Stat 412/512

Transforming to remove serial correlation

Feb 25 2015

Charlotte Wickham

stat512.cwick.co.nz

Reminders

DA #2 due today

Regression in your field due Friday

Quiz #3 this weekend

Last homework released today due, next Wednesday.

Study, study, study...

Final exam Thursday March 19th noon-2pm

Lastnames A-L WGR 153 (here)

Lastnames M-Z Kidder 350

Serial correlation

a.k.a autocorrelation

Positive serial correlation: an observation on one side of the mean tends to be followed by another observation on the same side of the mean.

Positive serial correlation tends to lead to standard errors that are too small.

Two solutions

1. Adjust standard errors to be more appropriate.

2. Filter variables to remove correlation.

Either way you need to estimate the extent of the correlation (and make an assumption about it’s structure).

More advanced methods explicitly model the correlation.

2. **Filter variables** to remove correlation.

If the AR(1) model is adequate and

$\mu\{ Y_t \mid X_t \} = \beta_0 + \boxed{\beta_1} X_t$

Then the **filtered** variables:

$V_t = Y_t - \boxed{\alpha} Y_{t-1}$

$U_t = X_t - \alpha X_{t-1}$

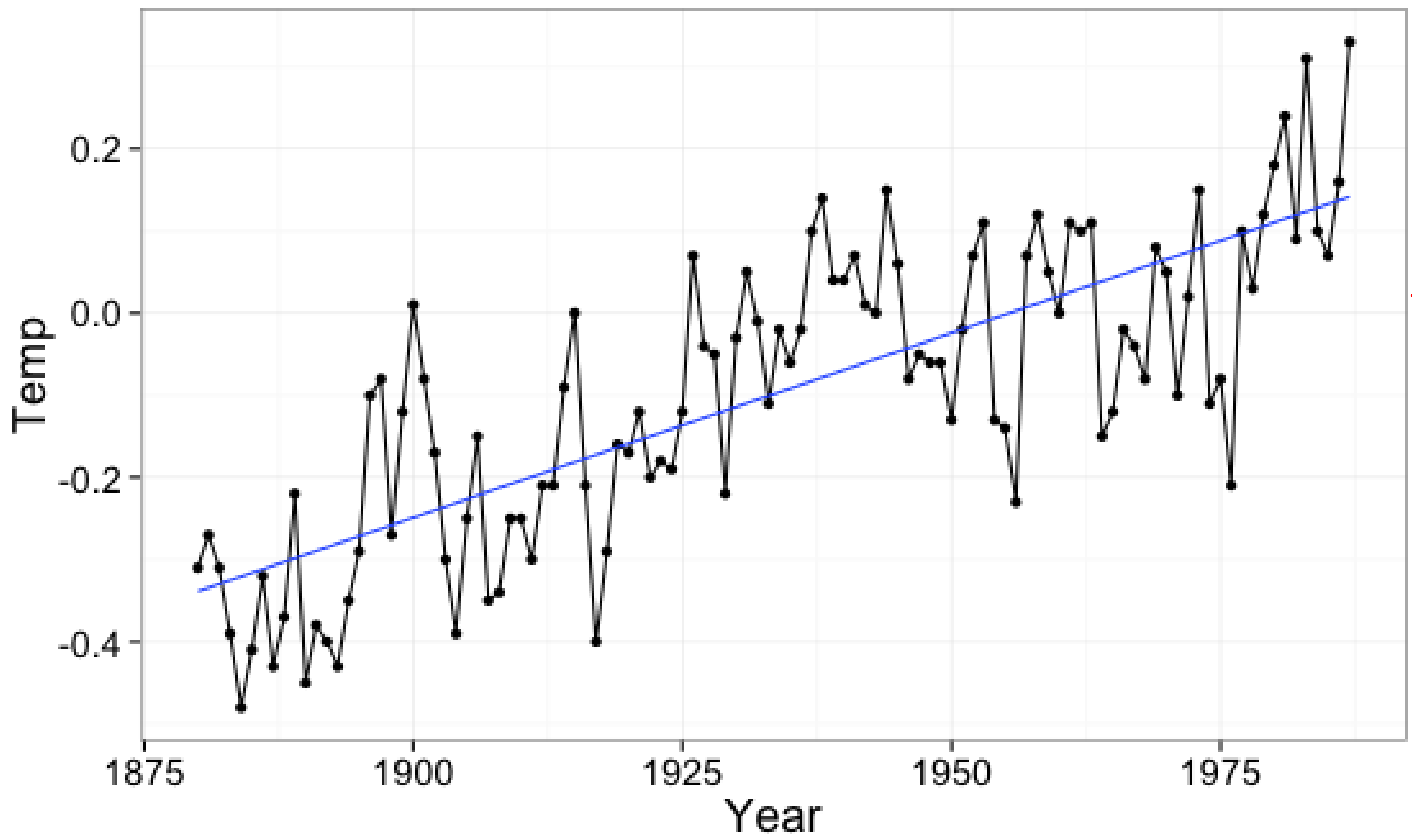
are related by the same slope:

with no serial correlation

$\mu\{ V_t \mid U_t \} = \beta_0(1 - \alpha) + \boxed{\beta_1} U_t$

Use $\boxed{r_1}$ as an estimate for α.
Filter response and explanatory.
Then regress filtered variables.

case1502: Global Temperature



The data are the temperatures (in degrees Celsius) averaged for the northern hemisphere over a full year, for years 1880 to 1987. The 108-year average temperature has been subtracted, so each observation is the temperature difference from the series average.

Your turn

Ignoring of the appropriateness of regression, how would you answer the question of interest?

Is the mean temperature increasing?

$$\mu\{Temp \mid Year\} = \beta_0 + \beta_1 Year$$

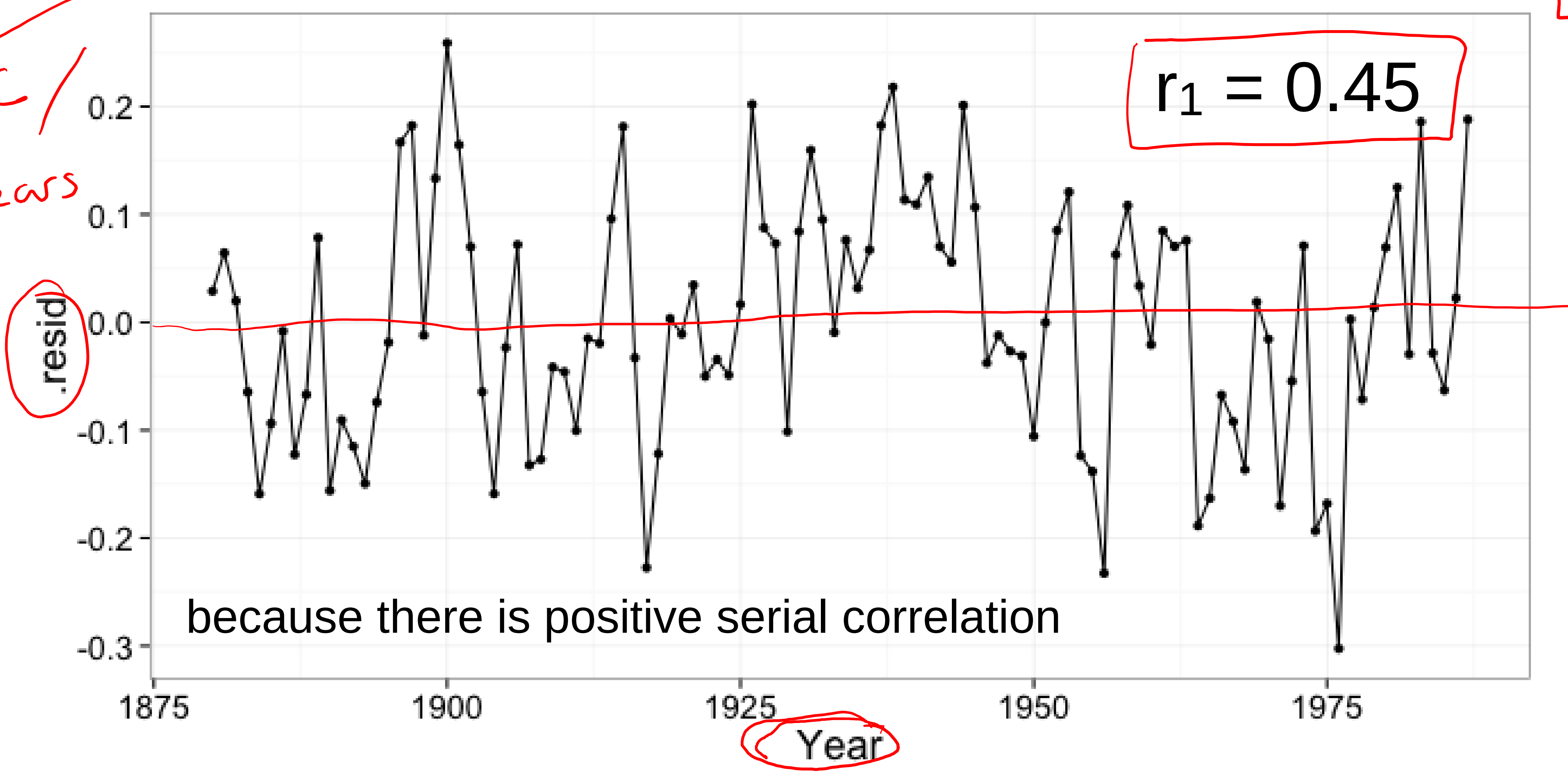
$$\beta_1 \neq 0 \, ?$$

Is serial correlation a problem?

```
> fit_slr <- lm(Temp ~ Year, data = case1502)
> summary(fit_slr)$coef
                Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -8.786714263 0.6795783683 -12.92966 1.592721e-23
Year         0.004493603 0.0003514301  12.78662 3.281042e-23
```

this will be an underestimate

0.4°C / 100 years



```
res <- residuals(fit_slr)
```


```
qplot(c(NA,res)    , c(res, NA))
```

Use filtering to get SE

If the AR(1) model is adequate and

$\mu\{ Temp_t \mid Year_t \} = \beta_0 + \beta_1 t$

Filtered variables:

$V_t = Temp_t - r_1 Temp_{t-1}$ response

$U_t = t - r_1 (t - 1)$ explanatory

Regress V_t on U_t

```
> case1502$lag_Year <- c(NA, case1502$Year[-nrow(case1502)])
> case1502$lag_Temp <- c(NA, case1502$Temp[-nrow(case1502)])
>
> # regress filtered variables
> fit_filt <- lm(I(Temp  - r1*lag_Temp) ~I(Year - r1*lag_Year) , data = case1502)
> summary(fit_filt)$coef
                              Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)                -4.92680691 0.6154922045 -8.004662 1.709667e-12
I(Year - r1 * lag_Year)     0.00460353 0.0005809344  7.924355 2.562586e-12
```

V_t U_t

0.46°C / century

Use adjustment to get SE

```
> summary(fit_slr)$coef
                Estimate   Std. Error   t value      Pr(>|t|)
(Intercept) -8.786714263 0.6795783683 -12.92966 1.592721e-23
Year         0.004493603 0.0003514301  12.78662 3.281042e-23
```

$$SE_{\beta_1} = \sqrt{(1 + r_1)/(1 - r_1)}\ SE_{\beta_{1slr}}$$

```
> sqrt((1+r1)/(1- r1)) * summary(fit_slr)$coef[, 2]
 (Intercept)         Year
1.1068682408 0.0005723943
```

Examine for serial correlation in the **residuals**. Not the raw response.

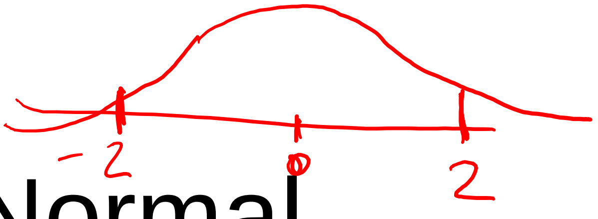
The filtering method extends to multiple explanatories.

Testing for serial correlation

$$\frac{0.45}{\sqrt{108}} \approx 0.04$$

Large sample test

$$Z = r_1/\sqrt{n}$$



If there is no serial correlation, Z has a Normal distribution.

only appropriate when n > 100

Runs test

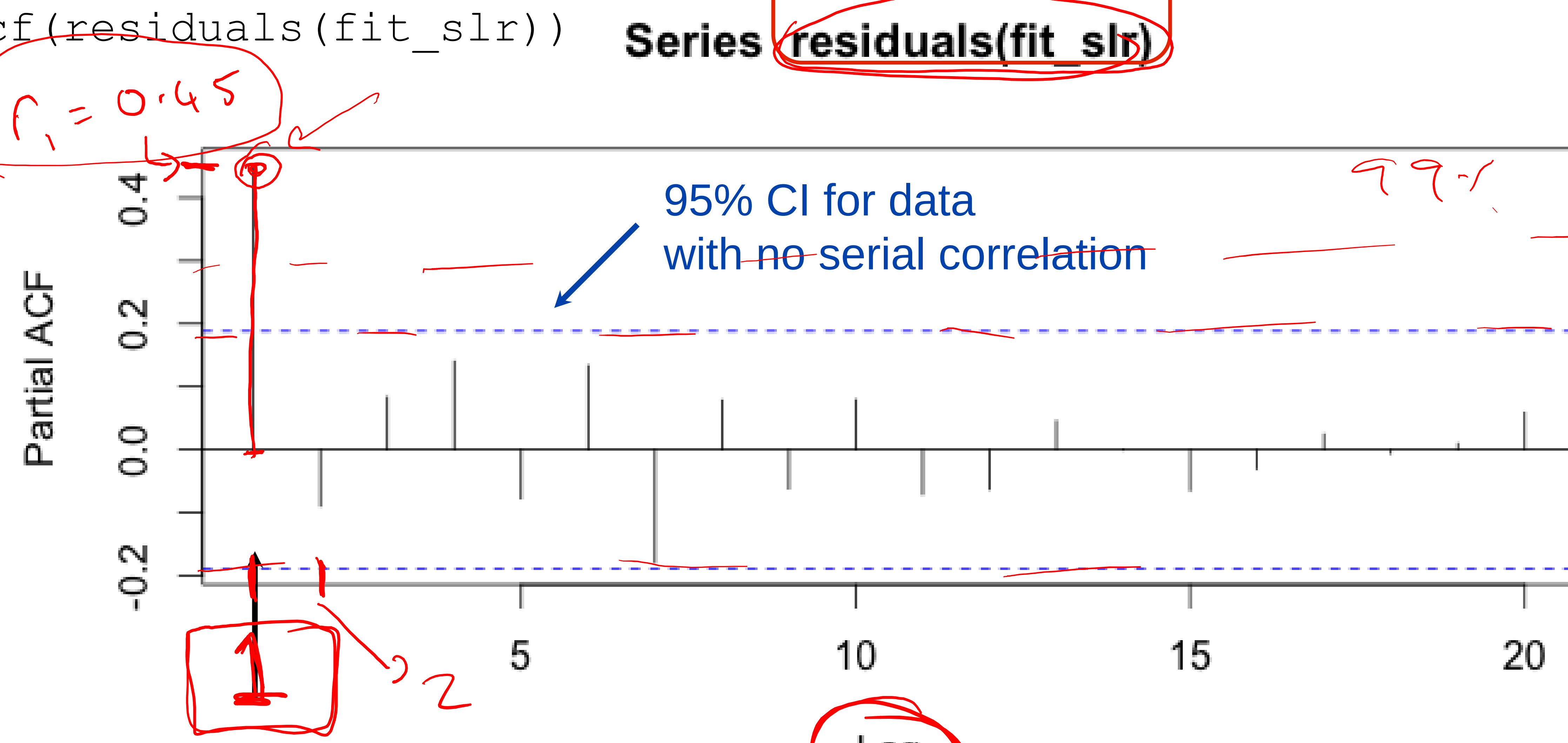
Count how many runs there and compare to how many we would expect by chance alone with no serial correlation. non-parametric

Is the AR(1) model adequate?

Partial AutoCorrelation Function

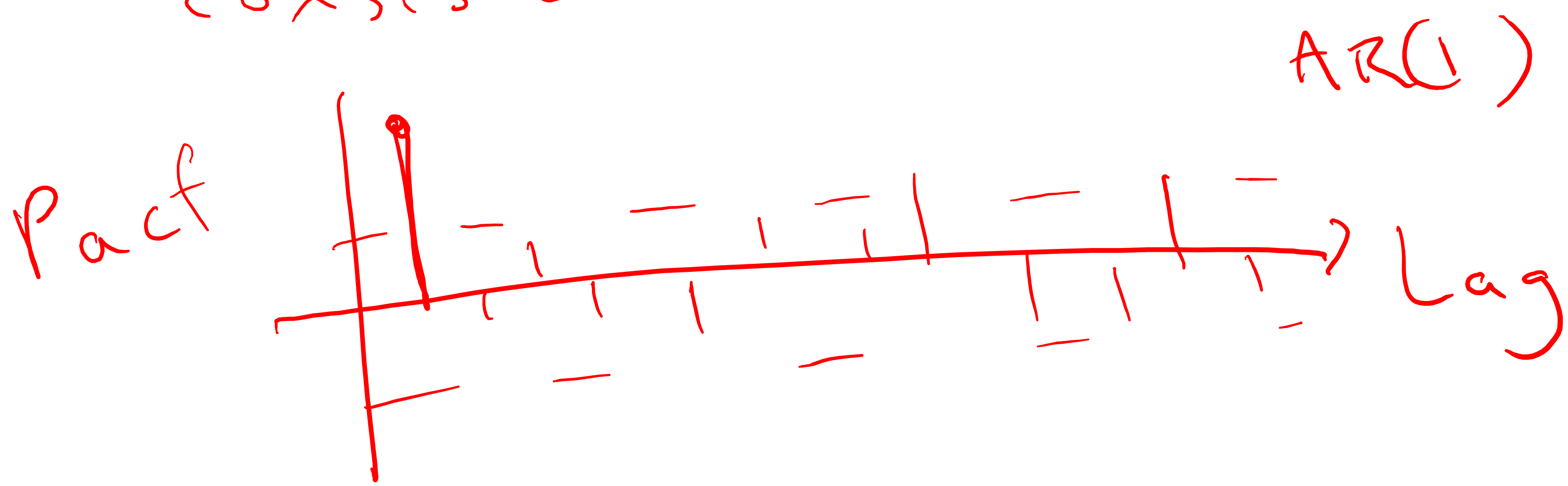
The primary tool is the PACF plot:

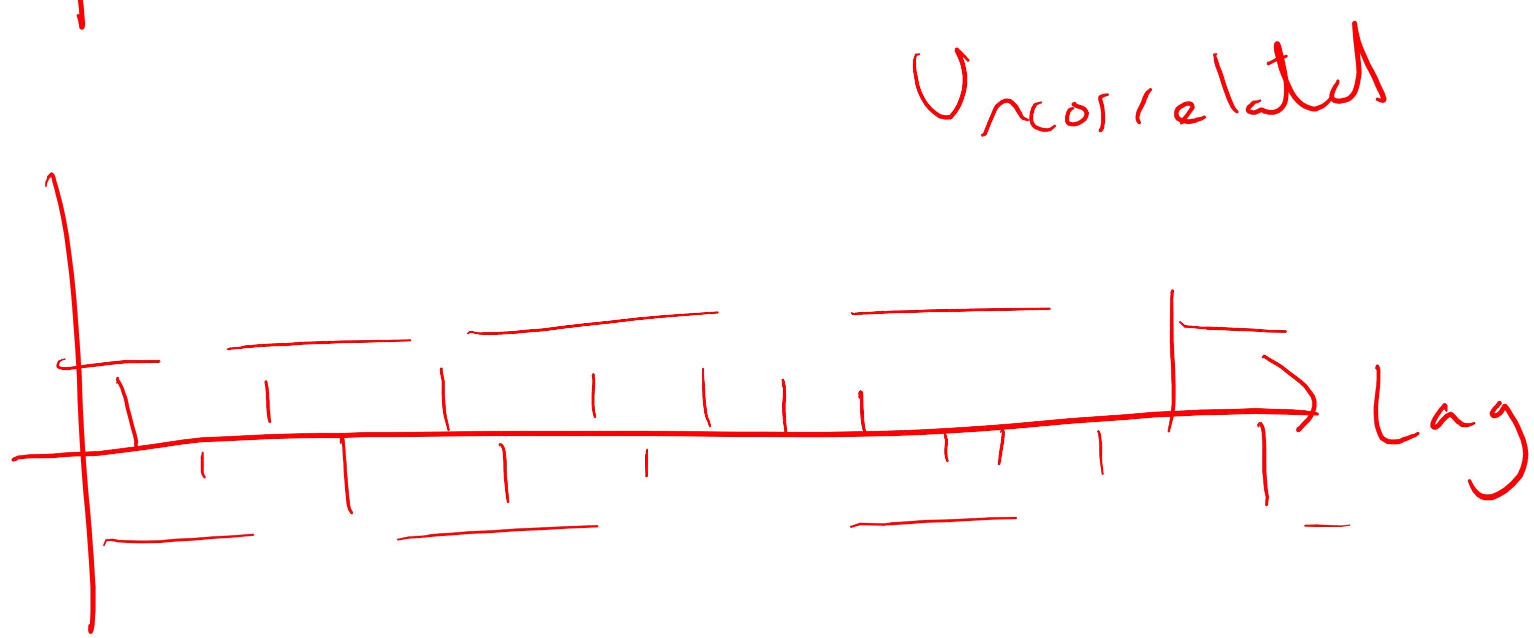
```
pacf(residuals(fit_slr))
```



Lag 1

For ~~as~~ series with AR(1) structure you tend to see lag 1 pac large, (outside 95% CI). And all other lags consistent with zero.





Partial autocorrelation functions for four different types of time series