

Extensions of Linear Correlation Analysis

MAR 599, Spring 2009

Daniel A. Birch

School for Marine Science and Technology
University of Massachusetts, Dartmouth

March 2009

Outline

- 1 Review
- 2 Geometric Mean Functional Regression
- 3 Spearman's Rank Correlation
- 4 Transformation of Variables
- 5 Autocovariance
- 6 Cross Covariance
- 7 Summary
- 8 Matlab Notes

- ① The **Sample Mean** $\hat{\mu}_x$ is an unbiased estimate of the process mean:

$$\hat{\mu}_x = \frac{1}{N} \sum_{i=1}^N x_i, \quad E(\hat{\mu}_x) = \mu_x$$

- ② The **Sample Variance** $\hat{\sigma}_x^2$ is an unbiased estimate of the process variance if the observations x_i are independent:

$$\hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu}_x)^2, \quad E(\hat{\sigma}_x^2) = \sigma_x^2$$

- ③ If x_i and y_j are independent for $i \neq j$, then the **Sample Covariance** \hat{c}_{xy} is an unbiased estimate of the Covariance:

$$\hat{c}_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y), \quad E(\hat{c}_{xy}) = c_{xy}$$

- The **Correlation Coefficient** ρ_{xy} is

$$\rho_{xy} \equiv \frac{c_{xy}(x, y)}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

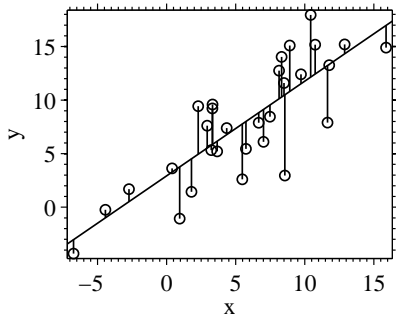
(Also called Pearson's Product-Moment Correlation and Pearson's r)

- If x and y are bivariate normal, then $\hat{\rho}_{xy}$ is an unbiased estimate of ρ_{xy} :

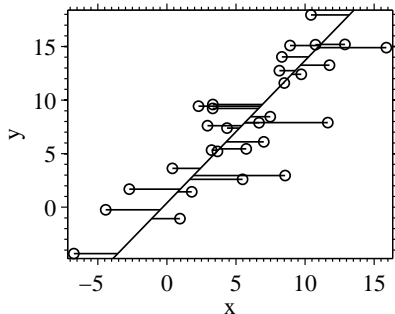
$$\hat{\rho}_{xy} = \frac{\hat{c}_{xy}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_y^2}}$$

Review — Linear Regression

$$\hat{y}_i = \hat{\mu}_y + \hat{\rho}_{xy} \frac{\hat{\sigma}_y}{\hat{\sigma}_x} (x_i - \hat{\mu}_x)$$

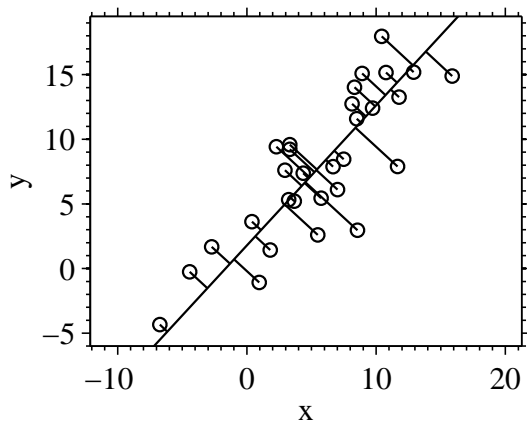


$$\hat{x}_i = \hat{\mu}_x + \hat{\rho}_{xy} \frac{\hat{\sigma}_x}{\hat{\sigma}_y} (y_i - \hat{\mu}_y)$$



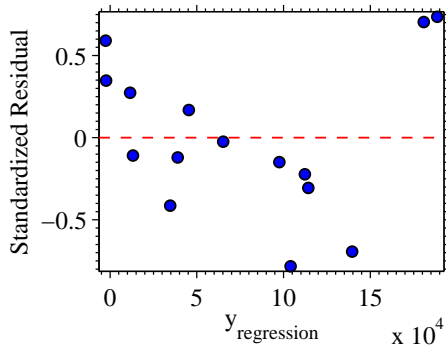
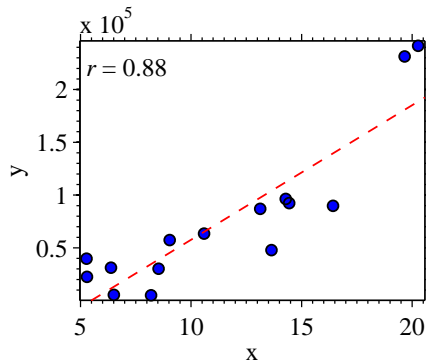
Geometric Mean Functional Regression

Neutral Regression

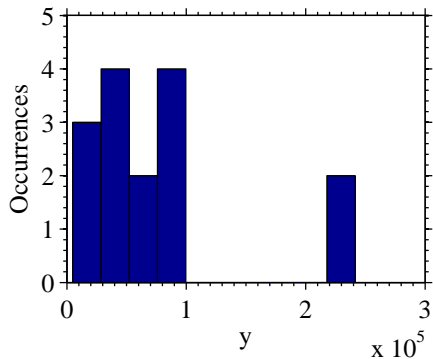
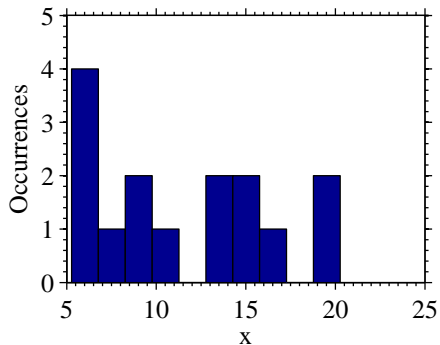


$$y_{\text{mgfr}} = \hat{\mu}_y + \frac{\hat{\sigma}_y}{\hat{\sigma}_x} (x_{\text{mgfr}} - \hat{\mu}_x)$$

Spearman's Rank Correlation — Introduction



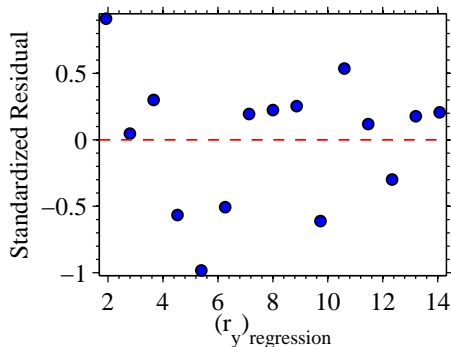
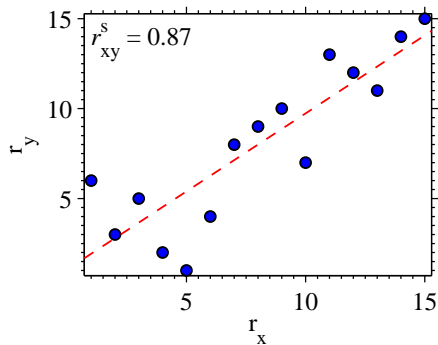
Spearman's Rank Correlation — Histograms



Spearman's Rank Correlation — Rank the Data

x	r_{x_i}	y	r_{y_i}
8.195	5	5003.649	1
5.271	1	39786.343	6
5.290	2	22593.309	3
14.437	12	92293.869	12
13.636	10	47832.689	7
6.382	3	31184.240	5
10.590	8	63428.344	9
19.662	14	231363.791	14
16.424	13	89830.505	11
9.033	7	57428.184	8
14.284	11	96320.207	13
6.508	4	5399.207	2
8.529	6	30266.736	4
13.129	9	86905.414	10
20.274	15	241451.277	15

Spearman's Rank Correlation — Calculate $\hat{\rho}_{xy}^s$



Calculate the linear correlation coefficient of the ranks:

$$\hat{\rho}_{xy}^s = \hat{\rho}_{r_{x_i} r_{y_i}}$$

Spearman's Rank Correlation — Hypothesis Testing

- Use a table for small N (e.g., von Storch (1999), Appendix K)

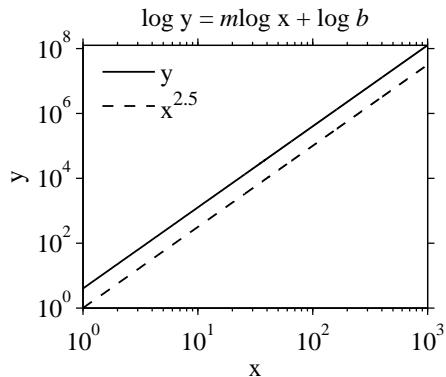
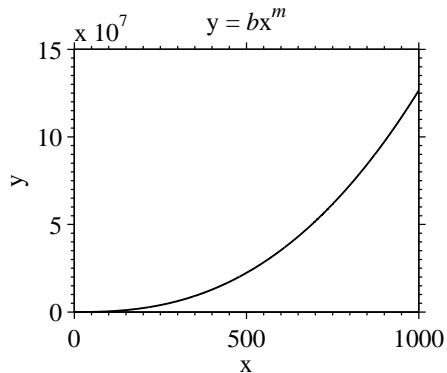
$N = 15$						
CDF	0.900	0.950	0.975	0.900	0.995	0.999
ρ^s	0.3500	0.4429	0.5179	0.600	0.6536	0.7464

- Use a normal distribution for large N

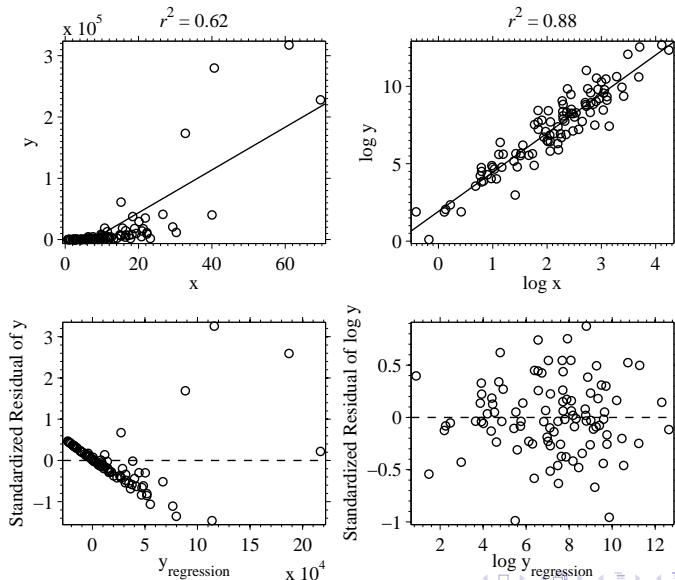
$$\rho_c^s \approx \frac{\Phi^{-1}(q)}{\sqrt{N-1}}$$

- Extra Credit: Find (or write) and demonstrate the use of an m-file that calculates the CDF for the Spearman rank correlation coefficient if the null hypothesis $\rho^s = 0$ is true.

Review — Power Law Scaling



Transformation of Variables — Log-Transform



A process $y(t)$ is **Stationary** if

- F_y does not change in time

Remarks:

- If y is stationary, then $F_{y(t)y(t+\tau)}$ depends only on $|\tau|$
- The autocovariance only makes sense if a quantity is stationary
- Many ocean properties are not truly stationary, but may be approximately stationary over the time scales of interest

Autocovariance (Autocorrelation)

- The **Autocovariance** is the covariance of a variable with itself at times separated by a lag τ :

$$c_{yy}(\tau) \equiv E((y(t) - \mu_y)(y(t + \tau) - \mu_y))$$

Autocovariance (Autocorrelation)

- The **Autocovariance** is the covariance of a variable with itself at times separated by a lag τ :

$$c_{yy}(\tau) \equiv E((y(t) - \mu_y)(y(t + \tau) - \mu_y))$$

- A biased estimate of the autocovariance is

$$\hat{c}_{yy}(k\Delta t) = \frac{1}{N - k} \sum_{i=1}^{N-k} (y_i - \hat{\mu}_y)(y_{i+k} - \hat{\mu}_y)$$

Autocovariance (Autocorrelation)

- The **Autocovariance** is the covariance of a variable with itself at times separated by a lag τ :

$$c_{yy}(\tau) \equiv E((y(t) - \mu_y)(y(t + \tau) - \mu_y))$$

- A biased estimate of the autocovariance is

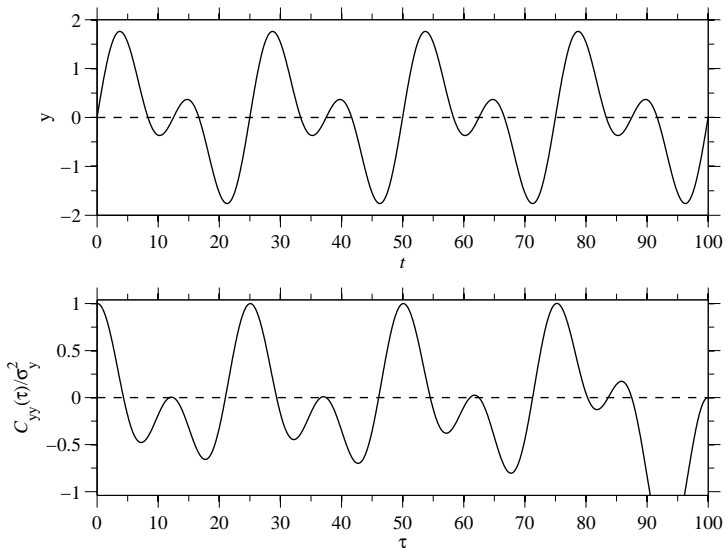
$$\hat{c}_{yy}(k\Delta t) = \frac{1}{N - k} \sum_{i=1}^{N-k} (y_i - \hat{\mu}_y)(y_{i+k} - \hat{\mu}_y)$$

- The **Normalized Autocovariance** is

$$\hat{\rho}_{yy}(k\Delta t) = \frac{\hat{c}_{yy}(k\Delta t)}{\hat{\sigma}_y^2}$$

where $\hat{\sigma}_y^2$ is the biased sample variance

Autocovariance — Example 1



Poulain and Niiler (1989)

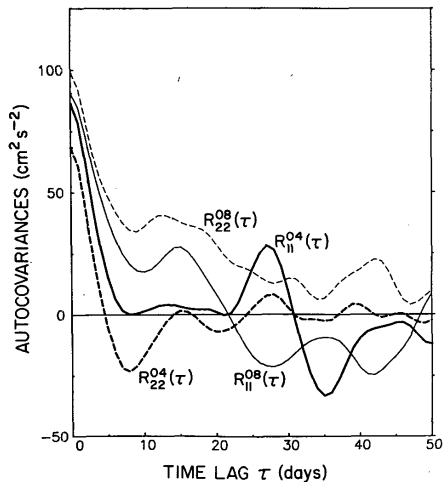


FIG. 5. Zonal (solid) and meridional (dashed) Lagrangian autocovariances versus time lag for drifters 04 and 08.

Cross Covariance (Cross Correlation)

The **Cross Covariance** of two random variables is the covariance of the variables at times separated by τ :

$$c_{xy}(\tau) = E((x(t + \tau) - \mu_x)(y(t) - \mu_y))$$

Cross Covariance (Cross Correlation)

The **Cross Covariance** of two random variables is the covariance of the variables at times separated by τ :

$$c_{xy}(\tau) = E((x(t + \tau) - \mu_x)(y(t) - \mu_y))$$

An estimate of the cross covariance is

$$\hat{c}_{xy}(k\Delta t) = \frac{1}{N-k} \sum_{i=1}^{N-k} (x_{i+k} - \hat{\mu}_x)(y_i - \hat{\mu}_y)$$

Cross Covariance (Cross Correlation)

The **Cross Covariance** of two random variables is the covariance of the variables at times separated by τ :

$$c_{xy}(\tau) = E((x(t + \tau) - \mu_x)(y(t) - \mu_y))$$

An estimate of the cross covariance is

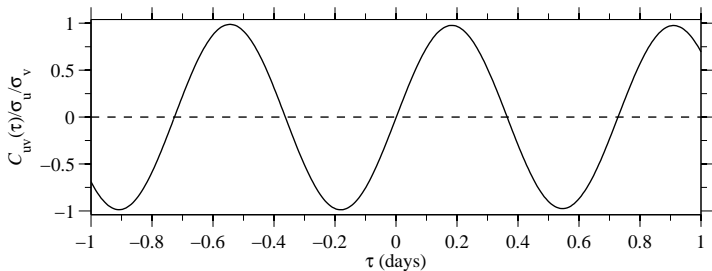
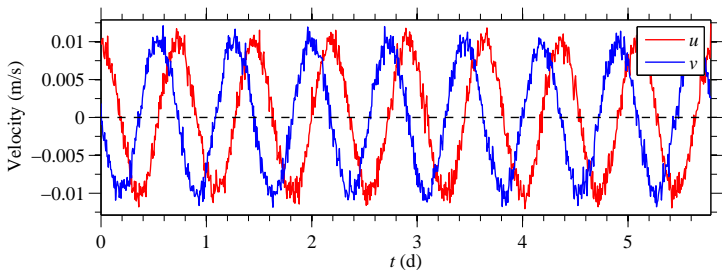
$$\hat{c}_{xy}(k\Delta t) = \frac{1}{N-k} \sum_{i=1}^{N-k} (x_{i+k} - \hat{\mu}_y)(y_i - \hat{\mu}_y)$$

The sample **Normalized Cross Covariance** (Cross-Correlation) is

$$\hat{\rho}_{xy}(k\Delta t) = \frac{\hat{c}_{xy}(k\Delta t)}{\hat{\sigma}_x \hat{\sigma}_y}$$

using the biased sample standard deviations

Cross Covariance — Example



Summary

- 1 Use the Spearman rank correlation coefficient if the data have a nonlinear, but one-to-one relationship
- 2 Transform variables if you suspect a specific functional relationship, but be aware of the assumptions you're making about the noise
- 3 Use the autocovariance to look for dominant periodicities or integral time scales/decorrelation times
- 4 Use the cross-covariance to see if one variable leads or lags another

Useful functions:

- ① 'tiedrank'
- ② 'xcov'

- ① Emery (2001): Data Analysis Methods in Physical Oceanography
- ② von Storch (1999): Statistical Analysis in Climate Research