Stat 565

Examining Residuals

Jan 14 2016

Charlotte Wickham

stat565.cwick.co.nz

So far…

$$x_t = m_t + s_t + z_t$$

Variable measured at time t

Trend

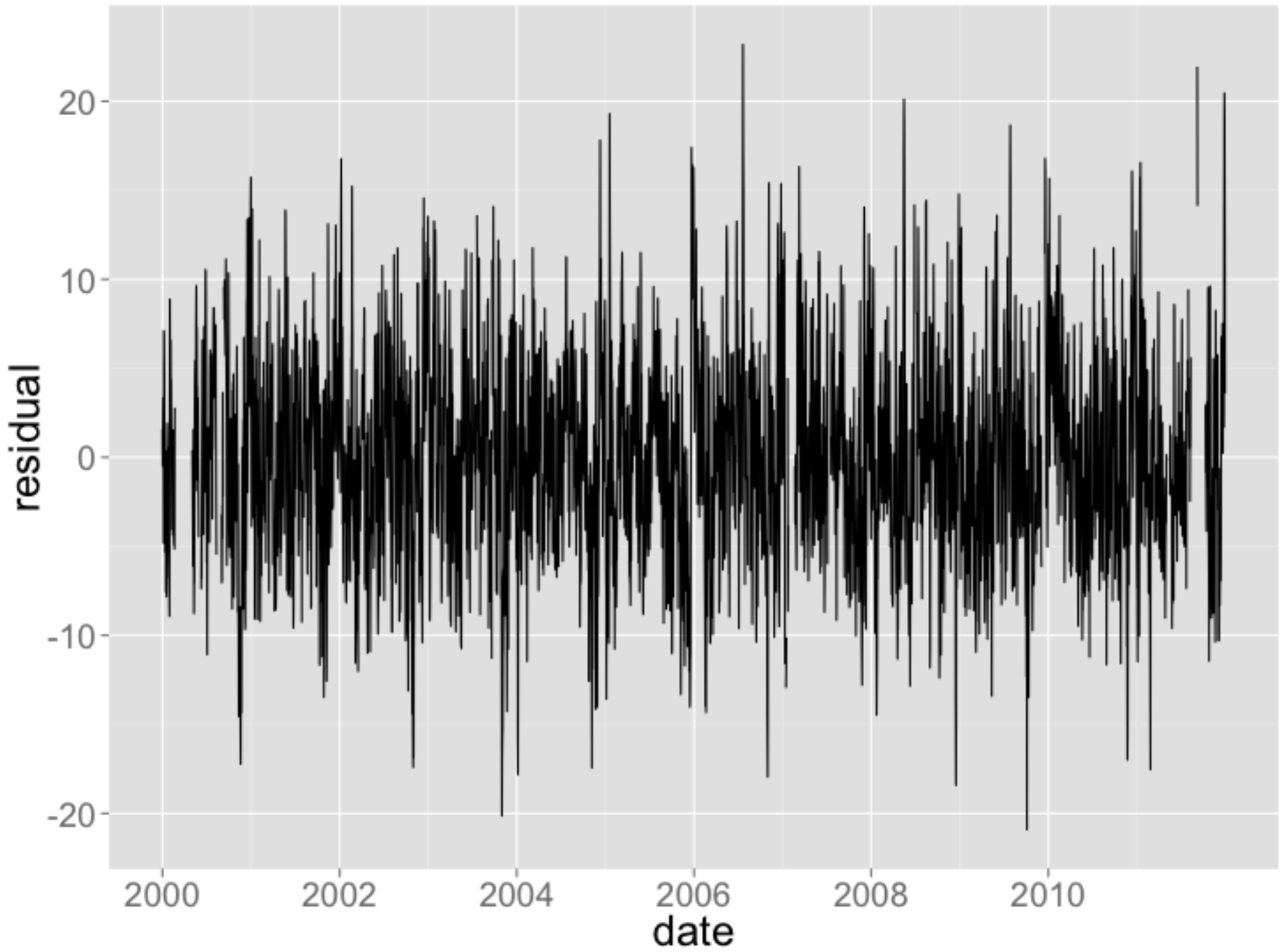
Seasonality

Noise

Estimate and subtract off

Should be left with this, stationary but probably has serial correlation

Residuals in Corvallis temperature series



Temp - loess smooth on day of year - loess smooth on date

Your turn

Now I have residuals, how could I check the variance doesn't change through time (i.e. is stationary)?

Is the variance stationary?

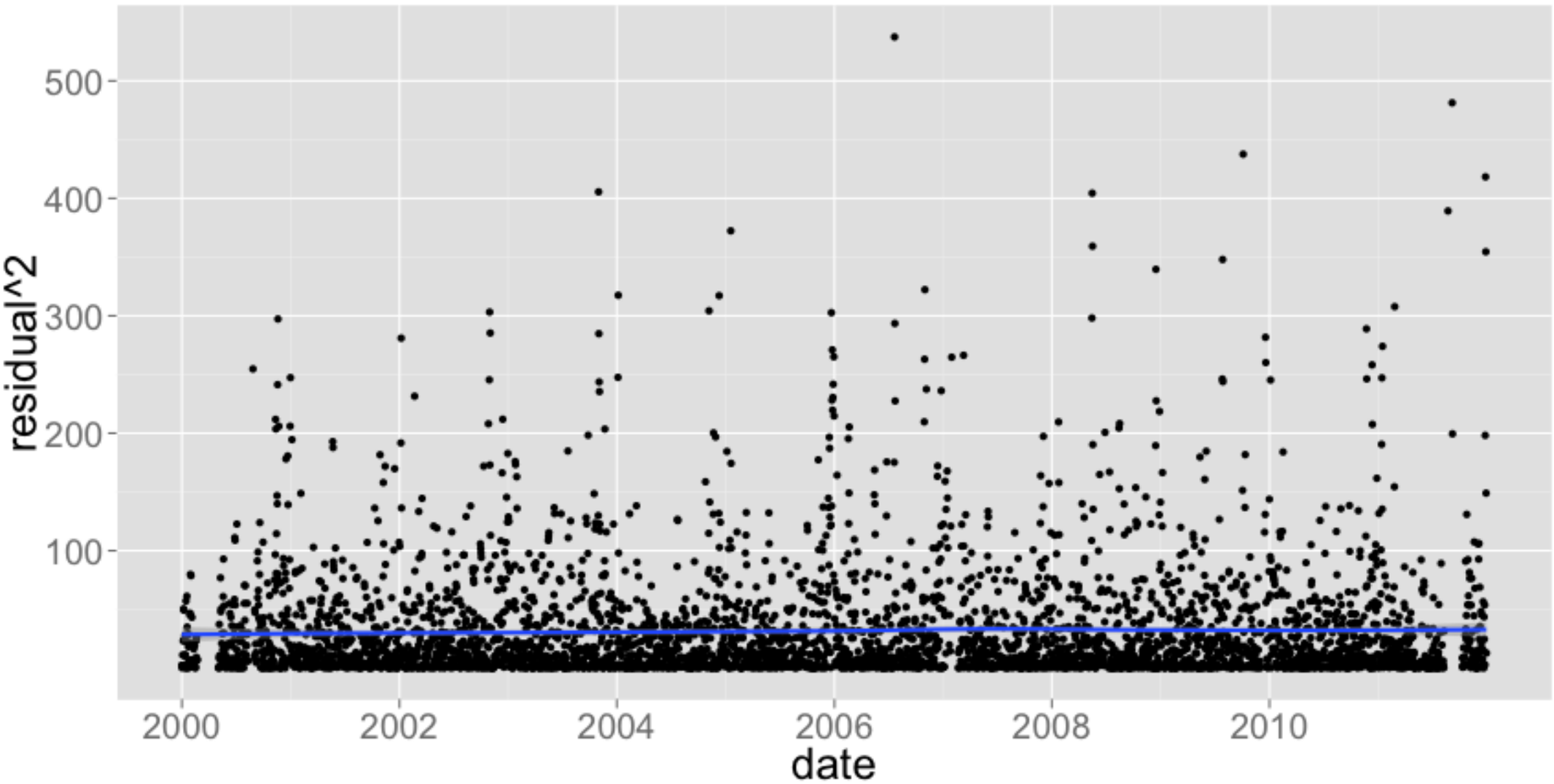
Same checks as for mean except using **squared residuals** or **absolute value of residuals**.

Why?

$var(x) = 1/n \sum (x - \mu)^2$

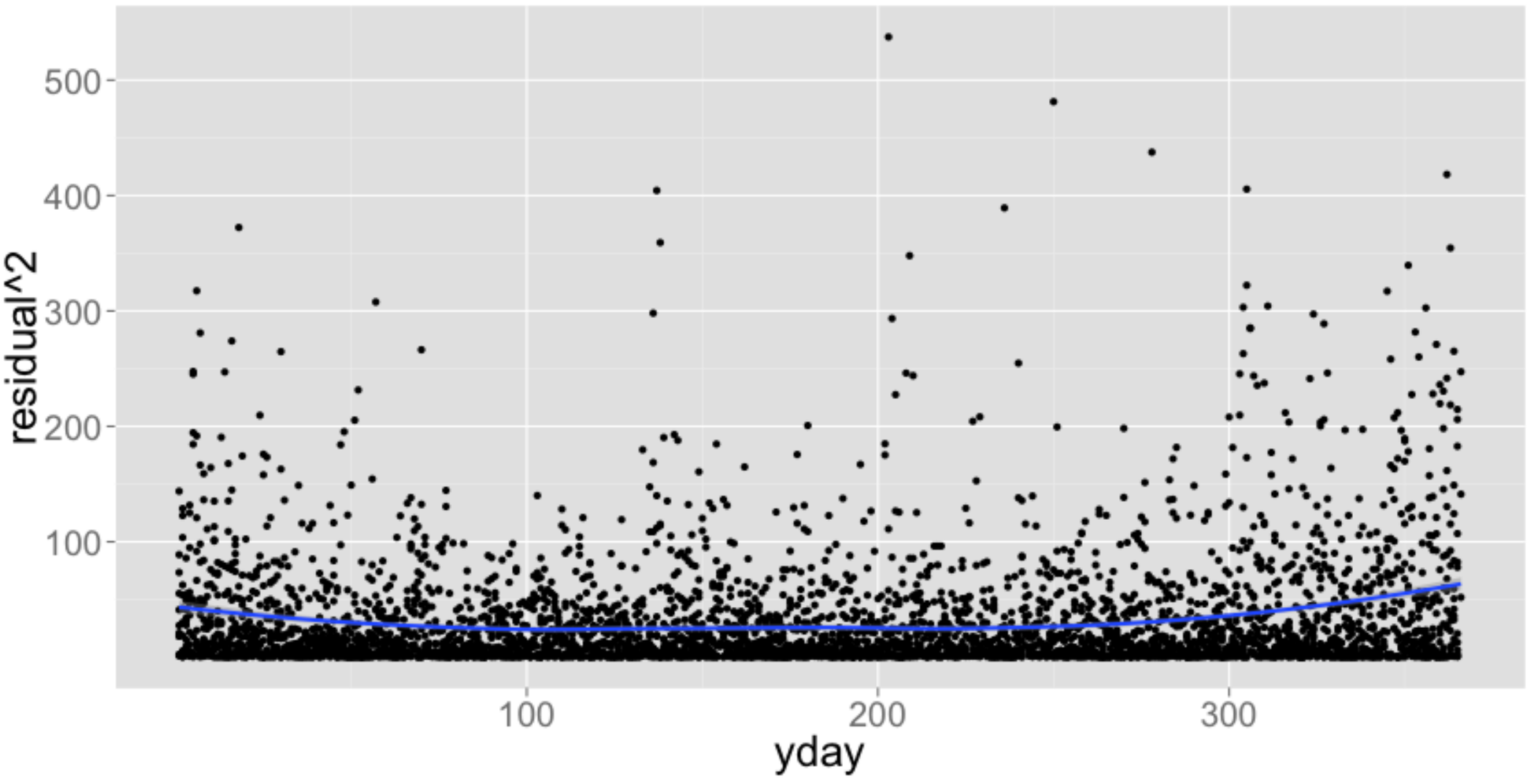
Converts a visual comparison of spread to a visual comparison of mean.

Plot squared residuals against time




```
qplot(date, residual^2, data = corv) +
  geom_smooth(method = "loess")
```

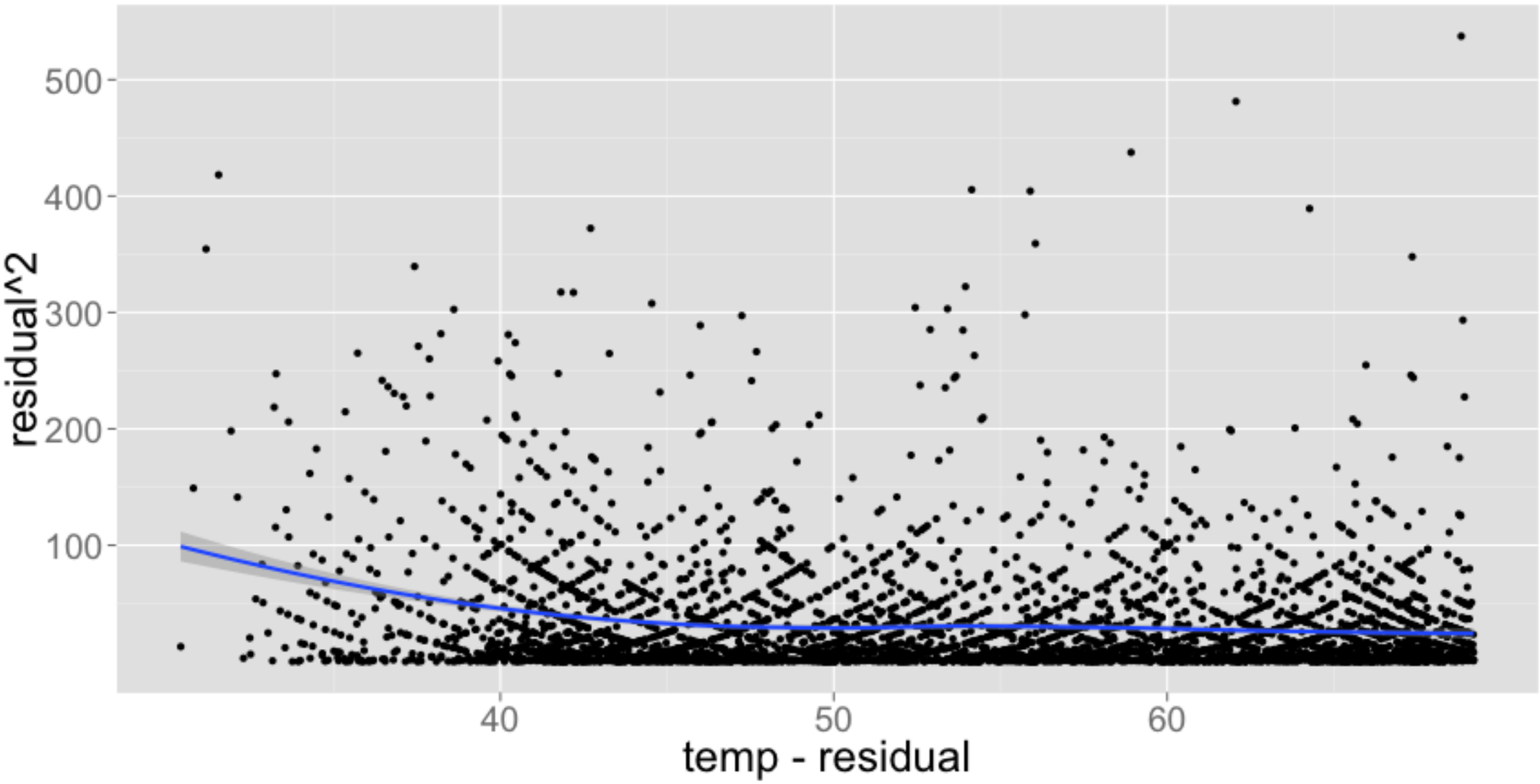
Plot squared residuals against season




```
qplot(yday, residual^2, data = corv) +
  geom_smooth(method = "loess", size = 1)
```

Fitted values against residuals

Looking for mean-variance relationship




```
qplot(temp - residual, residual^2, data = corv) +
  geom_smooth(method = "loess", size = 1)
```

Non-stationary variance

Just like the mean you can attempt to remove the non-stationarity in variance.

However, to remove non-stationary variance you **divide** by an estimate of the **standard deviation**.

Your turn

For the temperature series, serial dependence (a.k.a autocorrelation) means that today's residual is dependent on yesterday's residual.

Any ideas of how we could check that?

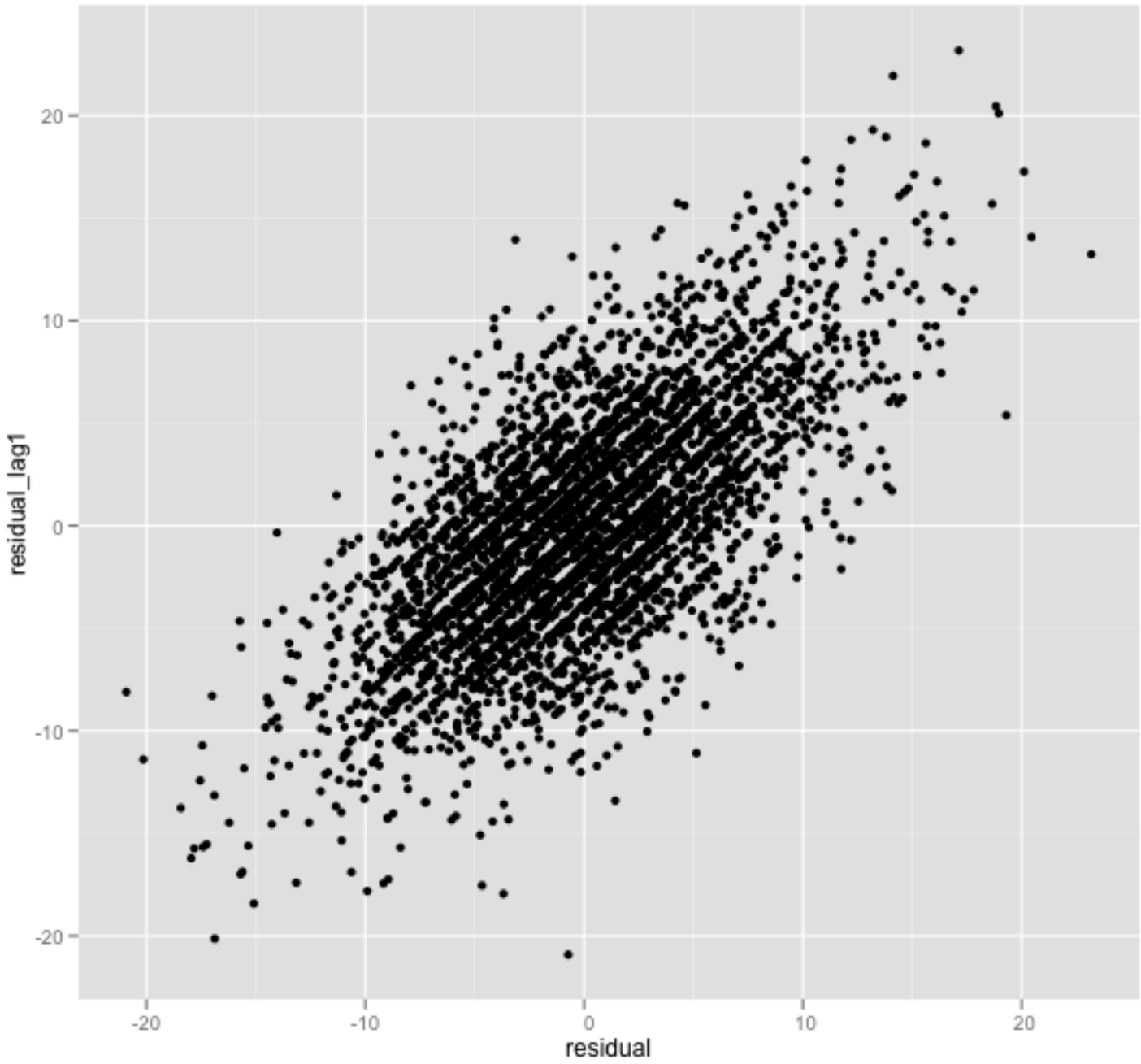
Is there autocorrelation in the residuals?

```
> corv$residual_lag1 <- c(NA, corv$residual[-nrow(corv)])

> head(corv)
.     residual   residual_lag1
.    1.5856663              NA
.   -0.4928295       1.5856663
.    1.4281641      -0.4928295
.    3.3486381       1.4281641
.    1.2685831       3.3486381
.   -4.8120101       1.2685831

> tail(corv)
.     residual  residual_lag1
.    1.705234       7.335494
.   14.077141       1.705234
.   20.451230      14.077141
.   18.827518      20.451230
.   12.206022      18.827518
.    3.586756      12.206022
```

x_{t-1} = lag 1 of x_t

Also see `?lag` for `ts` objects

```
qplot(residual, residual_lag1, data = corv)
```

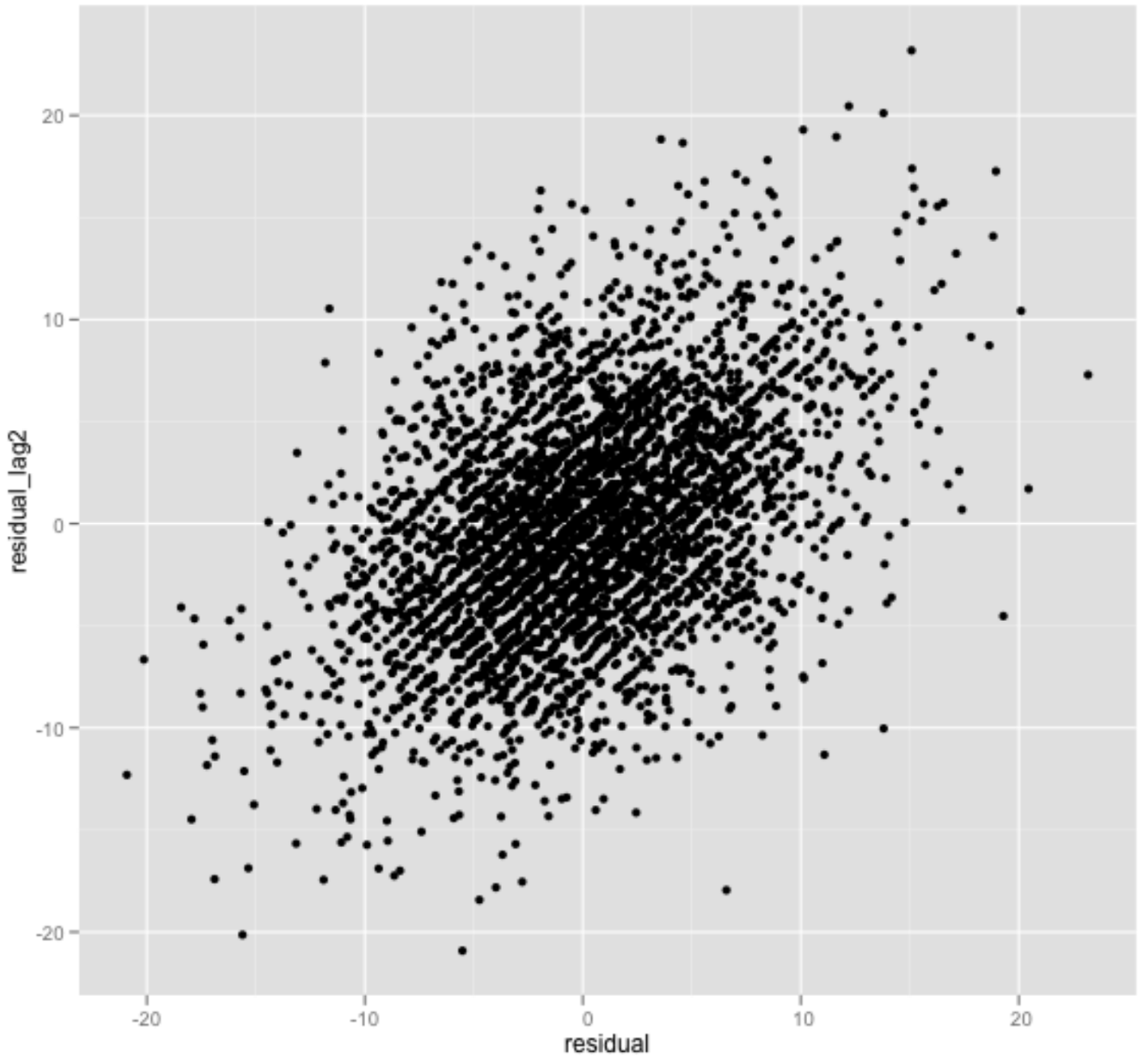



```
> with(corv, cor(residual, residual_lag1, use = "pairwise.complete.obs"))
[1] 0.6681828
```

0.67

```
corv$residual_lag2 <- c(NA, corv$residual_lag1[-nrow(corv)])
```

qplot(residual, residual_lag2, data = corv)




```
> with(corv, cor(residual, residual_lag2, use = "pairwise.complete.obs"))
[1] 0.4306014
```

0.44

Sample autocovariance

The sample autocovariance of x_t at lag h, is

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})$$

x\bar is sample mean overall sample series.

Almost the usual definition of sample covariance, apart from divisor.

Sample autocorrelation

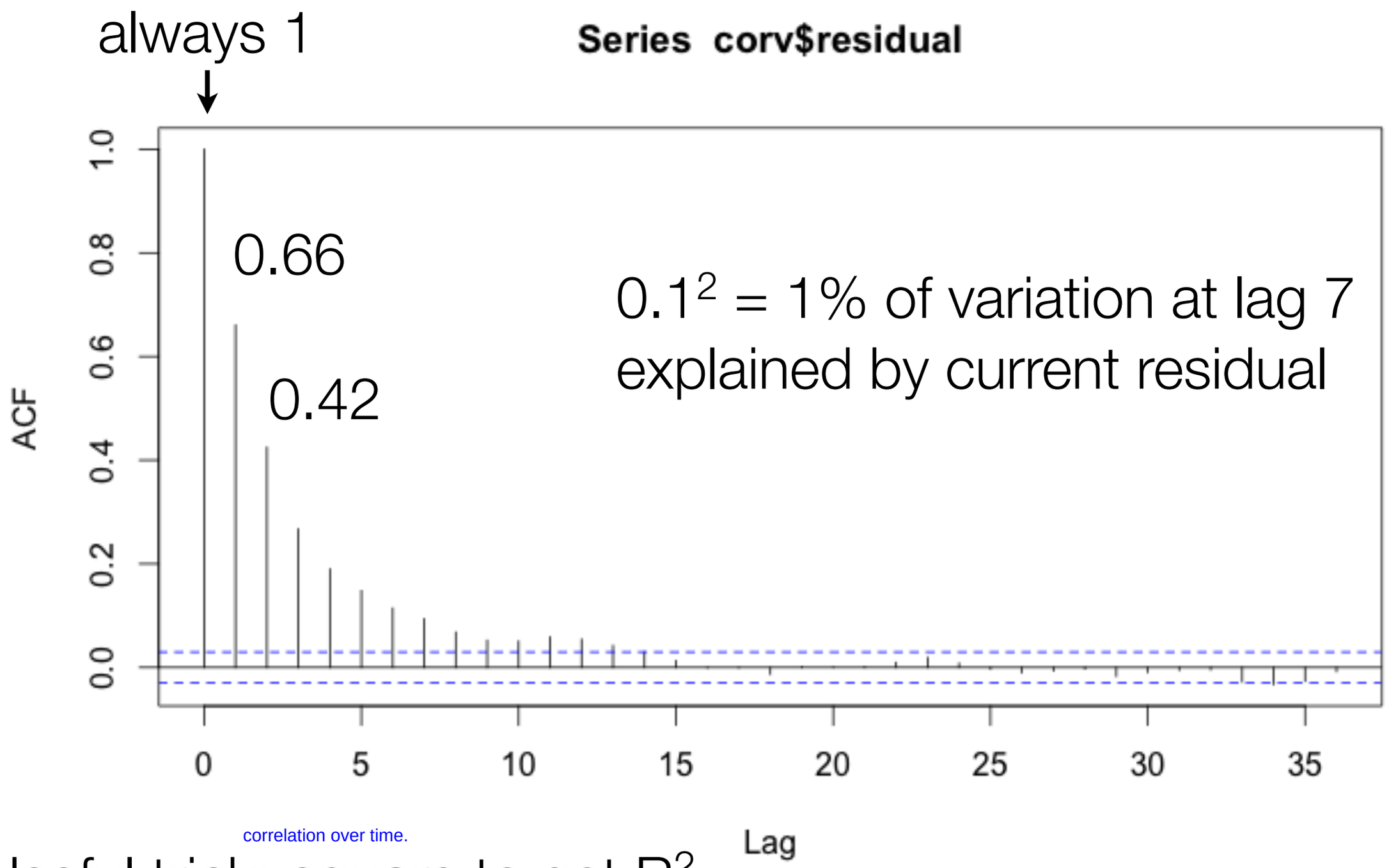
The sample autocorrelation of x_t at lag h, is the sample covariance at lag h, divided by the sample covariance at lag 0,

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

Usually displayed at a plot against h.

aka Correlogram

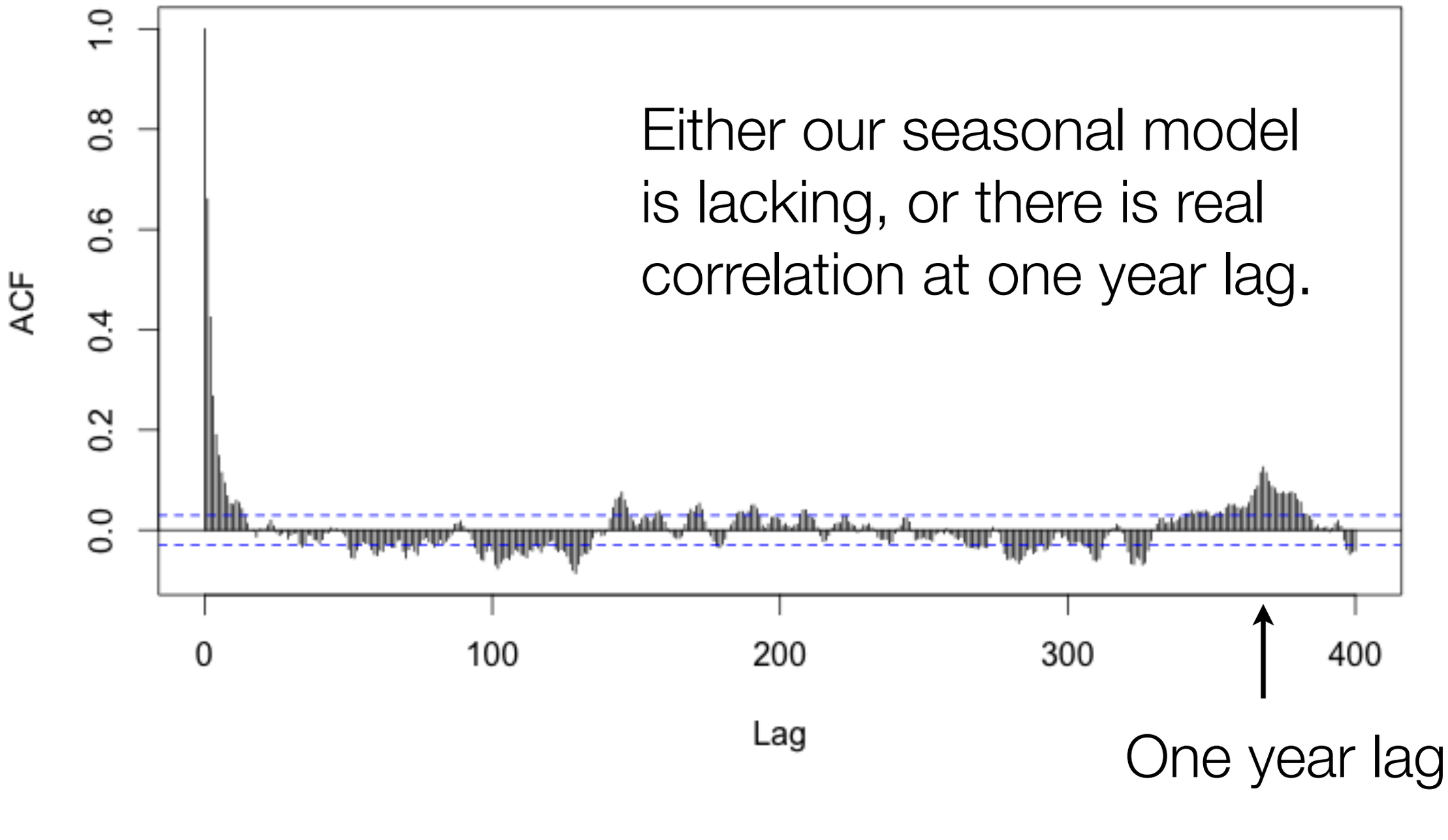
acf(corv$residual, na.action = na.pass)$acf



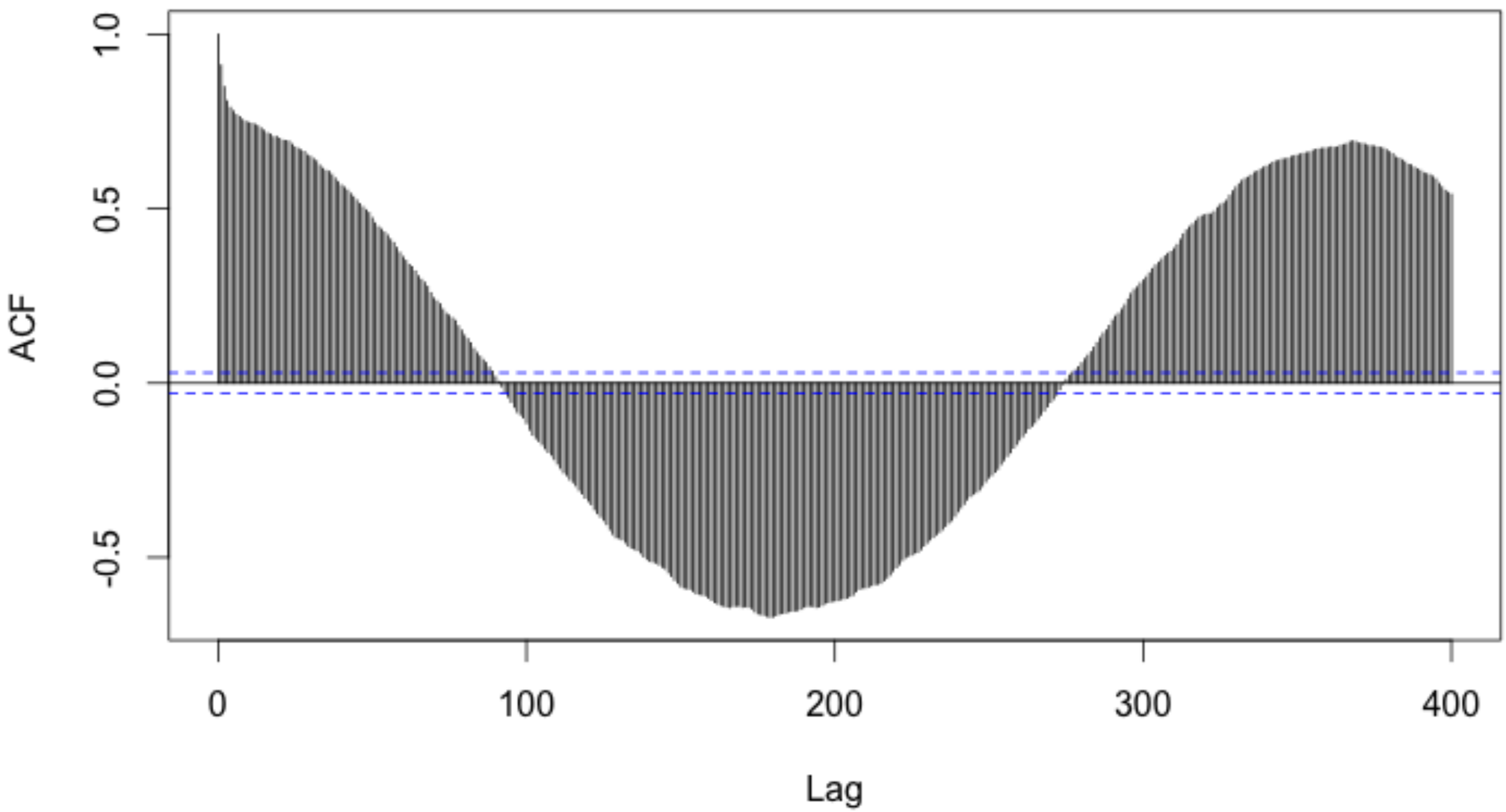
correlation over time.

Useful trick: square to get R^2

acf(corv$residual, **lag.max = 400**,
na.action = na.pass)



```
acf(corv$temp, lag.max = 400, na.action = na.pass)
```



You can find the sample acf on the raw series too,
but most of this pattern is explained by the seasonality

Autocorrelation function

The autocorrelation function and it's partner the partial autocorrelation function (coming soon) are central to time domain time series analysis.

We will learn the expected shape for some standard models of correlated time series, then use that knowledge to choose models for our data.

That ends the exploratory data analysis section....

We'll come back to these ideas when we get to regression, but for the next few weeks we will talk only about stationary series.

And now some math...

Notation

A time series $\{ x_t \}$ is a sequence of random variables indexed by t (time), t = 1,...,n.

Most of the time (at least in this class) we only have a single time series.

We assume that there are a population of possible time series generated by some time series model but we only observe one.

This is why stationarity is so central. A longer time series wouldn't give us any more information about the properties the series, if those properties were changing.

Basic moments

The mean function is,

$\mu_t = E[x_t]$

The autocovariance function is,

$\Upsilon(s, t) = E[(x_s - \mu_s) (x_t - \mu_t)]$

The variance function is,

$\sigma^2_t = \Upsilon(t, t) = E[(x_t - \mu_t)^2]$

Stationarity

A time series $\{ x_t \}$ is **strictly stationary** if the joint distribution function of

$\{ x_{t_1}, x_{t_2}, \ldots, x_{t_k}\}$ is identical to the joint distribution function of

$\{ x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}$

for all $k = 1, 2, \ldots$

all time points $\{t_1, \ldots t_n\}$ and

all $h = \pm 1, \pm 2, \ldots$

Weak Stationarity

A time series $\{x_t\}$ is **weakly stationary** if it's mean function doesn't depend on time, and it's autocovariance function only depends on the distance between the two time points,

$$\mu_t = E[\, x_t \,] = \mu$$

$$\Upsilon(s, t) = Cov(x_s, x_t) = \Upsilon(t - s)$$

Often rewrite as
$\Upsilon(h) = Cov(x_t, x_{t+h})$

x_t assumed to have finite variance