1.3 Forecast evaluation
Some simple forecasting methods

How would you forecast these data?
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 \( \{y_1, \ldots, y_T\} \).
- Forecasts: \( \hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T \).
### Some simple forecasting methods

#### Average method
- Forecast of all future values is equal to mean of historical data \( \{y_1, \ldots, y_T\} \).
- 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.
Some simple forecasting methods

**Average method**
- Forecast of all future values is equal to mean of historical data \( \{y_1, \ldots, y_T\} \).
- 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 \( k \) is integer part of \( (h - 1)/m \).
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

Forecasts for quarterly beer production

Year | Megalitres
--- | ---
1995 | 400
2000 | 450
2005 | 500
2010 | 550

Forecast
- Mean
- Naïve
- Seasonal naïve
Some simple forecasting methods

Dow Jones Index (daily ending 15 Jul 94)
Some simple forecasting methods

- **Mean**: `meanf(y, h=20)`
- **Naïve**: `naive(y, h=20)`
- **Seasonal naïve**: `snaive(y, h=20)`
- **Drift**: `rwf(y, drift=TRUE, h=20)`
Outline

1. Benchmark methods
2. Forecasting residuals
3. Lab session 4
4. Evaluating forecast accuracy
5. Lab session 5
6. Time series cross-validation
7. Lab session 6
8. Prediction intervals
Fitted values

- \( \hat{y}_{t|t-1} \) is the forecast of \( y_t \) based on observations \( y_1, \ldots, y_t \).
- 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.

For example:

- \( \hat{y}_t = \bar{y} \) for average method.
- \( \hat{y}_t = y_{t-1} + (y_T - y_1)/(T - 1) \) for drift method.
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.
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.
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: Google stock price

```r
autoplot(goog200) +
  xlab("Day") +
  ylab("Closing Price (US$)") +
  ggtitle("Google Stock (daily ending 6 December 2013)")
```
Naïve forecast:

\[ \hat{y}_{t|t-1} = y_{t-1} \]
Naïve forecast:

\[ \hat{y}_{t|t-1} = y_{t-1} \]

\[ e_t = y_t - y_{t-1} \]
Naïve forecast:

\[ \hat{y}_{t|t-1} = y_{t-1} \]

\[ e_t = y_t - y_{t-1} \]

Note: \( e_t \) are one-step-forecast residuals
Example: Google stock price

```r
fits <- fitted(naive(goog200))
autoplot(goog200, series="Data") +
  autolayer(fits, series="Fitted") +
  xlab("Day") + ylab("Closing Price (US$)") +
  ggtitle("Google Stock (daily ending 6 December 2013)")
```

![Google Stock (daily ending 6 December 2013)](image)
Example: Google stock price

```r
res <- residuals(naive(goog200))
autoplot(res) + xlab("Day") + ylab("") + ggtitle("Residuals from naïve method")
```

Residuals from naïve method
Example: Google stock price

```
gghistogram(res, add.normal=TRUE) +
ggtitle("Histogram of residuals")
```

Histogram of residuals
Example: Google stock price

`ggAcf(res) + ggtitle("ACF of residuals")`
ACF of residuals

- We assume that the residuals are white noise (uncorrelated, mean zero, constant variance). If they aren’t, then there is information left in the residuals that should be used in computing forecasts.

- So a standard residual diagnostic is to check the ACF of the residuals of a forecasting method.

- We expect these to look like white noise.
checkresiduals function

```
checkresiduals(naive(goog200))
```

Residuals from Naive method

ACF

Ljung-Box test

data: Residuals from Naive method
Q* = 11, df = 10, p-value = 0.4

Model df: 0. Total lags used: 10
Portmanteau tests

Test whether set of $r_k$ values are significantly different from a zero set.

### Ljung-Box test

\[
Q^* = T(T + 2) \sum_{k=1}^{h} (T - k)^{-1} r_k^2
\]

where $h$ is max lag being considered and $T$ is number of observations.

- My preferences: $h = 10$ for non-seasonal data, $h = 2m$ for seasonal data.
- If each $r_k$ close to zero, $Q$ will be small.
- p-value measures probability of results if residuals are WN.
checkresiduals function

checkresiduals(naive(goog200))

##
## Ljung-Box test
##
## data: Residuals from Naive method
## Q* = 11, df = 10, p-value = 0.4
## Model df: 0. Total lags used: 10
1 Benchmark methods
2 Forecasting residuals
3 Lab session 4
4 Evaluating forecast accuracy
5 Lab session 5
6 Time series cross-validation
7 Lab session 6
8 Prediction intervals
Lab Session 4
1 Benchmark methods
2 Forecasting residuals
3 Lab session 4
4 Evaluating forecast accuracy
5 Lab session 5
6 Time series cross-validation
7 Lab session 6
8 Prediction intervals
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.

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.
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 \( \{y_1, \ldots, y_T\} \)

- Unlike residuals, forecast errors on the test set involve multi-step forecasts.
- These are \textit{true} forecast errors as the test data is not used in computing \( \hat{y}_{T+h|T} \).
Measures of forecast accuracy

Forecasts for GOOG stock price

Forecast Method
- Mean
- Naive
- Drift
Measures of forecast accuracy

\[ y_{T+h} = (T + h)\text{th observation, } h = 1, \ldots, H \]

\[ \hat{y}_{T+h|T} = \text{its forecast based on data up to time } T. \]

\[ e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \]

\[
\begin{align*}
\text{MAE} & = \text{mean}(|e_{T+h}|) \\
\text{MSE} & = \text{mean}(e_{T+h}^2) \\
\text{RMSE} & = \sqrt{\text{mean}(e_{T+h}^2)} \\
\text{MAPE} & = 100\text{mean}(|e_{T+h}|/|y_{T+h}|)
\end{align*}
\]

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

\[ y_{T+h} = (T + h)\text{th observation, } h = 1, \ldots, H \]
\[ \hat{y}_{T+h|T} = \text{its forecast based on data up to time } T. \]
\[ e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \]

\[ \text{MAE} = \text{mean}(|e_{T+h}|) \]
\[ \text{MSE} = \text{mean}(e_{T+h}^2) \]
\[ \text{RMSE} = \sqrt{\text{mean}(e_{T+h}^2)} \]
\[ \text{MAPE} = 100\text{mean}(|e_{T+h}|/|y_{T+h}|) \]

- 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

\[
MASE = T^{-1} \sum_{t=1}^{T} \frac{|y_t - \hat{y}_{t|t-1}|}{Q}
\]

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,

\[
Q = (T - 1)^{-1} \sum_{t=2}^{T} |y_t - y_{t-1}|
\]

works well. Then MASE is equivalent to MAE relative to a naïve method.
# Measures of forecast accuracy

## Mean Absolute Scaled Error

\[
MASE = T^{-1} \sum_{t=1}^{T} \frac{|y_t - \hat{y}_{t|t-1}|}{Q}
\]

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,

\[
Q = (T - m)^{-1} \sum_{t=m+1}^{T} |y_t - y_{t-m}|
\]

works well. Then MASE is equivalent to MAE relative to a seasonal naïve method.
Measures of forecast accuracy

Forecasts for GOOG stock price

Forecast Method
- Mean
- Naive
- Drift
Measures of forecast accuracy

googtrain <- `window`(goog200, end=180)
googfc1 <- `meanf`(googtrain, h=20)
googfc2 <- `rwf`(googtrain, h=20)
googfc3 <- `rwf`(googtrain, h=20, drift=TRUE)
accuracy(googfc1, goog200)
accuracy(googfc2, goog200)
accuracy(googfc3, goog200)

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean method</td>
<td>82.89</td>
<td>82.43</td>
<td>15.93</td>
<td>21.61</td>
</tr>
<tr>
<td>Naïve method</td>
<td>18.29</td>
<td>16.04</td>
<td>3.08</td>
<td>4.21</td>
</tr>
<tr>
<td>Drift method</td>
<td>11.34</td>
<td>9.71</td>
<td>1.86</td>
<td>2.55</td>
</tr>
</tbody>
</table>
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.
Lab Session 5
1 Benchmark methods
2 Forecasting residuals
3 Lab session 4
4 Evaluating forecast accuracy
5 Lab session 5
6 Time series cross-validation
7 Lab session 6
8 Prediction intervals
Time series cross-validation

Traditional evaluation

Training data

Test data

↑ time
Time series cross-validation

Traditional evaluation

Time series cross-validation
Forecast accuracy averaged over test sets.

Also known as “evaluation on a rolling forecasting origin”
A good way to choose the best forecasting model is to find the model with the smallest RMSE computed using time series cross-validation.
Pipe function

Ugly code:

```r
e <- tsCV(goog200, rwf, drift=TRUE, h=1)
sqrt(mean(e^2, na.rm=TRUE))
sqrt(mean(residuals(rwf(goog200, drift=TRUE))^2, na.rm=TRUE))
```

Better with a pipe:

```r
goog200 %>%
  tsCV(forecastfunction=rwf, drift=TRUE, h=1) -> e
e^2 %>% mean(na.rm=TRUE) %>% sqrt
```

```r
goog200 %>%
  rwf(drift=TRUE) %>% residuals -> res
res^2 %>% mean(na.rm=TRUE) %>% sqrt
```
1. Benchmark methods
2. Forecasting residuals
3. Lab session 4
4. Evaluating forecast accuracy
5. Lab session 5
6. Time series cross-validation
7. Lab session 6
8. Prediction intervals
Lab Session 6
1. Benchmark methods
2. Forecasting residuals
3. Lab session 4
4. Evaluating forecast accuracy
5. Lab session 5
6. Time series cross-validation
7. Lab session 6
8. Prediction intervals
A forecast $\hat{y}_{T+h|T}$ is (usually) the mean of the conditional distribution $y_{T+h} \mid y_1, \ldots, y_T$.

A prediction interval gives a region within which we expect $y_{T+h}$ to lie with a specified probability.

Assuming forecast errors are normally distributed, then a 95% PI is

$$\hat{y}_{T+h|T} \pm 1.96\hat{\sigma}_h$$

where $\hat{\sigma}_h$ is the st dev of the $h$-step distribution.

When $h = 1$, $\hat{\sigma}_h$ can be estimated from the residuals.
Drift forecasts with prediction interval:

\[ \text{rwf}(\text{goog200}, \text{level}=95, \text{drift}=\text{TRUE}) \]

<table>
<thead>
<tr>
<th>#</th>
<th>Point Forecast</th>
<th>Lo 95</th>
<th>Hi 95</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>532.2</td>
<td>520.0</td>
<td>544.3</td>
</tr>
<tr>
<td>202</td>
<td>532.9</td>
<td>515.6</td>
<td>550.1</td>
</tr>
<tr>
<td>203</td>
<td>533.6</td>
<td>512.4</td>
<td>554.7</td>
</tr>
<tr>
<td>204</td>
<td>534.3</td>
<td>509.8</td>
<td>558.7</td>
</tr>
<tr>
<td>205</td>
<td>535.0</td>
<td>507.5</td>
<td>562.4</td>
</tr>
<tr>
<td>206</td>
<td>535.7</td>
<td>505.5</td>
<td>565.8</td>
</tr>
<tr>
<td>207</td>
<td>536.4</td>
<td>503.7</td>
<td>569.0</td>
</tr>
<tr>
<td>208</td>
<td>537.1</td>
<td>502.1</td>
<td>572.0</td>
</tr>
</tbody>
</table>
Prediction intervals

- Point forecasts are often useless without prediction intervals.
- Prediction intervals require a stochastic model (with random errors, etc).
- Multi-step forecasts for time series require a more sophisticated approach (with PI getting wider as the forecast horizon increases).
- Check residual assumptions before believing them.
- Usually too narrow due to unaccounted uncertainty.
Prediction intervals

Assume residuals are normal, uncorrelated, sd = $\hat{\sigma}$:

<table>
<thead>
<tr>
<th>Type of Forecasts</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean forecasts:</td>
<td>$\hat{\sigma}_h = \hat{\sigma}\sqrt{1 + 1/T}$</td>
</tr>
<tr>
<td>Naïve forecasts:</td>
<td>$\hat{\sigma}_h = \hat{\sigma}\sqrt{h}$</td>
</tr>
<tr>
<td>Seasonal naïve forecasts</td>
<td>$\hat{\sigma}_h = \hat{\sigma}\sqrt{k + 1}$</td>
</tr>
<tr>
<td>Drift forecasts:</td>
<td>$\hat{\sigma}_h = \hat{\sigma}\sqrt{h(1 + h/T)}$.</td>
</tr>
</tbody>
</table>

where $k$ is the integer part of $(h - 1)/m$.

Note that when $h = 1$ and $T$ is large, these all give the same approximate value $\hat{\sigma}$. 