .type winkler
  <chr>  <chr>   <dbl>
1 mean   Test    174.
2 snaive Test     86.3
```

Quantile score

Quantile score

$$Q_{p,t} = \begin{cases} 2(1-p)(f_{p,t} - y_t), & \text{if } y_t < f_{p,t} \\ 2p(y_t - f_{p,t}), & \text{if } y_t \geq f_{p,t} \end{cases}$$

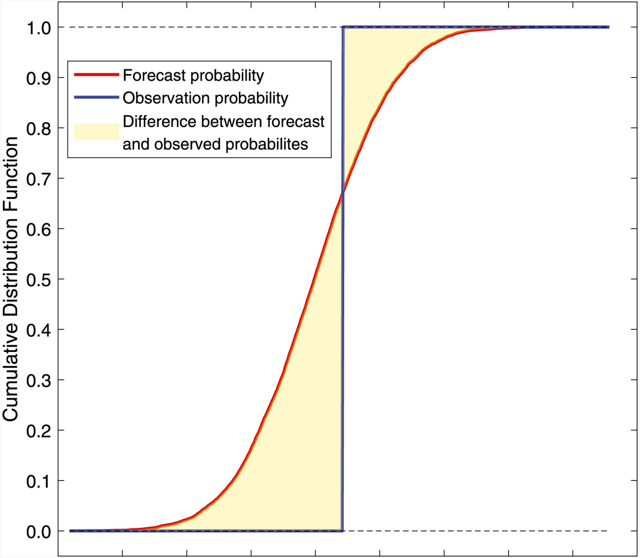
Continuous Ranked Probability Score (CRPS)

$$\text{CRPS} = \text{mean}(p_j),$$

where

$$p_j = \int_{-\infty}^{\infty} \left(G_j(x) - F_j(x)\right)^2 dx,$$

Continuous Ranked Probability Score (CRPS)



Quantile score and CRPS

```
beer_fc |>
  accuracy(aus_production, list(measures=distribution_accuracy_measures))
```

```
# A tibble: 2 x 4
  .model .type percentile  CRPS
  <chr>  <chr>      <dbl> <dbl>
1 mean   Test       22.7  22.4
2 snaive Test        8.80  8.71
```

Quantile score

```
beer_fc |>
  accuracy(aus_production, list(qs=quantile_score), probs = .9)
```

```
# A tibble: 2 x 3
  .model .type    qs
  <chr>  <chr> <dbl>
1 mean   Test  14.1
2 snaive Test   4.60
```