Stat 412/512

Serial Correlation

Feb 23 2015

Charlotte Wickham

stat512.cwick.co.nz

Overview of regression

A model for the mean:

$$\mu\{Y \mid X_1, \ldots X_p\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

+ assumptions:

There is a Normally distributed subpopulation at each combination of explanatory variables values.

The means of the subpopulations fall on the line/surface defined above ($\mu\{Y \mid X_1, \ldots X_p\}$)

The subpopulation standard deviations are all equal to σ

The selection of an observation from one subpopulation is independent of the selection of any other observation.

The deviation of an observation from the mean, is independent of the deviation from the mean for any other observation.

Ch. 15 &16

equivalent

Serial Correlation

The multiple regression tools rely on the observations being independent (after accounting for the effects of the explanatory variables).

Often when measurements are made at adjacent points in time or space the observations are correlated.

case1501: Patch-cut logging

Clear cutting (stripping the land of all vegetation) is one method of logging Douglas Fir.

Water quality in streams is adversely affected by clear cutting.

An alternative is patch cutting.

Observe two watersheds, one from patch-cut and one undisturbed.

Measure water quality by nitrates.

Is the mean nitrate level higher for the patch cut watershed?

Your turn

Watershed	Nitrate
PC	10
PC	12
PC	13
UD	10
UD	11

X many

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

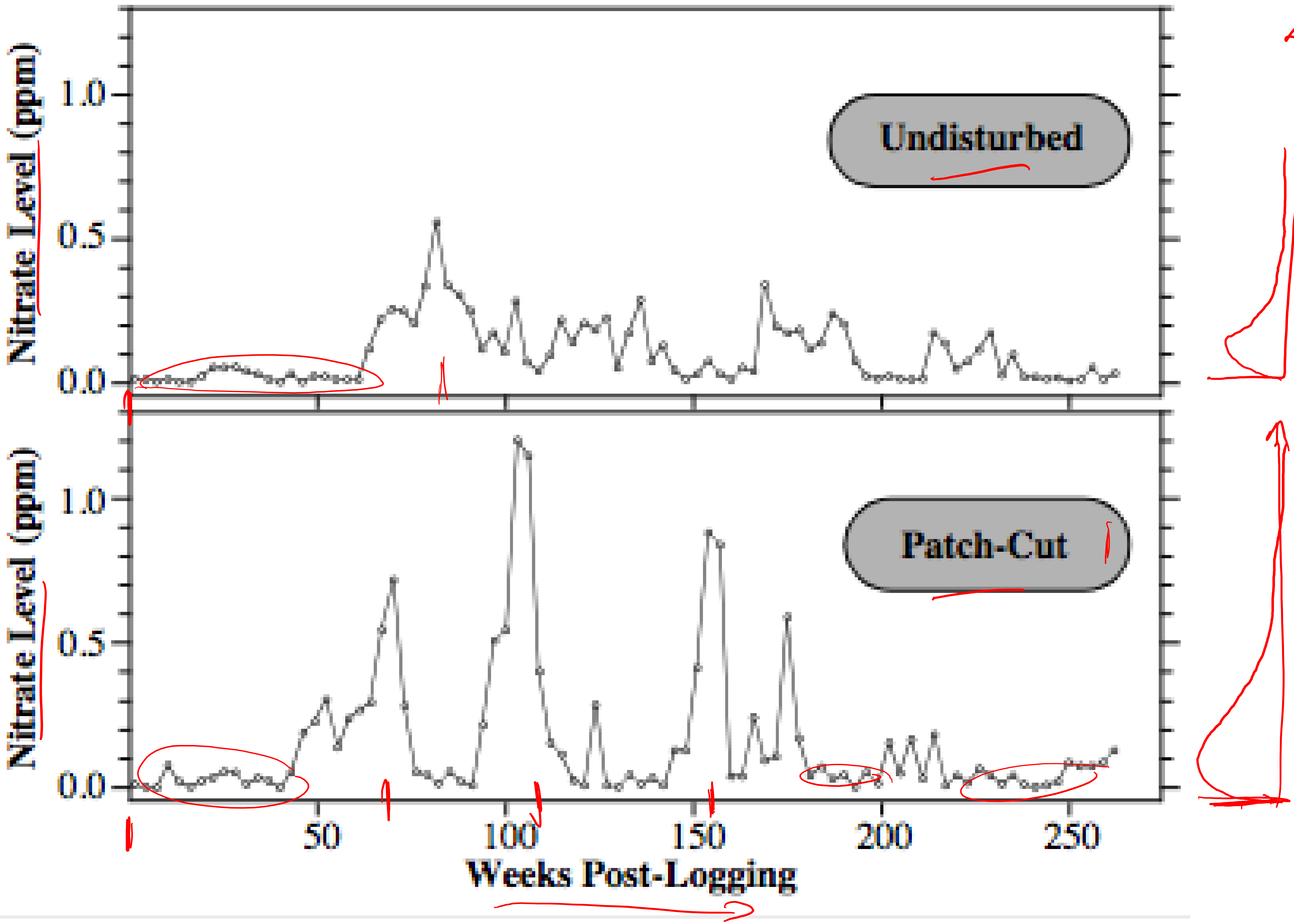
Is the mean nitrate level higher for the patch cut watershed compared to undisturbed watershed?

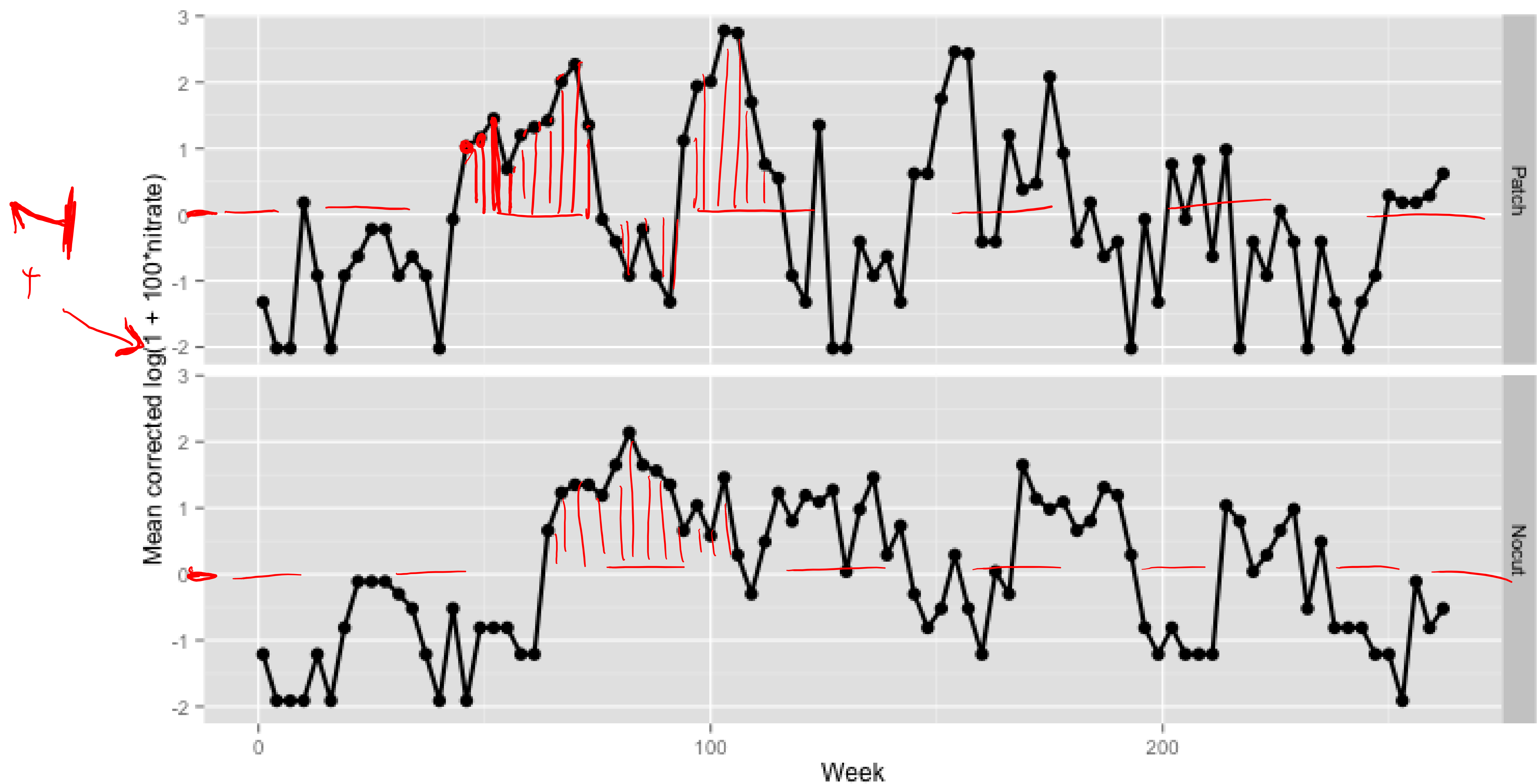
two sample t-test

$\mu\{\text{Nitrate} \mid \quad\} = \beta_0 + \beta_1 PC$

$\beta_1 = 0$

Nitrates (NO3-N) in runoff from patch-cut and undisturbed watersheds, for five years after logging





After transformation, and subtracting the sample average from each.

Both are centered around 0.

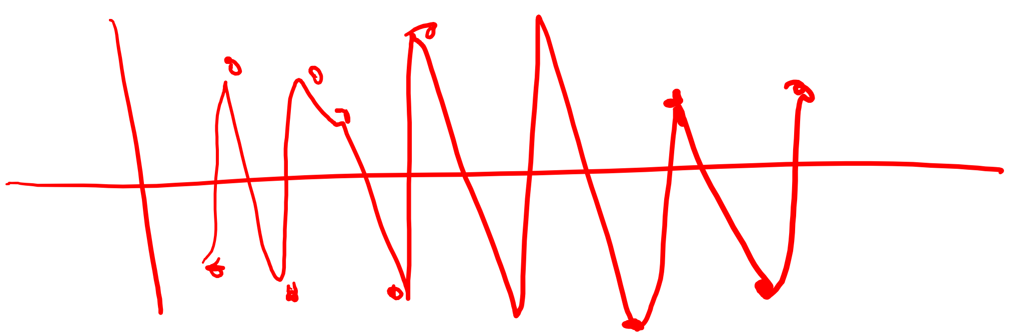
Notice the “runs” of observations above or below the mean.

Serial correlation

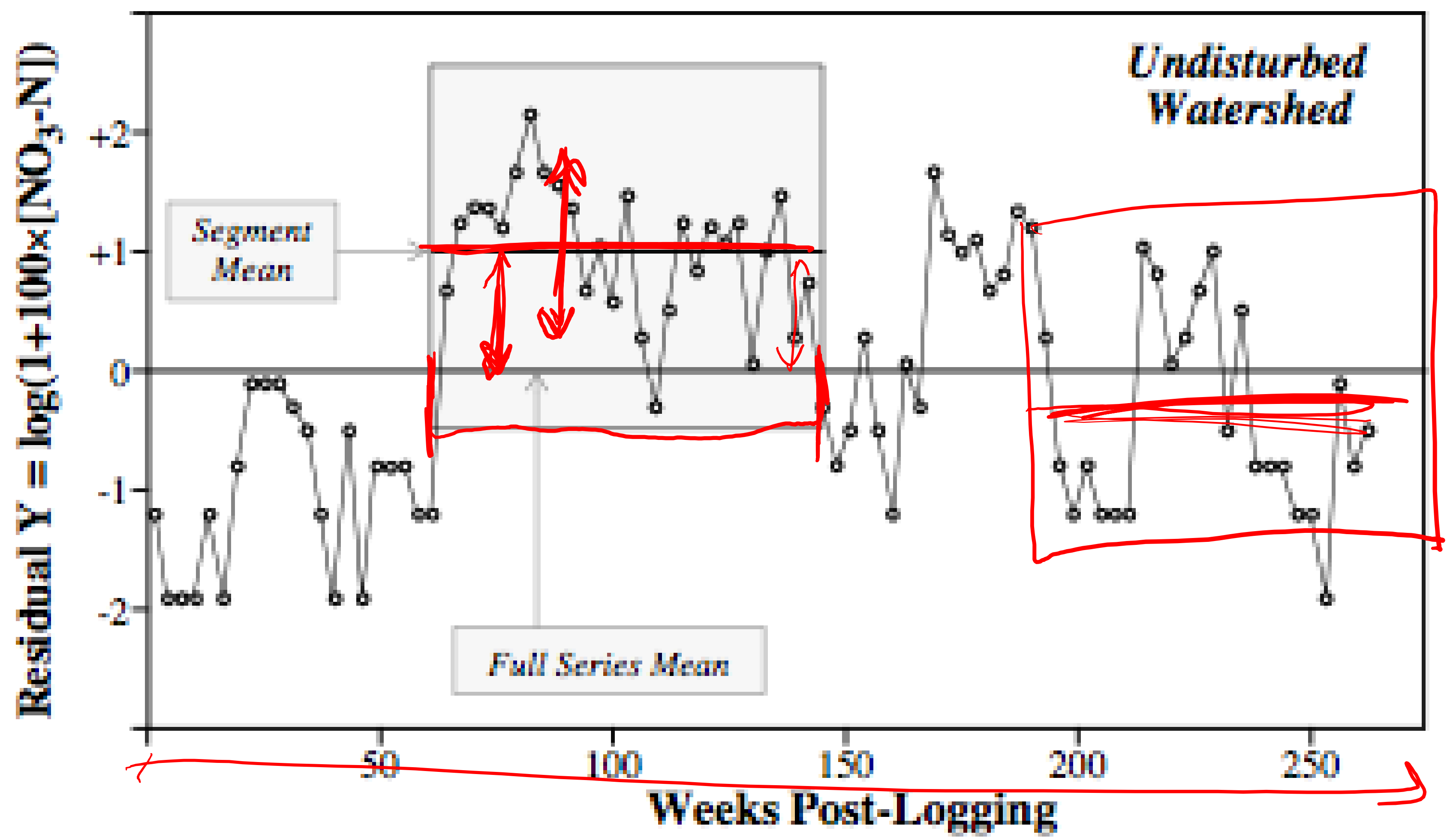
a.k.a autocorrelation

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.

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

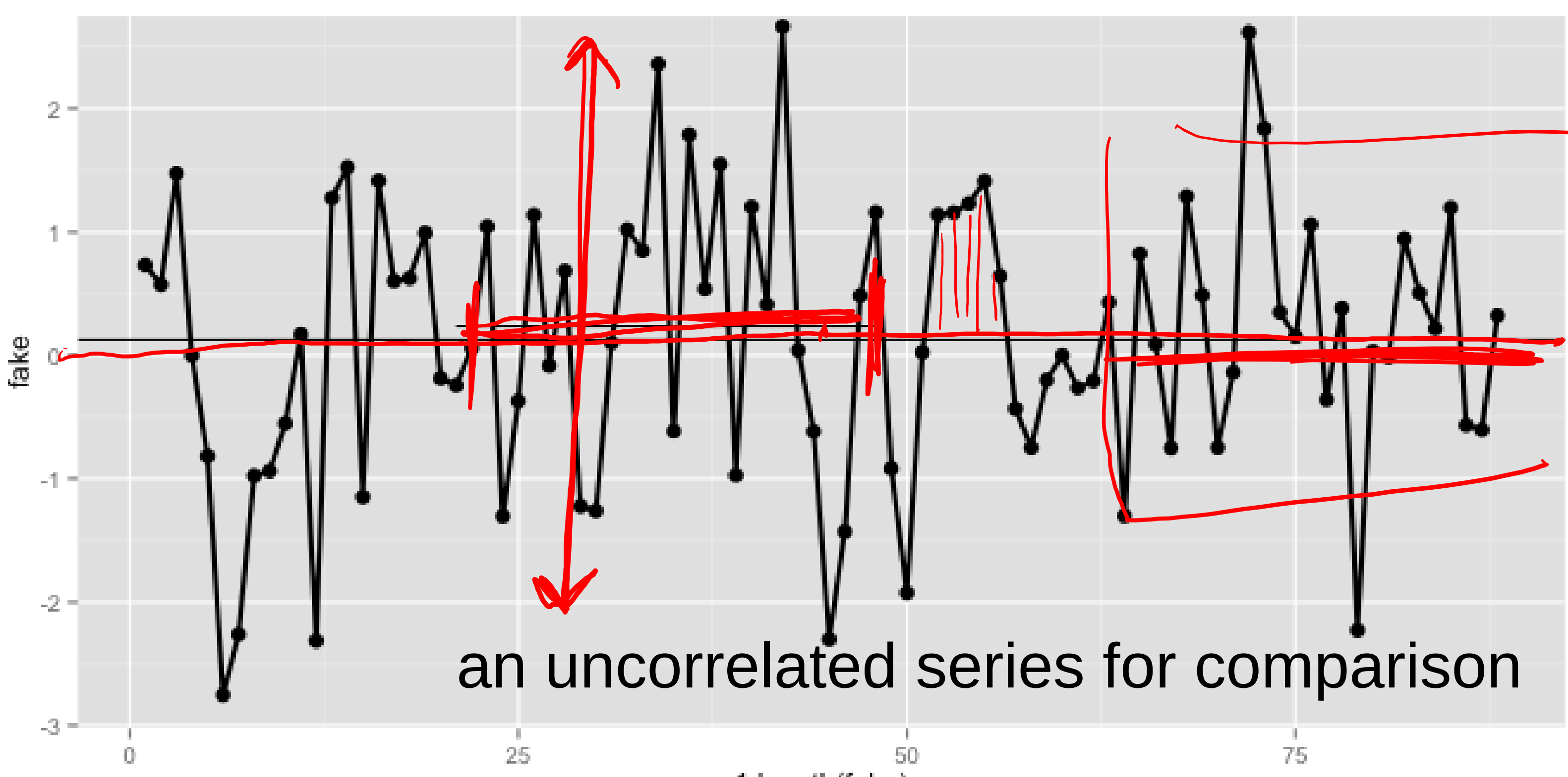


Mean-corrected nitrate concentrations after transformation, and a demonstration that the average of a segment of a time series may grossly misrepresent the full series mean



The "runs" make averages of subsamples much more variable about the mean than for uncorrelated series.

The observations also exhibit less variability than expected without correlation.



The usual SE on the average formula, $s/\sqrt{n}$ ⟶ too small will overestimate the precision when there is positive autocorrelation.

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.

Time Series

Longitudinal Data

Examining serial correlation

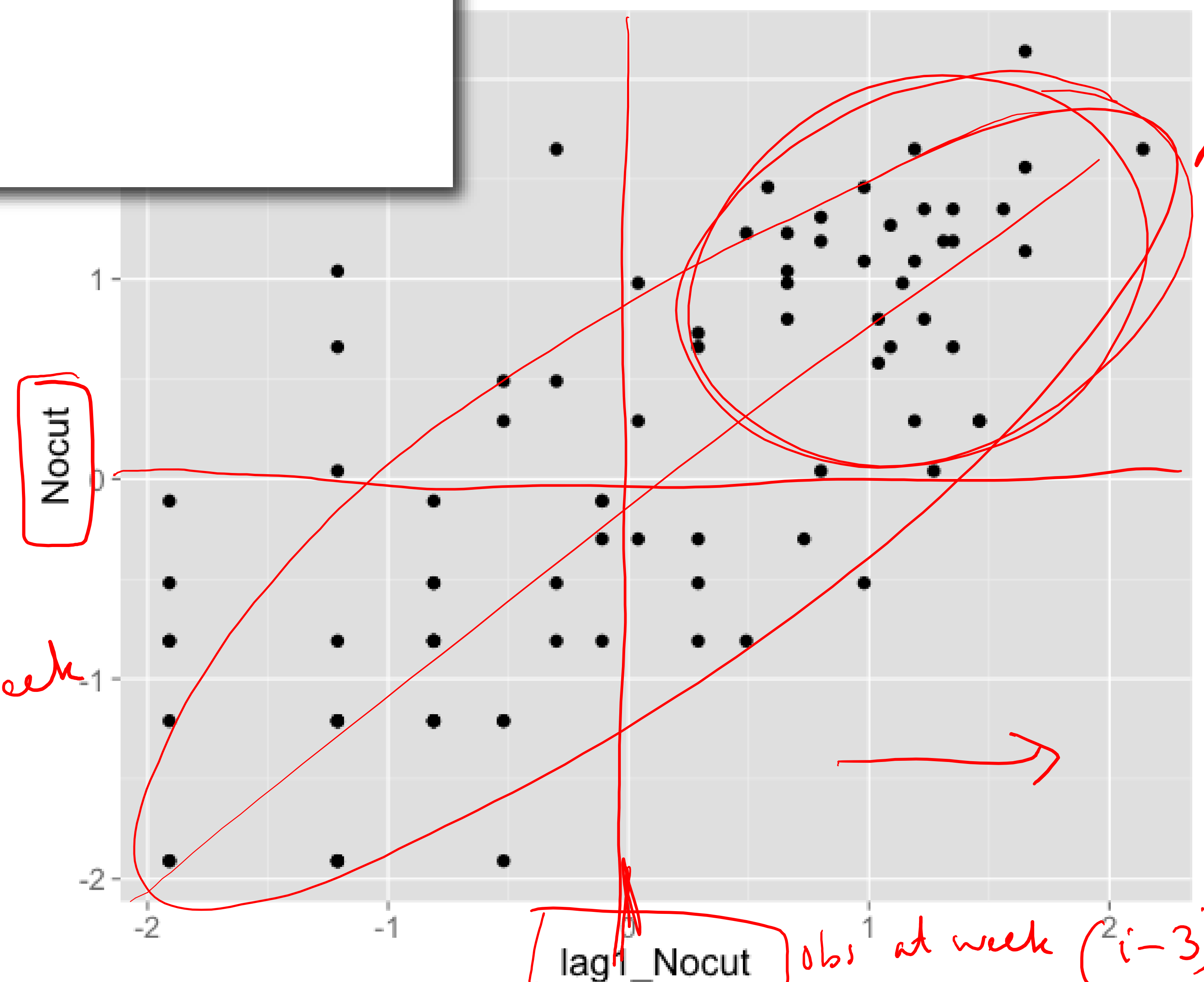
Week	Patch	Nocut	lag1_Patch	lag1_Nocut
1	1	-1.32	-1.21	NA	NA
2	4	-2.02	-1.91	-1.32	-1.21
3	7	-2.02	-1.91	-2.02	-1.91
4	10	0.18	-1.91	-2.02	-1.91
5	13	-0.92	-1.21	0.18	-1.91
6	16	-2.02	-1.91	-0.92	-1.21

the transformed nitrate concentration from the previous week

plotted in slides

We see a positive correlation!

obs at week i



Estimating serial correlation

essentially, the sample correlation between current and previous residual – $r_1 = \boxed{\frac{c_1}{c_0}}$

$$c_1 = \frac{1}{n-1}\sum_{t=2}^{n} res_t \times res_{t-1}$$

transformed response

covariance of current and previous residual – avg watershed

$$c_0 = \frac{1}{n-1}\sum_{t=1}^{n} res_t^2$$

variance of residuals

Or in R:

```
with(case1501, acf(Nocut))$acf
```

residuals

```
      [,1]
[1,]  1.000000000
[2,]  0.744034541
[3,]  0.622066930
[4,]  0.493194040
```

r_1 1st serial correlation coefficient

r_2 2nd serial correlation coefficient

correlation between now & two observations ago

1. An adjusted SE on the sample average

$$\mathrm{SE}_{\overline{Y}} = \sqrt{\frac{1+r_1}{1-r_1}}\,\frac{s}{\sqrt{n}}$$

sample avg. 1st correlation

where r_1 is the **first serial correlation coefficient**.

Appropriate under the autoregressive model of order 1 (AR(1)):

- The series is measured at equally spaced times
- Let v be the long run series mean, then

mean

μ { Y_t - v | past history } = α (Y_{t-1} - v)

$-1 < a < 1$

$0 < a < 1$ +ve

where α is the first order autocorrelation coefficient.

A two sample comparison

Do the usual two sample procedure, but adjust the standard error:

$$\overline{Y}_C - \overline{Y}_U = 2.016 - 1.905 = \boxed{0.111}$$

two sample SE calc.

$$\mathrm{SE}_{\overline{Y_c}-\overline{Y_u}} = \sqrt{\frac{1+r_1}{1-r_1}}\, s_p \sqrt{\frac{1}{n_c}+\frac{1}{n_u}}$$

adjustment

$$= \sqrt{\frac{1+0.644}{1-\boxed{0.644}}}\, 1.183 \sqrt{\frac{1}{88}+\frac{1}{88}} = \boxed{0.383}$$

+ve correlation

>1

pooled correlation coefficient

pooled s.d.

no correlation

pooled = assume the same for both watersheds, and use both sets of data to estimate them

2. Filter variables to remove correlation.

If the AR(1) model is adequate and

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

Then the **filtered** variables:

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

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

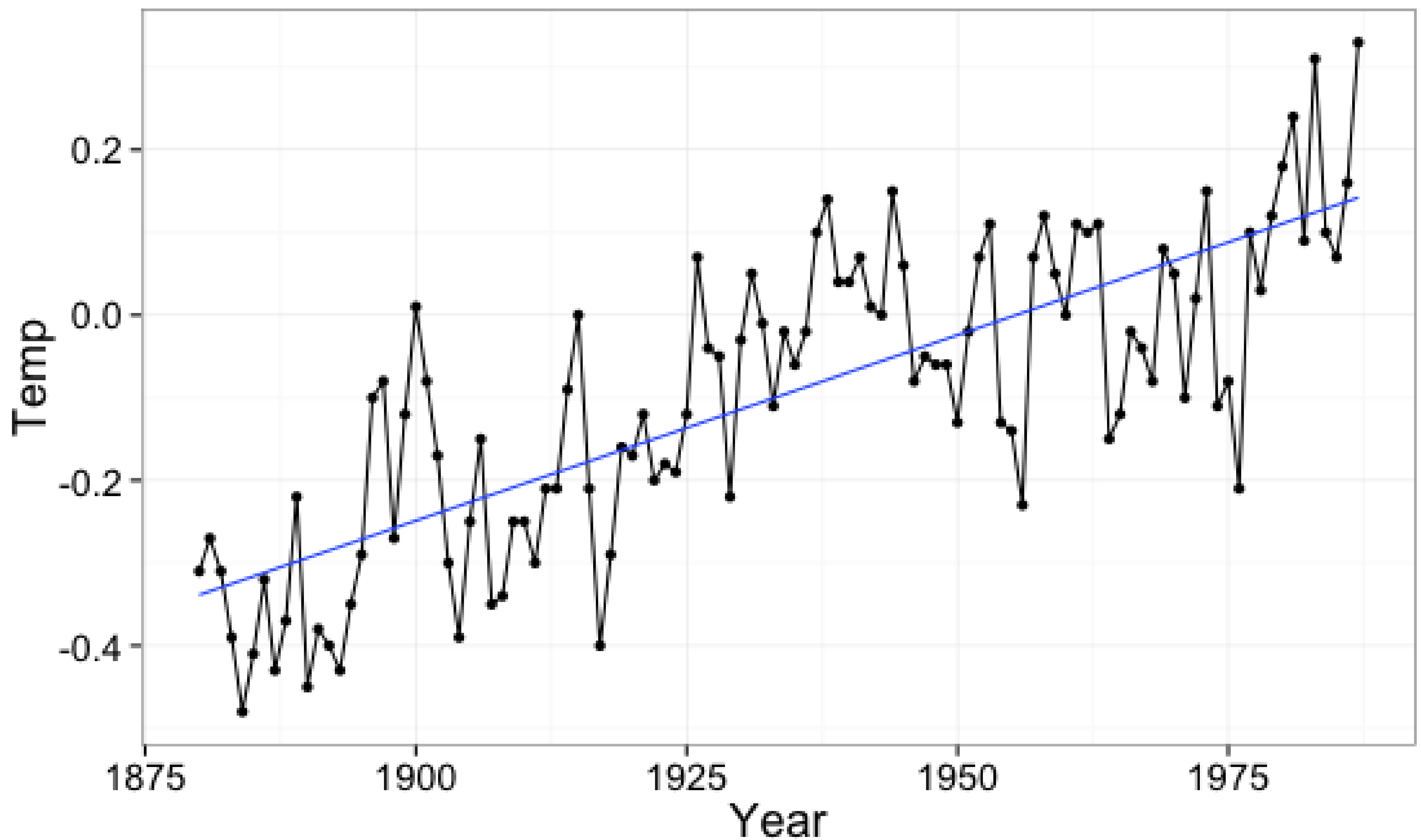
are related by the same slope:

with no serial correlation

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

Use 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

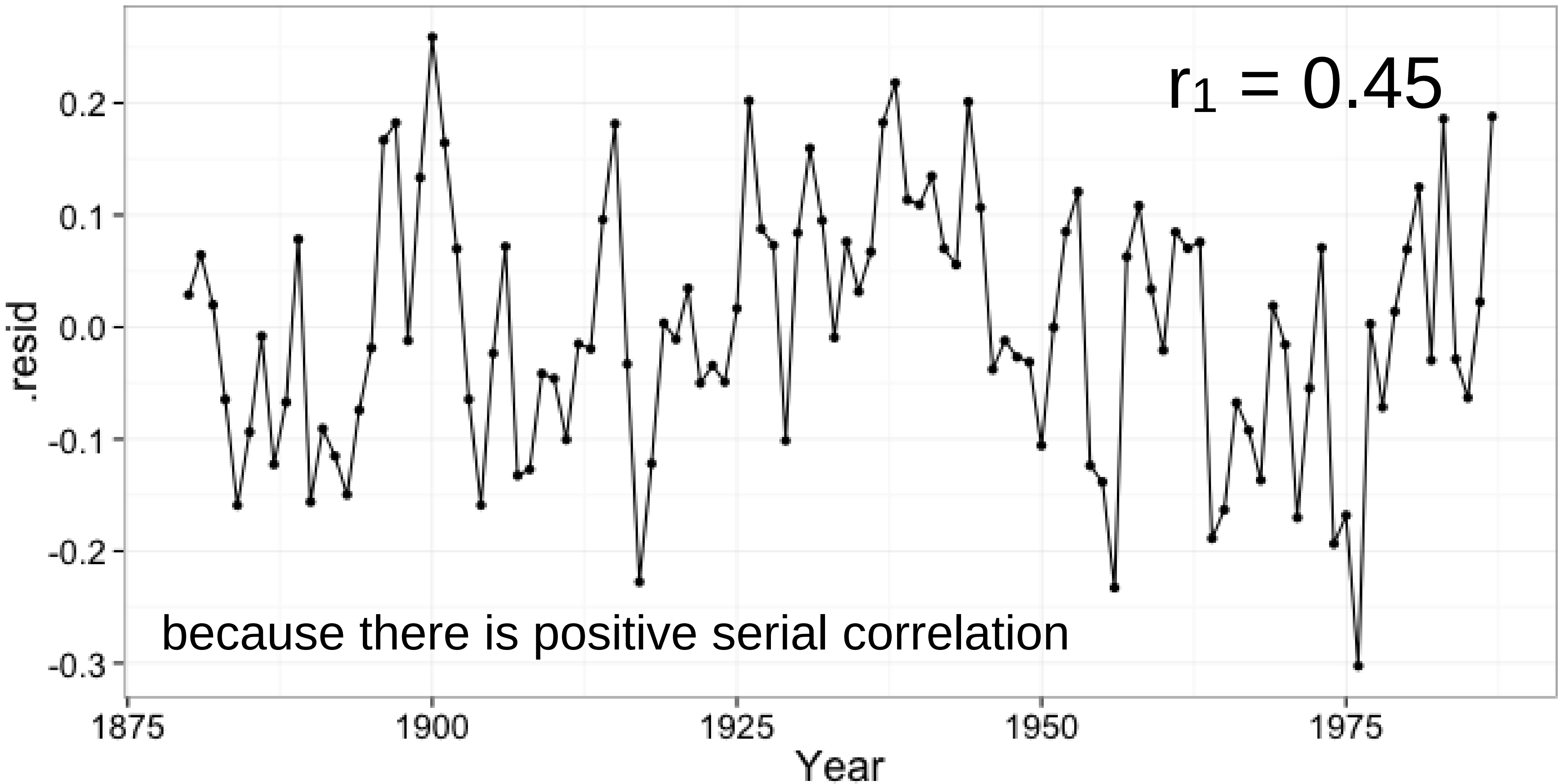
Ignoring of the appropriateness of regression, how would you answer the question of interest?

Is the mean temperature increasing?

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



2. 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:

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

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

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
```

1. 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

Large sample test

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

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

only appropriate when n > 100

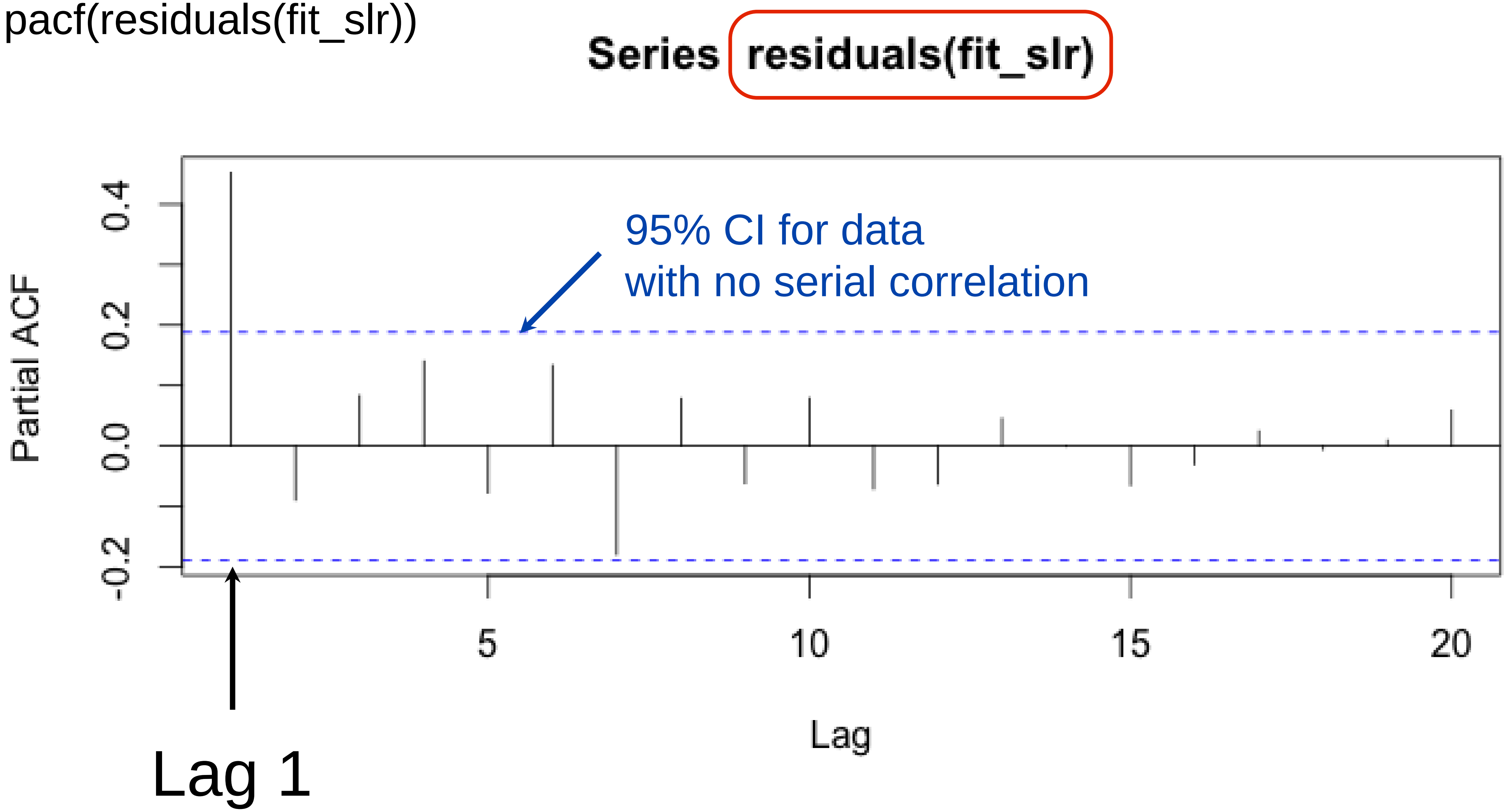
Runs test

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?

The primary tool is the PACF plot:

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



Partial autocorrelation functions for four different types of time series