MONASH University

MONASH BUSINESS SCHOOL

2017 Beijing Workshop on Forecasting

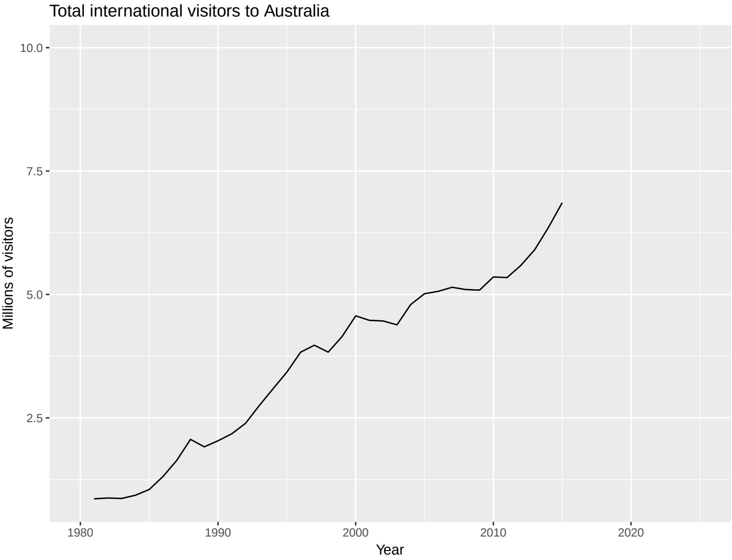
Forecast Accuracy and Evaluation

Rob J Hyndman

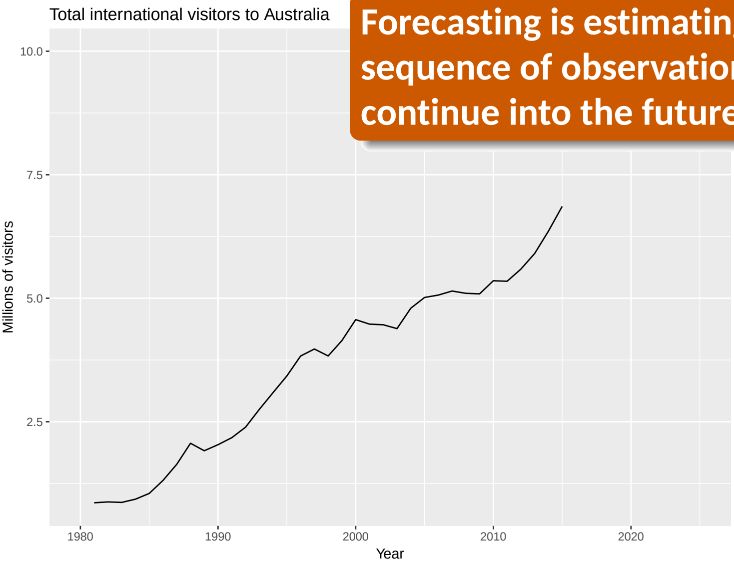
robjhyndman.com/beijing2017

Outline

The statistical forecasting perspective

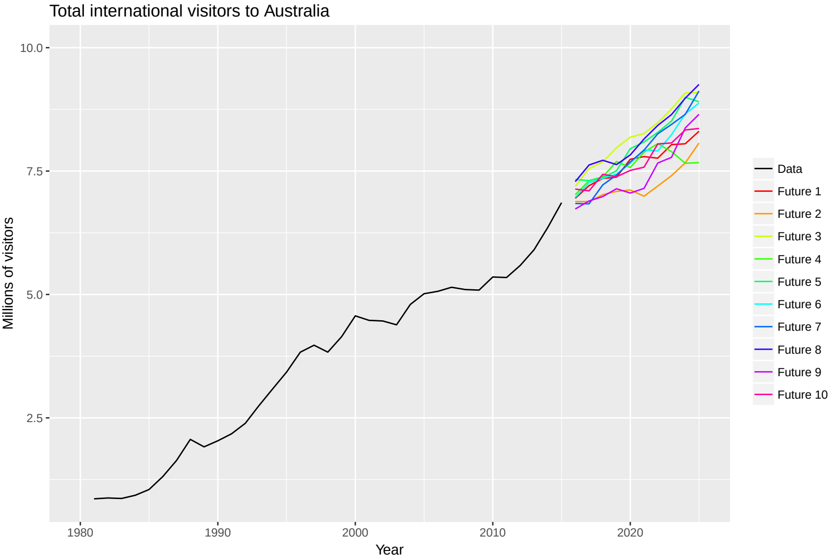


The statistical forecasting perspective

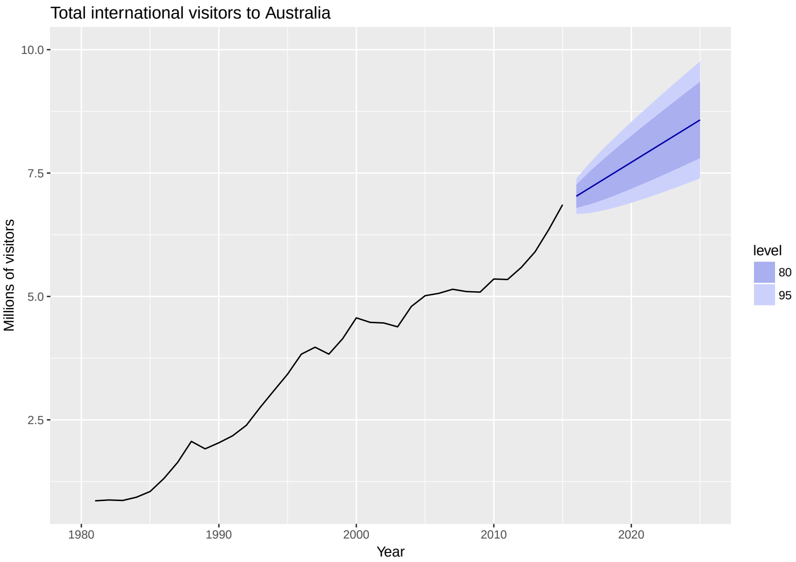


Forecasting is estimating how the sequence of observations will continue into the future.

Sample futures



Forecast intervals



Statistical forecasting

- Thing to be forecast: a random variable, y_t.

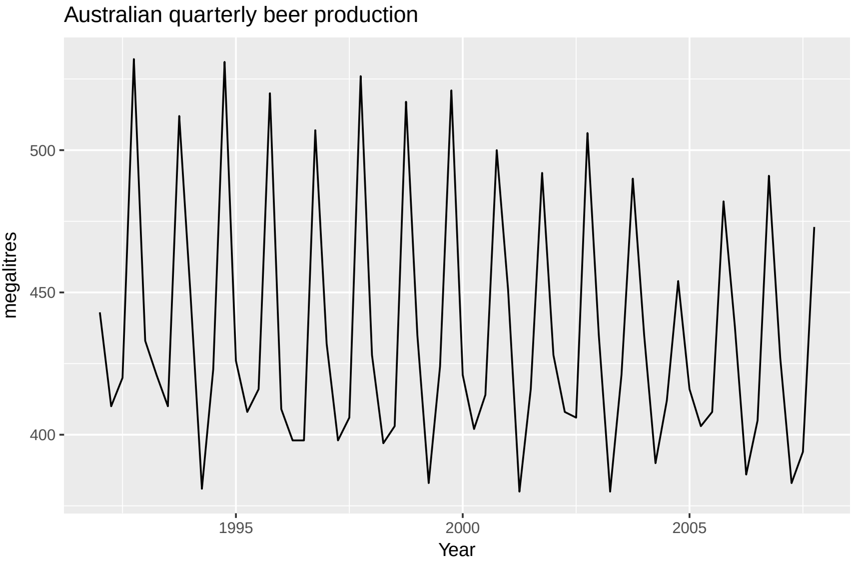
Forecast distributions:

$$y_{t|t-1} = y_t \big| \{y_1, y_2, \dots, y_{t-1}\}$$
$$y_{T+h|T} = y_{T+h} \big| \{y_1, y_2, \dots, y_T\}$$

- The "point forecast" is the mean (or median) of $y_{T+h|T} = y_{T+h} \big| \{y_1, y_2, \dots, y_T\}$
- The "forecast variance" is $\text{Var}[y_{T+h} | y_1, y_2, \dots, y_T]$
- A prediction interval or "interval forecast" is a range of values of y_t with high probability.

Outline

Some simple forecasting methods

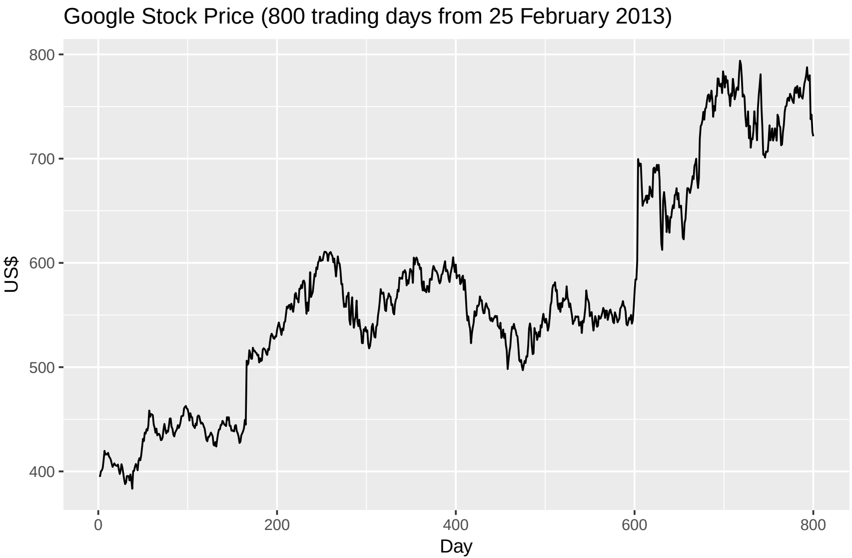


How would you forecast these data?

Some simple forecasting methods

How would you forecast these data?

Some simple forecasting methods



How would you forecast these data?

Some simple forecasting methods

Average method

- Forecast of all future values is equal to mean of historical data $\{y_1, \dots, y_T\}$.
- Forecasts: $\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T$

Naïve method

- Forecasts equal to last observed value.
- Forecasts: $\hat{y}_{T+h|T} = y_T$.
- Consequence of efficient market hypothesis.

Seasonal naïve method

- Forecasts equal to last value from same season.
- Forecasts: $\hat{y}_{T+h|T} = y_{T+h-km}$ where m = seasonal period and $k = \lfloor (h-1)/m \rfloor + 1$.

Some simple forecasting methods

Average method

- Forecast of all future values is equal to mean of historical data $\{y_1, \dots, y_T\}$.
- Forecasts: $\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T$

Naïve method

- Forecasts equal to last observed value.
- Forecasts: $\hat{y}_{T+h|T} = y_T$.
- Consequence of efficient market hypothesis.

Seasonal naïve method

- Forecasts equal to last value from same season.
- Forecasts: $\hat{y}_{T+h|T} = y_{T+h-km}$ where m = seasonal period and $k = \lfloor (h-1)/m \rfloor + 1$.

Some simple forecasting methods

Average method

- Forecast of all future values is equal to mean of historical data $\{y_1, \dots, y_T\}$.
- Forecasts: $\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T$

Naïve method

- Forecasts equal to last observed value.
- Forecasts: $\hat{y}_{T+h|T} = y_T$.
- Consequence of efficient market hypothesis.

Seasonal naïve method

- Forecasts equal to last value from same season.
- Forecasts: $\hat{y}_{T+h|T} = y_{T+h-km}$ where m = seasonal period and $k = \lfloor (h-1)/m \rfloor + 1$.

Some simple forecasting methods

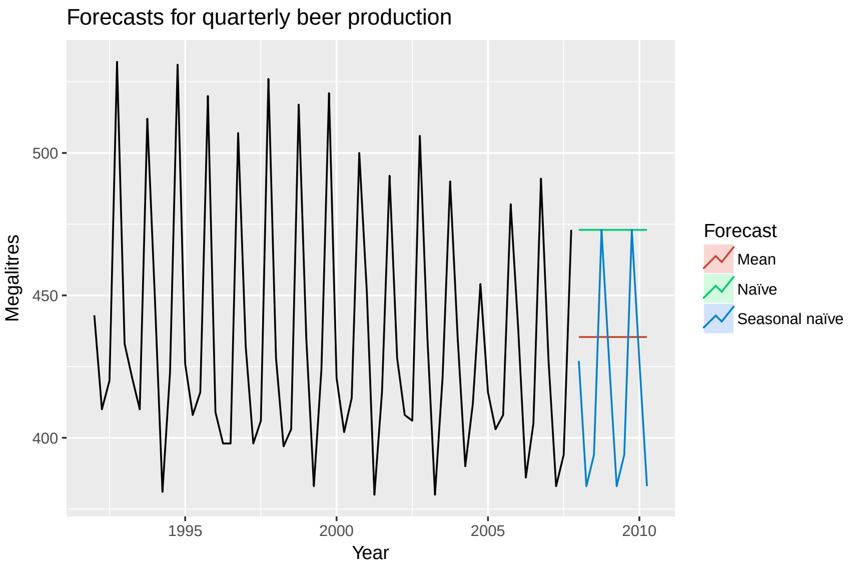
Drift method

- Forecasts equal to last value plus average change.
- Forecasts:

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1})$$
$$= y_T + \frac{h}{T-1}(y_T - y_1).$$

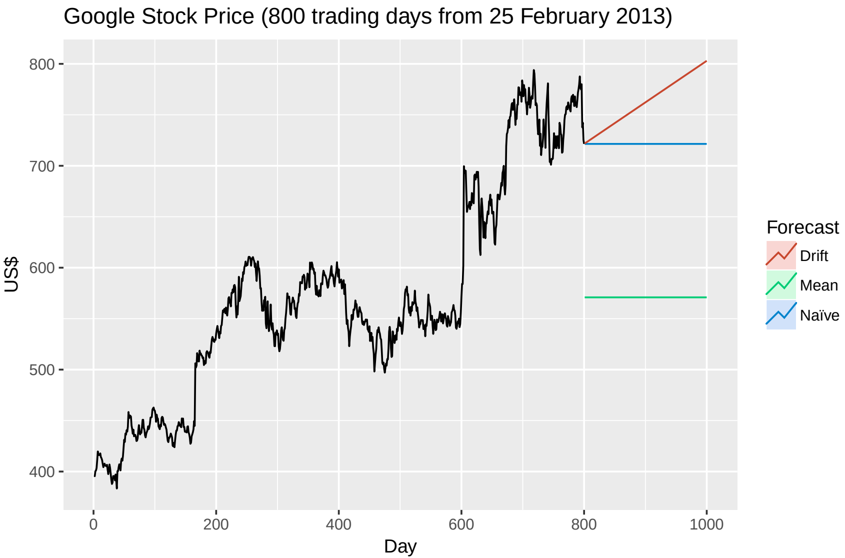
- Equivalent to extrapolating a line drawn between first and last observations.

Some simple forecasting methods



Some simple forecasting methods

Some simple forecasting methods



Outline

Fitted values

- $\hat{y}_{t|t-1}$ is the forecast of y_t based on observations $y_1, \dots, y_{t-1}$.
- We call these "fitted values".
- Sometimes drop the subscript: $\hat{y}_t \equiv \hat{y}_{t|t-1}$.
- Often not true forecasts since parameters are estimated on all data.

Examples:

1. $\hat{y}_t = \bar{y}$ for average method.
2. $\hat{y}_t = y_{t-1}$ for naive method
3. $\hat{y}_t = y_{t-1} + (y_T - y_1)/(T-1)$ for drift method.
4. $\hat{y}_t = y_{t-m}$ for seasonal naive method

Fitted values

- $\hat{y}_{t|t-1}$ is the forecast of y_t based on observations $y_1, \dots, y_{t-1}$.
- We call these "fitted values".
- Sometimes drop the subscript: $\hat{y}_t \equiv \hat{y}_{t|t-1}$.
- Often not true forecasts since parameters are estimated on all data.

Examples:

1. $\hat{y}_t = \bar{y}$ for average method.
2. $\hat{y}_t = y_{t-1}$ for naive method
3. $\hat{y}_t = y_{t-1} + (y_T - y_1)/(T - 1)$ for drift method.
4. $\hat{y}_t = y_{t-m}$ for seasonal naive method

Forecasting residuals

Residuals in forecasting: difference between observed value and its fitted value: $e_t = y_t - \hat{y}_{t|t-1}$.

Assumptions

1. $\{e_t\}$ uncorrelated. If they aren't, then information left in residuals that should be used in computing forecasts.
2. $\{e_t\}$ have mean zero. If they don't, then forecasts are biased.

Useful properties (for prediction intervals)

3. $\{e_t\}$ have constant variance.
4. $\{e_t\}$ are normally distributed.

Forecasting residuals

Residuals in forecasting: difference between observed value and its fitted value: $e_t = y_t - \hat{y}_{t|t-1}$.

Assumptions

1. $\{e_t\}$ uncorrelated. If they aren't, then information left in residuals that should be used in computing forecasts.
2. $\{e_t\}$ have mean zero. If they don't, then forecasts are biased.

Useful properties (for prediction intervals)

3. $\{e_t\}$ have constant variance.
4. $\{e_t\}$ are normally distributed.

Forecasting residuals

Residuals in forecasting: difference between observed value and its fitted value: $e_t = y_t - \hat{y}_{t|t-1}$.

Assumptions

1. $\{e_t\}$ uncorrelated. If they aren't, then information left in residuals that should be used in computing forecasts.
2. $\{e_t\}$ have mean zero. If they don't, then forecasts are biased.

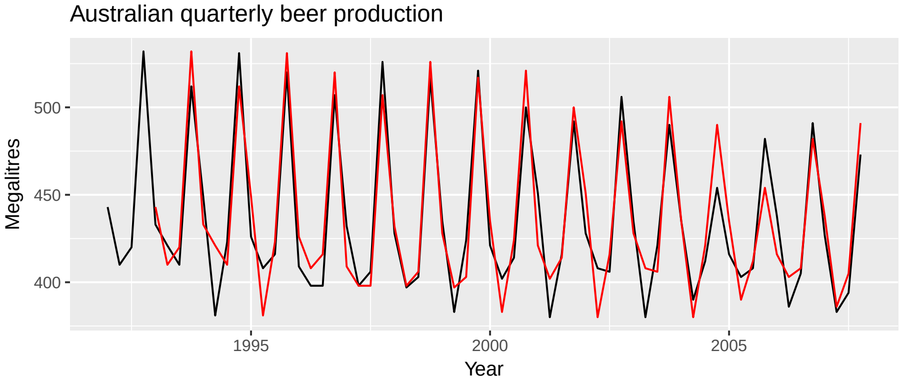
Useful properties (for prediction intervals)

3. $\{e_t\}$ have constant variance.
4. $\{e_t\}$ are normally distributed.

Example: Australian beer production

Seasonal naïve forecast:

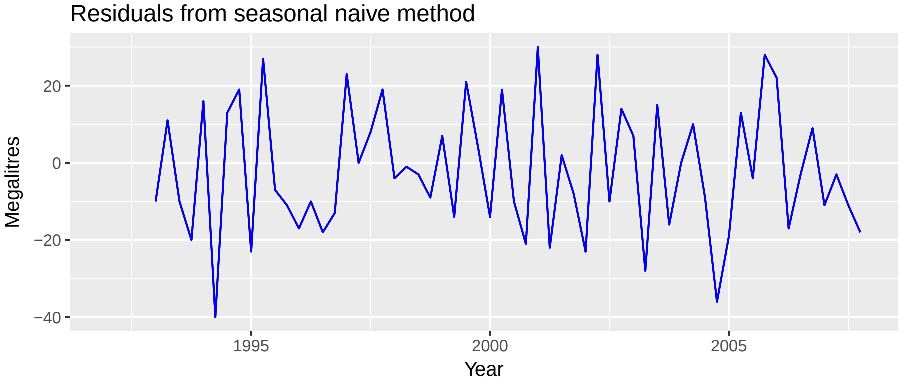
$$\hat{y}_{t|t-1} = y_{t-12} \qquad e_t = y_t - y_{t-12}$$



Example: Australian beer production

Seasonal naïve forecast:

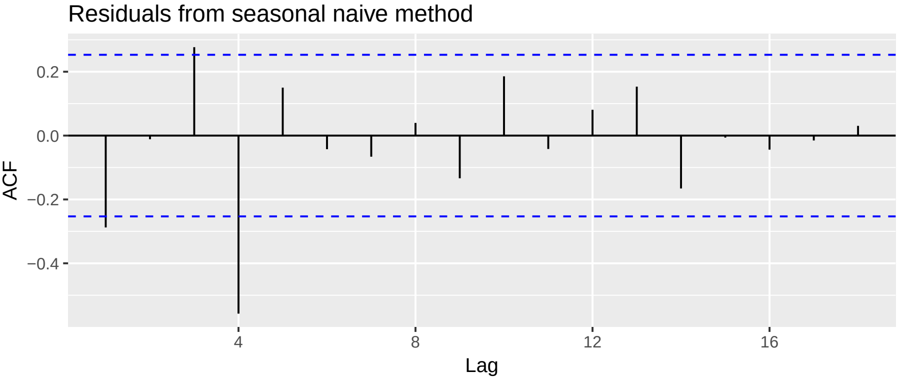
$$\hat{y}_{t|t-1} = y_{t-12} \qquad e_t = y_t - y_{t-12}$$



Example: Australian beer production

Seasonal naïve forecast:

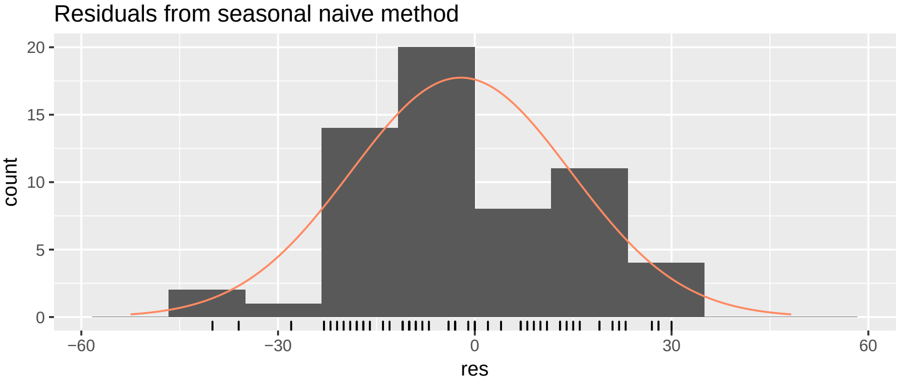
$$\hat{y}_{t|t-1} = y_{t-12} \qquad e_t = y_t - y_{t-12}$$



Example: Australian beer production

Seasonal naïve forecast:

$$\hat{y}_{t|t-1} = y_{t-12} \qquad e_t = y_t - y_{t-12}$$

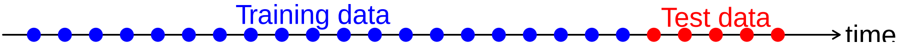


Forecasting residuals

- Minimizing the size of forecasting residuals is used for estimating model parameters (e.g., minimizing MSE or maximizing likelihood.
- In general, forecasting residuals cannot be used (directly) for estimating forecast accuracy.
- Forecast accuracy can only be measured using *genuine* forecasts; i.e., on different data.
- Forecasting residuals can help suggest model improvements.

Outline

Training and test sets



- The test set must not be used for *any* aspect of model development or calculation of forecasts.
- Forecast accuracy is based only on the test set.
- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters. (Compare R^2)
- Problems can be overcome by measuring true *out-of-sample* forecast accuracy. Training set used to estimate parameters. Forecasts are made for test set.

Measures of forecast accuracy

Training set: T observations

Test set: H observations

$$\text{MAE} = \frac{1}{H}\sum_{h=1}^{H} |y_{T+h} - \hat{y}_{T+h|T}|$$

$$\text{MSE} = \frac{1}{H}\sum_{h=1}^{H} (y_{T+h} - \hat{y}_{T+h|T})^2 \qquad \text{RMSE} = \sqrt{\frac{1}{H}\sum_{h=1}^{H} (y_{T+h} - \hat{y}_{T+h|T})^2}$$

$$\text{MAPE} = \frac{100}{H}\sum_{h=1}^{H} |y_{T+h} - \hat{y}_{T+h|T}|/|y_{T+h}|$$

- MAE, MSE, RMSE are all scale dependent.
- MAPE is scale independent but is only sensible if $y_{T+h} \gg 0$ for all h, and y has a natural zero.

Measures of forecast accuracy

Training set: T observations

Test set: H observations

$$\text{MAE} = \frac{1}{H}\sum_{h=1}^{H} |y_{T+h} - \hat{y}_{T+h|T}|$$

$$\text{MSE} = \frac{1}{H}\sum_{h=1}^{H} (y_{T+h} - \hat{y}_{T+h|T})^2 \qquad \text{RMSE} = \sqrt{\frac{1}{H}\sum_{h=1}^{H} (y_{T+h} - \hat{y}_{T+h|T})^2}$$

$$\text{MAPE} = \frac{100}{H}\sum_{h=1}^{H} |y_{T+h} - \hat{y}_{T+h|T}|/|y_{T+h}|$$

- MAE, MSE, RMSE are all scale dependent.
- MAPE is scale independent but is only sensible if $y_{T+h} \gg 0$ for all h, and y has a natural zero.

Measures of forecast accuracy

Mean Absolute Scaled Error

$$\text{MASE} = \frac{1}{H}\sum_{h=1}^{H} |y_{T+h} - \hat{y}_{T+h|T}|/Q$$

where Q is a stable measure of the scale of the time series $\{y_t\}$.

Proposed by Hyndman and Koehler (IJF, 2006).

For non-seasonal time series,

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

works well. Then MASE is equivalent to MAE relative to a naïve method.

Measures of forecast accuracy

Mean Absolute Scaled Error

$$\text{MASE} = \frac{1}{H}\sum_{h=1}^{H} |y_{T+h} - \hat{y}_{T+h|T}|/Q$$

where Q is a stable measure of the scale of the time series $\{y_t\}$.

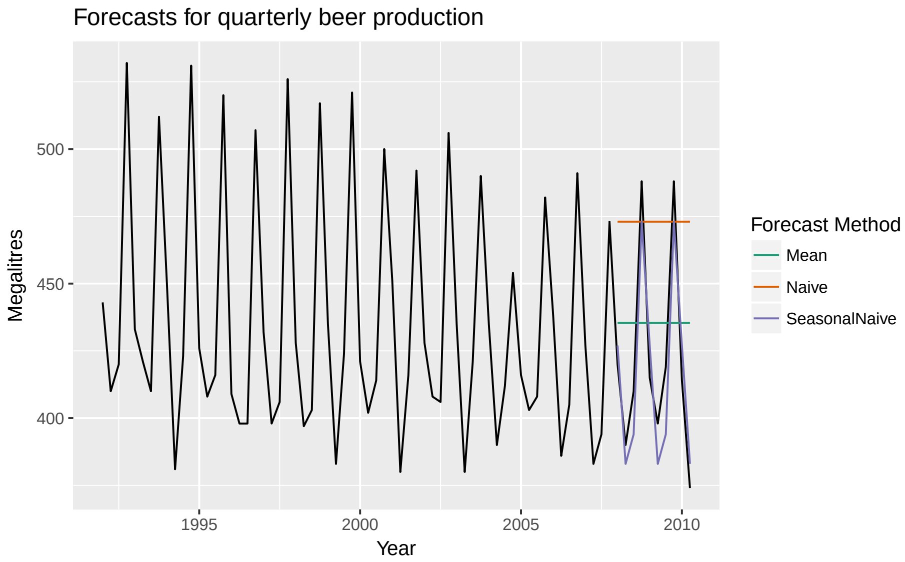
Proposed by Hyndman and Koehler (IJF, 2006).

For seasonal time series,

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

works well. Then MASE is equivalent to MAE relative to a seasonal naïve method.

Measures of forecast accuracy



Measures of forecast accuracy

	RMSE	MAE	MAPE	MASE
Mean method	38.5	34.8	8.28	2.44
Naïve method	62.7	57.4	14.18	4.01
Seasonal naïve method	14.3	13.4	3.17	0.94

Measures of forecast accuracy

Scaling can be used with any measure, and with different scaling statistics.

Mean Squared Scaled Error

$$\text{MASE} = \frac{1}{H}\sum_{h=1}^{H}(y_{T+h} - \hat{y}_{T+h|T})^2/Q$$

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

- Assumes $\{y_t\}$ is difference stationary.
- Minimizing MSSE leads to conditional mean forecasts.
- MSSE < 1 : out-of-sample multi-step forecasts are more accurate than in-sample one-step forecasts.

Measures of forecast accuracy

- Many suggested scale-free measures of forecast accuracy are degenerate due to infinite variance.
- The denominator must be positive with probability one.
- Distribution of most measures are highly skewed when applied to real data.

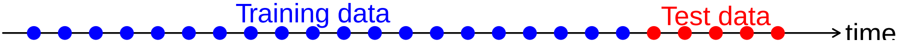
Poll: true or false?

1. Good forecast methods should have normally distributed residuals.
2. A model with small residuals will give good forecasts.
3. The best measure of forecast accuracy is MAPE.
4. If your model doesn’t forecast well, you should make it more complicated.
5. Always choose the model with the best forecast accuracy as measured on the test set.

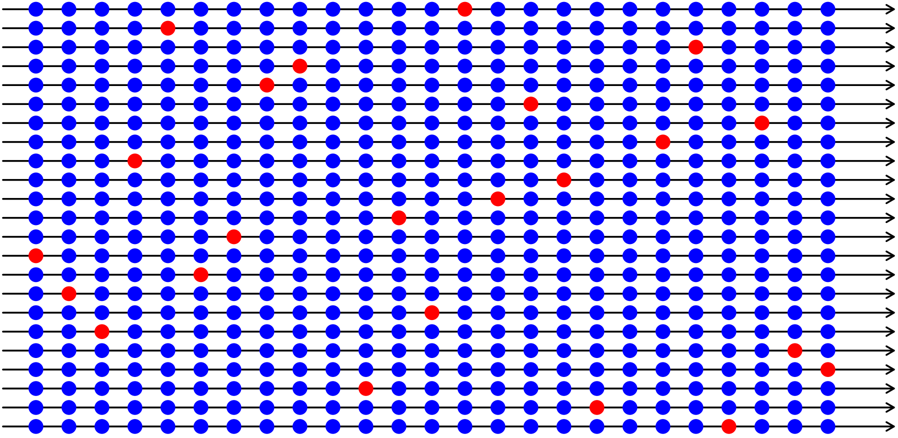
Outline

Cross-validation

Traditional evaluation

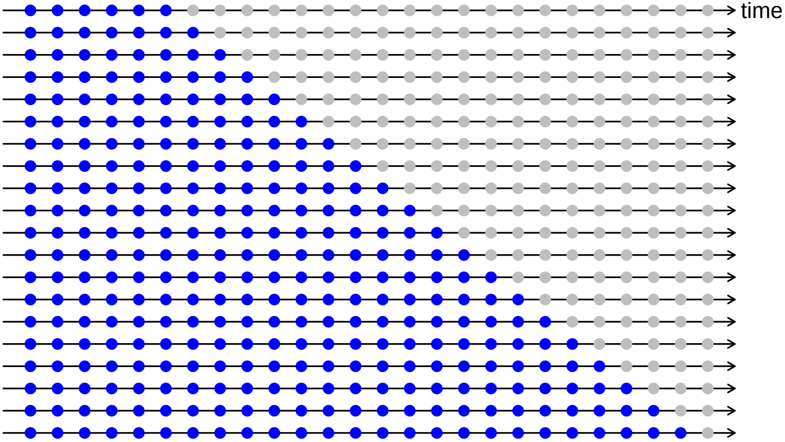


Leave-one-out cross-validation



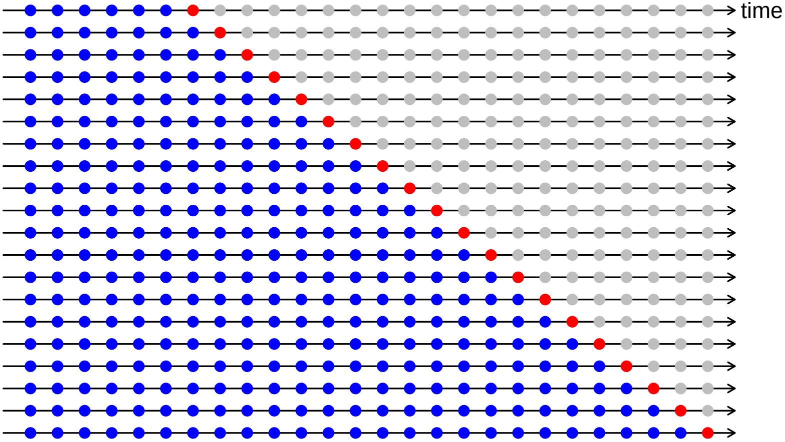
Cross-validation

Time series cross-validation



Cross-validation

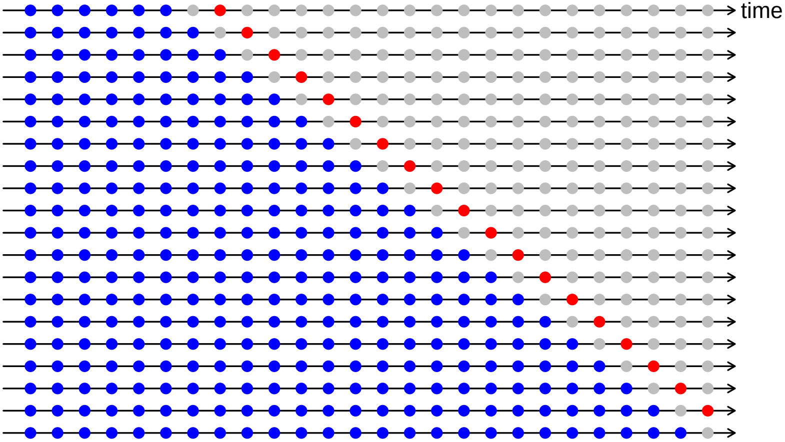
Time series cross-validation



$h = 1$

Cross-validation

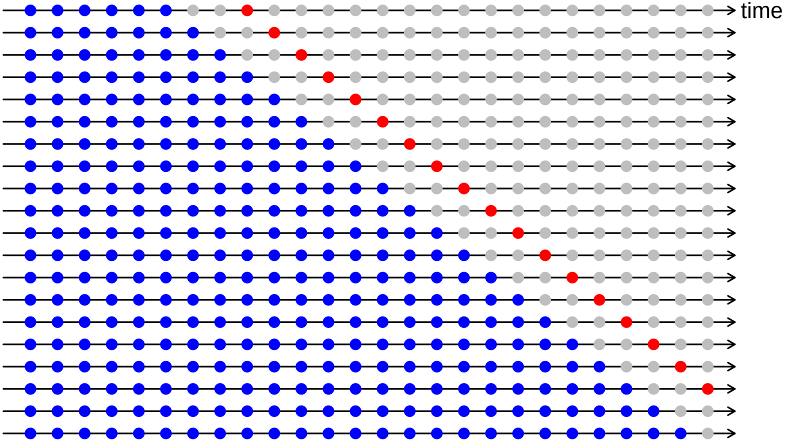
Time series cross-validation



$h = 2$

Cross-validation

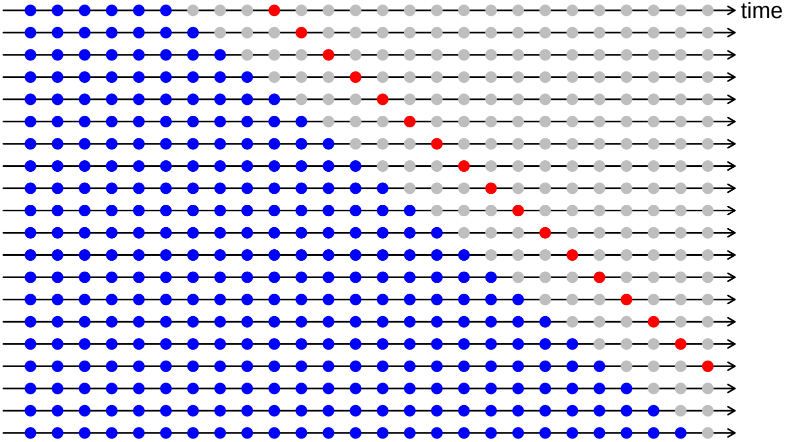
Time series cross-validation



$h = 3$

Cross-validation

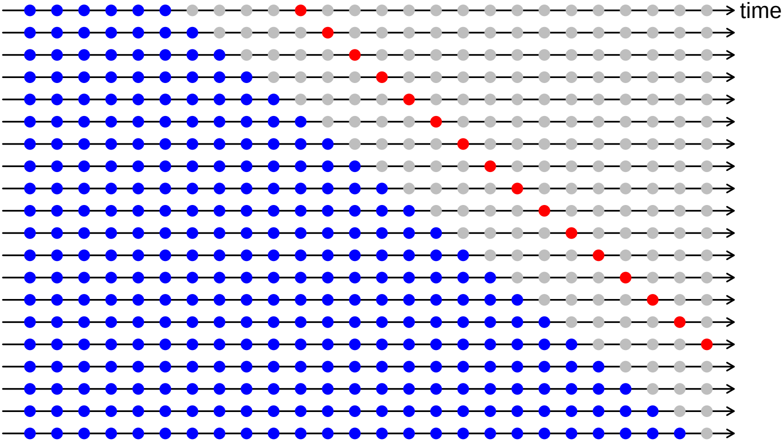
Time series cross-validation



$h = 4$

Cross-validation

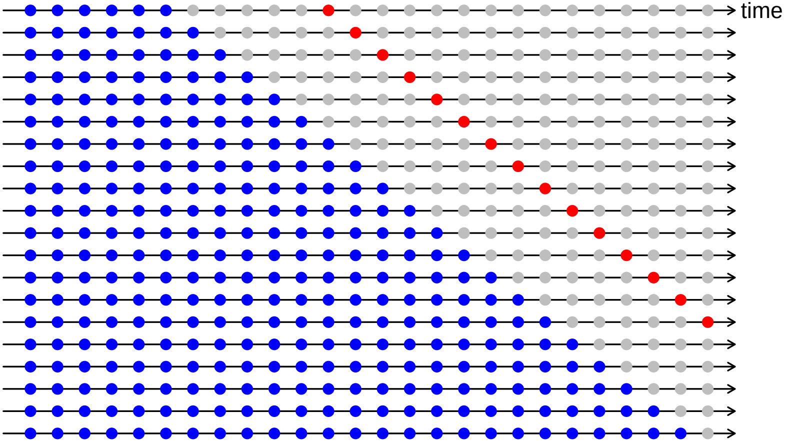
Time series cross-validation



$h = 5$

Cross-validation

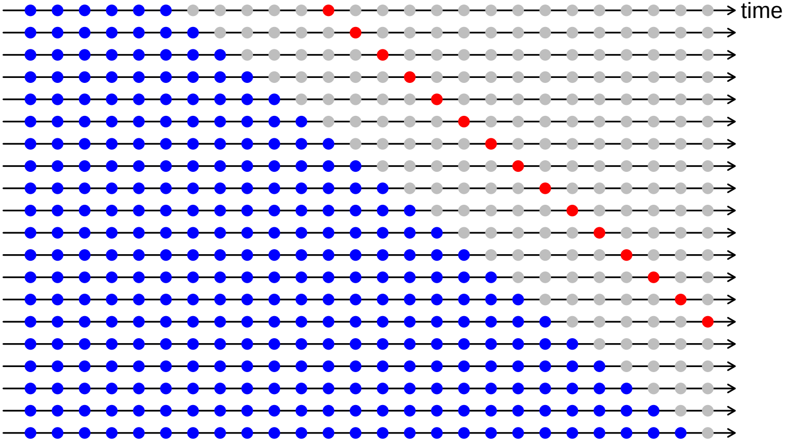
Time series cross-validation



$h = 6$

Cross-validation

Time series cross-validation



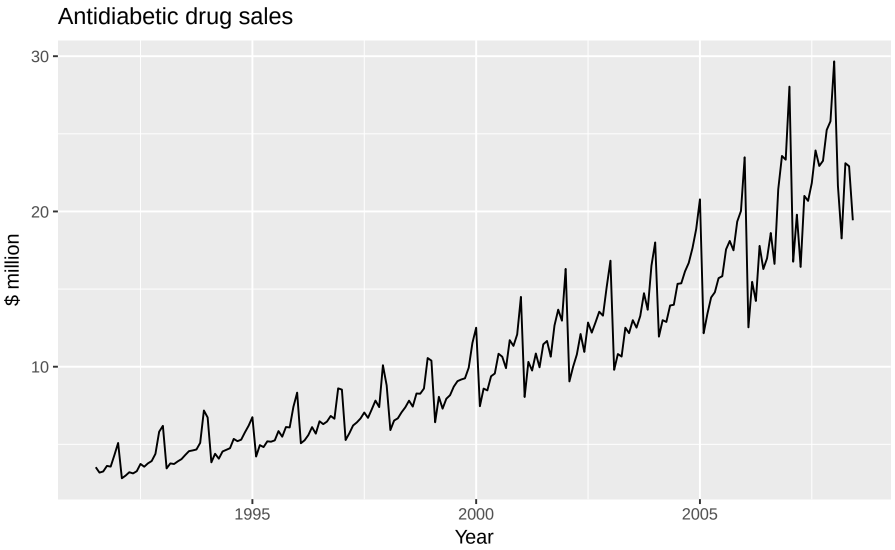
Also known as “Evaluation on a rolling forecast origin”

Time series cross-validation

Assume k is the minimum number of observations for a training set.

- Select observation $t + h$ for test set, and use observations at times $1, 2, \dots, t$ to estimate model.
- Compute error on forecast for time $t + h$.
- Repeat for $t = k, k + 1, \dots, T - h$ where T is total number of observations.
- Compute accuracy measure over all errors.

Example: Pharmaceutical sales



Example: Pharmaceutical sales

Which of these models is best?

- Linear model with trend and seasonal dummies applied to log data.
- ARIMA model applied to log data
- ETS model applied to original data

- Forecast $h = 12$ steps ahead based on data to time t for $t = 1, 2, \ldots$.
- Compare MAE values for each forecast horizon.

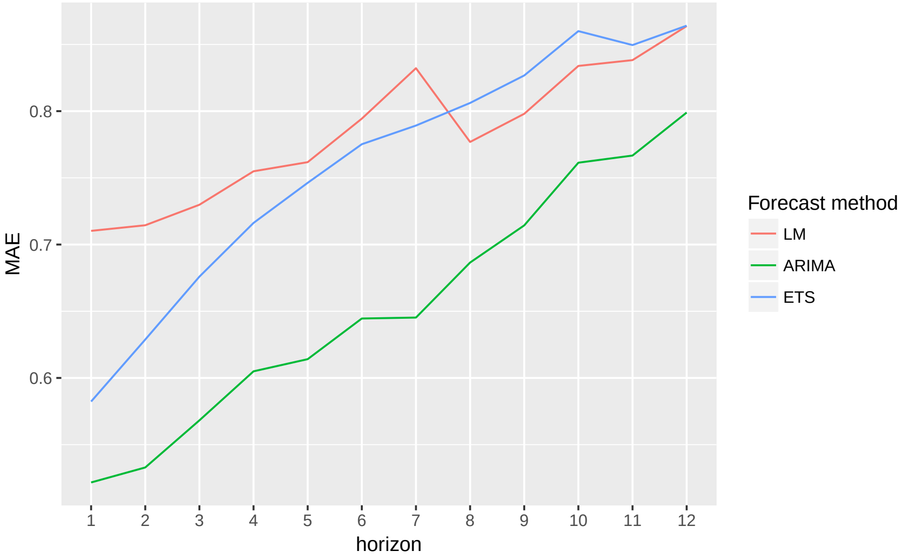
Example: Pharmaceutical sales

Which of these models is best?

- Linear model with trend and seasonal dummies applied to log data.
- ARIMA model applied to log data
- ETS model applied to original data

- Forecast $h = 12$ steps ahead based on data to time t for $t = 1, 2, \ldots$.
- Compare MAE values for each forecast horizon.

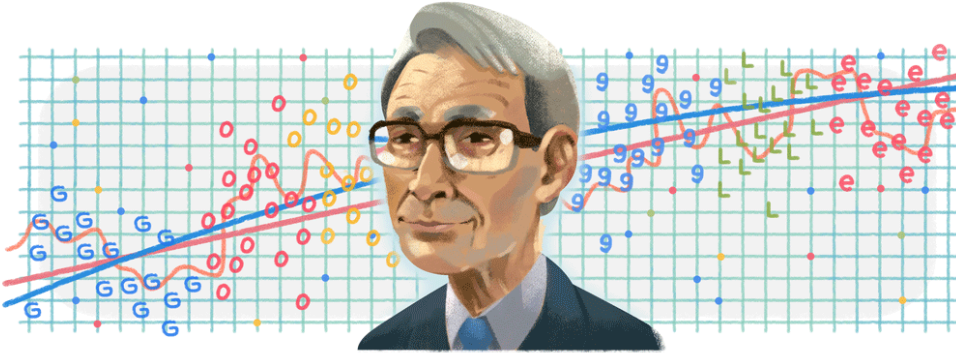
Example: Pharmaceutical sales



Example: R code

```r
f1 <- function(y,h) {
  forecast(tslm(y ~ trend + season, lambda=0), h=h)
}
f2 <- function(y,h) {
  forecast(auto.arima(y, D=1, lambda=0), h=h)
}
f3 <- function(y,h) {
  forecast(ets(y), h=h)
}
e1 <- tsCV(a10, f1, h=12)
e2 <- tsCV(a10, f2, h=12)
e3 <- tsCV(a10, f3, h=12)
mae <- data.frame(
    LM = colMeans(abs(e1), na.rm=TRUE),
    ARIMA = colMeans(abs(e2), na.rm=TRUE),
    ETS = colMeans(abs(e3), na.rm=TRUE))
```

Hirotugu Akaike (1927–2009)



Akaike, H. (1974), “A new look at the statistical model identification”, *IEEE Transactions on Automatic Control*, **19**(6): 716–723.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2k$$

where L is the model likelihood and k is the number of estimated parameters in the model.

- If L is Gaussian, then $\text{AIC} \approx c + T \log \text{MSE} + 2k$ where c is a constant, MSE is from one-step forecasts on **training set**, and T is the length of the series.

Minimizing the Gaussian AIC is asymptotically equivalent (as $T \to \infty$) to minimizing MSE from one-step forecasts on **test set** via time series cross-validation.

- AICc a bias-corrected small-sample version.
- AIC/AICc *much* faster than CV

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2k$$

where L is the model likelihood and k is the number of estimated parameters in the model.

- If L is Gaussian, then $\text{AIC} \approx c + T \log \text{MSE} + 2k$ where c is a constant, MSE is from one-step forecasts on **training set**, and T is the length of the series.

Minimizing the Gaussian AIC is asymptotically equivalent (as $T \to \infty$) to minimizing MSE from one-step forecasts on **test set** via time series cross-validation.

- AICc a bias-corrected small-sample version.
- AIC/AICc *much* faster than CV

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2k$$

where L is the model likelihood and k is the number of estimated parameters in the model.

- If L is Gaussian, then $\text{AIC} \approx c + T \log \text{MSE} + 2k$ where c is a constant, MSE is from one-step forecasts on **training set**, and T is the length of the series.

Minimizing the Gaussian AIC is asymptotically equivalent (as $T \to \infty$) to minimizing MSE from one-step forecasts on **test set** via time series cross-validation.

- AICc a bias-corrected small-sample version.
- AIC/AICc *much* faster than CV

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2k$$

where L is the model likelihood and k is the number of estimated parameters in the model.

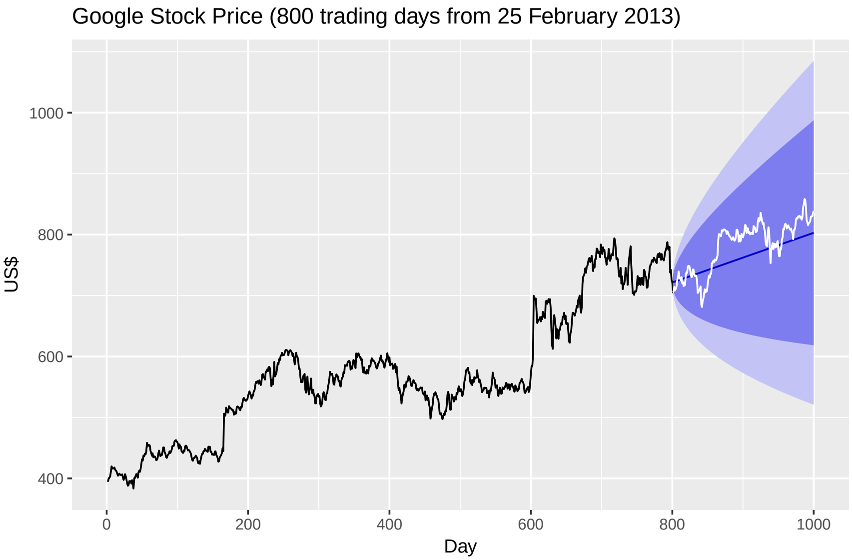
- If L is Gaussian, then $\text{AIC} \approx c + T \log \text{MSE} + 2k$ where c is a constant, MSE is from one-step forecasts on **training set**, and T is the length of the series.

Minimizing the Gaussian AIC is asymptotically equivalent (as $T \to \infty$) to minimizing MSE from one-step forecasts on **test set** via time series cross-validation.

- AICc a bias-corrected small-sample version.
- AIC/AICc *much* faster than CV

Outline

Probablistic forecasting



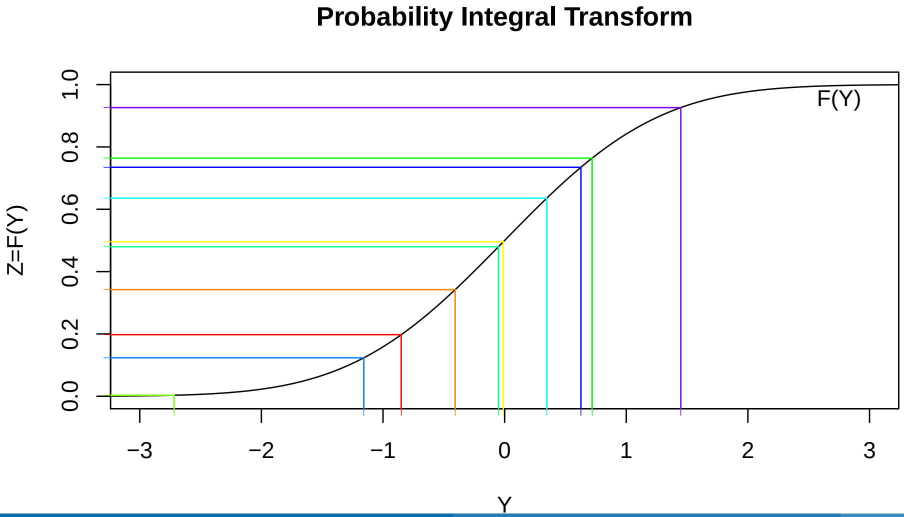
Probabilistic forecasting

How to evaluate a forecast probability distribution?

- Forecast intervals: percentage of observations covered compared to nominal percentage.
- Density forecasting
- Quantile forecasting
- Distribution forecasting

Probability Integral Transform

Let F = cdf of Y and $Z = F(Y)$. If F is continuous, then Z is standard uniform.



Calibration

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf.

Calibration

(a) $\hat{F}$ is marginally calibrated if $E[\hat{F}(y)] = P(Y \leq y)\ \forall y \in \mathbb{R}$.

(b) $\hat{F}$ is probabilistically calibrated if $Z = \hat{F}(Y)$ has a standard uniform distribution.

- ➥ We could plot a histogram of $Z = \hat{F}(Y)$ and check that it looks uniform.
- ➥ This is a more sophisticated version of testing if prediction intervals have the correct coverage.

Sharpness

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf.

Sharpness

➥ A "sharp" forecast distribution has narrow prediction intervals.

- A good probabilistic forecast is both calibrated and sharp.
- Scoring rules combine calibration and sharpness in a single measure.

Scoring rules

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf. A scoring rule assigns numerical score $S(\hat{F}_{T+h|T}, y_{T+h})$.

Dawid-Sebastiani score:

$$\text{DSS}(\hat{F}, y) = \frac{(y - \mu_{\hat{F}})^2}{\sigma_{\hat{F}}^2} + 2\log \sigma_{\hat{F}}$$

Generalization of MSE assuming normality.

A “proper” scoring rule has the property:

$$\mathrm{E}_F[S(F, Y)] < \mathrm{E}_F[S(\hat{F}, Y)]$$

Scoring rules

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf. A scoring rule assigns numerical score $S(\hat{F}_{T+h|T}, y_{T+h})$.

Dawid-Sebastiani score:

$$\text{DSS}(\hat{F}, y) = \frac{(y - \mu_{\hat{F}})^2}{\sigma_{\hat{F}}^2} + 2 \log \sigma_{\hat{F}}$$

Generalization of MSE assuming normality.

A “proper” scoring rule has the property:

$$\text{E}_F[S(F, Y)] < \text{E}_F[S(\hat{F}, Y)]$$

Scoring rules

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf. A scoring rule assigns numerical score $S(\hat{F}_{T+h|T}, y_{T+h})$.

Dawid-Sebastiani score:

$$\text{DSS}(\hat{F}, y) = \frac{(y - \mu_{\hat{F}})^2}{\sigma^2_{\hat{F}}} + 2\log \sigma_{\hat{F}}$$

Generalization of MSE assuming normality.

A "proper" scoring rule has the property:

$$\text{E}_F[S(F, Y)] < \text{E}_F[S(\hat{F}, Y)]$$

Scoring rules

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf. A scoring rule assigns numerical score $S(\hat{F}_{T+h|T}, y_{T+h})$.

Continuous Ranked Probability Score:

$$\text{CRPS}(\hat{F}, y) = \int [\hat{F}(x) - 1_{\{y \le x\}}]^2 dx = \text{E}_{\hat{F}}|Y - y| - \tfrac{1}{2}E_{\hat{F}}|Y - Y'|$$

where Y and Y' have cdf $\hat{F}$. *Generalization of MAE.*

Continuous Ranked Probability Score:

Let $\hat{Q}_{T+h|T} = \hat{F}^{-1}_{T+h|T}$ be the forecast quantile function

$$\text{CRPS}(\hat{Q}, y) = 2\int_0^1 \left[\hat{Q}(p) - y\right]\left[1_{\{y < \hat{Q}(p)\}} - p\right] dp$$

Scoring rules

$Y_{T+h|T}$ has cdf $F_{T+h|T}$. $\hat{F}_{T+h|T}$ is our forecast cdf. A scoring rule assigns numerical score $S(\hat{F}_{T+h|T}, y_{T+h})$.

Energy Score

$$\text{ES}(\hat{F}, y) = \text{E}_{\hat{F}}|Y - y|^{\alpha} - \tfrac{1}{2}E_{\hat{F}}|Y - Y'|^{\alpha}$$

where Y and Y' have cdf $\hat{F}$ and $\alpha \in (0, 2]$.

Log Score

$$\text{logS}(\hat{F}, y) = -\log \hat{f}(y)$$

where $\hat{f} = d\hat{F}/dy$ is density corresponding to $\hat{F}$.

R packages

https://github.com/earowang/tsibble

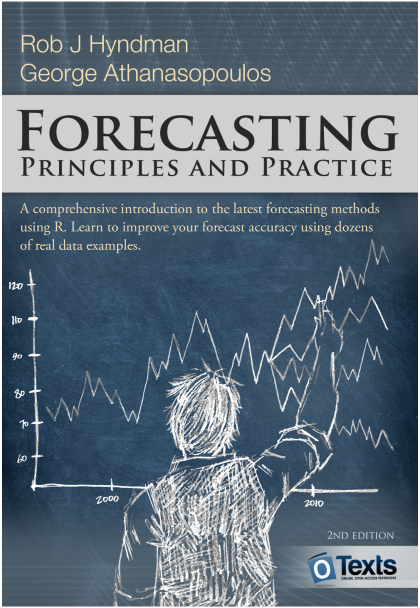
http://pkg.earo.me/sugrrants

https://github.com/mitchelloharawild/fasster

http://pkg.robjhyndman.com/forecast

http://pkg.earo.me/hts

Textbook



OTexts.org/fpp2