Miscellaneous I - Lagged Regressions

STAT 1321/2320

Kiran Nihlani

Department of Statistics
University of Pittsburgh

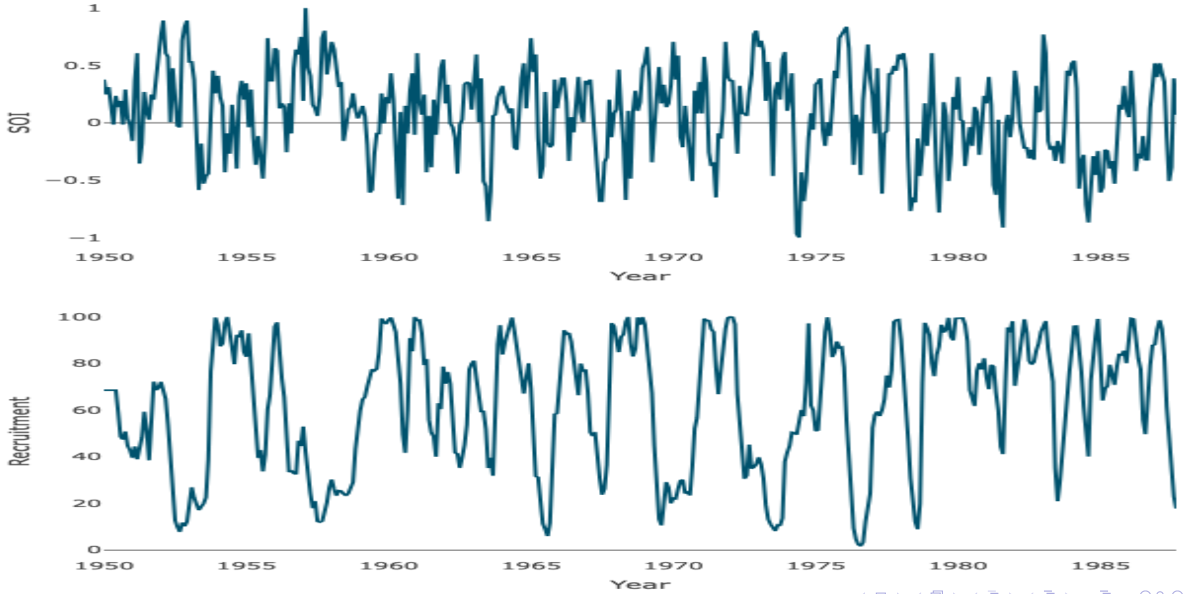
Motivation

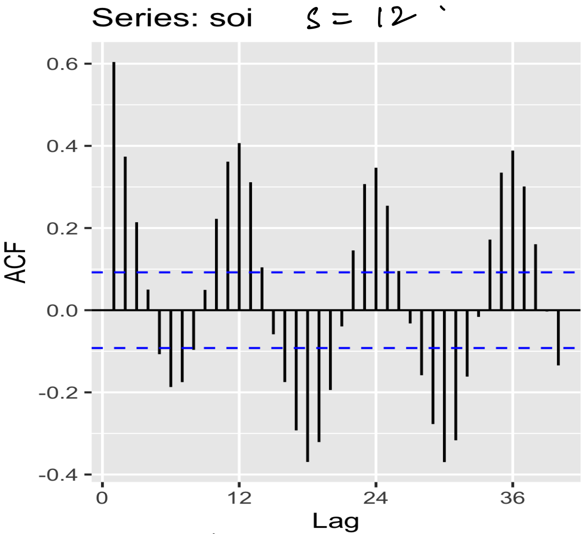
- So far we have only analyzed one time series at a time.
- We can also try to analyze two or more time series together to understand and isolate
  - the effect of lagged observations of the same series
  - the "actual" effect of time series on each other.

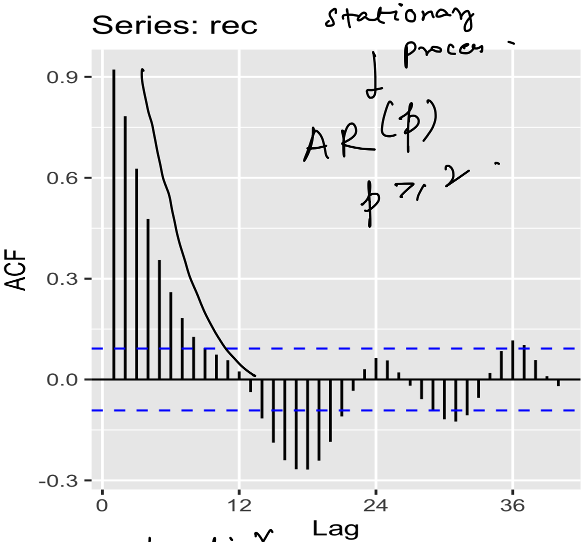
spurious correlation / regression.

Example - Southern Oscillation Index (SOI) and Recruitment

- The SOI measures changes in air pressure, related to sea surface temperatures in the central Pacific Ocean.
- The recruitment series records the number of new fish.
- Question of interest - can SOI inform/predict recruitment?
- This data is available for 434 months as “soi” and “rec” in the astsa package.


1
0.5
0
−0.5
−1
SOI
1950
1955
1960
1965
1970
1975
1980
1985
Year
100
80
60
40
20
0
Recruitment
1950
1955
1960
1965
1970
1975
1980
1985
Year





trend: ×
seasonality: ✓

trend: ×
seasonality: ×

Cross-correlation Function (CCF)

$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$.

- Just like the ACF and PACF, we can define another correlation function that calculates the linear, serial dependence between two time series.
- The sample CCF can be calculated as:

$\hat{\gamma}_{xy}(h) = \text{ACov}(X_{t+h}, Y_t)$

$$\widehat{\rho}_{xy}(h) = \frac{\widehat{\gamma}_{xy}(h)}{\sqrt{\widehat{\gamma}_x(0)\widehat{\gamma}_y(0)}}$$

where $\widehat{\gamma}_{xy}(h)$ is the sample cross-covariance between the two series given as:

$$\widehat{\gamma}_{xy}(h) = n^{-1}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y})$$

CCF

- You may think of sample CCF as the set of sample correlations between X_{t+h} and Y_t for $h = 0, \pm 1, \pm 2, \ldots$
- When one or more X_{t+h}, with negative h, are predictors of Y_t, it is sometimes said that X leads Y.
- When one or more X_{t+h} , with positive h, are predictors of Y_t, it is sometimes said that X lags Y.
- You may use the CCF plot to
  - identify which variable is leading or lagging, and
  - which lags are relevant to capture the relationship between the two series.

$$\mathrm{corr}(X_{t+h}, y_t)$$

$$h = \pm 1$$

$$\mathrm{corr}(X_{t+1}, y_t)$$

$$\mathrm{corr}(X_{t-1}, y_t).$$

Large-Sample Distribution of Cross-Correlation

The large sample distribution of $\widehat{\rho}_{xy}(h)$ is normal with mean zero and

$$\sigma_{\widehat{\rho}_{xy}} = \frac{1}{\sqrt{n}}$$

if at least one of the processes is **independent white noise**.

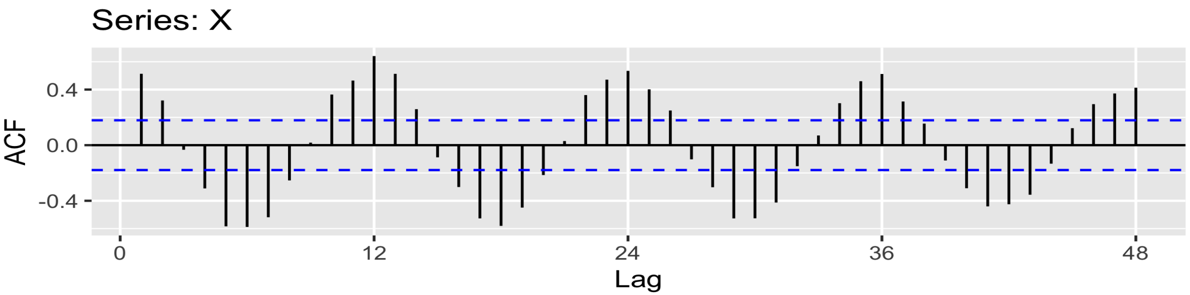
This implies that while we can use the asymptotic distribution to create confidence bounds for non-white noise series, they cannot be used to assess significance of correlations.

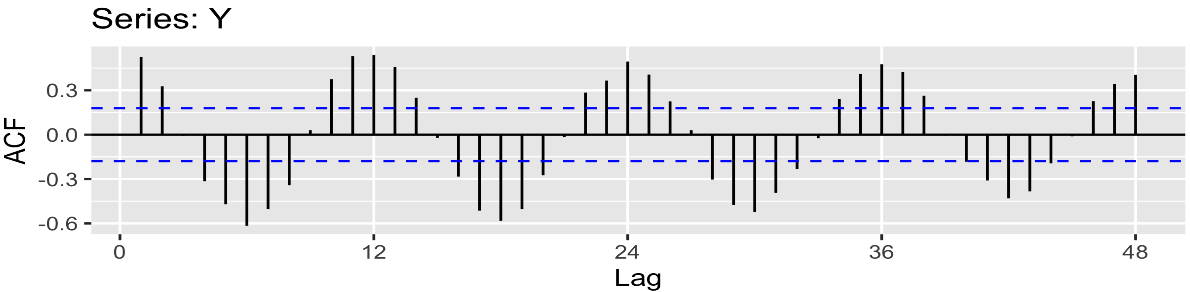
Large-Sample Distribution of Cross-Correlation

- Consider two time series

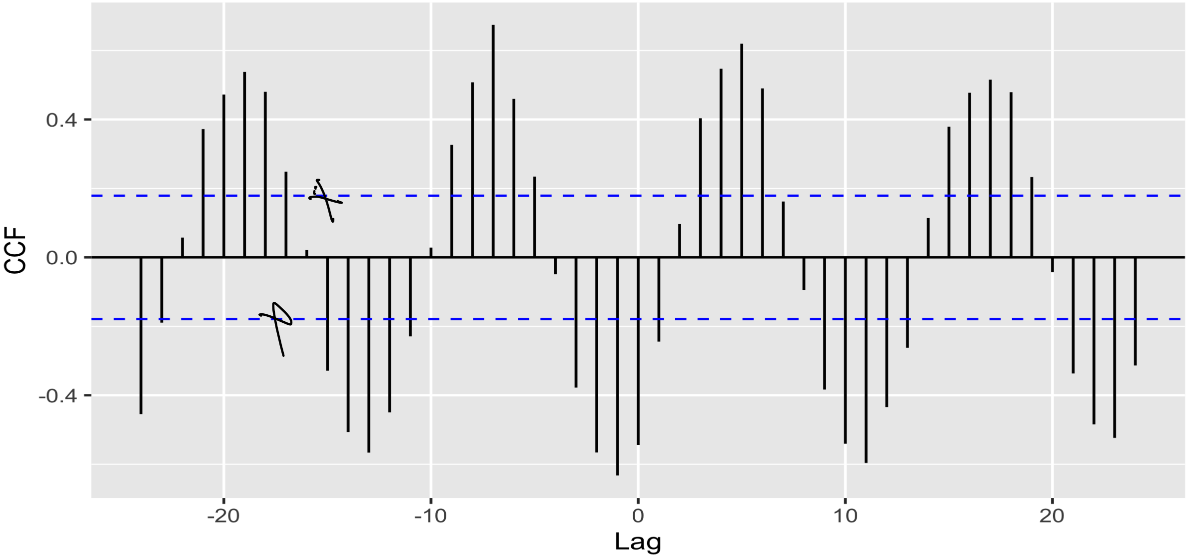
$$X_t = 2\cos\left(\frac{2\pi t}{12}\right) + W_t \quad \text{and} \quad Y_t = 2\cos\left(\frac{2\pi(t+5)}{12}\right) + W_t^*$$

- The white noise series are iid N(0,1).
- This implies that both X_t and Y_t have seasonal components and they are independent.
- The sample CCF should be close to 0 in this case.


Series: X
ACF
0.4
0.0
-0.4
0
12
24
36
48
Lag


Series: Y
ACF
0.3
0.0
-0.3
-0.6
0
12
24
36
48
Lag

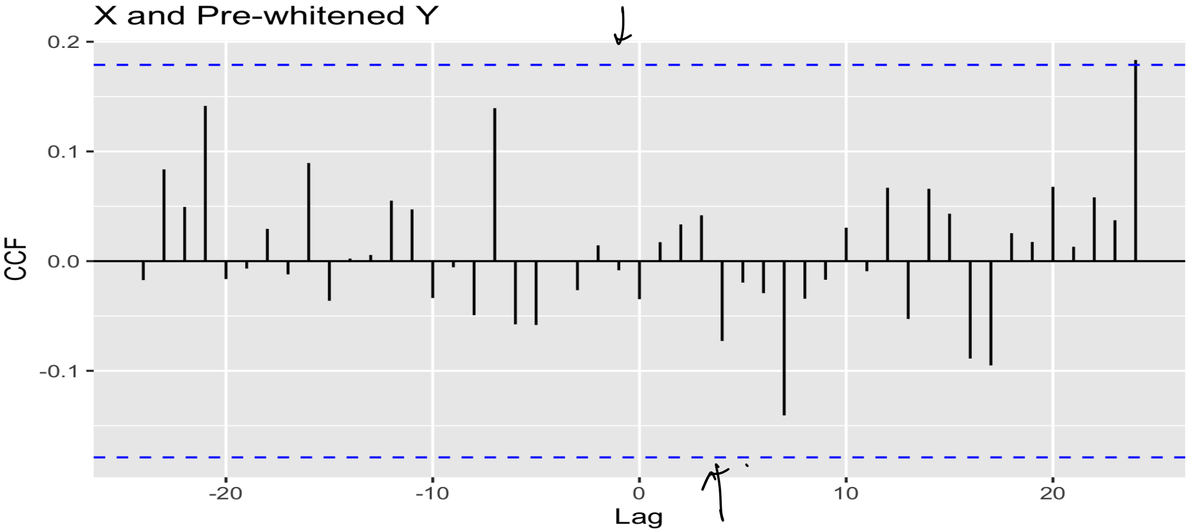
Series: X & Y



CCF - Pre-whitening

fit a model to y.

calculate residuals.



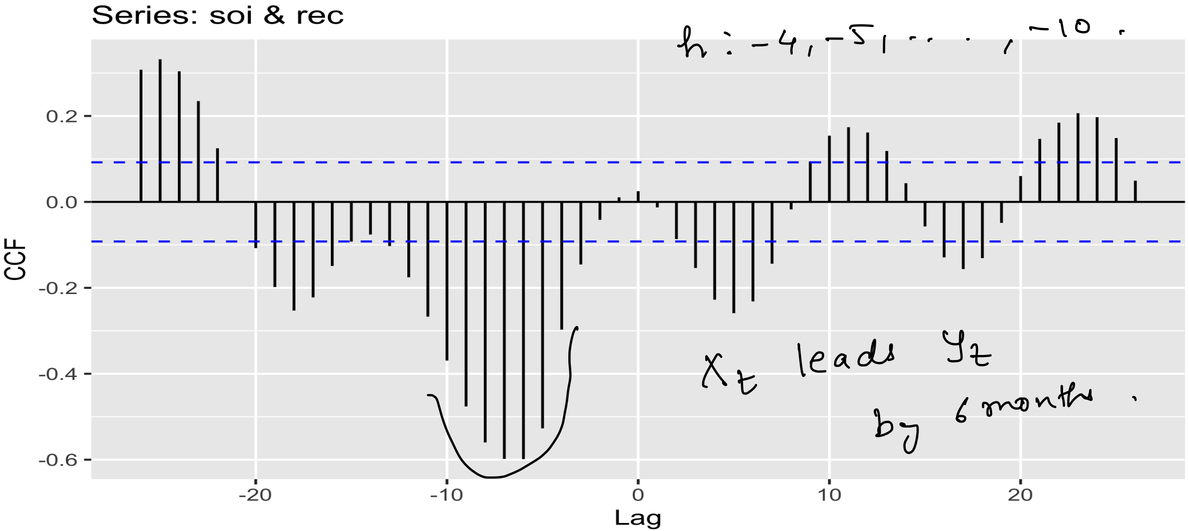
CCF - SOI and Recruitment

soi : X_t

rec : y_t

$corr(X_{t+h}, y_t)$

$h: -4, -5, \ldots, -10$



X_t leads y_t by 6 months.

REC $\sim$ SOI

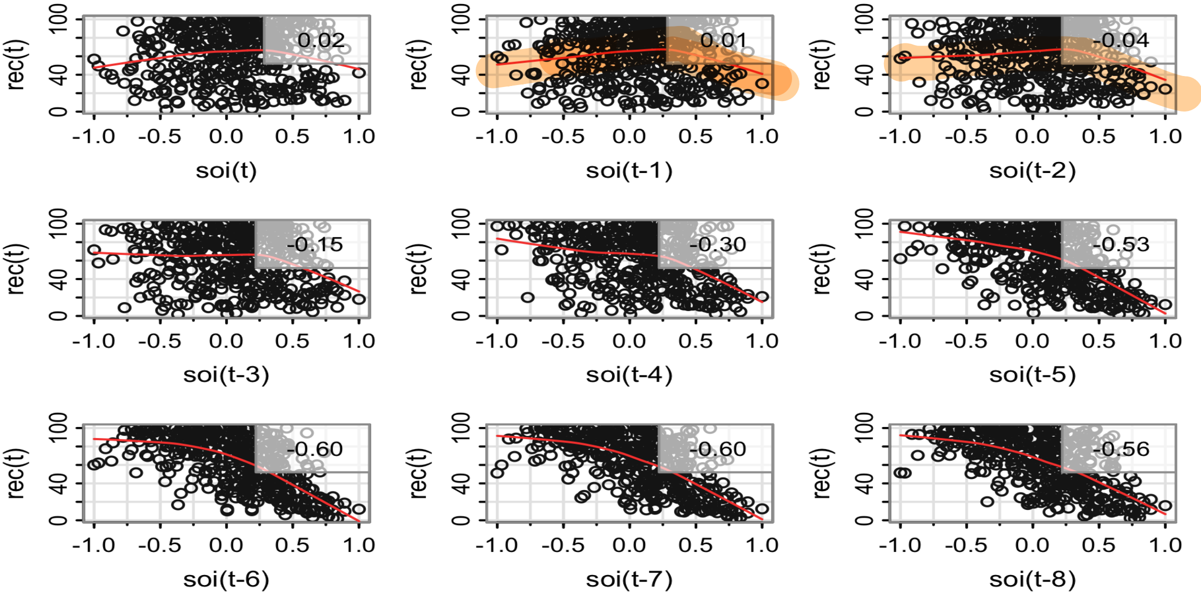
$\downarrow$

lags -

$SOI_{t-4} + SOI_{t-5} + \cdots + SOI_{t-10}$

Observations from CCF

- The plot is created with SOI and Recruitment at X_t and Y_t respectively.
- Both series exhibit trend and/or seasonality (hence, not white noise) so the confidence bounds have no relevance.
- We see significant cross-correlations for both positive and negative lags but high, significant correlations are between $h = -4$ to -10.
- The highest, negative cross - correlations are at lags $-6, -7$.
- This indicates that the SOI series leads the recruitment series by 6-7 months.



rec(t)
0
40
100
-1.0 -0.5 0.0 0.5 1.0
soi(t)
0.02
soi(t-1)
0.01
soi(t-2)
-0.04
soi(t-3)
-0.15
soi(t-4)
-0.30
soi(t-5)
-0.53
soi(t-6)
-0.60
soi(t-7)
-0.60
soi(t-8)
-0.56

Modelling - SOI vs Recruitment

- Model 1: We know that lags $h = -5, -6, \ldots, -10$ of SOI are relevant for predicting Recruitment
- We can fit a regression model as follows:

y_{t-1} y_{t-2} y_{t-3}

$$Y_t = \beta_0 + \beta_1 X_{t-5} + \beta_1 X_{t-6} + \beta_1 X_{t-7} + \beta_1 X_{t-8} + \beta_1 X_{t-9} + \beta_1 X_{t-10}$$

- We can evaluate the residuals of the model for statioanrity/white noise to see if needs further modelling.

Modelling - SOI vs Recruitment

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.2743     0.8703  79.601  < 2e-16 ***
soilag5     -23.8255     2.7657  -8.615  < 2e-16 ***
soilag6     -15.3775     3.1651  -4.858 1.65e-06 ***
soilag7     -11.7711     3.1665  -3.717 0.000228 ***
soilag8     -11.3008     3.1664  -3.569 0.000398 ***
soilag9      -9.1525     3.1651  -2.892 0.004024 **
soilag10    -16.7219     2.7693  -6.038 3.33e-09 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.42 on 436 degrees of freedom
Multiple R-squared:  0.6251,    Adjusted R-squared:   0.62
F-statistic: 121.2 on 6 and 436 DF,  p-value: < 2.2e-16
```

Modelling - SOI vs Recruitment

- Model 2: We know that lags $h = -5, -6, \ldots, -10$ of SOI are relevant for predicting Recruitment
- Using a technique called pre-whitening, we can show that lags 1 and 2 of the Recruitment series are informative towards understanding the relationship.
- We can fit a regression model as follows:

$$Y_t = \beta_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \beta_1 X_{t-5} + \beta_1 X_{t-6} + \beta_1 X_{t-7} + \beta_1 X_{t-8} + \beta_1 X_{t-9} + \beta_1 X_{t-10}$$

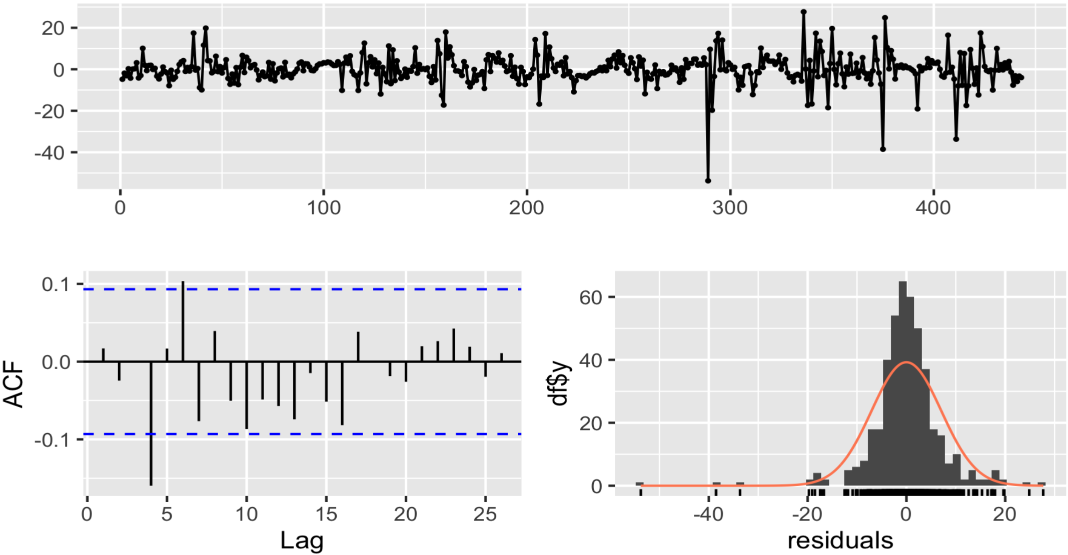
Modelling - SOI vs Recruitment

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.43047    1.33384   8.570  < 2e-16 ***
reclag1        1.25702    0.04316  29.128  < 2e-16 ***
reclag2       -0.41946    0.04120 -10.182  < 2e-16 ***
soilag5      -21.19210    1.11838 -18.949  < 2e-16 ***
soilag6        9.77648    1.56238   6.257  9.4e-10 ***
soilag7       -1.19189    1.32247  -0.901   0.3679
soilag8       -2.17345    1.30806  -1.662   0.0973 .
soilag9        0.56520    1.30035   0.435   0.6640
soilag10      -2.58630    1.19529  -2.164   0.0310 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.034 on 434 degrees of freedom
Multiple R-squared:  0.9392,    Adjusted R-squared:  0.938
F-statistic: 837.5 on 8 and 434 DF,  p-value: < 2.2e-16
```

Modelling - SOI vs Recruitment

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.76807    1.00640   8.712  < 2e-16 ***
reclag1       1.24694    0.04336  28.759  < 2e-16 ***
reclag2      -0.37251    0.03864  -9.639  < 2e-16 ***
soilag5     -20.83104    1.10577 -18.838  < 2e-16 ***
soilag6       8.63164    1.43779   6.003 4.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.089 on 438 degrees of freedom
Multiple R-squared:  0.9376,    Adjusted R-squared:  0.9371
F-statistic:  1647 on 4 and 438 DF,  p-value: < 2.2e-16
```

Residuals


20
0
-20
-40
0
100
200
300
400
ACF
0.1
0.0
-0.1
0
5
10
15
20
25
Lag
df$y
60
40
20
0
-40
-20
0
20
residuals

$$D_i = \begin{cases} 0 & \text{if } Soi < 0 . \\ 1 & \geq 0 . \end{cases}$$

$$y_t = \beta_0 + \beta_1 \, Soi_{t-6} + \varepsilon_t .$$