

# Stat 565

## Spurious Regression

Feb 16 2016

Charlotte Wickham

[stat565.cwick.co.nz](http://stat565.cwick.co.nz)

## **Last time:**

We can extend regression to deal with correlated errors.

## **Today:**

Explore the concept of spurious correlation/regression

What is it?

How can we deal with it?

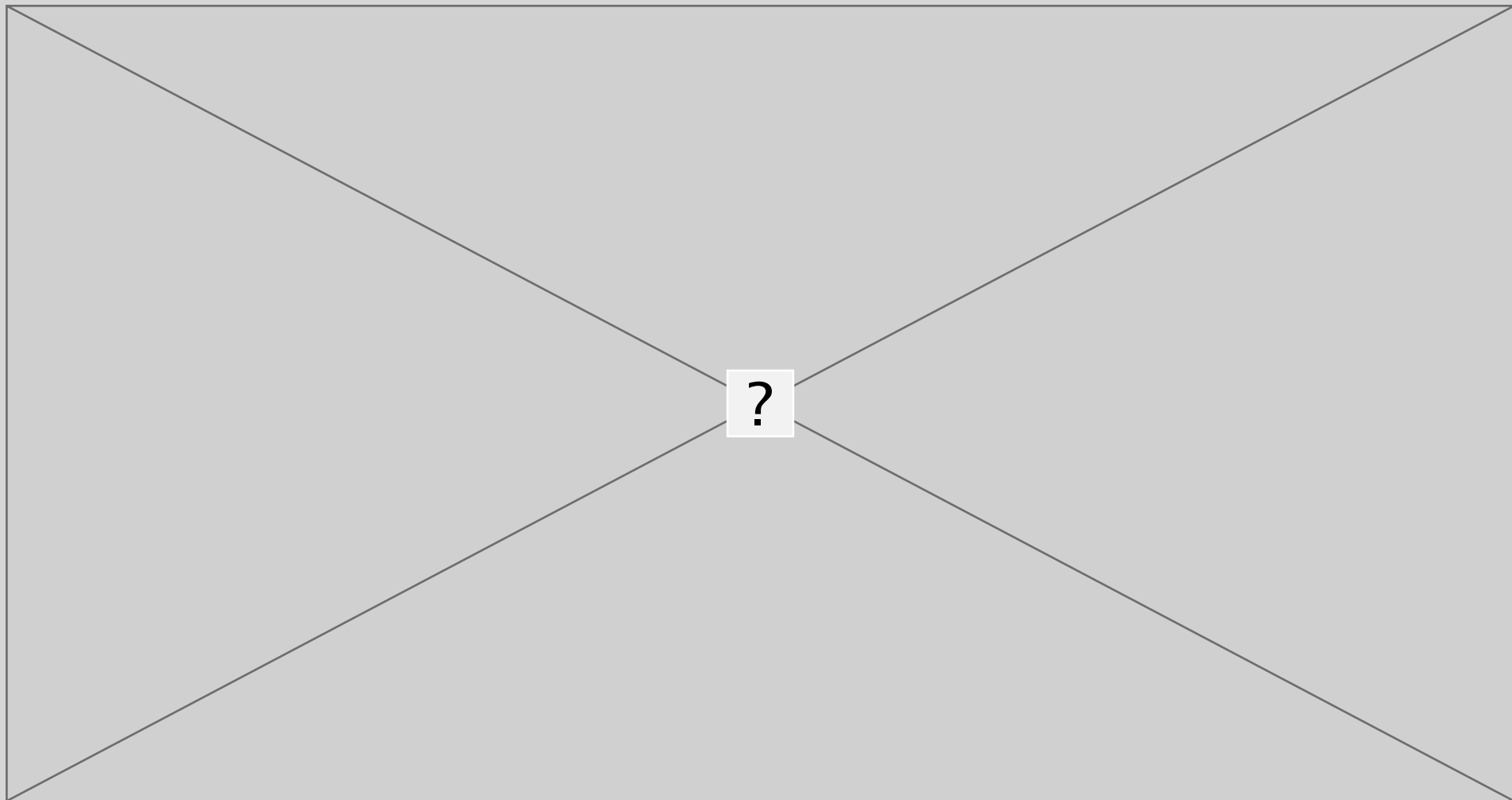
Practice with some time series regression analyses.

# Your turn

Where does most of the variation in temperature come from?

What about particulates?

What about mortality?



# Spurious correlation

With **completely independent** series:

Positive correlation can arise:

If trend dominates the series, and they are both trending in the same direction.

If seasonality dominates the series, the cycle length is roughly the same, and the cycles are in phase

If the cycles are out of phase the relationship will appear non-linear.

# Simulated examples

$$x_t = 5 \cos(2\pi t / 10) + z_t$$

$$y_t = 10 \cos(2\pi t / 10) + w_t$$

?

0.96

?

$$x_t = 10 \cos(2\pi t / 10) + z_t$$

$$y_t = 10 \cos(2\pi (t+2) / 10) + w_t$$

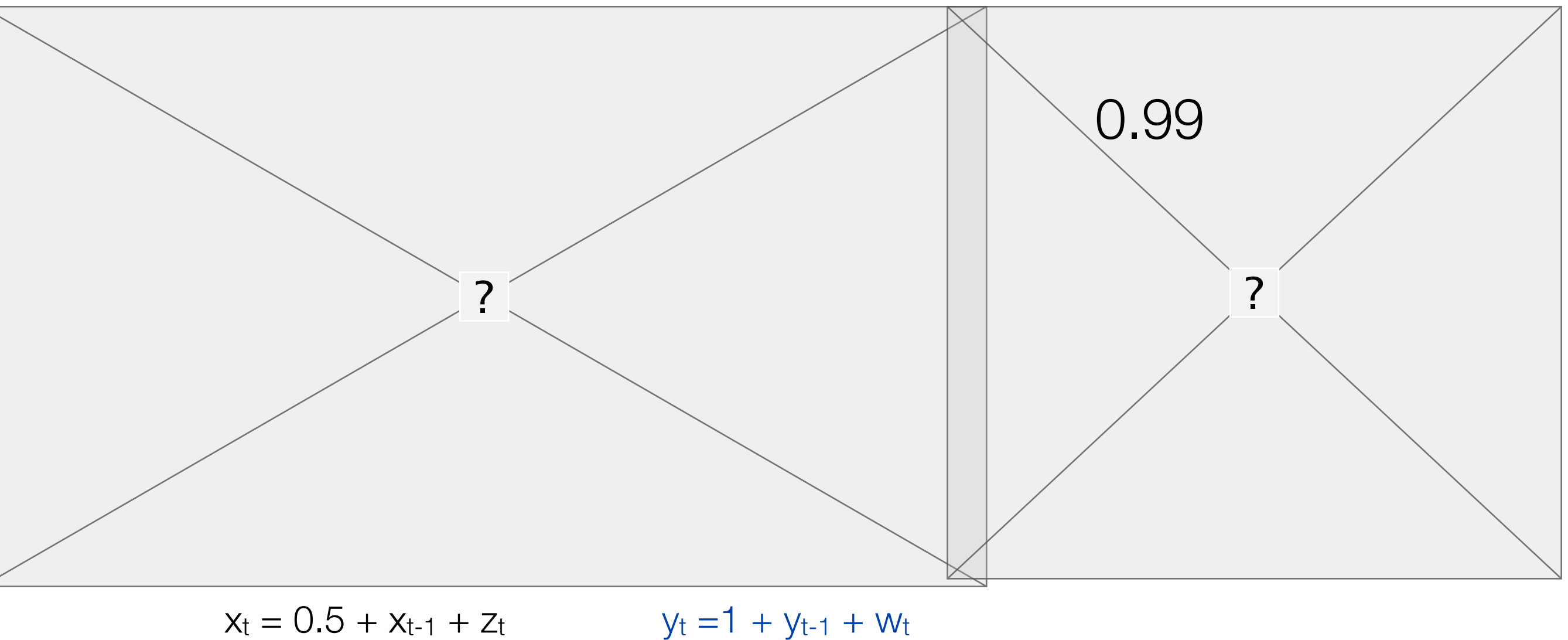
0.80

?

?

$w_t, z_t$  independent white noise

# Simulated examples



$w_t, z_t$  independent white noise

Real world examples

# Your turn

What do you think is driving the correlation?

Snake bites are positively associated with ice cream sales.

Electricity usage is positively associated with chocolate consumption in Australia over the last 50 years.

The stock market is positively correlated with sunspot activity.

# What might be surprising...

Is that this problem also arises with stationary series...

Consider the regression model:

$$y_t = 1 + \beta x_t + z_t$$

where  $x_t$  is white noise and  $z_t$  is white noise.

Imagine, I observe 100 observations each of  $y_t$  and  $x_t$ ,

and estimate the above model,

then test the null hypothesis  $\beta = 0$ .

**If in reality  $\beta = 0$ , how often will I reject the null hypothesis at the 5% level?**



```
one_sim <- function(){
```

```
  n <- 100
```

```
  x <- arima.sim(list(), n = n)
```

```
  z <- arima.sim(list(), n = n)
```

```
  y <- 1 + 0*x + z
```

```
  fit <- arima(y, order = c(0, 0, 0), xreg = x)
```

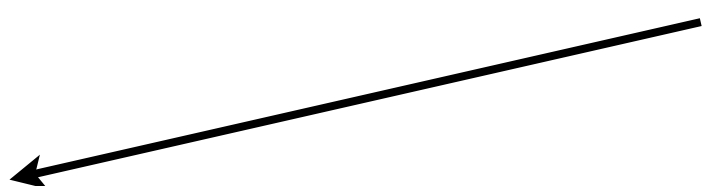
```
  abs( fit$coef["x"] /  
        sqrt(diag(fit$var.coef)["x"]) ) > 1.96
```

```
}
```

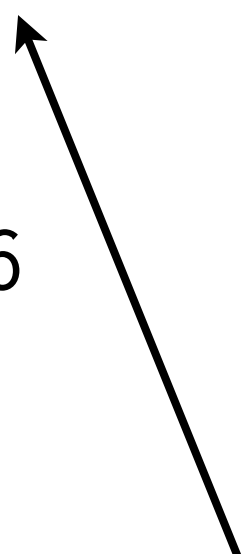
```
reject <- replicate(1000, one_sim())
```

```
table(reject)
```

an empty list  
gives white  
noise



from Tues, fits  
a regression  
on x with  
ARMA errors



In pairs

<http://stat565.cwick.co.nz/code/12-sim-code.r>

**But what if  $x_t$  or  $z_t$  are stationary time series?**

Confirm:

- if **only one** of  $x$  or  $z$  is AR(1) the error rate is  $\sim 5\%$
- if **both**  $x$  and  $z$  are AR(1) the error rate is  $> 5\%$

How does the error rate depend on  $\alpha$ ? Fix  $\alpha$  for  $x$  and vary  $\alpha$  for  $z$ .

Examine the residuals for one fit when both  $x$  and  $z$  are  $AR(1)$  with parameter  $\alpha = 0.9$ .

Make a suggestion for fixing our regression.

Try it and see if the error rate is  $\sim 5\%$ .

# How about stochastic trends?

We saw that as  $\alpha$  increased, the error rate increased. If  $\alpha=1$ , we have a random walk.

Charlotte demonstrates that the error rate gets even worse.

But... differencing both series before regressing them solves the problem. This is equivalent to fitting an ARIMA model to the errors.

Things get really bad if you have an integrated random walk, error  $\sim 95\%$ .

# Strategy for regression

Difference to achieve stationarity first. Difference both series by the same amount, convention is use  $x$  to decide the differencing.

Fit regression model, use differenced series, or specify  $ARIMA(0, d, 0)$  errors.

Examine residuals to pick an ARMA model.

Fit regression model with ARIMA errors.

Check diagnostics.

Interpret.

# Paired data analysis

Choose one from:

<http://stat565.cwick.co.nz/code/12-pair-da.R>

Bluebird chip prices and sales

Public transit boardings and gas prices

US personal income and consumption