Getting Started II: *Sample Stats & Optimization*

- Sample Statistics: data characteristics; measures of …
  - Central tendency: Sample mean ($\bar{x}$)
  - Variability/dispersion: Sample variance (S_{xx}); Sample standard deviation (S_x)
  - Association/relationship: Sample covariance (S_{xy}); Sample correlation (r_{xy} or ρ_{xy})
- Standardization: $z_i = (x_i - \bar{x})/S_x$ ($\bar{z} = 0$; $S_{zz}=S_z=1$)
- Example: *Anscombe's Quartet*
- Optimization: *OLS = min SSRs*
  - *FOCs*: *First Order Conditions* - identify solution candidates
  - *SOCs*: *Second Order Conditions* - evaluate candidates (identify min's and max's)
- Example: min SSRs (Sum Squared Residuals)

Sample Statistics: Sample Mean

- You have a dataset consisting of n observations of two variables (x, y): $\{(x_i, y_i)\}\ i = 1, 2, \ldots n$.
- The ***sample mean*** (*average*):
  - $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$. Note that $\sum x_i = n\bar{x}$.

- Deviations from means:
  - $dx_i = (x_i - \bar{x})$ and $dy_i = (y_i - \bar{y})$
- By construction, the total/sum of the deviations from the means for any variable will be zero
  - $\sum dx_i = \sum (x_i - \bar{x}) = \left(\sum x_i\right) - n\bar{x} = 0$ and $\sum dy_i = \sum (y_i - \bar{y}) = 0$.

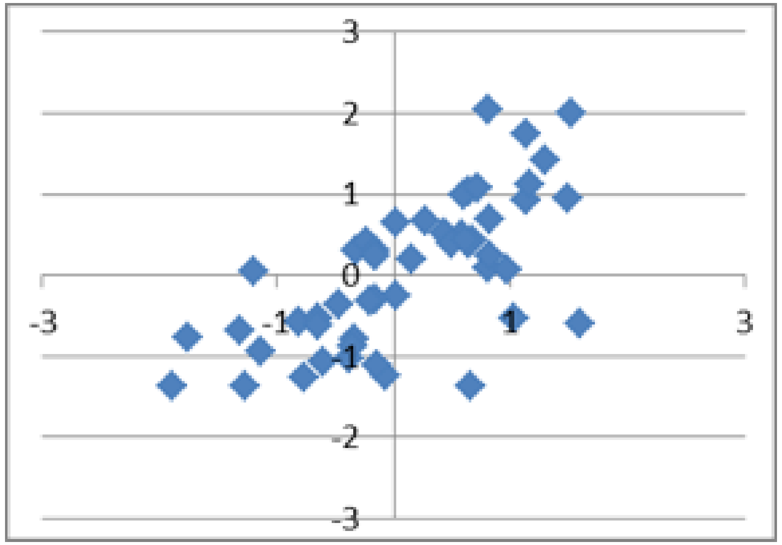
… Sample Variance, Standard Deviation & Covariance

- The ***sample variance***:
  - $S_{xx} = S_x^2 = \frac{1}{n-1}\sum (dx_i)^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$ and likewise for the y's.
  - Since $\sum x_i = n\overline{x}$, $S_{xx} = \frac{1}{n-1}\sum x_i^2 - \frac{n}{n-1}\overline{x}^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\overline{x}^2\right)$.
- The ***sample standard deviation***:
  - $S_x = \sqrt{S_{xx}} = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum (dx_i)^2} = \sqrt{\frac{1}{n-1}\sum (x_i - \overline{x})^2}$, and likewise for the y's.
- The ***sample covariance***:
  - $\mathrm{cov}\left(x, y\right) = S_{xy} = \frac{1}{n-1}\sum (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}\sum x_i y_i - \frac{n}{n-1}\overline{xy}$

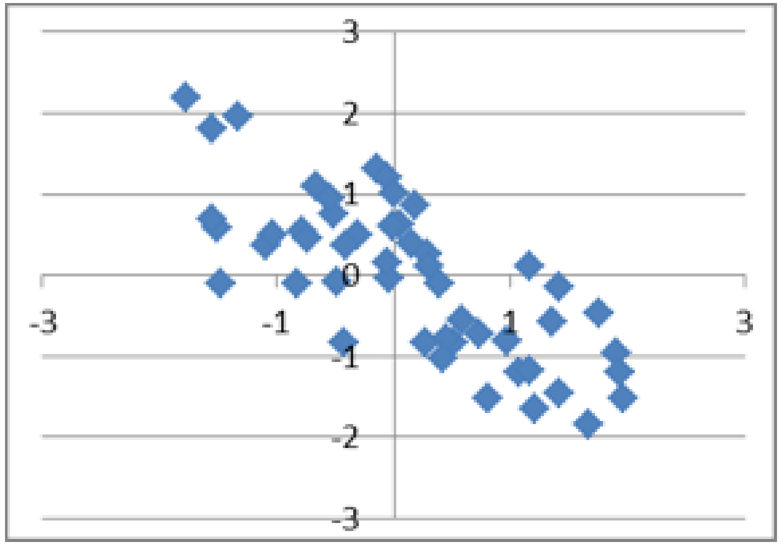
  (since $\sum x_i = n\overline{x}$ and $\sum y_i = n\overline{y}$.)

Sample Covariance: Some Intuition

- In the following examples, $\bar{x} = 0$ and $\bar{y} = 0$
- On the left, most of the data are in quadrants I and III, where $(x_i - \bar{x})(y_i - \bar{y}) > 0$; when you sum those positive products (and divide by n-1) you get a positive sample covariance.
- Most of the action on the right is in quadrants II and IV where $(x_i - \bar{x})(y_i - \bar{y}) < 0$; those products sum to a negative number, and we have a negative covariance.



Positive Covariance



Negative Covariance

Sample Covariance: A Few Properties

- The sample variance of x is the sample covariance of x with itself

$$\operatorname{cov}(x,x) = \frac{1}{n-1}\sum(x_i-\overline{x})(x_i-\overline{x}) = \frac{1}{n-1}\sum(x_i-\overline{x})^2 = S_{xx}$$

- The covariance of a sum is the sum of the variances plus twice the covariance:

$$\operatorname{var}(x+y) = S_{xx} + 2S_{xy} + S_{yy}$$

- If $S_{xy} = 0$, then $\operatorname{var}(x+y) = S_{xx} + S_{yy} = \operatorname{var}(x) + \operatorname{var}(y)$

- The covariance of linear transformations of the x's and y's:

$$\operatorname{cov}(a+bx, c+dy) = bdS_{xy} = bd\operatorname{cov}(x,y)$$

- The covariance of x with sums of variables:

$$\operatorname{cov}(x, y+z) = \operatorname{cov}(x,y) + \operatorname{cov}(x,z)$$

… the sum of the covariances of x with each other variable.

- We can drop either $\overline{x}$ or $\overline{y}$ (but not both!) from the equation for the sample covariance.

$$S_{xy} = \frac{1}{n-1}\sum(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{n-1}\sum x_i(y_i-\overline{y})$$

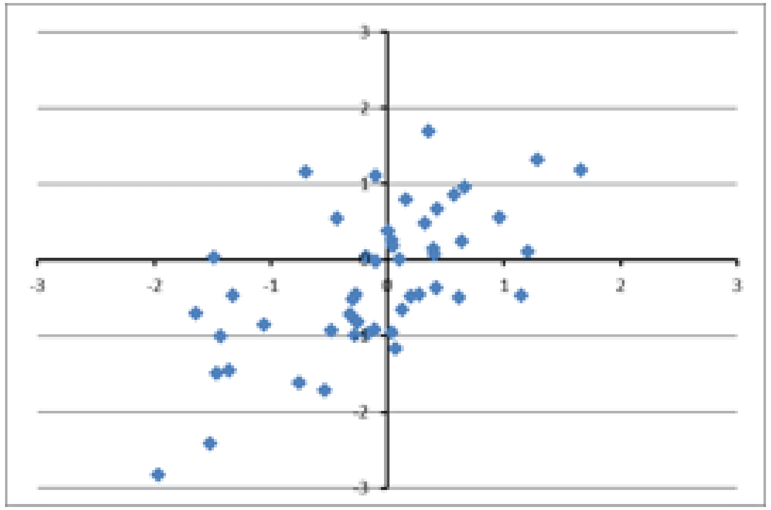
$$= \frac{1}{n-1}\sum(x_i-\overline{x})y_i$$

Sample Correlation

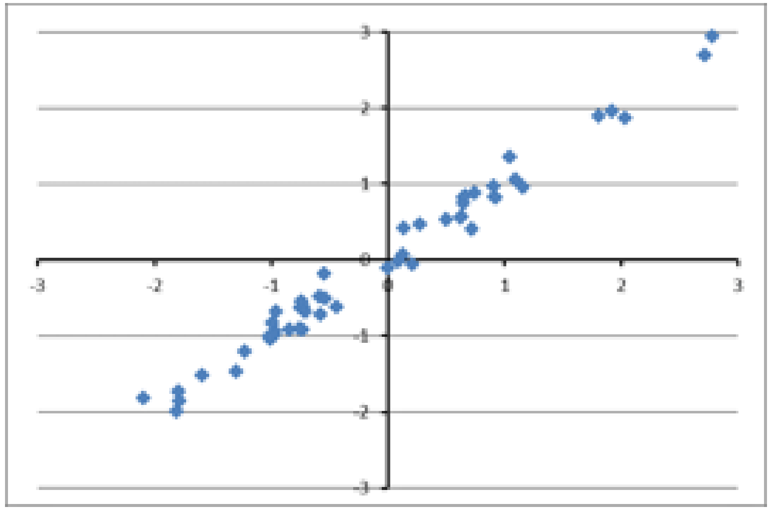
- The ***sample correlation*:** $\rho_{xy} = \dfrac{S_{xy}}{S_x S_y}$
  - The ratio of the sample covariance to the product of the sample standard deviations (some use r_{xy} for this sample statistic).
  - By construction, $\left|\rho_{xy}\right| \leq 1, \; or \; -1 \leq \rho_{xy} \leq 1$ (this is the Cauchy–Schwarz inequality).
- If $S_{xy} = 0$, the sample covariance is 0 and the sample correlation is also 0 … and if the sample covariance is negative (positive), then so is the sample correlation.
- If $\left|\rho_{xy}\right|$ is close to 1 then the relationship between x and y will look quite linear (with a positive slope if $\rho_{xy} \sim 1$, and a negative slope if $\rho_{xy} \sim -1$.

Correlation captures the extent to which x and y are moving together *in a linear fashion*.

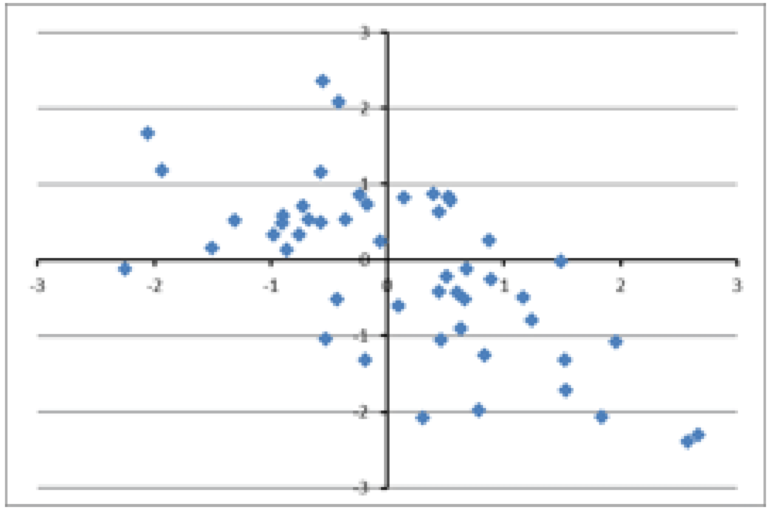
Sample Correlation: Examples



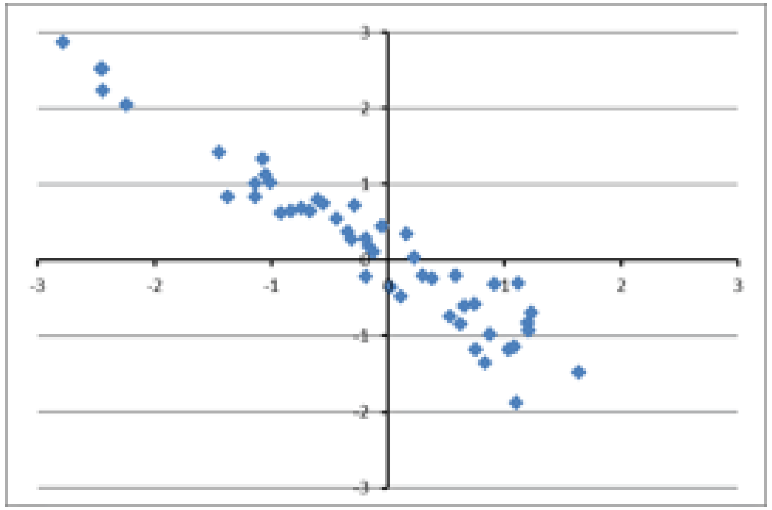
$\rho_{xy} = .635$



$\rho_{xy} = .991$



$\rho_{xy} = -.652$



$\rho_{xy} = -.959$

Standardize/Normalize Data

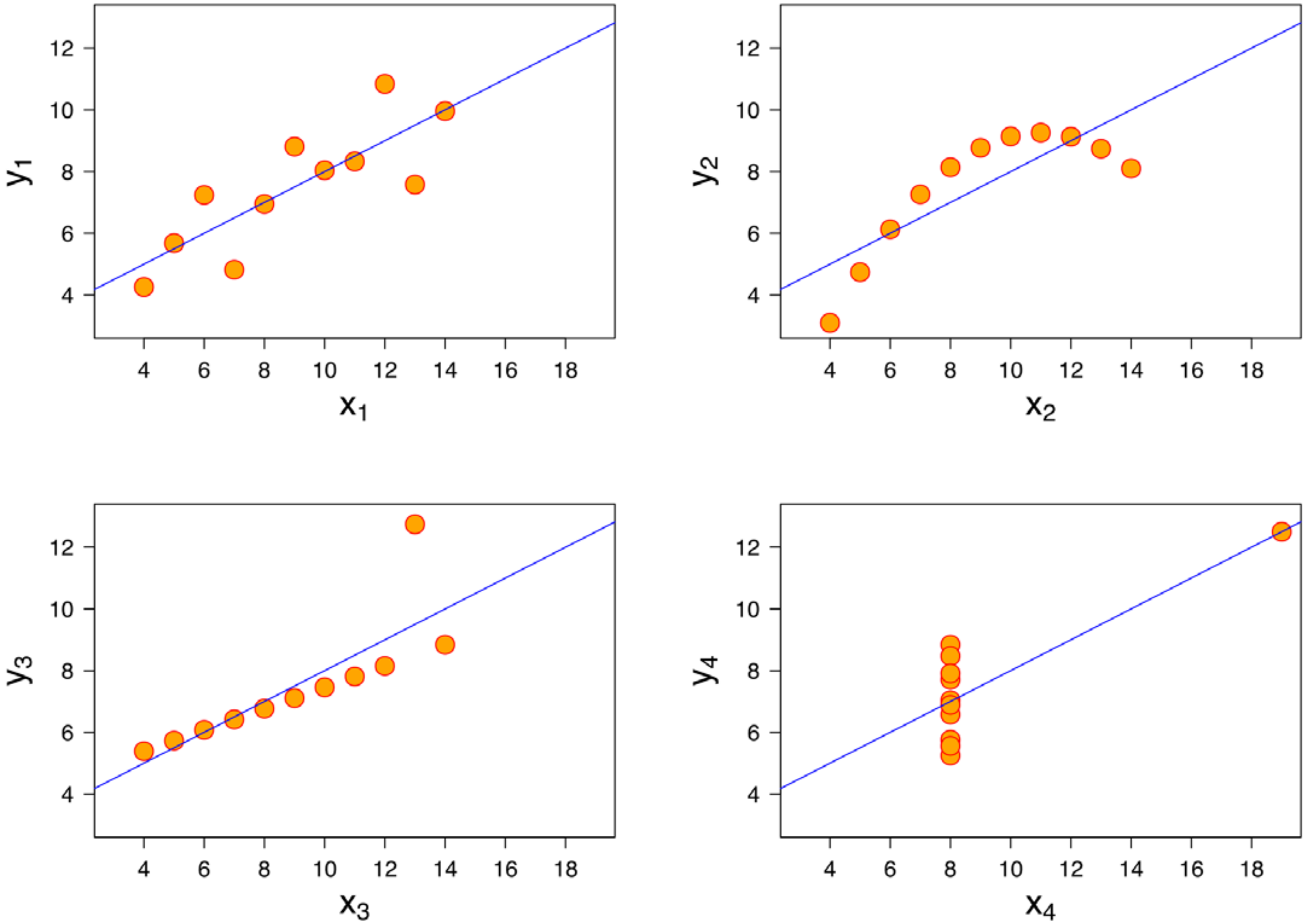
$$z_i = \frac{x_i - \bar{x}}{S_x}$$

- To *standardize*, or *normalize*, a variable:
  - first subtract the variable's mean from each observation,
  - then divide each new value by the variable's standard deviation.
- ***Means and variances***: The result is a transformed variable, z, with mean 0 and variance 1:
  - Sample Mean: $\bar{z} = 0$; Sample Variance (Std Dev): $S_{zz} = S_z = 1$
- ***Covariances and correlations***: * indicates standardized, so: $x_i^* = \dfrac{x_i - \bar{x}}{S_x}$ and $y_i^* = \dfrac{y_i - \bar{y}}{S_y}$.
  - Sample Covariance: $S_{x^*y^*} = \rho_{xy}$; Sample Correlations: $\rho_{x^*y^*} = S_{x^*y^*} = \dfrac{S_{xy}}{S_x S_y} = \rho_{xy}$.

Standardization will typically affect sample means, variances and covariances of variables… but it does not impact sample correlations.

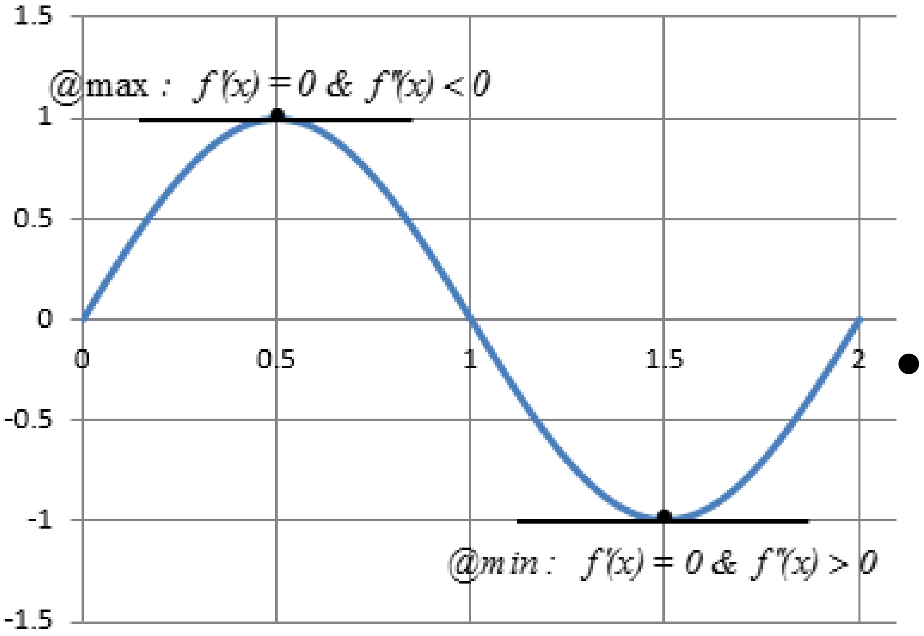
Anscombe's Quartet: Sample stats tell you a lot… but…

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : σ^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : σ^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places



Optimization: FOCs and SOCs

OLS ≡ min SSRs



- *Ordinary Least Squares (OLS)*: min SSRs to estimate unknown parameter values.
  - Which coefficients minimize the sum of the squared differences between predicted and actual values?
  - We call these differences between predicted and actual values *residuals*… and the sum of the squared residuals *SSR*s, for, well, *Sum of Squared Residuals*.
- To solve optimization problems (assume differentiability):
  - ***First Order Conditions (FOCs)***: Identify solution candidates
  - ***Second Order Conditions (SOCs)*:** Which candidates minimize SSRs?

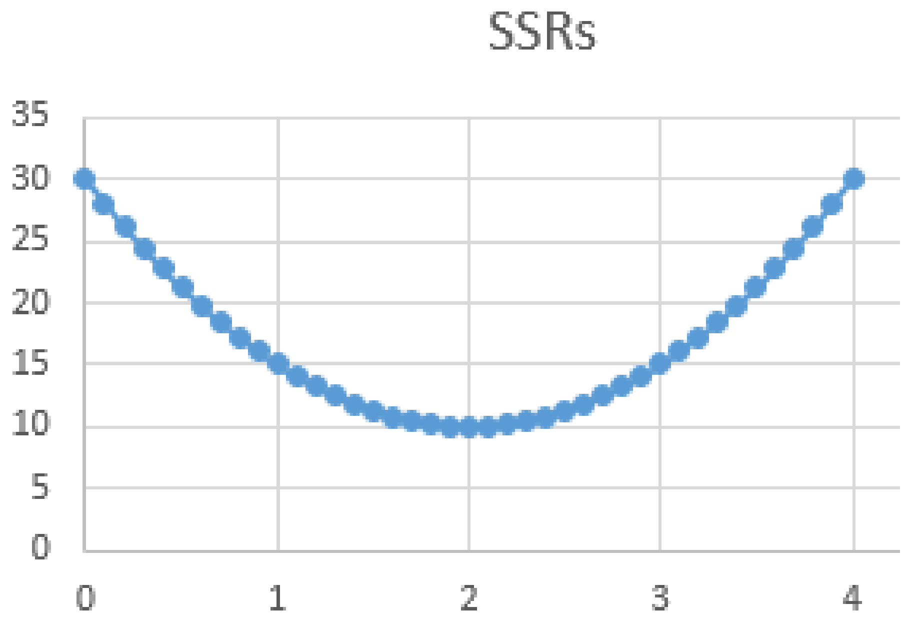
An Example: Min SSRs (to estimate unknown mean)

- Your sample data:: $\{y_i\}$ $i = 1, 2, \ldots n$.
- Estimate the mean μ : find the number m* that is *closest* to the observed sample
- m* minimizes Sum Squared Residuals (SSR): $SSR = \sum (y_i - m)^2$
- FOC: $\dfrac{dSSR}{dm} = \sum 2(y_i - m)(-1) = 0$; $\sum y_i = \sum m^* = nm^*$;

  $m^* = \dfrac{1}{n} \sum y_i = \bar{y}$.

- SOC: $\dfrac{d^2 SSR}{dm^2} = \sum 2(-1)(-1) = 2n > 0$

  So we have a minimum @ $m^* = \bar{y}$.



Onwards… to *SLR Analytics*