Lecture 3
Autocorrelation and Model selection

Autocorrelation Function

Autocorrelation

Consider a time series Y_t

- Mean: $E(Y_t) = \mu_t$
- Variance: $\mathrm{var}(Y_t)$.

The **autocovariance** of Y_t at lag j is

$$\mathrm{cov}(Y_t, Y_{t-j}) = E[\,(Y_t - \mu_t)\,(Y_{t-j} - \mu_{t-j})\,]$$

Autocorrelation

Consider a time series Y_t

- Mean: $E(Y_t) = \mu_t$
- Variance: $\mathrm{var}(Y_t)$.

The **autocorrelation** of Y_t at lag j is

$$\mathrm{cor}(Y_t, Y_{t-j}) = \frac{\mathrm{cov}(Y_t, Y_{t-j})}{\sqrt{\mathrm{var}(Y_t)\,\mathrm{var}(Y_{t-j})}}$$

Autocorrelations in data

For any j:

$$\widehat{\text{cov}}(Y_t, Y_{t-j}) = \frac{1}{n} \sum_{t=j+1}^{n} (Y_t - \widehat{\mu}_t) \, (Y_{t-j} - \widehat{\mu}_{t-j})$$

where typically

$$\widehat{\mu}_t = X_t' \widehat{\beta}$$

Note $\widehat{\text{var}}(Y_t) = \widehat{\text{cov}}(Y_t, Y_{t-j})$ with $j = 0$.

Autocorrelations in data

For any j:

$$\widehat{\text{cov}}(Y_t, Y_{t-j}) = \frac{1}{n} \sum_{t=j+1}^{n} \widehat{Z}_t \widehat{Z}_{t-j}$$

where $\widehat{Z}_t = Y_t - X_t' \widehat{\beta}$ are residuals from regression on trend, seasonals etc as appropriate.

Autocorrelations in retail sales

```
Y <- window(log(Retail_q), start=c(2000,1),
            end=c(2018,4))
Time <- time(Y)
gfc <- 2008.5
Time_postgfc <- 1*(Time>gfc)*(Time-gfc)
QD <- seasonaldummy(Y)
X <- cbind(Time, Time_postgfc, QD)
AR0 <- Arima(Y, order=c(0,0,0), xreg=X)
Z <- AR0$residuals
```

Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
print(round(acfZ$acf[1:5],3))
```

```
[1] 1.000 0.730 0.619 0.379 0.258
```

Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
print(round(acfZ$acf[1:5],3))
```

```
[1] 1.000 0.730 0.619 0.379 0.258
```

$\uparrow$

$\mathrm{cor}(\widehat{Z}_t, \widehat{Z}_t) = 1$

(obviously...)

Autocorrelations in retail sales

```r
Z <- AR0$residuals
acfZ <- acf(Z)
print(round(acfZ$acf[1:5],3))
```

```
[1] 1.000 0.730 0.619 0.379 0.258
```

$$\mathrm{cor}(\widehat{Z}_t, \widehat{Z}_{t-1}) = \mathbf{0.730}$$

Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
print(round(acfZ$acf[1:5],3))
```

```
[1] 1.000 0.730 0.619 0.379 0.258
```

$\uparrow$

$$\mathrm{cor}(\widehat{Z}_t, \widehat{Z}_{t-2}) = 0.619$$

Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
print(round(acfZ$acf[1:5],3))
```

```
[1] 1.000 0.730 0.619 0.379 0.258
```

$$\mathrm{cor}(\widehat{Z}_t, \widehat{Z}_{t-3}) = \mathbf{0.379}$$

Autocorrelations in retail sales

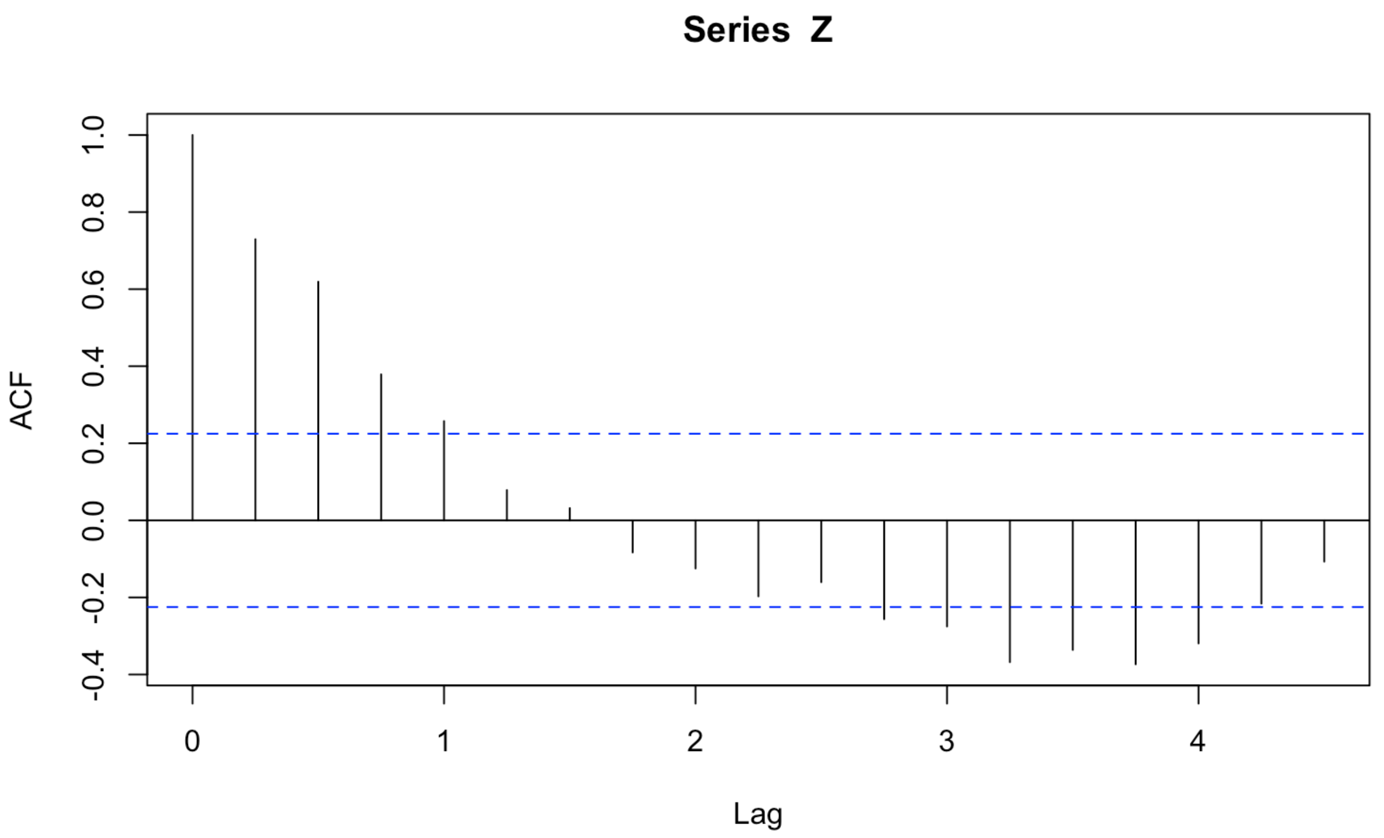
```
Z <- AR0$residuals
acfZ <- acf(Z)
print(round(acfZ$acf[1:5],3))
```

```
[1] 1.000 0.730 0.619 0.379 0.258
```

$$\mathrm{cor}(\widehat{Z}_t, \widehat{Z}_{t-4}) = 0.258$$

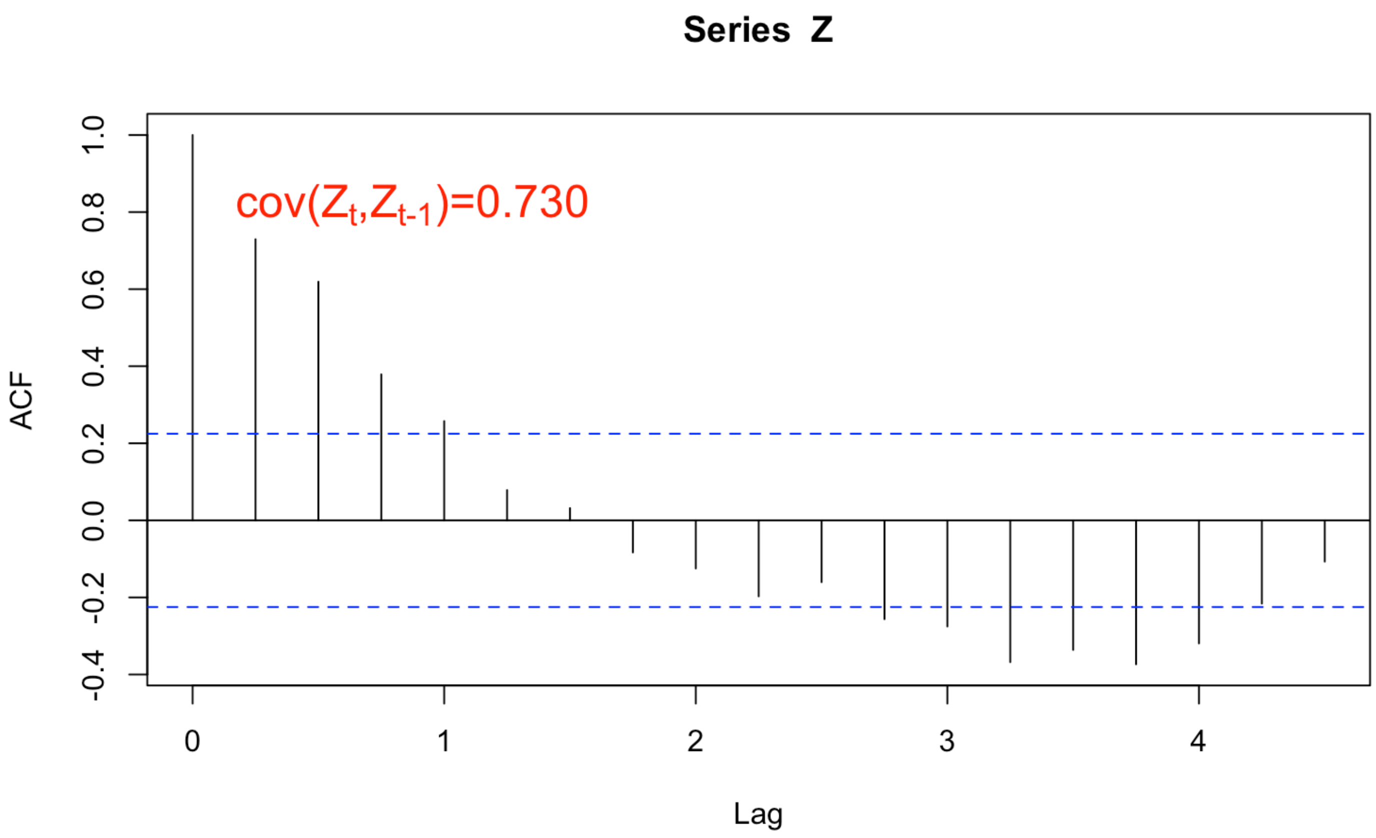
Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
```



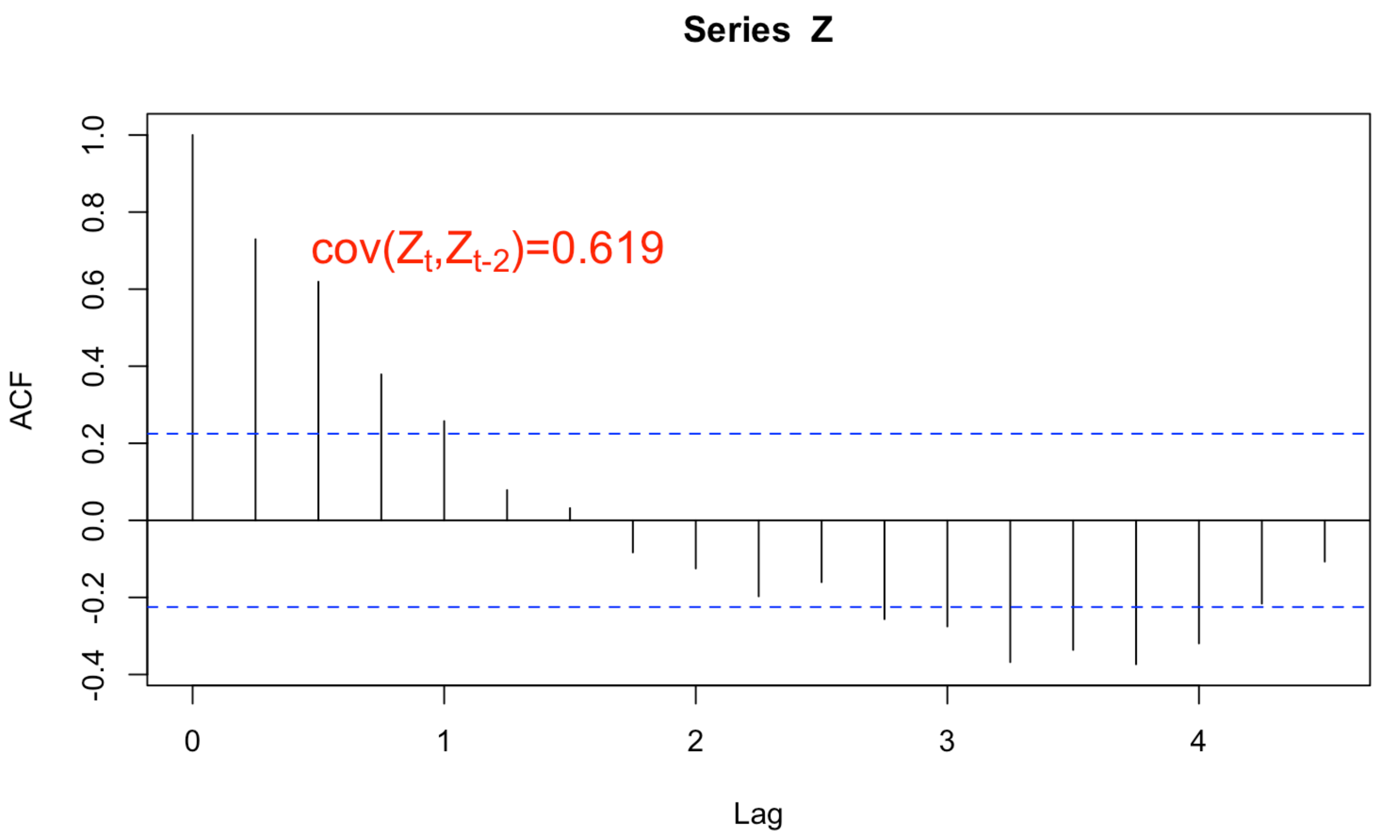
Autocorrelations in retail sales

```r
Z <- AR0$residuals
acfZ <- acf(Z)
```



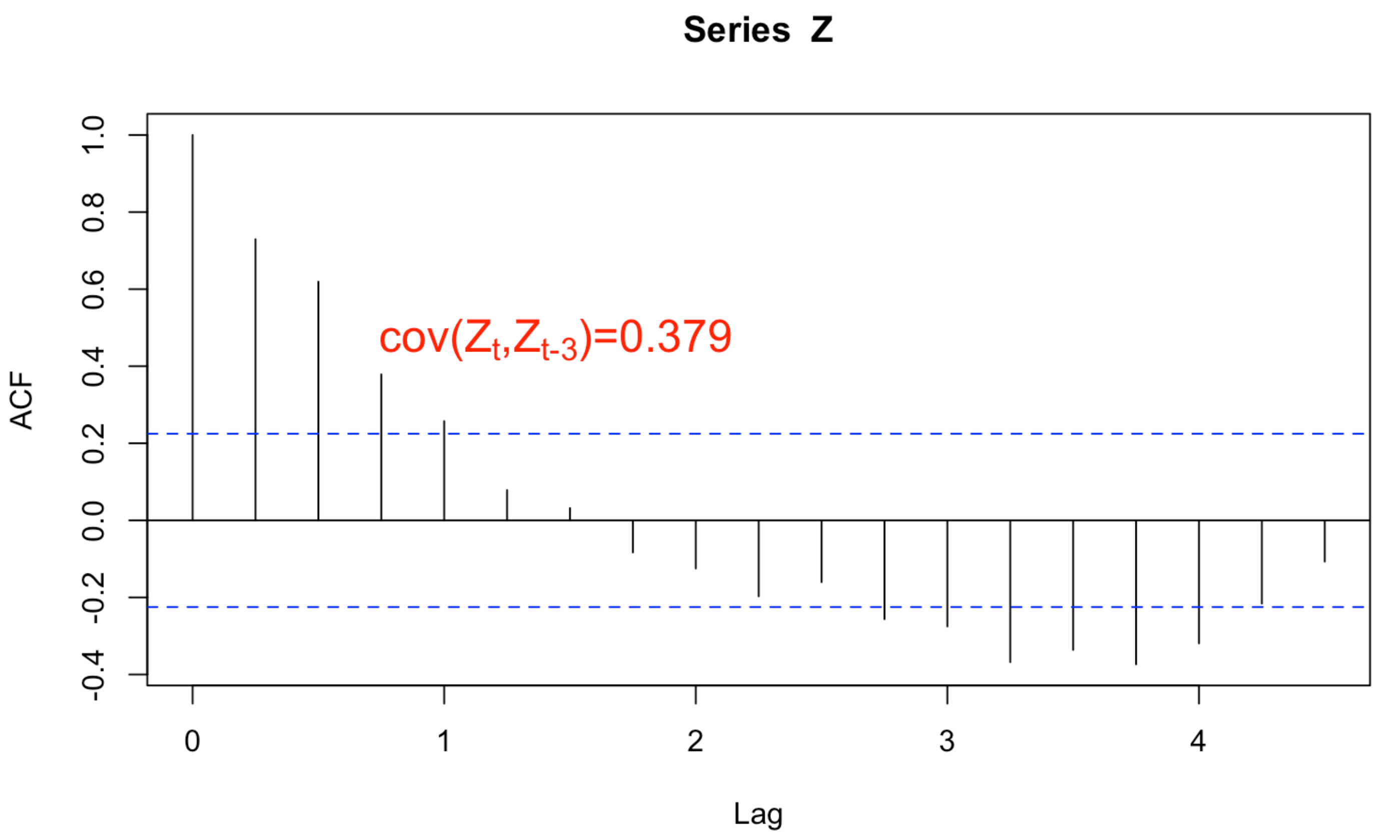
Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
```



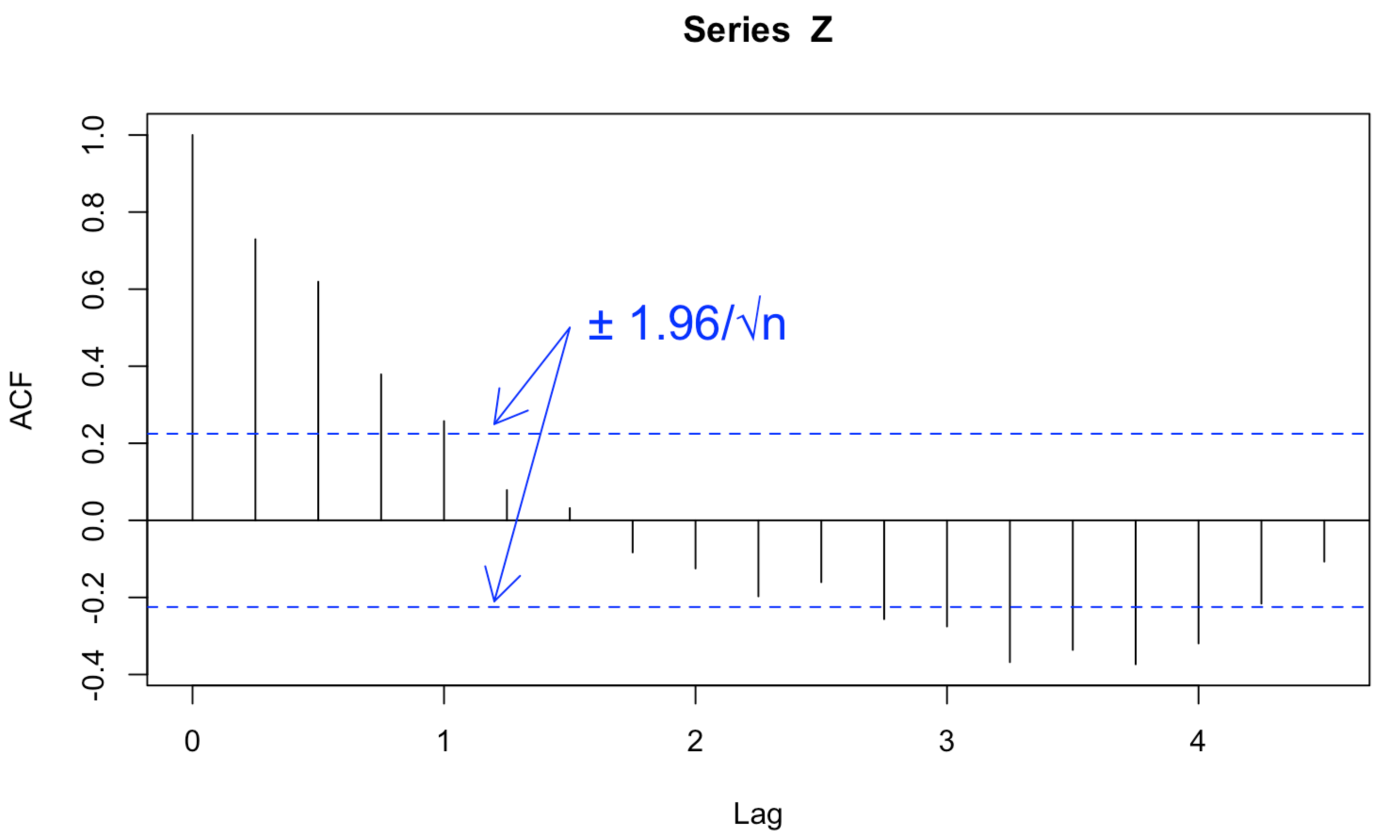
Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
```



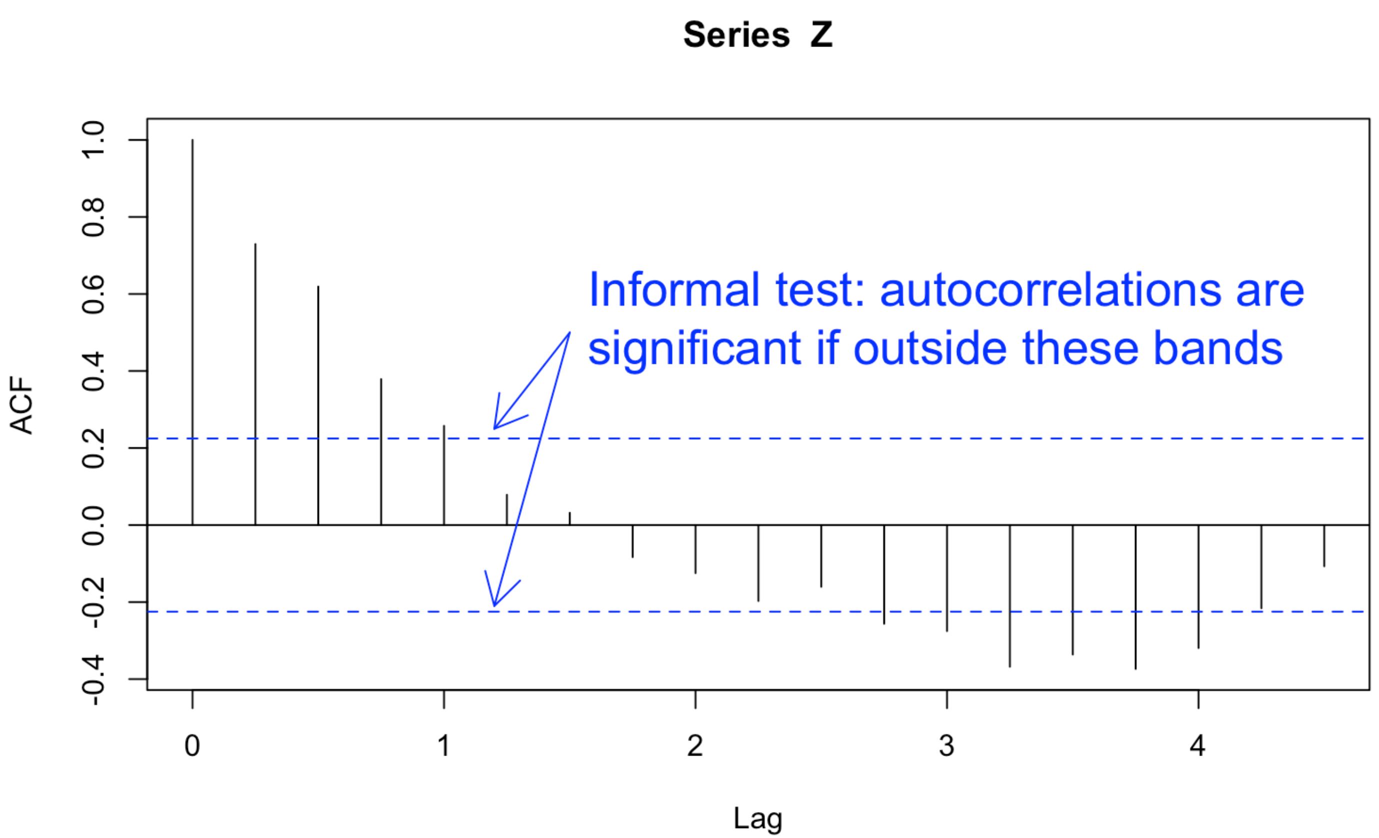
Autocorrelations in retail sales

```r
Z <- AR0$residuals
acfZ <- acf(Z)
```



Autocorrelations in retail sales

```
Z <- AR0$residuals
acfZ <- acf(Z)
```



A formal autocorrelation test

```
checkresiduals(AR0)
```

```

	Ljung-Box test

data:  Residuals from Regression with
ARIMA(0,0,0) errors
Q* = 92.465, df = 8, p-value < 2.2e-16

Model df: 0.   Total lags used: 8
```

A formal autocorrelation test

Ljung-Box test statistic:

$$Q^* = n(n+2)\sum_{k=1}^{l}\frac{{r_k}^2}{n-k}$$

$$r_k = \widehat{\mathrm{cor}}(\widehat{Z}_t, \widehat{Z}_{t-k})$$

$l = \min(8, n/5)$ for quarterly data.

A formal autocorrelation test

Ljung-Box test statistic:

$$Q^* = n(n+2)\sum_{k=1}^{l}\frac{{r_k}^2}{n-k}$$

H_0 : all autocorrelations are zero

H_1 : at least one autocorrelation not zero

Reject H_0 for $p < 0.05$, where the p-value for Q^* uses a χ^2_{l-p} distribution.

A formal autocorrelation test

```r
checkresiduals(AR0)
```

```
	Ljung-Box test

data:  Residuals from Regression with
ARIMA(0,0,0) errors
Q* = 92.465, df = 8, p-value < 2.2e-16

Model df: 0.   Total lags used: 8
```

$\Rightarrow$ reject H_0

Why does autocorrelation matter?

Recall the one-step-ahead forecast errors

$$U_t = Y_t - E(Y_t|\mathscr{F}_{t-1})$$

satisfy $E(U_t|\mathscr{F}_{t-1}) = 0$, and hence

$$\text{cor}(U_t, U_{t-k}) = 0 \text{ for every } k > 0$$

$\Rightarrow$ if residuals from a model are autocorrelated then that model is *misspecified* for the conditional expectation.

A formal autocorrelation test

```
checkresiduals(AR0)
```

```
    Ljung-Box test

data:  Residuals from Regression with
ARIMA(0,0,0) errors
Q* = 92.465, df = 8, p-value < 2.2e-16
```

AR0 is misspecified

```
Model df: 0.   Total lags used: 8
```

AR(1) model

```
AR1 <- Arima(Y, order=c(1,0,0), xreg=X)
checkresiduals(AR1)
```

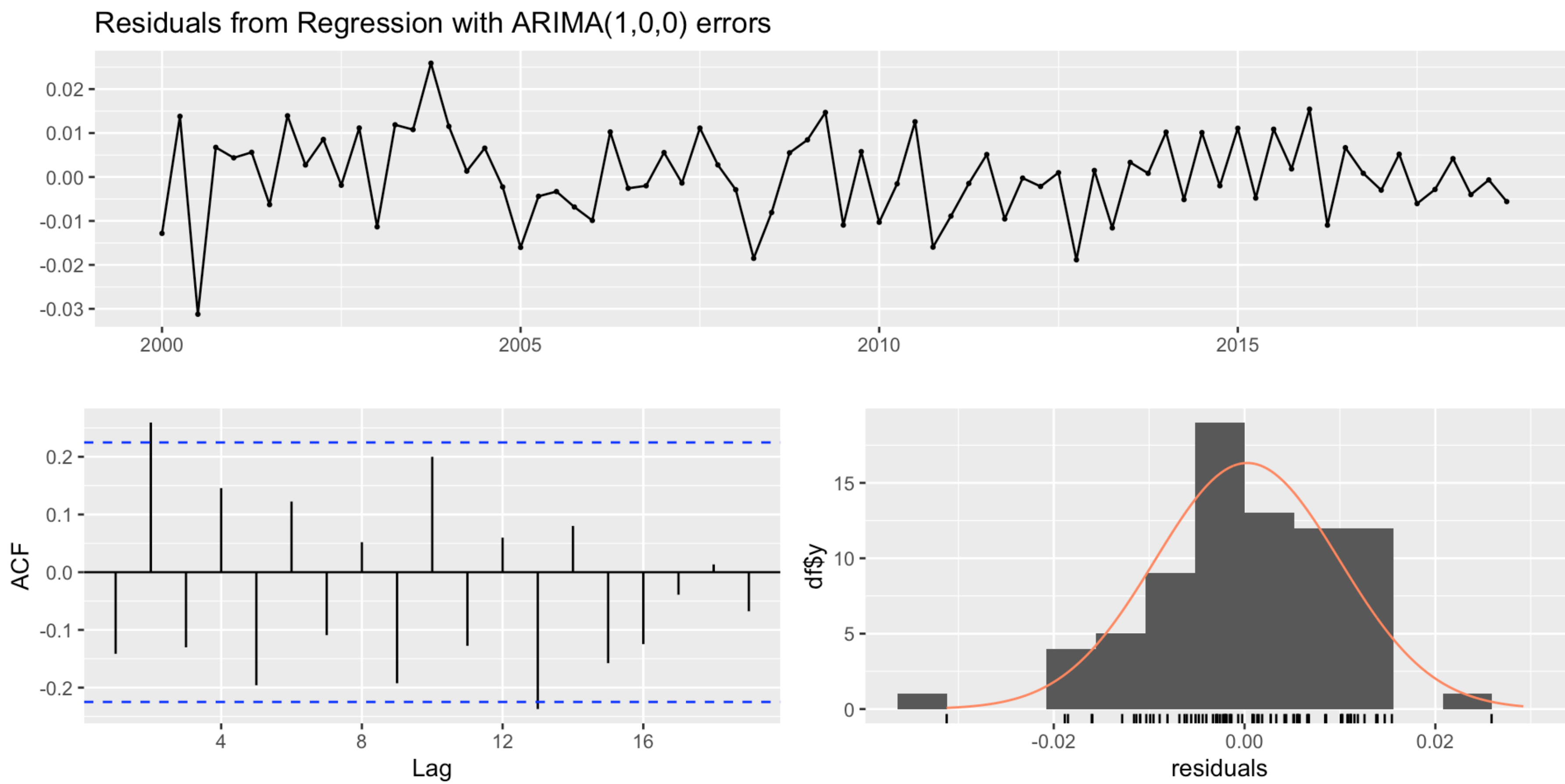
```
	Ljung-Box test

data:  Residuals from Regression with
ARIMA(1,0,0) errors
Q* = 15.822, df = 7, p-value = 0.02679
```

AR1 is misspecified

```
Model df: 1.   Total lags used: 8
```

AR(1) model

```
AR1 <- Arima(Y, order=c(1,0,0), xreg=X)
checkresiduals(AR1)
```



AR(2) model

```
AR2 <- Arima(Y, order=c(2,0,0), xreg=X)
checkresiduals(AR2)
```

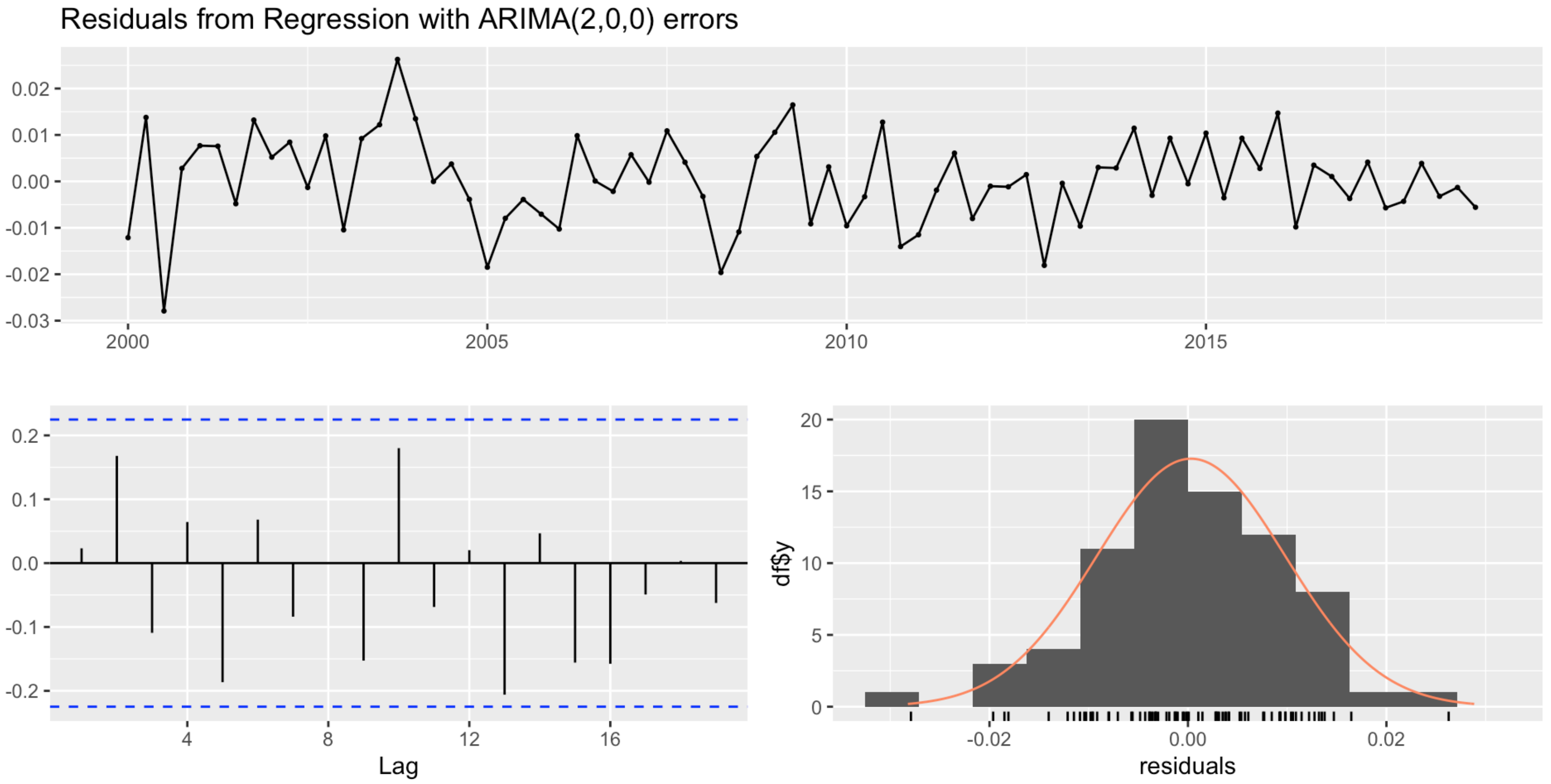
```
	Ljung-Box test

data:  Residuals from Regression with
ARIMA(2,0,0) errors
Q* = 7.5095, df = 6, p-value = 0.2763
```

AR2 passes the test!

```
Model df: 2.   Total lags used: 8
```

AR(2) model

```
AR2 <- Arima(Y, order=c(2,0,0), xreg=X)
checkresiduals(AR2)
```



Model selection

Akaike Information Criterion (AIC)

Consider a model $m(Y_{t-1}, Y_{t-2}, \ldots; \theta)$ of $E(Y_t | \mathscr{F}_{t-1})$.

Example. AR(p):

$$m(Y_{t-1}, Y_{t-2}, \ldots; \theta) \\ = \theta_1 Y_{t-1} + \ldots + \theta_p Y_{t-p}$$

Akaike Information Criterion (AIC)

Consider a model $m(Y_{t-1}, Y_{t-2}, \ldots ; \theta)$ of $E(Y_t | \mathscr{F}_{t-1})$.

Define residuals

$$\widehat{U}_t = Y_t - m(Y_{t-1}, Y_{t-2}, \ldots ; \hat{\theta})$$

and residual variance

$$\hat{\sigma}_U^2 = \frac{1}{n} \sum_{t=1}^{n} \widehat{U}_t^2.$$

Akaike Information Criterion (AIC)

The AIC is

$$\mathrm{AIC} = n \log(\hat{\sigma}_U^2) + 2(M + 1)$$

where M is the number of parameters in θ.

Eg. $M = p$ for the AR(p) model.

Akaike Information Criterion (AIC)

The AIC is

$$\mathrm{AIC} = n \log(\hat{\sigma}_U^2) + 2(M+1)$$

where M is the number of parameters in θ.

Model selection:

Choose a model to make AIC as *small* as possible.

Akaike Information Criterion (AIC)

The AIC is

$$\text{AIC} = n\log(\hat{\sigma}_U^2) + 2(M+1)$$

where M is the number of parameters in θ.

Small $\sigma_U^2 \Rightarrow$ Model fits well.

Small $M \Rightarrow$ Model is *parsimonious*.

Akaike Information Criterion (AIC)

The AIC is

$$\mathrm{AIC} = n \log(\hat{\sigma}_U^2) + 2(M + 1)$$

where M is the number of parameters in θ.

Technically: AIC is an estimate of the "Kullback-Leibler distance" of the model from the true data distribution.

Corrected Akaike Information Criterion

The AICc is

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2M^2 + 2M}{n - M - 1}$$

AICc generally has superior accuracy in smaller samples.

Illustration of model search

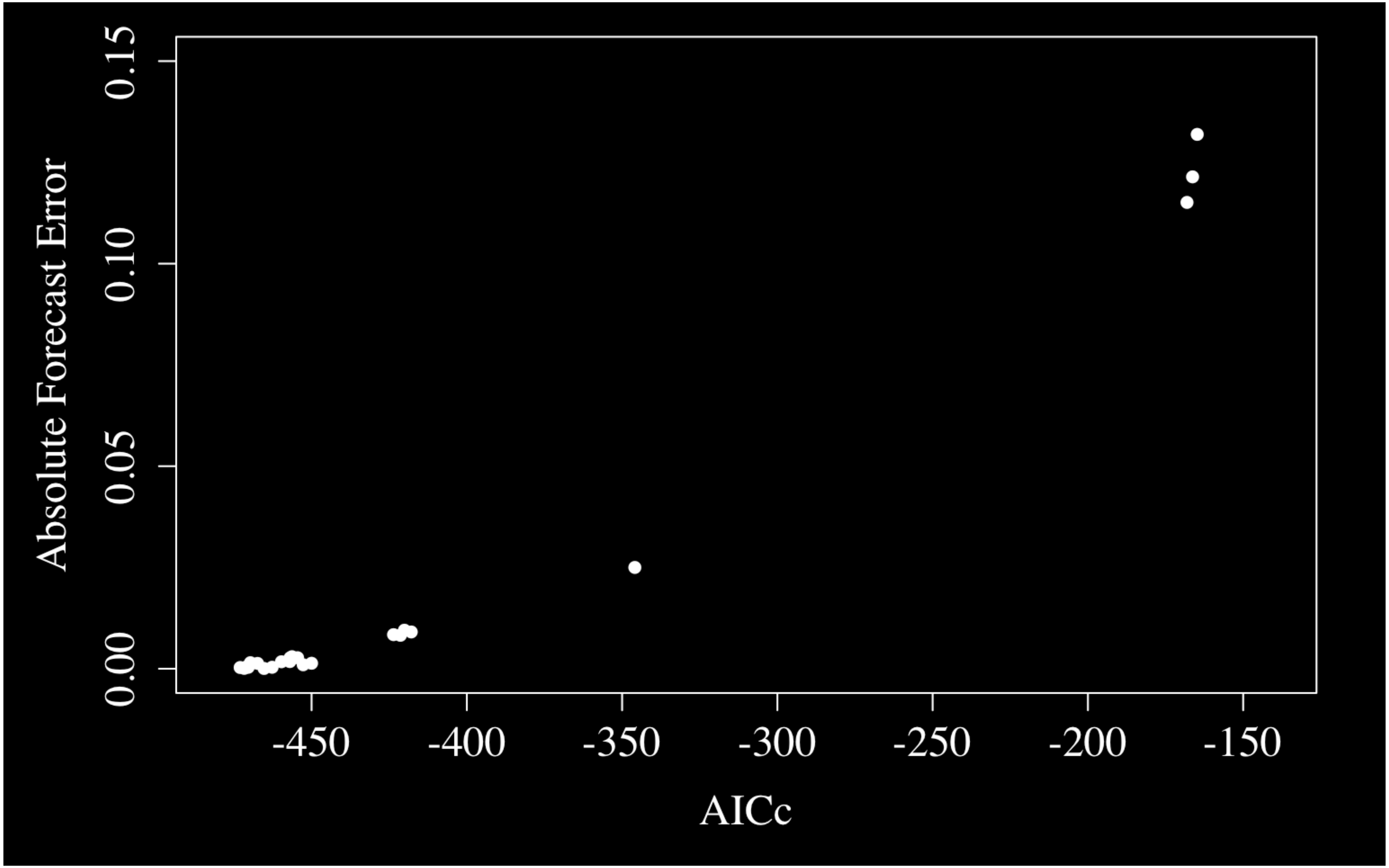
We can compute the autocorrelation test and AIC for each model specification combining:

- X1: linear trend only
- X2: linear trend and quarterly dummies
- X3: linear trend, GFC trend break, quarterly dummies
- AR(p), $p = 0, 1, 2, \ldots$.

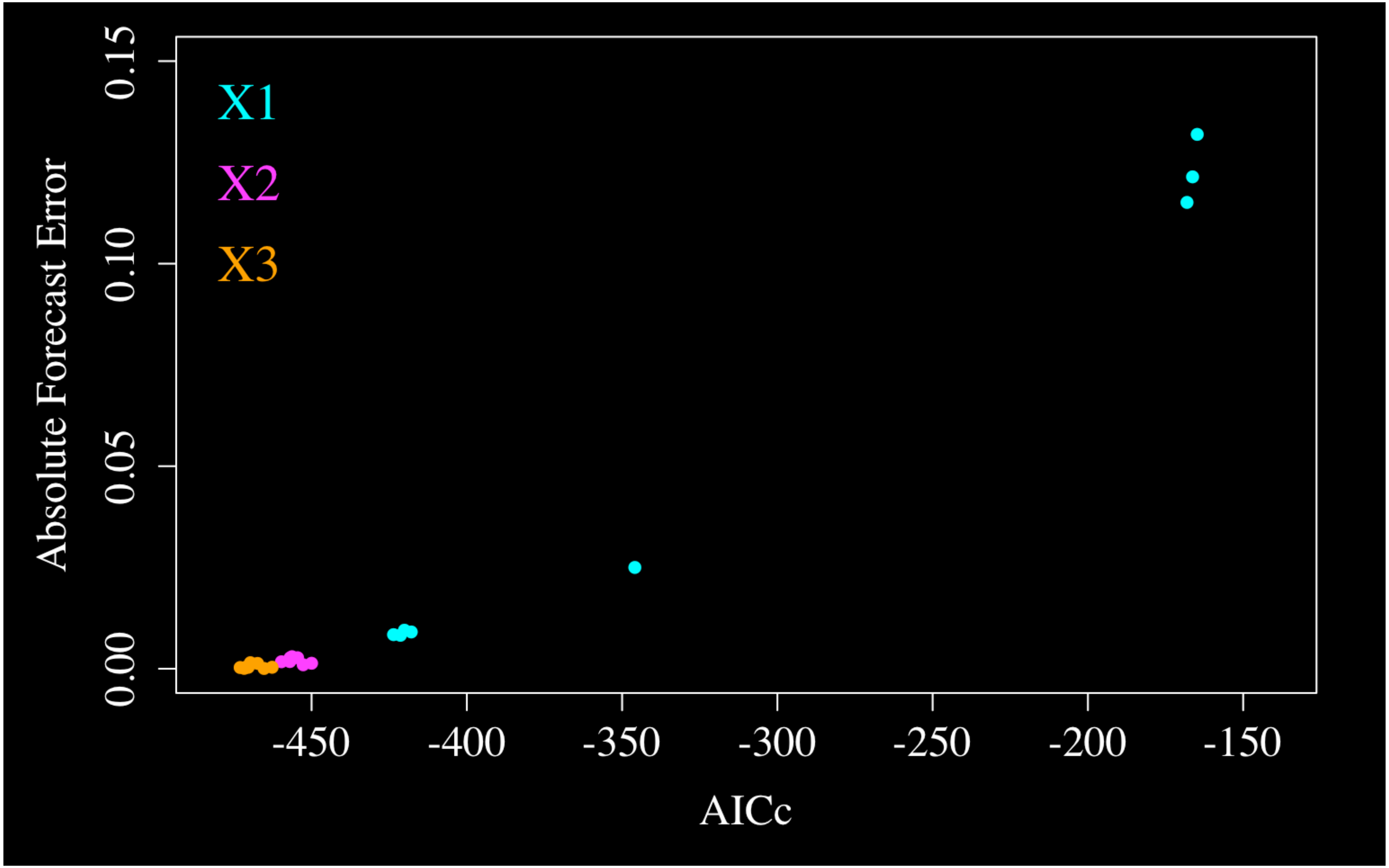
Ljung-Box p-values				AICc values		
	X1	X2	X3	X1	X2	X3
AR0	0.000	0.000	0.000	−170.2	−277.5	−412.5
AR1	0.000	0.007	0.027	−168.1	−457.0	−471.7
AR2	0.000	0.056	0.276	−166.3	−456.3	−471.2
AR3	0.000	0.661	0.749	−164.8	−459.7	−473.0
AR4	0.000	0.499	0.588	−345.9	−457.0	−470.4
AR5	0.005	0.589	0.572	−423.6	−456.9	−469.7
AR6	0.005	0.557	0.398	−421.4	−454.5	−467.4
AR7	0.019	0.100	0.110	−420.1	−452.7	−465.3

Preferred model: AR(3) with linear trend, GFC trend break, quarterly dummies (X3)

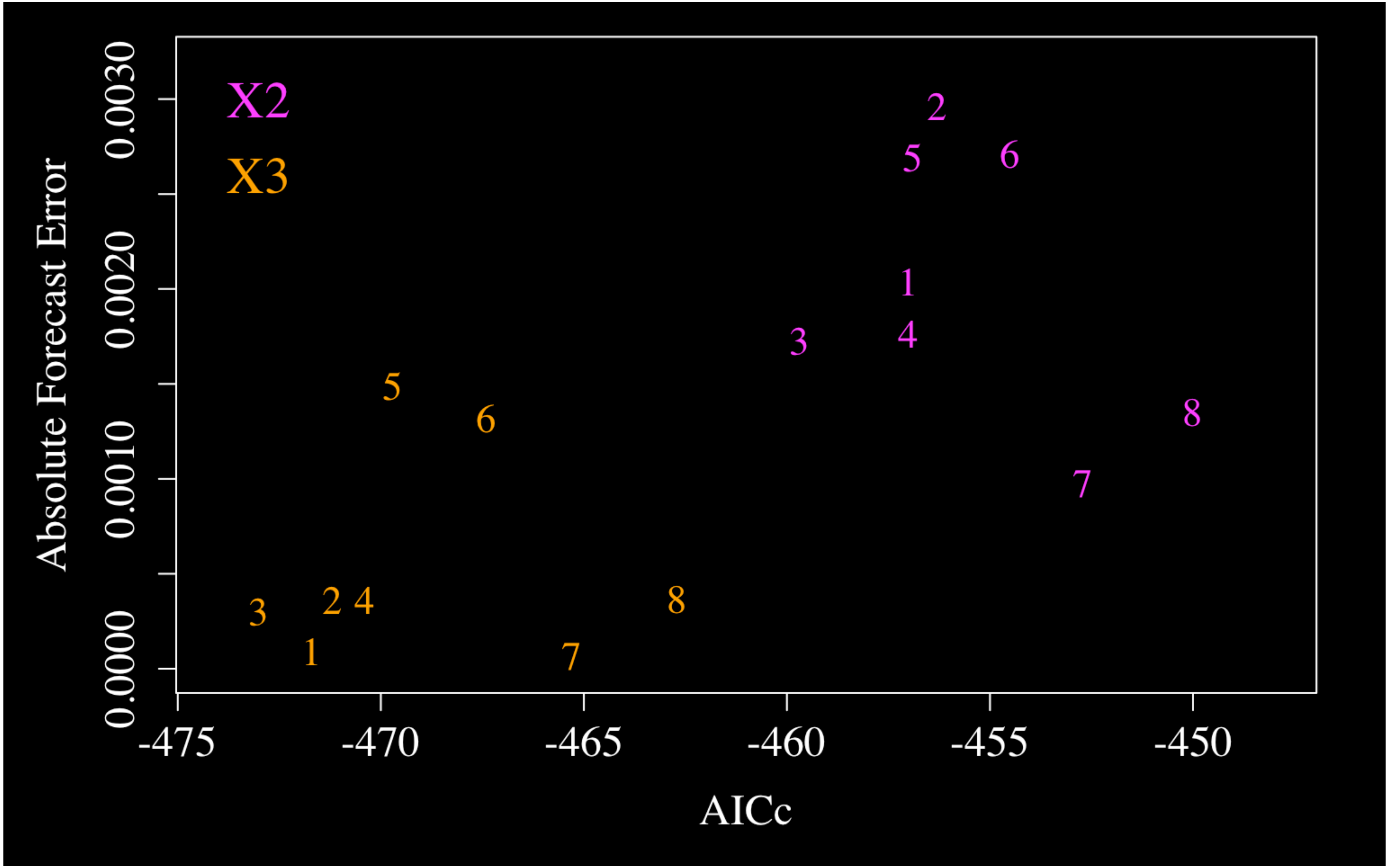
AICc vs Forecasting Accuracy (2019q1)



AICc vs Forecasting Accuracy (2019q1)



AICc vs Forecasting Accuracy (2019q1)



Lecture 3 Summary

Lecture 3 Summary

- Autocorrelation function describes autocorrelation properties of time series and residuals.
- The Ljung-Box autocorrelation test is used to test model residuals
- The AIC(c) is used to choose amongst forecasting models.