uncertainty and statistics from Hartmann notes…
some food for thought ☺

Uncertainty: What are statistics good for?

Statistics test whether something could have occurred by chance, subject to some assumptions and prior assumptions. Statistics test only one kind of uncertainty, of which we can define three.

Aleatory Uncertainty: This is random uncertainty that we can measure with statistics.

Epistemic Uncertainty: This is uncertainty associated with lack of knowledge about things we could in principle know about. We know the physics, but are uncertain about the parameters.

Structural Uncertainty: This is uncertainty arising from things we don't know about. Unknown unknowns. We thought the world was flat, but it is really spherical.

	Tuesday	Thursday
January	**January 17** Complete pre-class survey, Set up python environment, Read syllabus.	**January 19** 1. Introductions/Basic statistics/Bayes Theorem (Barnes 1.1-1.2)
	January 24 2. Statistical Significance Testing /Hypothesis testing/Resampling/Monte Carlo (Barnes 1.3-1.5) HW#1 assigned	**January 26** Application LAB #1 Basic Statistics and Hypothesis testing
	January 31 Applications LAB #1 cont.	**February 2** 3. Compositing/Other distributions/Non-parametric tests (Barnes 1.6-1.8)
February	**February 7** 4. Regression (Barnes 2.1-2.2) HW#1 due HW#2 assigned	**February 9** 5. Autocorrelation/Autoregressive model/Sample Size/Multiple Regression (Barnes 2.3-2.4)
	February 14 Applications LAB #2 Regression/AR1	**February 16** Applications LAB #2 cont.
	February 21 6. EOFs via Eigenanalysis/SVD (Barnes 3.1.1-3.1.4) HW#2 due, HW#3 assigned	**February 23** 7. EOFs with actual data (Barnes 3.1.5)
	February 28 8. More EOFs with Dr. Nicola Maher	**March 2** Applications LAB #3 – Matrix methods/EOFs
March	**March 7** Applications LAB #3 cont.	**March 9** 9. Harmonic analysis; power spectra (Barnes 4.1.1-4.1.2) HW#3 due; HW#4 assigned
	March 14 10. Fourier Transforms/Significance testing of spectral peaks/Data windows (Barnes 4.1.3-4.1.5)	**March 16** Applications LAB #4 – Timeseries analysis/Power spectra
	March 21 Applications LAB #4 cont.	**March 23** PRESENTATIONS on Homework #2, #3 HW#4 due
	SPRING BREAK – NO CLASS	
April/May	**April 4** 11. Filtering (Barnes 4.1.6; Hartmann 7), HW#5 assigned	**April 6** 12. Finish Filtering (Barnes 4.1.6; Hartmann 7)
	April 11 Applications LAB #5 – Timeseries analysis/Filtering	**April 13** Applications LAB #5 cont. HW#5 due, HW#6 assigned
	April 18 13. Machine Learning Overview	**April 20** 14. Machine Learning – SOMs, clustering
	April 25 Applications LAB #6: Machine Learning	**April 27** Applications LAB #6 cont.
	May 2 15. More machine learning with Dr. Nicola Bodini Homework #6 due	**May 4** PRESENTATIONS: Homework #4, #5, #6

Spring 2023: ATOC5860 Objective Data Analysis

Classes in yellow are entirely "learning by doing" application labs in small groups

Last updated: January 13, 2022

Today's plan…. Barnes Chapter 2.1.1 and 2.1.2

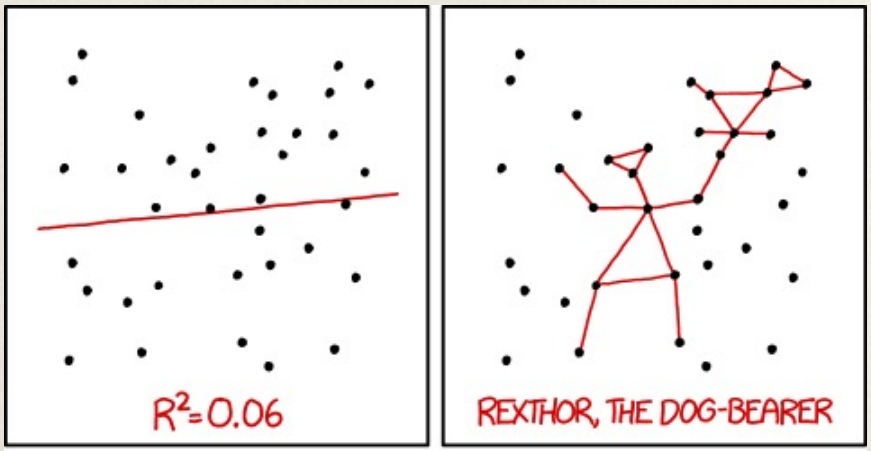
1. Least Squares Linear Regression, linear of X and Y. Goal: minimize cost function to find the regression coefficient (a1) and y-intercept (a0) and predict Y using X.
2. How good is the fit? Find the Pearson correlation coefficient (r) and estimating the variance explained (r^2)
3. Statistical significance testing of correlations including Fisher-Z transformation
4. Spearman's rank correlation (tests whether a set of paired data monotonically vary together)

Linear Regression: What is it?

In linear regression, the goal is to determine

- the linear fit of X and Y (the regression coefficient)
- the robustness of the fit (the correlation coefficient)

Regression is simple but powerful, however, this also makes it easily misused. In a sense, the entire class (EOFs; Fourier analysis) is all based on regression.



https://imgs.xkcd.com/comics/linear_regression.png

Linear Regression: Finding the "best-fit" to the observed data

How do we find the slope and y-intercept of the line that best fits the observed data? Let's assume $x(t)$ and $y(t)$ are time series sampled at N time steps (so that each point represents a time step).

First, we have to define what "best-fit" means. For now, we will use the conventional definition which means that we want to reduce the sum of the squared errors of y.

Using the method of least squares:

$$\hat{y}(t) = a_1 x(t) + a_0 \tag{1}$$

where

- $\hat{y(t)}$ denotes the estimate of $y(t)$ based on the linear relationship with $x(t)$
- a_1 denotes the slope, a.k.a. the regression coefficient
- a_0 denotes the y-intercept

Linear Regression: What is the error of the fit?

Define the error of the fit as the sum of squares of the $y(\mathtt{estimate}) - y(\mathtt{actual})$:

$$Q = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{N} (a_1 x_i + a_0 - y_i)^2 \tag{2}$$

where the subscript i denotes the time step.

The error is squared so that

- the error is positive definite (don't want positive and negative errors canceling out)
- the minimization of Q (the derivative of Q) is a linear problem

Note: the square causes larger errors to be more heavily weighted

Linear Regression: Minimize error = Minimize Q (cost function)

We now follow steps from our college Calculus I course and find the a_1 and a_0 that minimize Q (sometimes called the cost function):

$$\frac{dQ}{da_0} = 0 \tag{3}$$

$$0 = 2\sum_{i=1}^{N}(a_1 x_i + a_0 - y_i) \tag{4}$$

$$0 = a_1\sum_{i=1}^{N} x_i + a_0 N - \sum_{i=1}^{N} y_i \tag{5}$$

$$\frac{dQ}{da_1} = 0 \tag{6}$$

$$0 = 2\sum_{i=1}^{N}(a_1 x + a_0 - y_i)x_i \tag{7}$$

$$0 = a_1\sum_{i=1}^{N} x_i^2 + a_0\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i y_i \tag{8}$$

Divide through by N and move the y terms to the left hand side, where overbars denote the mean and primes denote departures from the mean:

$$\overline{y} = a_1\overline{x} + a_0 \tag{9}$$

$$\overline{xy} = a_1\overline{x^2} + a_0\overline{x} \tag{10}$$

Two equations, two unknowns.

Linear Regression: Solutions that Minimize Q (cost function)

The solutions are:

$$a_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \tag{11}$$

Note that,

$$\overline{xy} = \overline{x} \cdot \overline{y} + \overline{x'y'} \tag{12}$$

and

$$\overline{x^2} = \overline{x}^2 + \overline{x'^2} \tag{13}$$

Hence,

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{14}$$

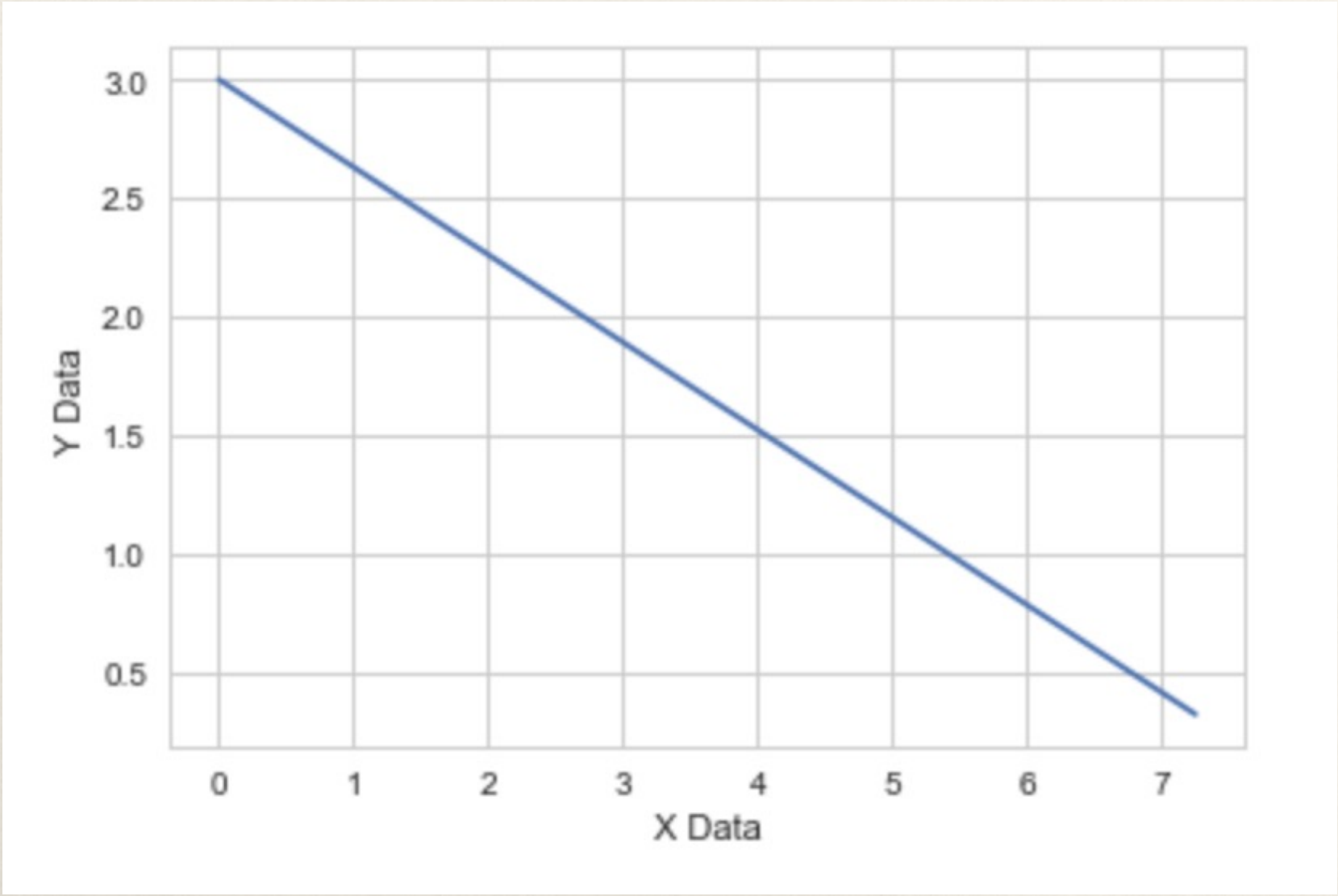
$$a_0 = \overline{y} - a_1\overline{x} \tag{15}$$

In other words:

1) The slope (a_1) is the covariance(X,Y) divided by the variance(X).

2) The y-intercept (a_0) is the average(Y) minus a_1 multiplied by the average(X). (a_0=0 if Y and X are anomalies)

Regression: Example in a perfect world.



Try the python code – linear_regression.ipynb

Linear Regression:
How do I place confidence intervals on the slope (a_1)?

One can put confidence limits on the slope a_1 in a manner of ways. For example, a jackknife approach can be used to determine the sensitivity to removing a single point. Alternatively, the standard error σ_{a_1} (standard deviation) of the slope is given by

$$\sigma_{a_1}^2 = \frac{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{16}$$

where we assume that x is known exactly. Looking at this equation, you can see intuitively that it is somewhat like the error in our y estimate divided by the variance in our x values - sort of like a slope of errors or variances.

Then, the confidence bounds on the slope follow the normal distribution, and the confidence interval for the true slope b is given by

$$a_1 - t_\alpha^{N-2} \cdot \sigma_{a_1} < b < a_1 + t_\alpha^{N-2} \cdot \sigma_{a_1} \tag{17}$$

The $N-2$ comes from the fact that two degrees of freedom were used to estimate a_1 and a_0.

Try the python code, also discussed in lecture #2 – jackknife_example.ipynb

Linear Regression: How good is the fit? Spread of the data given by the correlation coefficient (r)

How much we "believe" the regression coefficient (a_1) depends on the spread of the dots about the best fit line. If the dots are closely packed about the regression line, then the fit is good. The spread of the dots is given by the correlation coefficient r.

Here is one way to derive the correlation coefficient:

By definition, the total variance of $y(t)$ is

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{18}$$

And by definition, the total variance of the fit of $x(t)$ to $y(t)$ (i.e. the variance of $\hat{y}(t)$) is

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \hat{\overline{y}})^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2 \tag{19}$$

where we have used the fact that

$$\hat{\overline{y}} = a_1\overline{x} + a_0 = \overline{y} \tag{20}$$

Linear Regression: How good is the fit? Spread of the data given by the correlation coefficient (r) and variance explained = r^2

The percent of the total variance in $y(t)$ explained by the fit $\hat{y}(t)$ is thus given by the ratio

$$r^2 = \frac{\text{explained variance}}{\text{total variance}} \tag{21}$$

$$= \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{22}$$

(for all steps see Barnes notes)

$$= \frac{(\overline{x'y'})^2}{\overline{x'^2} \cdot \overline{y'^2}} \tag{30}$$

Hence,

$$r = \frac{\overline{x'y'}}{\sigma_x \sigma_y} \tag{31}$$

Recall that $\sigma_x = (\overline{x'^2})^{1/2}$.

Linear Regression: correlation coefficient (r) and variance explained (r^2)

- r^2 is the fraction of variance explained by the linear least-squares fit between the two variablces
- r^2 always lies between 0 and 1

The correlation coefficient r

- $r = \sqrt{r^2} = \left(\frac{\overline{x'y'}}{\sigma_x \sigma_y}\right)$
- r varies between -1 and 1

Relationships between r and r^2:

$$r = 0.99 \Rightarrow r^2 = 0.98 \quad (32)$$
$$r = 0.90 \Rightarrow r^2 = 0.81 \quad (33)$$
$$r = 0.70 \Rightarrow r^2 = 0.49 \quad (34)$$
$$r = 0.50 \Rightarrow r^2 = 0.25 \quad (35)$$
$$r = 0.25 \Rightarrow r^2 = 0.06 \quad (36)$$

Linear Regression: Relationship between slope (a_1) and correlation coefficient (r)

1.1.2 Relationship between the slope of the regression line and the correlation coefficient:

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{37}$$

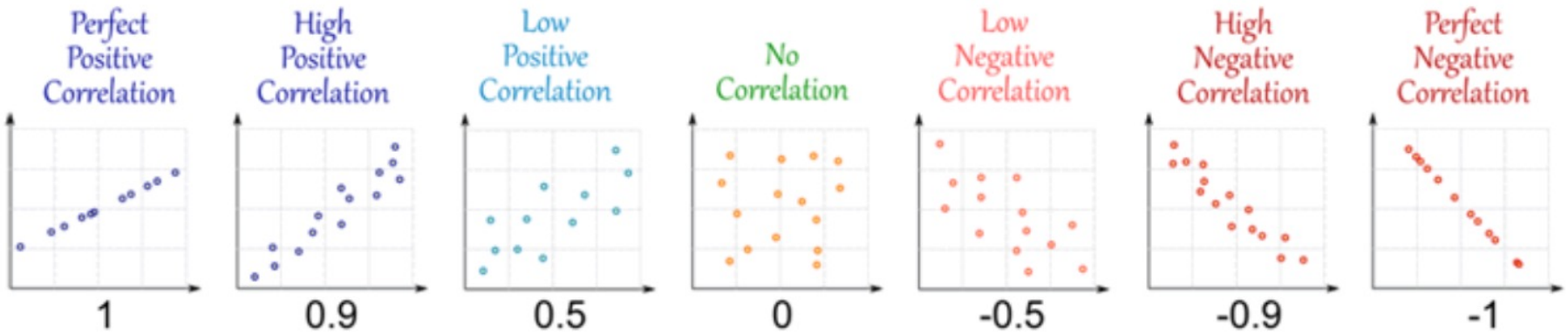
so, it follows that

$$a_1 = r\frac{\sigma_y}{\sigma_x} \tag{38}$$

- the regression coefficient (slope) of $y(t)$ on $x(t)$ can be thought of as the correlation coefficient multiplied by the ratio of the standard deviations of y and x
- regression coefficients give you information about the correlation coefficient and the relative amplitudes of variations of y and x
- in the special case where x and y are standardized, the correlation coefficient and the regression coefficient are equal

Linear Regression: General Comments

- only works for linear relationships
- does not reveal relationships that are lagged or out of phase
- need to be careful about estimating the true sample size (more on this later)
- correlation DO NOT reveal cause and effect
- flipping the x and y will not give you the same results, so very important to physically justify your choice of x and y!



Remember – Correlation here implies a linear fit, which may or may not be a good approximation.

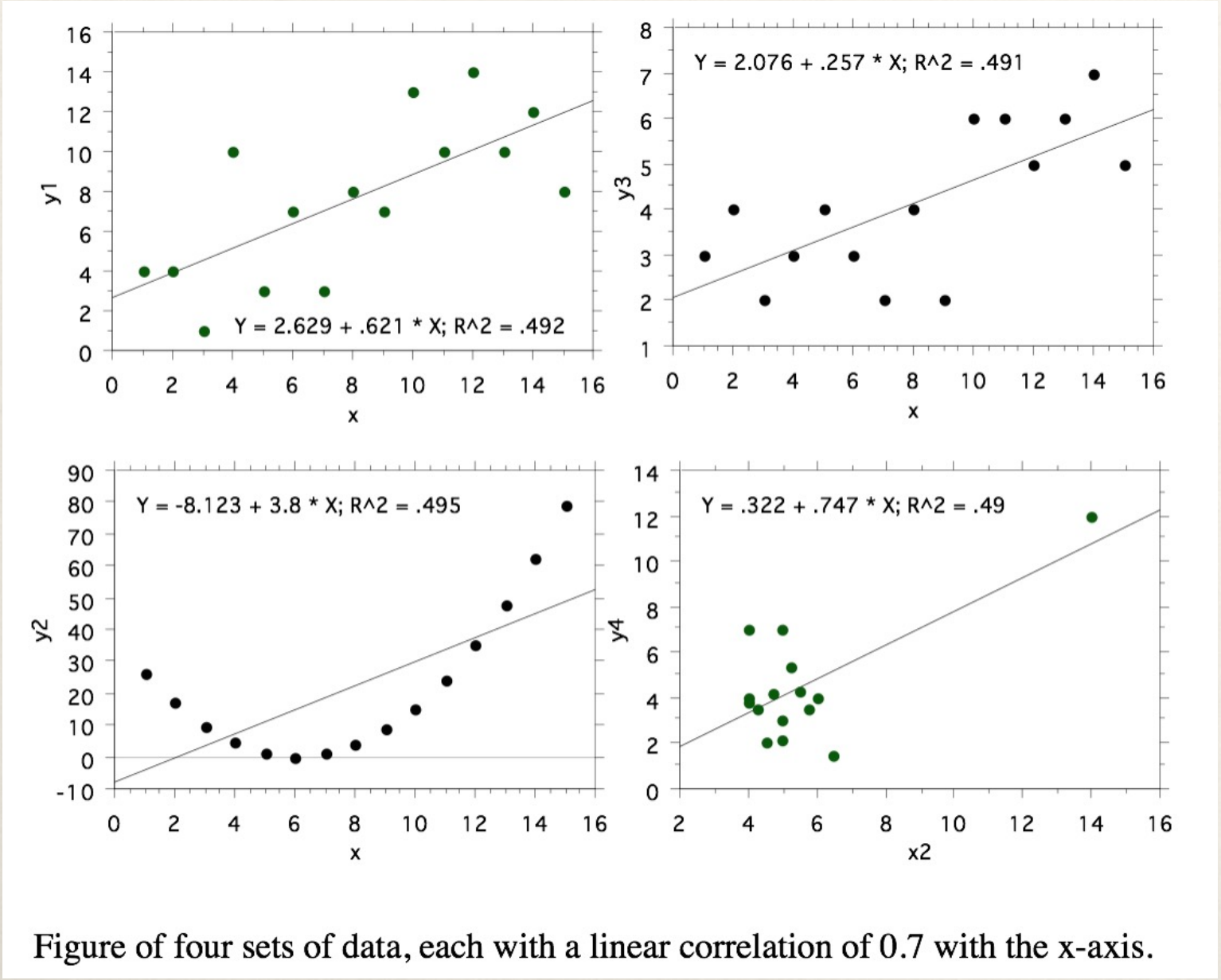
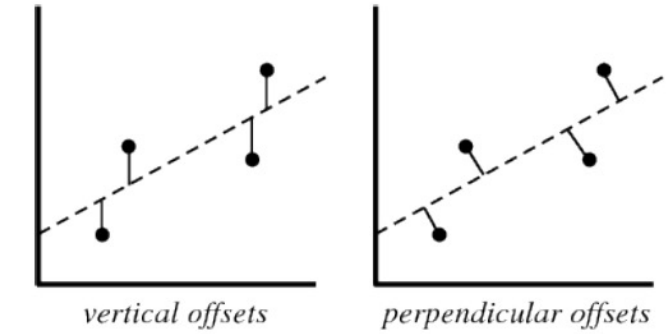


Figure of four sets of data, each with a linear correlation of 0.7 with the x-axis.

Orthogonal Least Square Regression

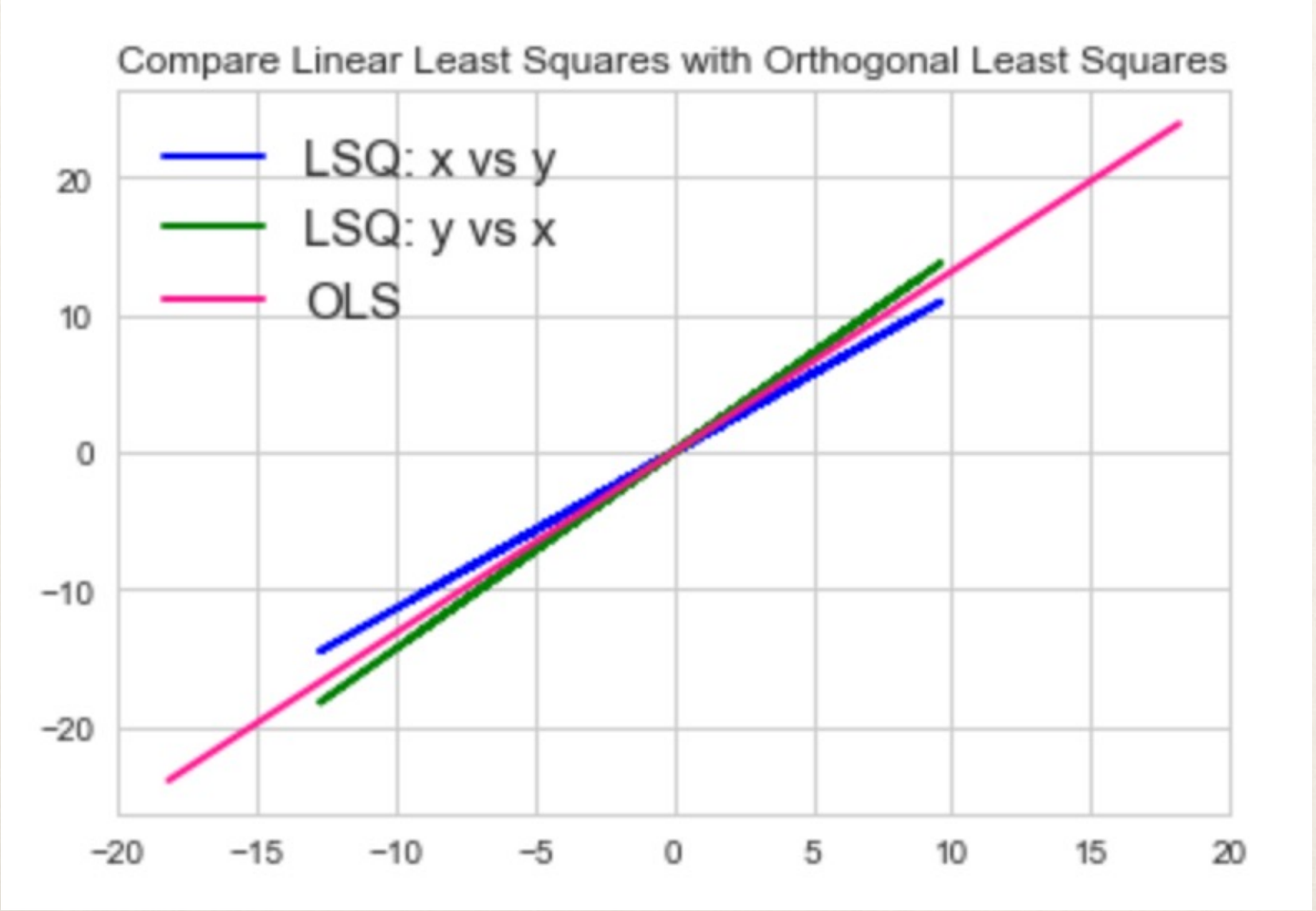
If you can't justify which of your data is dependent and which is independent, or it doesn't make sense to do so, orthogonal least squares may actually be what you want. In this case, you minimize the orthogonal distances, rather than the vertical distances:



It just so happens that in 2-dimensions, EOF analysis (to be discussed later) will give you the orthogonal least squares fits. So, in a few weeks, you will be capable of calculating this too.

Try the python code – LSQ_OLS.ipynb

Linear Least Squares (LSQ) vs. Orthogonal Least Squares (OLS)



Try the python code – LSQ_OLS.ipynb

REVIEW ON THE BOARD
LECTURE #4 – PART 1

WRITING OFTEN HELPS CONCEPTS SINK IN…

GROUP WORK

- *jackknife_example.ipynb*
- *linear_regression.ipynb*
- *LSQ_OLS.ipynb*
- *Explain "Lecture #4 on the board" notes to each other.*

Theory of correlation – Testing the statistical significance

The correlation, r, between two time series, $x(t)$ and $y(t)$ gives a measure of how well the two time series vary linearly together (or don't).

$-1 \leqslant r \leqslant 1$, with numbers closer to 1 implying that the time series vary linearly with one another.

Now, we will discuss techniques for testing the statistical significance of correlations. We will denote the sample correlation as r and the theoretical true value as ρ.

If $\rho = 0$, we can use a form of the z-statistic and t-statistic:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{44}$$

Note: The above statistic only applies if the underlying distributions are normal. Generally if N is big (>30…), the central limit theorem applies and normality can be assumed.

Example – Test the hypothesis that the true correlation (rho) is zero

2.1 Testing the hypothesis that $\rho = 0$

We have two time series, each of length 20, and they are correlated at $r = 0.6$. Does this correlation exceed the 95% confidence interval under the null hypothesis that $\rho = 0$? You can assume both time series are sampled from underlying normal distributions.

We had no prior knowledge (before getting the samples) that the correlation would be positive or negative, so we will use a two-tailed t-test.

$t_c = 2.1$ for $\nu = N - 2 = 18$, so we want to know if the sample statistic $t > 2.1$.

$$t = \frac{0.6\sqrt{20-2}}{\sqrt{1-.6^2}} = 3.18. \tag{45}$$

Since $t > 2.1$, we can reject the null hypothesis.

Theory of correlation:
Using Fisher-Z when the true correlation is not 0.

If $\rho \neq 0$, we must use a test called the Fisher-Z Transformation. When the true correlation is not zero, the underlying distribution is not symmetric, and so we cannot use the normal distribution (t/z-test). However, the Fisher-Z Transformation "transforms" the distribution of r into something that is normally distributed.

$$\text{Fisher-Z} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{51}$$

Turns out, the Fisher-Z statistic is normally distributed with a mean and standard deviation of:

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{52}$$

$$\sigma_z = \frac{1}{\sqrt{N-3}} \tag{53}$$

Theory of correlation – Confidence intervals using Fisher-Z

Thus, the confidence bounds for Z become:

$$Z - t_c \sigma_Z \leqslant \mu_Z \leqslant Z + t_c \sigma_Z \tag{54}$$

If you have μ_Z and want the corresponding actual correlation ρ, you can use

$$\rho = \frac{e^{2\mu_Z} - 1}{e^{2\mu_Z} + 1} = \tanh(\mu_Z) \tag{55}$$

Example using Fisher-Z

2.2 Testing the hypothesis that ρ is not equal to 0

What are the confidence limits on the true correlation if you drew 21 samples from a normal distribution and obtained $r = 0.8$?

$$Z = \frac{1}{2}\ln\left(\frac{1+0.8}{1-0.8}\right) = 1.0986 \tag{46}$$

$$\sigma_Z = \frac{1}{\sqrt{21-3}} = .235 \tag{47}$$

Calculating $t_{0.025} = 2.1$ (using $\nu = 21 - 3$) leads to:

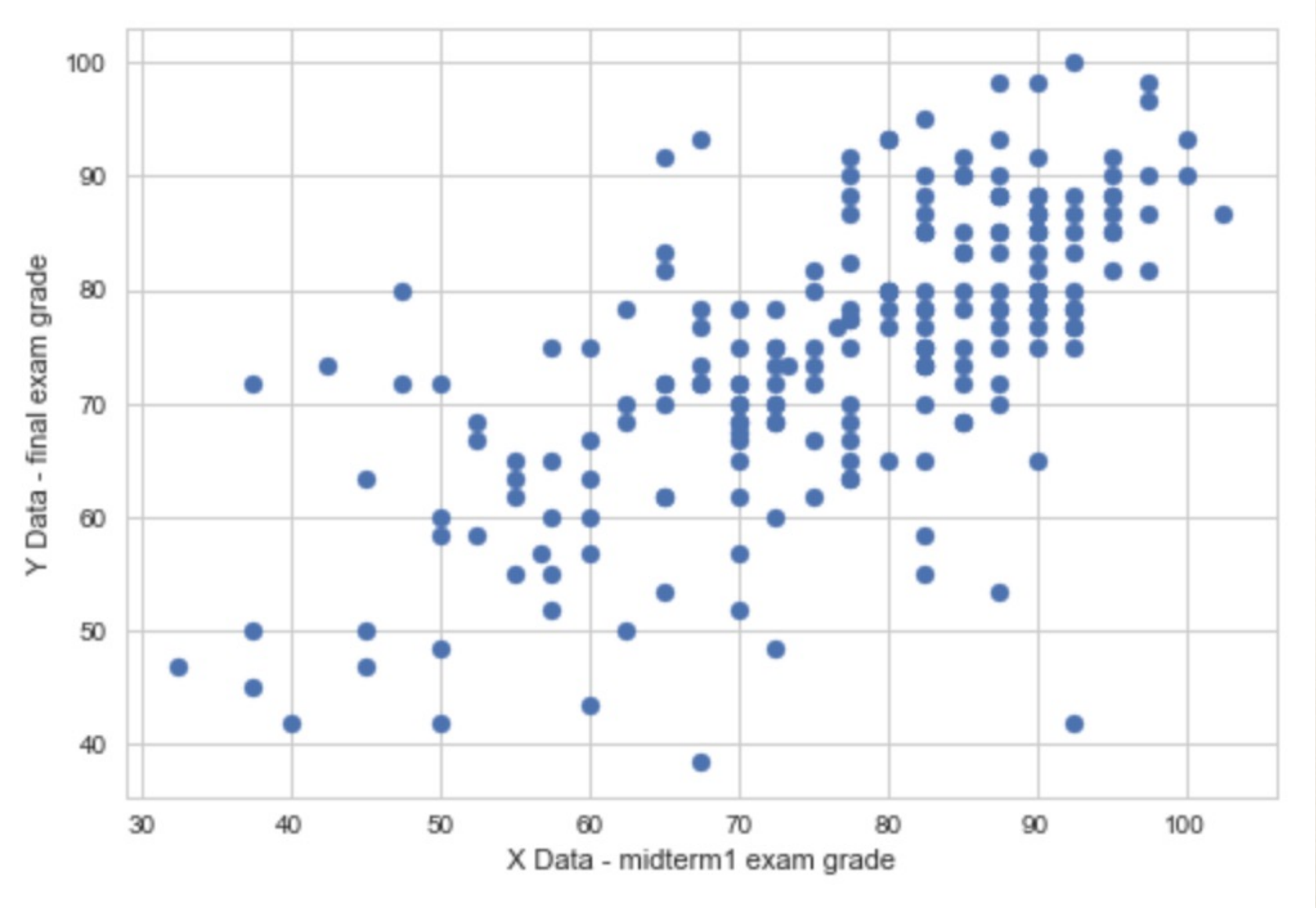
$$Z - 2.1\sigma_Z \leqslant \mu_Z \leqslant Z + 2.1\sigma_Z \tag{48}$$

$$0.61 \leqslant \mu_Z \leqslant 1.59 \tag{49}$$

We still need to get this back into correlation form, so plugging into the equation for ρ gives

$$0.54 \leqslant \rho \leqslant 0.92 \tag{50}$$

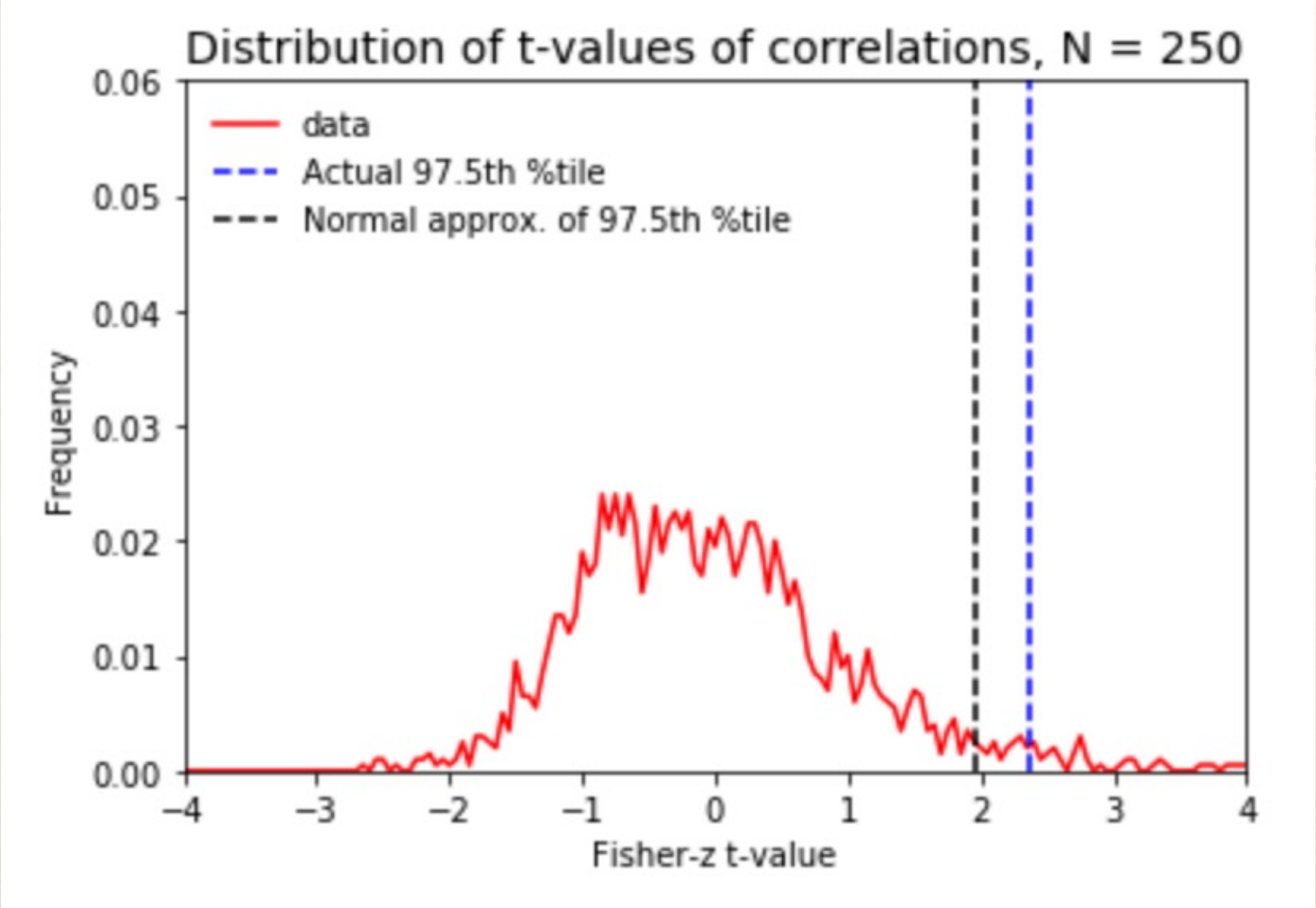
Let's do another example: Can a student's grade on the first midterm (X) be used to predict their final exam grade (Y)?



Try the python code – linear_regression_grades.ipynb

What happens if you don't use Fisher-z?

You will incorrectly assess the statistical confidence limits on your regression…



Try the python code –testing_normality_of_correlations.ipynb

Theory of correlation – Comparing two non-zero sample correlations (r1, r2) using Fisher-Z

If we want to test the difference between two correlations that are non-zero, we can once again use the Fisher transformation for each and use the fact that Z is normally distributed. Suppose we have two samples of size N_1 and N_2 which give correlation coefficients of r_1 and r_2. We test for a significant difference between these correlations by first calculating the Fisher-Z transformation for each:

$$Z_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) \tag{58}$$

$$Z_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right) \tag{59}$$

$$\tag{60}$$

and then calculating our normal z-score from the difference of means:

$$z = \frac{Z_1 - Z_2 - \Delta_{1,2}}{\sigma_{1,2}} \tag{61}$$

where

$$\Delta_{1,2} = \mu_1 - \mu_2 \tag{62}$$

is the *transformed* difference you expect (your null hypothesis). If your null hypothesis is that the true correlations of the two samples are equal ($\rho_1 = \rho_2$), then

$$\Delta_{1,2} = \mu_1 - \mu_2 = 0 \tag{63}$$

$\sigma_{1,2}$ is given in the following way

$$\sigma_{1,2} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \tag{64}$$

Spearman's rank correlation – What is it?

Spearman's rank correlations is a nonparametric test that tests whether a set of paired data monotonically vary together (when one goes up, the other goes down), but it is not concerned with how much it goes up or down. Since this is a nonparametric test, no assumption about normality needs to be made.

The idea is very simple, your original paired data x_i and y_i get converted into ranks X_i and Y_i and

$$\rho = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2 (Y_i - \overline{Y_i})^2}} \tag{65}$$

When there are duplicate values, the ranks are equal to the average position.

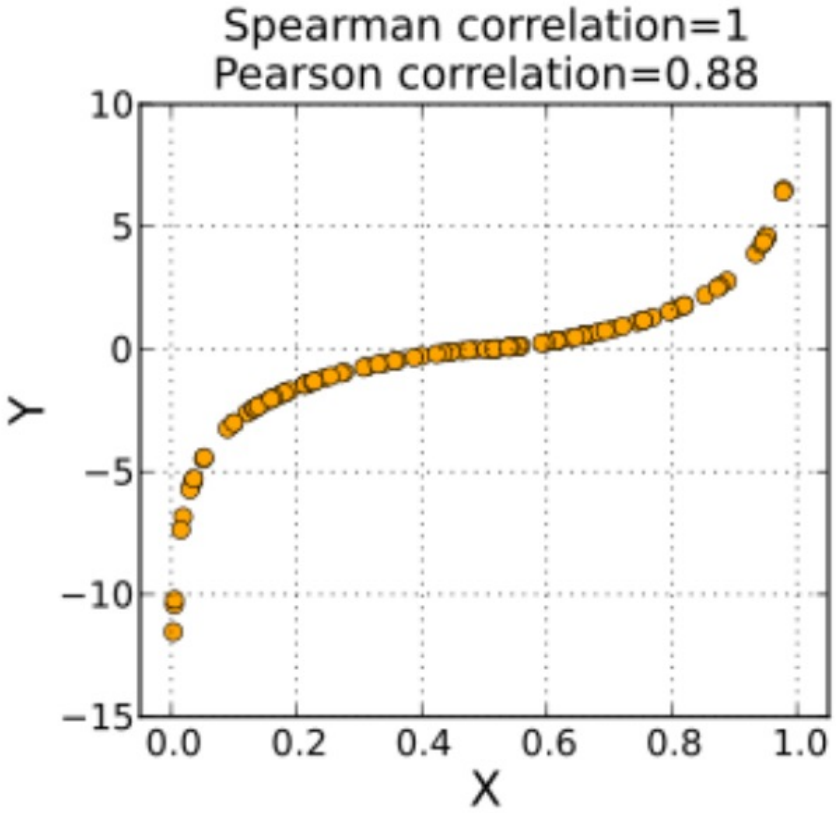
The standard error of Spearman's rank correlation ρ is given by

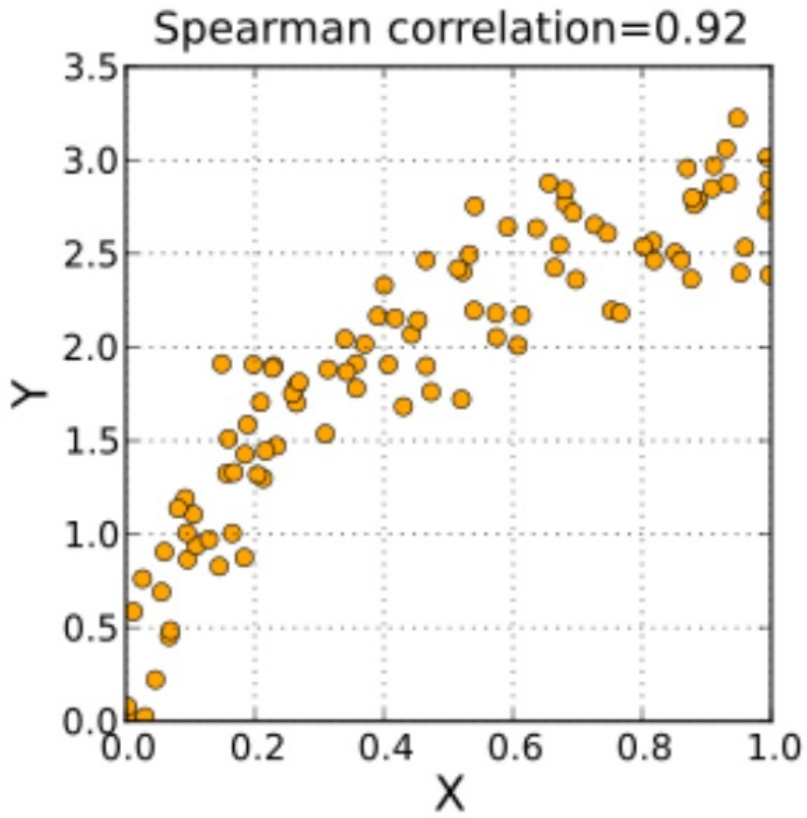
$$\sigma_\rho = \frac{0.6325}{(N-1)^{1/2}} \tag{66}$$

To determine significance, you can use the Fisher-Z test and the t-test (for a null hypothesis that $\rho = 0$) as for the Pearson correlation.

Spearman's rank correlation – Example

Remember: Spearman tests whether X and Y monotonically vary together. Pearson tests whether X and Y are linear.





REVIEW ON THE BOARD
LECTURE #4 – PART 2

WRITING OFTEN HELPS CONCEPTS SINK IN…

GROUP WORK

- *linear_regression_grades.ipynb*
- *testing_normality_of_correlations.ipynb*
- *Explain "Lecture #4 on the board" notes to each other.*

Review Notes on the board and Jupyter Notebooks from lecture #4:

jackknife_example.ipynb
linear_regression.ipynb
LSQ_OLS.ipynb
linear_regression_grades.ipynb
testing_normality_of_correlations.ipynb

Skim Barnes Chapter 2

See Thursday, if not before. ☺