Orthogonal Least Squares

Paired Correlation

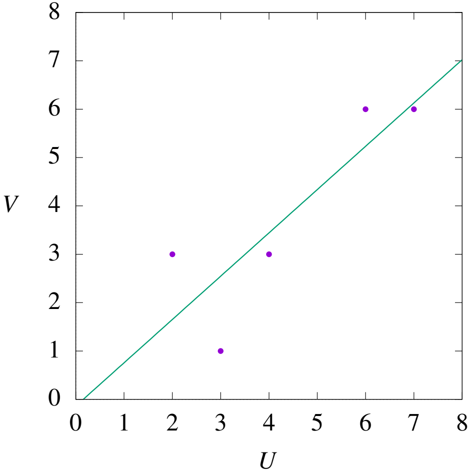
Serial Correlation & Auto Correlation

Let's Regress a Bit

This is the **least squares linear regression line** for the (u_i, v_i) data plotted.

$$V = mU + b \qquad m = r\frac{s_v}{s_u} \qquad b = \bar{v} - m\bar{u}$$

Where r is Pearson's correlation coefficient and s_u, s_v are the sample standard deviations of the u_i's and the v_i's.

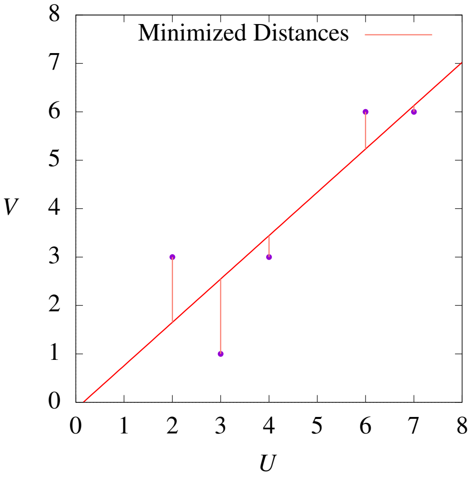


How are these equations derived? Specifically, what **assumptions** are made about the bivariate (u_i, v_i) data pairs?

Least Squares Regression

The assumption behind least squares regression is that there is little or no measurement error of the independent variable (u_i), but **there is** measurement error or experimental "noise" in the dependent variable (v_i).

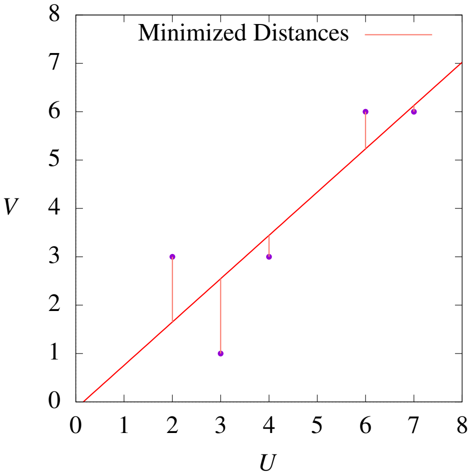
Least squares regression equations determine the "best fit line" that minimizes the vertical distances from data points to line.



Least Squares Regression

The assumption behind least squares regression is that there is little or no measurement error of the independent variable (u_i), but **there is** measurement error or experimental "noise" in the dependent variable (v_i).

Least squares regression equations determine the "best fit line" that minimizes the vertical distances from data points to line.



But in the same spirit that we discount the notion of **outliers** in simulation results, we must recognize these assumptions are inappropriate for simulation results.

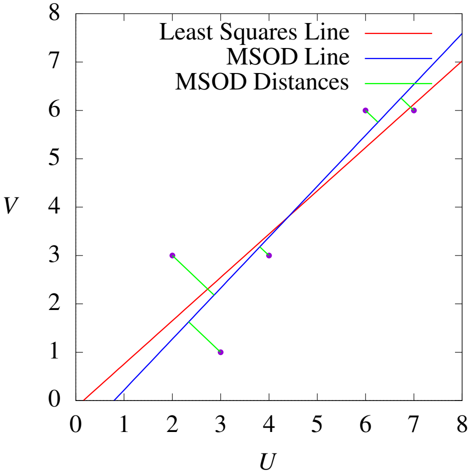
Simulations don't have "measurement error" in their outputs.

Orthogonal Least Squares Regression (MSOD)

Instead we want to minimize the distance to the best fit line across the full 2d plane, in both u_i and v_i directions.

It is still a **M**inimization process,
it still minimizes the sum of **S**quared **D**istances between points and line,
but now these "distance" line segments intersect the line at an **O**rthogonal 90°.

Hence **MSOD** regression or "best fit" lines.



Orthogonal Least Squares

Given the sample averages $\bar{u}$, $\bar{v}$, $Cov(u,v)$ and

$$\theta = \frac{1}{2}\tan^{-1}\left(s_u^2 - s_v^2, 2Cov(u,v)\right)$$

for the bivariate sample (u_i, v_i)

The **orthogonal least squares (MSOD) regression line** is

$$V = (\tan\theta)U + (\bar{v} - \bar{u}\tan\theta)$$

Where by convention $-\pi < \tan^{-1}(x,y) \leq \pi$ and $-\frac{\pi}{2} < \theta \leq \frac{\pi}{2}$ (**hint: use `atan2`**).

And what, pray tell, is $Cov(u,v)$?

The Covariance of a Bivariate Set (u_i, v_i)

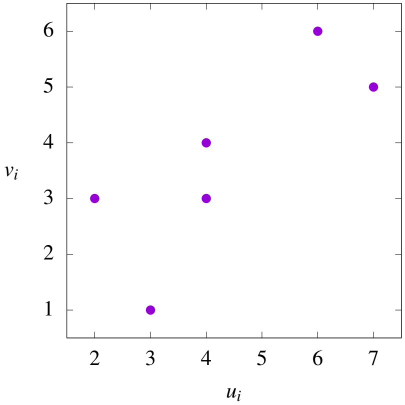
The conventional equation for *Covariance*

$$Cov(u_i, v_i) = \frac{1}{n}\Sigma(u_i - \bar{u})(v_i - \bar{v})$$

What does the **covariance** tell us about a data set?

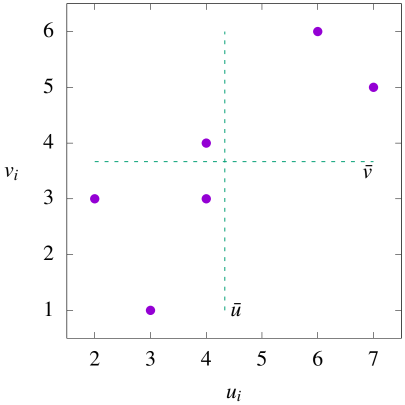
How does it "work?"

i	u_i	v_i
1	2	3
2	3	1
3	4	3
4	4	4
5	6	6
6	7	5



The Covariance of a Bivariate Set (u_i, v_i)

i	u_i	v_i	$(u_i - \bar{u})$	$(v_i - \bar{v})$
1	2	3	$-$	$-$
2	3	1	$-$	$-$
3	4	3	$-$	$-$
4	4	4	$-$	$+$
5	6	6	$+$	$+$
6	7	5	$+$	$+$



A large $+$ covariance $\rightarrow$ most of the pairs in mean-relative quadrants I & III

A large $-$ covariance $\rightarrow$ most of the pairs in mean-relative quadrants II & IV

Welford's Equation for $Cov(u_i, v_i)$

The conventional equation for *Covariance*

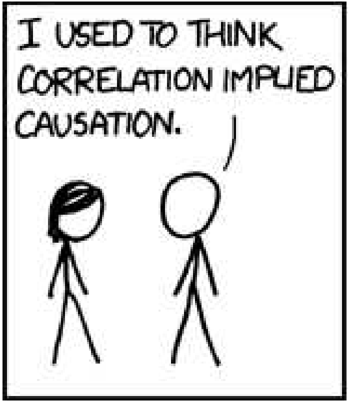
$$Cov(u_i, v_i) = \frac{1}{n}\Sigma(u_i - \bar{u})(v_i - \bar{v})$$

But we (of course) want to use a **Welford** styled iterative approach: w_i is $i \cdot Cov(u_i, v_i)$ for the first i pairs of data points.

$$w_i = w_{i-1} + \left(\frac{i-1}{i}\right)(u_i - \bar{u}_{i-1})(v_i - \bar{v}_{i-1})$$

here $\bar{u}_i$ and $\bar{v}_i$ are the (Welford maintained) averages of the first i data points.

Correlation



r, Pearson's Correlation Coefficient

Linear correlation is covariance "normalized" by the spread in the data — a more universal measure:

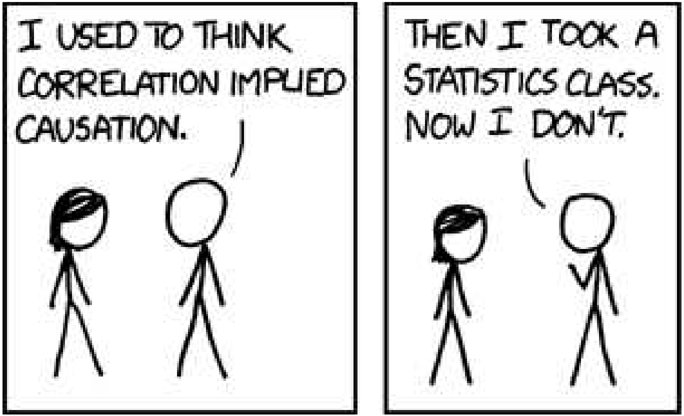
$$Cov(u_i, v_i) = \frac{1}{n}\Sigma(u_i - \bar{u})(v_i - \bar{v}) \qquad r = \frac{Cov(u_i, v_i)}{s_u s_v}$$

Welford equations (r_i is the value of r after i data points):

$$w_i = w_{i-1} + \left(\frac{i-1}{i}\right)(u_i - \bar{u}_{i-1})(v_i - \bar{v}_{i-1}) \qquad r_i = \frac{w_i}{i \cdot s_{u_i} \cdot s_{v_i}}$$

r, **the linear correlation coefficent**, measures the **linearly** predictive nature of some variable set (u_i) to its pairwise "dependent" set (v_i).

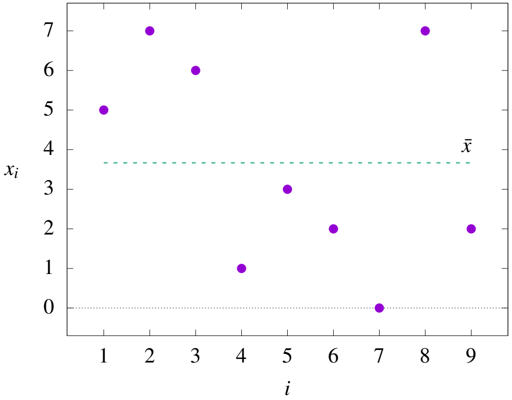
Serial Correlation & Auto Correlation



Serial Correlation

Serial correlation (aka "autocorrelation") uses the tools of bivariate (u_i, v_i) data sets on a **lagged** version of one data set (x_i).

i	x_i
1	5
2	7
3	6
4	1
5	3
6	2
7	0
8	7
9	2



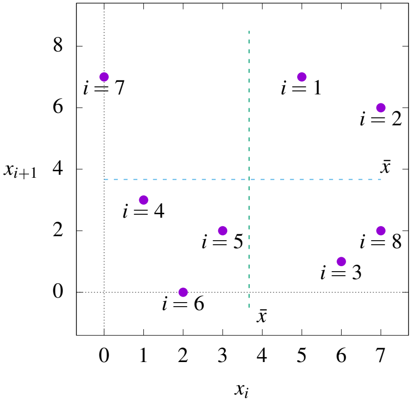
Serial Correlation

Serial correlation (aka "autocorrelation") uses the tools of bivariate (u_i, v_i) data sets on a **lagged** version of one data set (x_i).

For a **lag** of j, we think of x_i as the independent data set, and the x_{i+j}s as their dependent data set, forming bivariate data points (x_i, x_{i+j}).

For a lag of $j = 1\ldots$

i	x_i	(x_i, x_{i+1})
1	5	$(5,7)$
2	7	$(7,6)$
3	6	$(6,1)$
4	1	$(1,3)$
5	3	$(3,2)$
6	2	$(2,0)$
7	0	$(0,7)$
8	7	$(7,2)$
9	2	



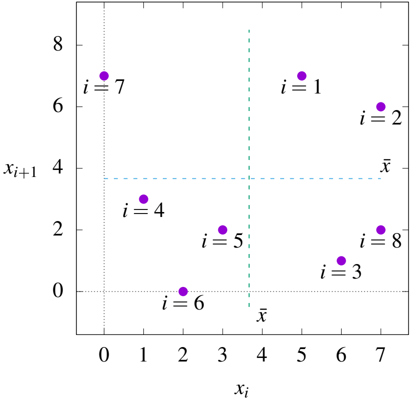
Serial Correlation

Like covariance and correlation, most of the pairs in mean-relative quadrants I & III suggests **postive serial correlation**.

Due to the nature of this formulation and the implicit time related ordering of the x_i, we **only** consider sequenced data for this type of analysis.

For example, we never apply serial correlation analysis to Monte Carlo estimates or aggregate statistics of many individual simulations.

We apply serial correlation analysis on metrics generated **from within our simulation** as the simulation **progresses in (simulated) time**.

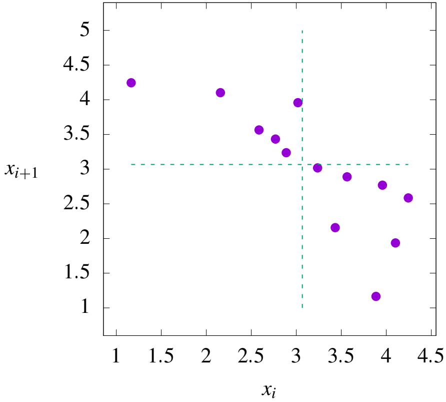


Negative Serial Correlation?

Discuss: what would the (i, x_i) plot look like for a data set with **negative serial correlation** (assume a lag of $j = 1$, so correlation among (x_i, x_{i+1}))?

Negative Serial Correlation?

Discuss: what would the (i, x_i) plot look like for a data set with **negative serial correlation** (assume a lag of $j = 1$, so correlation among (x_i, x_{i+1}))?



Work backwards!

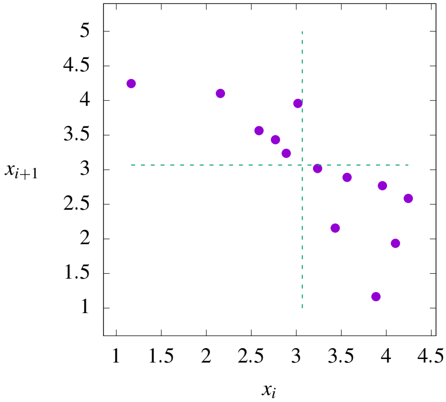
For negative serial correlation, we want the majority of points to be in *mean-relative* quadrants II & IV.

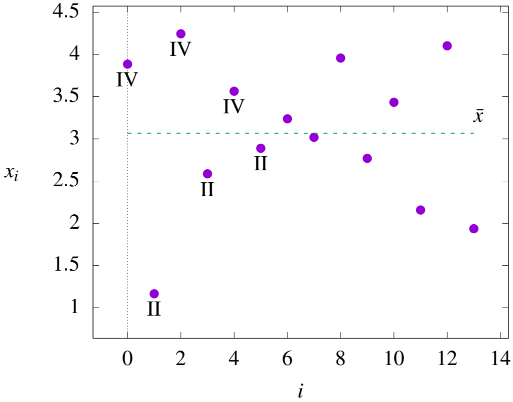
The points in II have an independent coordinate (x_i) **below** $\bar{x}$, and the dependent coordinate (x_{i+1}) **above** $\bar{x}$.

What's the relationship for points in IV?

Negative Serial Correlation?

Discuss: what would the (i, x_i) plot look like for a data set with **negative serial correlation** (assume a lag of $j = 1$, so correlation among (x_i, x_{i+1}))?





This creates a pattern of flip-flops across $\bar{x}$ as i (sample number) increases.

$\pm$ Serial Correlation — Why do we care?

Why do we need to know about **positive** or **negative serial correlation** for the art of computer simulation?

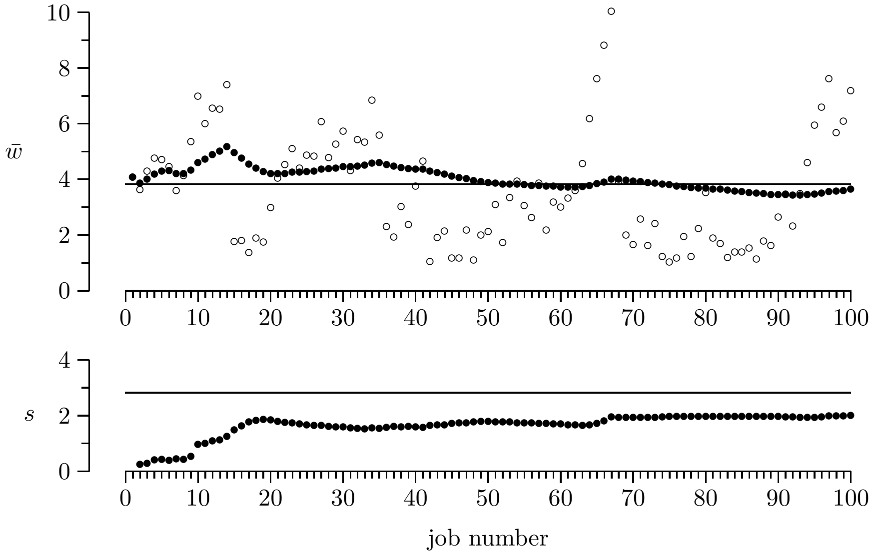
We often need to generate averages **and confidence intervals** for averages emanating from our simulation. For instance we want to determine a 95% confidence interval for the average **sojourn time** ($\bar{w}$) of jobs through an SSQ.

The confidence interval techniques learned in an intro statistics course use the *Central Limit Theorem*, the the CLT requires *iid* samples!

iid = **independent** and **identically distributed**

While the w_i of each SSQ job might be sampled from the same probability distribution, the samples are clearly not **independent**!

Example 4.1.7: Serial Correlation



The authors show that a **positive serial correlation** creates an "underestimated" (bias) s compared to the true theoretical σ for the system

$\rightarrow$ **confidence intervals will be too small**

As it turns out **negative serial correlation** creates an "overestimated" s (again compared to the true σ of the system)

$\rightarrow$ **confidence intervals will be too big**

Bias in either direction can exist — the critical point is that we no longer have independent x_i, so our simple statistical conclusions (eg: confidence intervals) may not work out so well.

Important: s is the correct value for each data set![1] It is when we go from s to CLT **confidence intervals** that the flaw creeps in. With CLT we are assuming *iid* data points! s does not require *iid* (it's just an equation), the **standard CLT construction of confidence intervals does.**

[1] Which is why I used "quotes" in the first two statements of this slide!
There are ways around this CI generating limitations, see Appendix F for all the intricate details…Not a course requirement, I'm just saying it's there.

Correlation Exhaustion

Serial or Auto-correlation Conventional pairwise analysis but with one data set $(u_i = x_i, v_i = x_{i+j})$, a "$j$-lagged" pairing to itself.
Does the i-th value "predict" the j-th subsequent value?

Positive Serial Correlation When clusters (plural!) of data points fall above or below $\bar{x}$.
The canonical example in computer simulation is sojourn times of jobs through an SSQ with traffic intensity ≈ 1.

Negative Serial Correlation When the components of the j-lagged pairings (x_i, x_{i+j}) consistently lie on either side of $\bar{x}$. In the case of $j = 1$, the x_i data points consistently fall on **alternating** sides of $\bar{x}$.

For all of these j usually small.

Approximate One-Pass Autocovariance

Given that the two-pass equation for the sample autocovariance of x with lag j is:

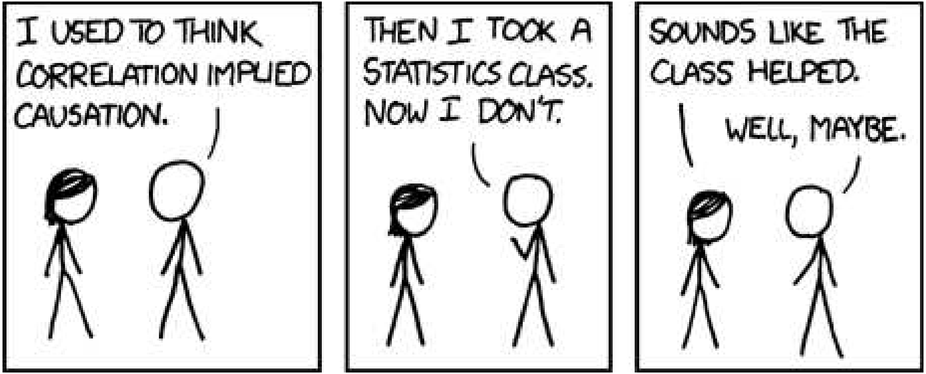
$$c_j = \frac{1}{n-j} \sum_{i=1}^{n-j} (x_i - \bar{x})(x_{i+j} - \bar{x})$$

where $\bar{x}$ is the sample mean of all x_i. The natural one-pass analogue is

$$\hat{c}_j = \left[\frac{1}{n-j} \sum_{i=1}^{n-j} x_i \, x_{i+j} \right] - \bar{x}^2$$

Notes:

1. These are not algebraically equivilant,
2. Better to use the Welford equations for $w_i = i \cdot Cov(u, v)$ (which is **also** not algebraically equivilant but at least numerically stable).


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

fini