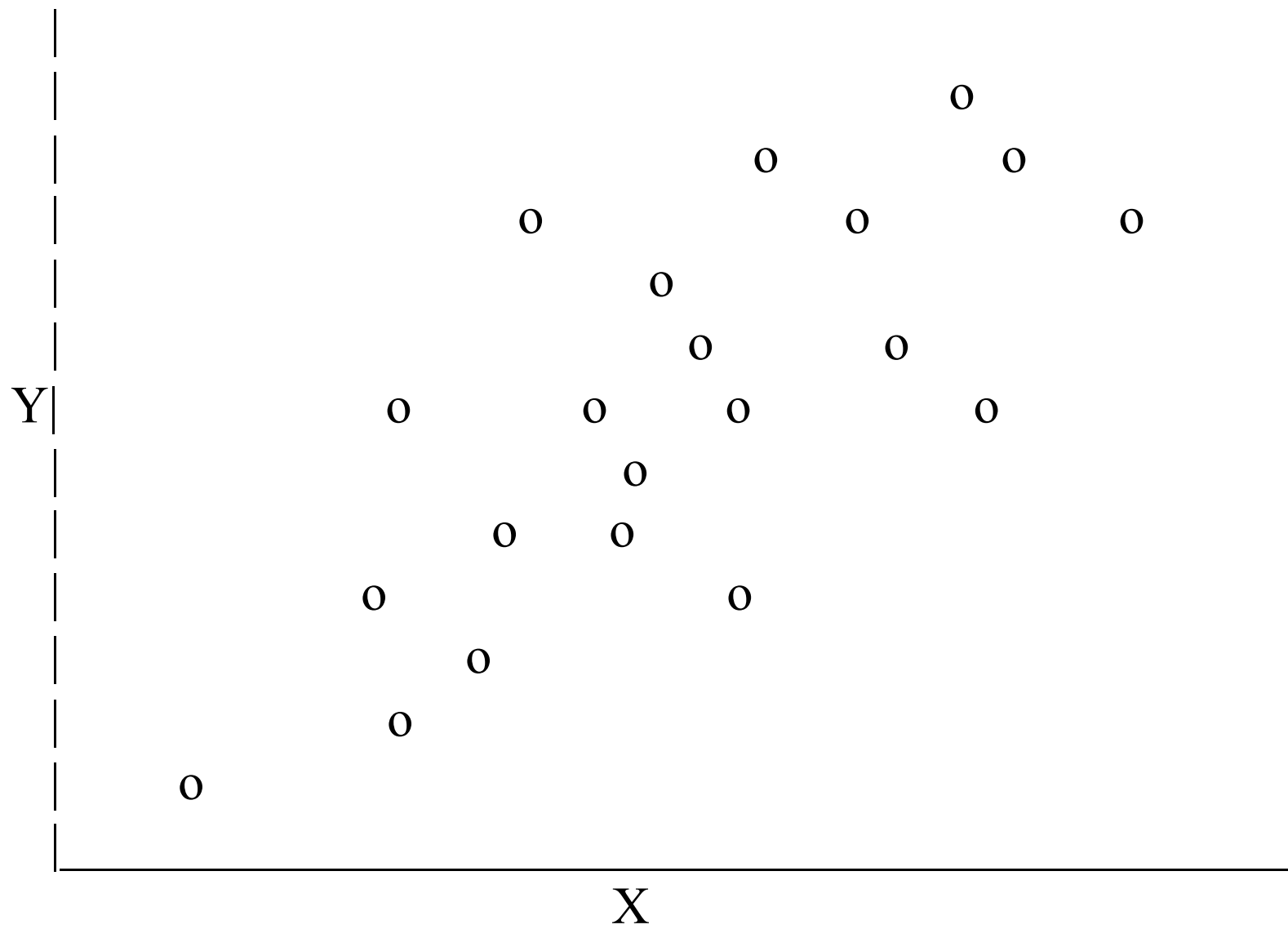


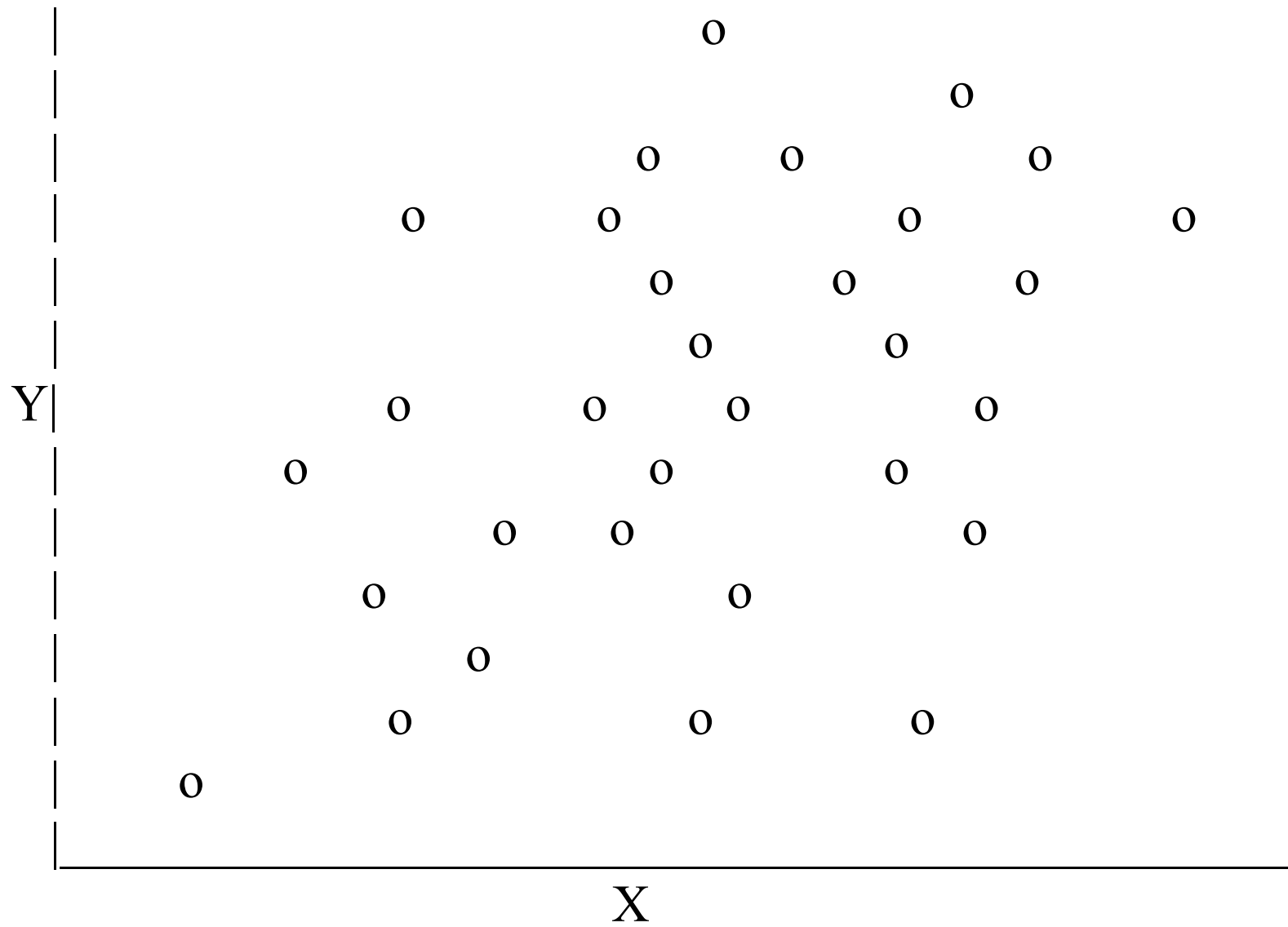
# Correlation, Simple Regression, and Low-order Multiple Regression.

Pearson product-moment correlation  
...is what we will usually mean by “correlation”.

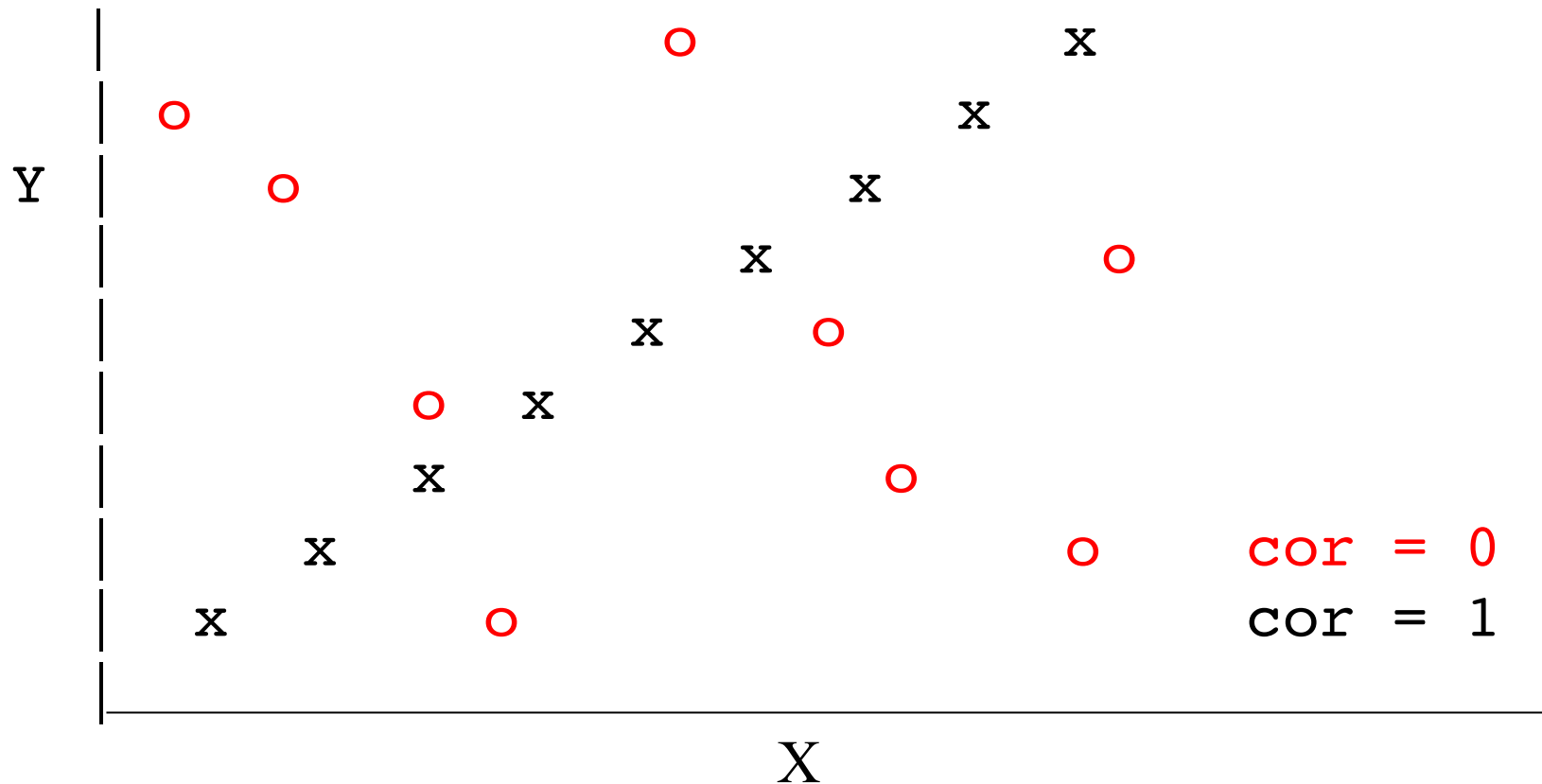
It describes the strength of a **linear** relationship  
between  $x$  and  $y$ .



correlation = 0.8

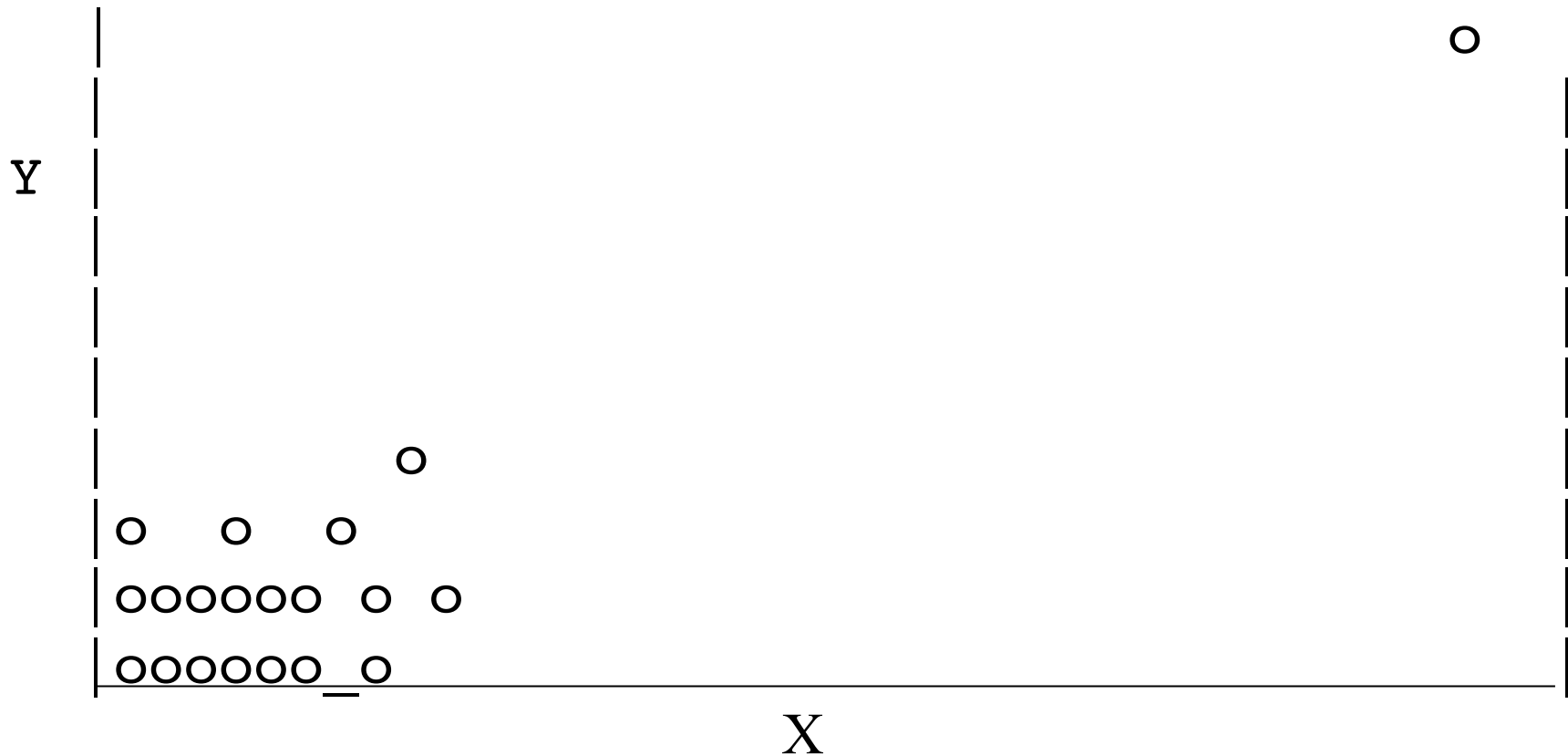


correlation = 0.55



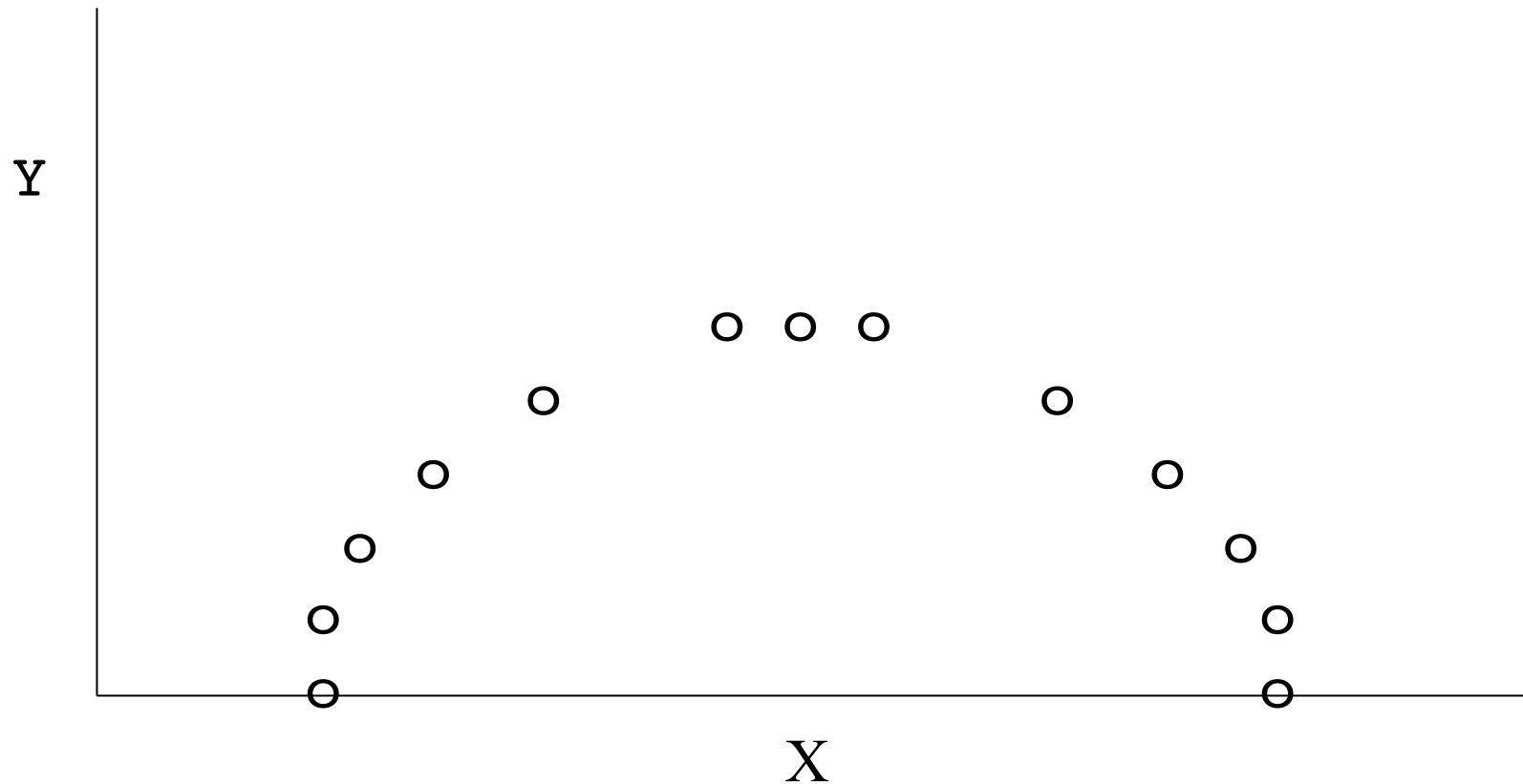
correlation for all 18 points = 0.707    correlation squared = 0.5

When points having a **perfect correlation** are mixed with an equal number of points having **no correlation**, and the two sets have same mean and variance for X and Y, correlation is 0.707. Correlation squared (“amount of variance accounted for”) is 0.5.



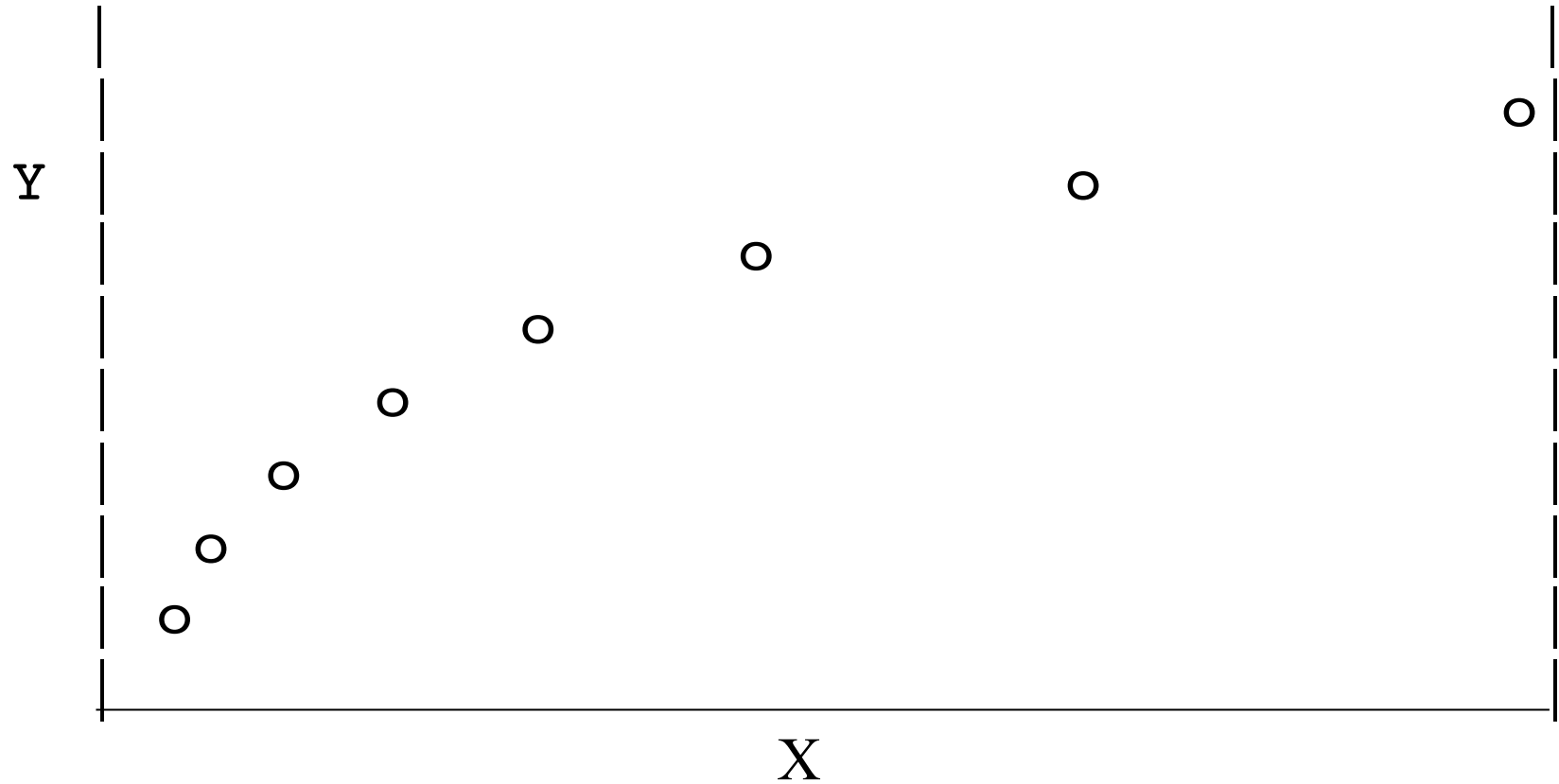
correlation = 0.87 (due to one outlier in upper right)

If domination by one case is not desired, can use the Spearman rank correlation (correlation among ranks instead of actual values).



correlation = 0 **but there is a strong nonlinear relationship**

The Pearson correlation only detects linear relationships.



correlation = 0.9 **but there is an exact nonlinear relationship**  
such as  $y = \sqrt{x}$

## Standardizing X and Y to equalize their units

$$SD_x = \sigma_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

find SD for the  
x data set

$$z_x = \frac{(x - \bar{x})^2}{SD_x}$$

Convert each  
x element to  
to its standardized  
value (z)

Then do the same for y. Now their units are on an equal footing, with mean = 0, SD = 1.



## Covariance and correlation between two variables

$$\text{Variance}_x = (SD_x)^2 = \sigma_x^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \quad \text{x itself}$$

(can do  
same for y)

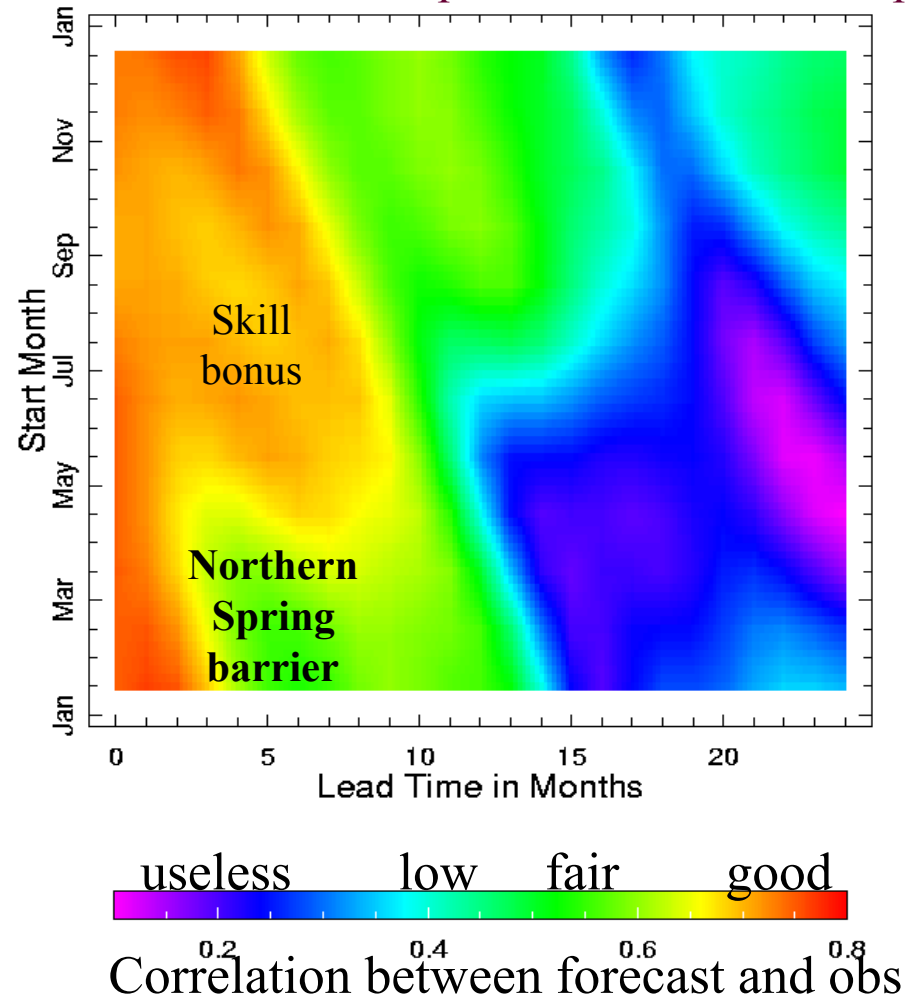
$$\text{Co variance}_{x,y} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n} \quad \text{x vs. y}$$

$$\text{Correlation}_{x,y} = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{n} \sum_{i=1}^n z(x_i)z(y_i)$$

This is the **Pearson** product-moment correlation (the “standard” correlation)

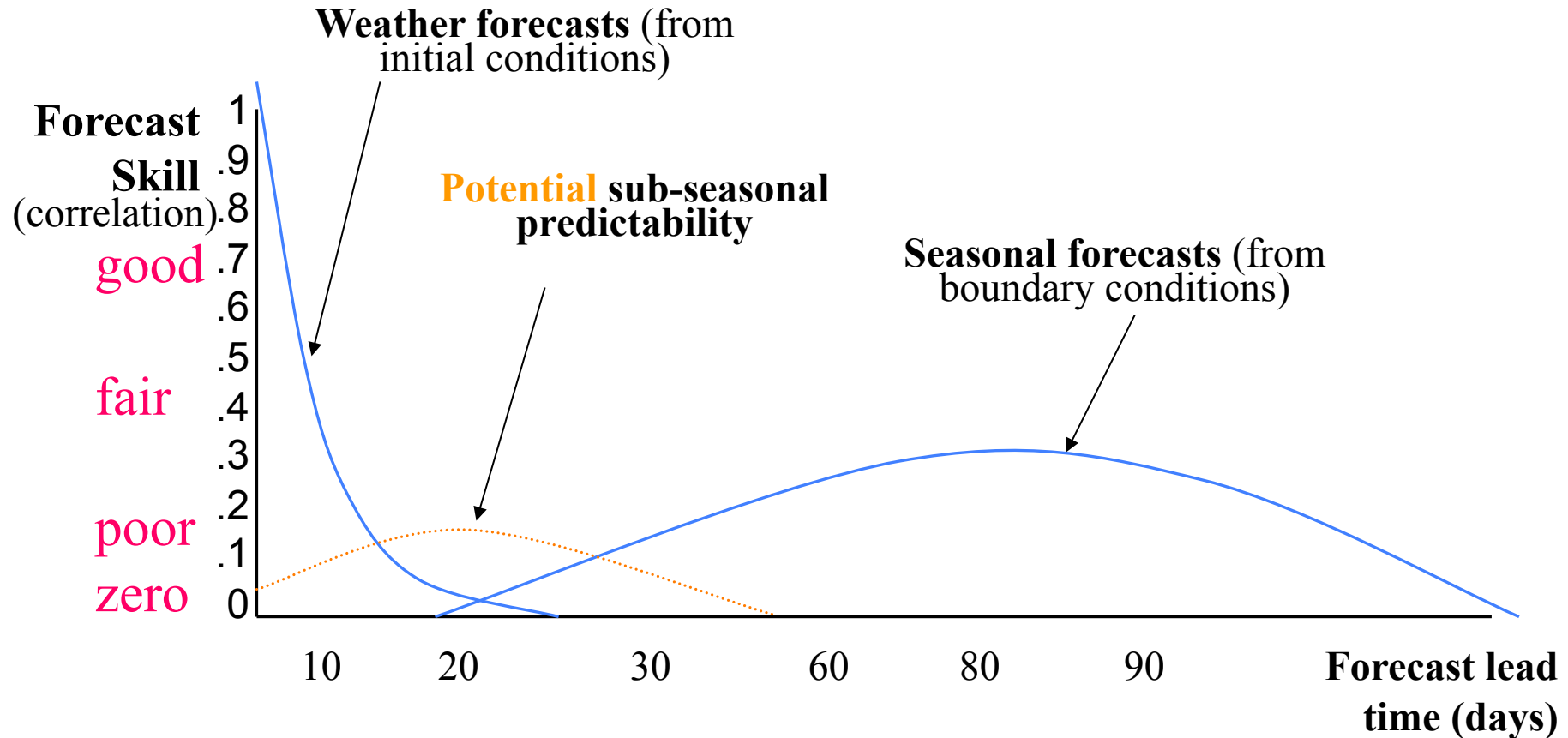
Basis of climate predictability lies in predictability of ENSO

Skill of Cane-Zebiak model in prediction of SST in tropical Pacific



# Lead time and forecast skill

Correlation between temperature and precipitation forecasts and their subsequent corresponding observations



## Skill of forecasts at different time ranges:

1-2 day weather	good
3-7 day weather	fair
Second week weather	poor, but not zero
Third week weather	virtually zero
Fourth week weather	virtually zero
1-month climate (day 1-31)	poor to fair
1-month climate (day 15-45)	poor, but not zero
3-month climate (day 15-99)	poor to fair

At shorter ranges, forecasts are based on initial conditions and skill deteriorates quickly with time.

Skill gets better at long range for ample time-averaging, due to consistent boundary condition forcing

## Approximate\* Standard Error of a Zero Correlation Coefficient

(as would be expected if X and Y are independent **random** data)

$$\sigma_{0-cor} = \frac{1}{\sqrt{n-1}}$$

Examples of  $\sigma_{0-cor}$  and critical values for **2-sided significance at 0.05 level** for various sample sizes n

n	$\sigma_{0-cor}$	$COR_{critical.025}$
10	0.33	0.65
20	0.23	0.45
50	0.14	0.28
100	0.10	0.20
400	0.05	0.10

Note: For significance of a correlation, z-distribution is used, rather than t-distribution, for any sample size.

\*For small n, true values of  $\sigma_{0-cor}$  are slightly smaller.

Confidence intervals for a nonzero correlation ( $r$ ) are smaller than those for zero correlation, and are asymmetric such that the interval toward lower absolute values of  $r$  is larger.

For example: for  $n=100$  and  $r = 0.35$ , 95% confidence interval is  $0.17$  to  $0.51$ . That is  $0.35$  minus  $0.18$ , but  $0.35$  plus  $0.16$ . (For  $r = 0$ , it is  $0$  plus  $0.20$  and  $0$  minus  $0.20$  – a larger span.)

Sampling distribution around a population correlation is computed using the Fisher  $r$ -to- $Z$  transformation, then finding a symmetric confidence interval in  $Z$ , then finally converting back to  $r$ .

The use of linear correlation for prediction:

## **Simple Linear Regression**

(“simple” implies just one predictor;  
if more than one, is Multiple Linear Regression)

Determination of a regression line  
to fit points on the  $x$  vs.  $y$  scatterplot,  
so that if given a value of  $x$ , a “best  
prediction” can be made for  $y$ .

A line in the x vs. y coordinate system has the form

$$y = a + bx \quad a \text{ is y-intercept} \quad b \text{ is slope}$$

Regression line is defined such that the sum of the squares of the errors (the predicted y vs. true y) is minimized.

Such a line predicts y from x such that:  $Z_y = COR_{xy}Z_x$

For example, if  $COR_{xy} = .5$  then y will be predicted to be half as many SDs away from its mean as x.



**Proof that  $z_y = COR_{xy}z_x$  minimizes the squared errors.**

That is, proof that the slope (the “b” in  $y=bx+a$ ) should be set to be the correlation coefficient between y and x when y and x are in standardized (z) form where their means are zero and SDs are 1.

The squared error to be minimized, where i ranges from 1 to n pairs of predicted versus actual values of y, is  $\frac{1}{n} \sum_i [\hat{z}_{y_i} - z_{y_i}]^2$

where  $\hat{z}_y$  refers to the **predicted** standardized value of y, and  $z_y$  the **actual** (observed) standardized value of y.

Substituting  $bz_{x_i}$  for  $\hat{z}_{y_i}$  leads to  $\frac{1}{n} \sum_i [bz_{x_i} - z_{y_i}]^2$

Expanding the square in  $\frac{1}{n} \sum_i [bz_{x_i} - z_{y_i}]^2$  we get

$$\frac{1}{n} b^2 \sum_i z_{x_i}^2 - \frac{1}{n} 2b \sum_i z_{x_i} z_{y_i} + \frac{1}{n} \sum_i z_{y_i}^2$$

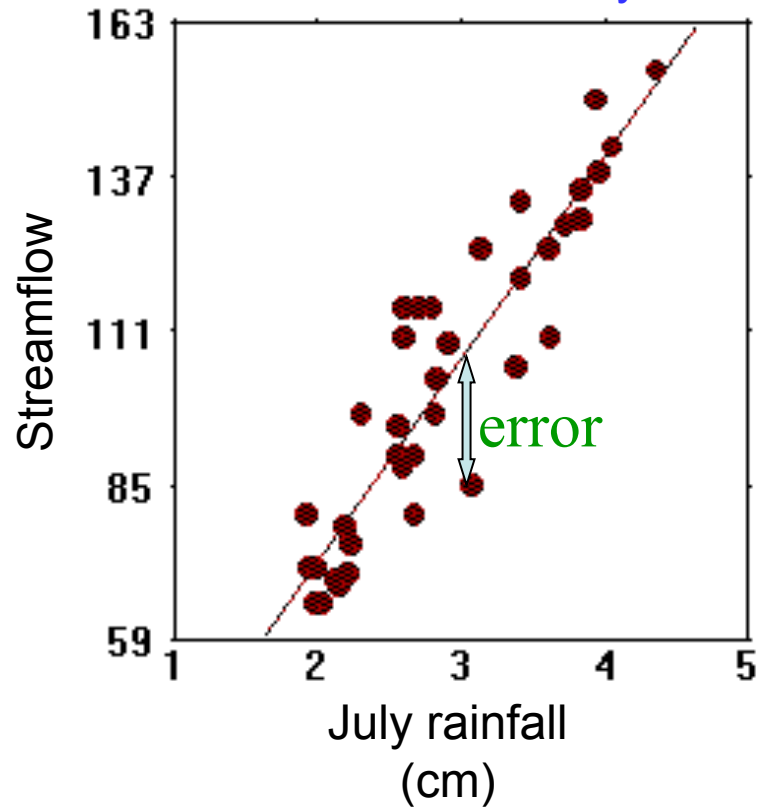
Because  $\frac{1}{n} \sum_i z^2 = 1$  for any variable, and  $\frac{1}{n} \sum_i z_{x_i} z_{y_i} = cor_{xy}$ ,

the expression to be minimized reduces to  $b^2 - 2b(cor_{xy}) + 1$  .

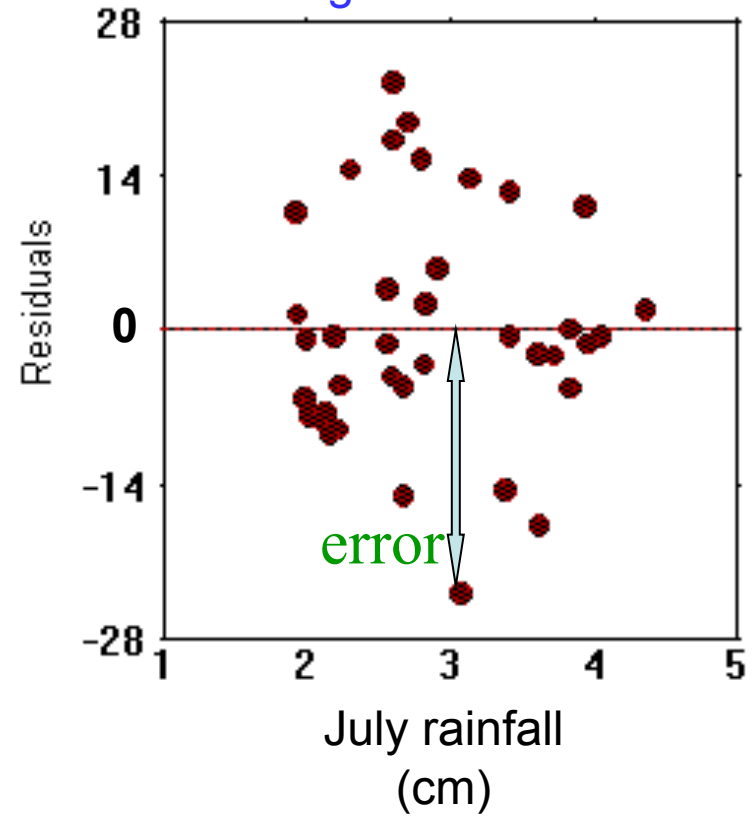
To find what value of b minimizes the expression, **set its derivative to zero**:  $2b - 2(cor_{xy}) = 0$

We then see that  $b = cor_{xy}$

Simple Regression Prediction  
of Streamflow from July Rainfall



Deviations of Observations  
from Regression Predictions



Simple regression prediction, **standardized units**:  $Z_y = cor_{xy}Z_x$

If we incorporate the **physical units of x and y** rather than the standardized (z) version in SD units, we get:

$$y = \bar{y} + (cor_{xy}) \frac{SD_y}{SD_x} (x - \bar{x})$$

The above equation “tailors” the basic z relationship by adjusting for (1) **ratio of SD of y to SD of x**, and (2) the **difference between the mean of y and the mean of x**.

$(cor_{xy}) \frac{SD_y}{SD_x}$  is the slope (b) of the regression line

$\bar{y} - b\bar{x}$  is the y-intercept

## Standard error of estimate of regression forecasts

....is the **standard deviation of the error distribution**,  
where the errors are  $y_{predicted} - y_{actual}$

St Error of Estimate (of standardized y data, or  $z_y$ ) =

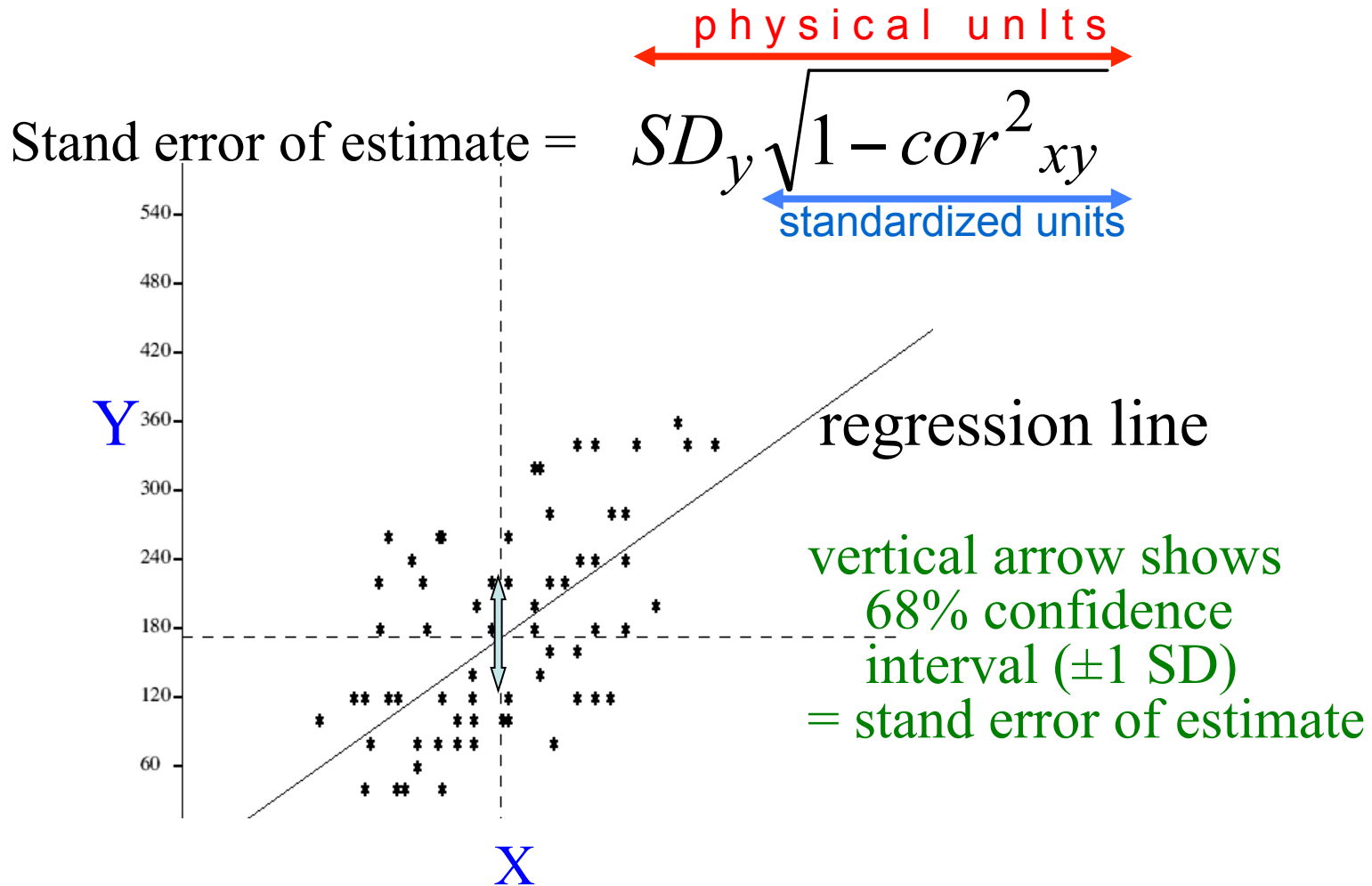
$$\sqrt{1 - cor^2_{xy}}$$

St Error of Estimate (of actual y data in physical units) =

$$SD_y \sqrt{1 - cor^2_{xy}}$$

When  $cor = 0$ , Stand Error of Estimate is same as the SD of y.  
When  $cor = 1$ , Stand Error of Estimate is 0 (all errors are zero).

# Standard error of estimate for a regression forecast



The linear regression model can lead to probability forecasts for any result, given the exact prediction and the correlation, and an assumption that the variables are normally distributed.

## Correlation vs. Standard Error of Estimate

Standard Error  
of Estimate

$$\sqrt{1 - cor^2_{xy}}$$

Correlation (as a fraction of SD  
of the predictand [y])

1.00	0.00
0.90	0.44
0.80	0.60
0.70	0.71
0.60	0.80
0.50	0.87
0.40	0.92
0.30	0.95
0.20	0.98
0.10	0.99
0.00	1.00

half →

We need quite a high correlation to get a low standard error of estimate: need  $cor = 0.866$  to get an SD of the error down to **half** of the SD of the predicted variable (y).

Standard error of estimate (in standardized units) for the prediction model as a whole (generalized for any possible values of x) is

$$\sqrt{1 - cor^2_{xy}}$$

But this can be defined more accurately if we know the x value. Let  $z_o$  be the standardized value of the predictor (x). Then standard error of estimate as function of  $z_o$  is

$$\sqrt{1 - cor^2_{xy}} \sqrt{\frac{1}{n} + \frac{z_o^2}{n}}$$

If we are dealing with a single case, 1 is added to the content under the second square root term.

**Standard error is larger when x value is farther away from mean.** There is also an “unbiasing” adjustment, even if x is at its mean. Both of these effects are smaller when the sample size is larger.



## Simple Linear Regression Problem: Coupled GCM forecasts for Fiji for next Jan-Feb-Mar

Suppose we know that the correlation between a coupled GCM rainfall forecast for parts of Fiji in Jan-Feb-Mar (made at beginning of December), and the actual rainfall, is **0.52**. This does not come as a surprise, because we know that Fiji is sensitive to the ENSO state and that climate models are able to reproduce this relationship to a moderate extent. By early December the ENSO state is usually stable.

Suppose we want to issue a rainfall forecast for the station of Nadi on the north side of the main Fiji island, using the forecast from this model. We have the following historical data:

Model Predictions (JFM):

Mean: **1140 mm**

SD: **700 mm**

Observations (JFM):

Mean: **935 mm**

SD: **500 mm**

If the model forecast for the current year is **1890 mm**, what would be our regression-based best forecast for the actual precipitation?

JFM season in Nadi, Fiji:

Model Predictions (JFM):                      Observations (JFM):

Mean: **1140 mm**                                      Mean: **935 mm**

SD: **700 mm**    SD: **500 mm**

Correl (forecast vs. observations) = **0.52**      Model predicts **1890 mm**

We use:                       $z_y = COR_{xy}z_x$       and                       $z_x = \frac{x - \bar{x}}{SD_x}$

z value for predictor ( $z_x$ ) is  $(1890 - 1140) / 700 = 1.07$

Then z value for forecast of precip ( $z_y$ ) is  $(0.52)(1.07) = 0.56$   
(forecast of precip is 0.56 SDs above its mean.)

Forecast of precip = mean of y +  $(0.56)(SD_y)$

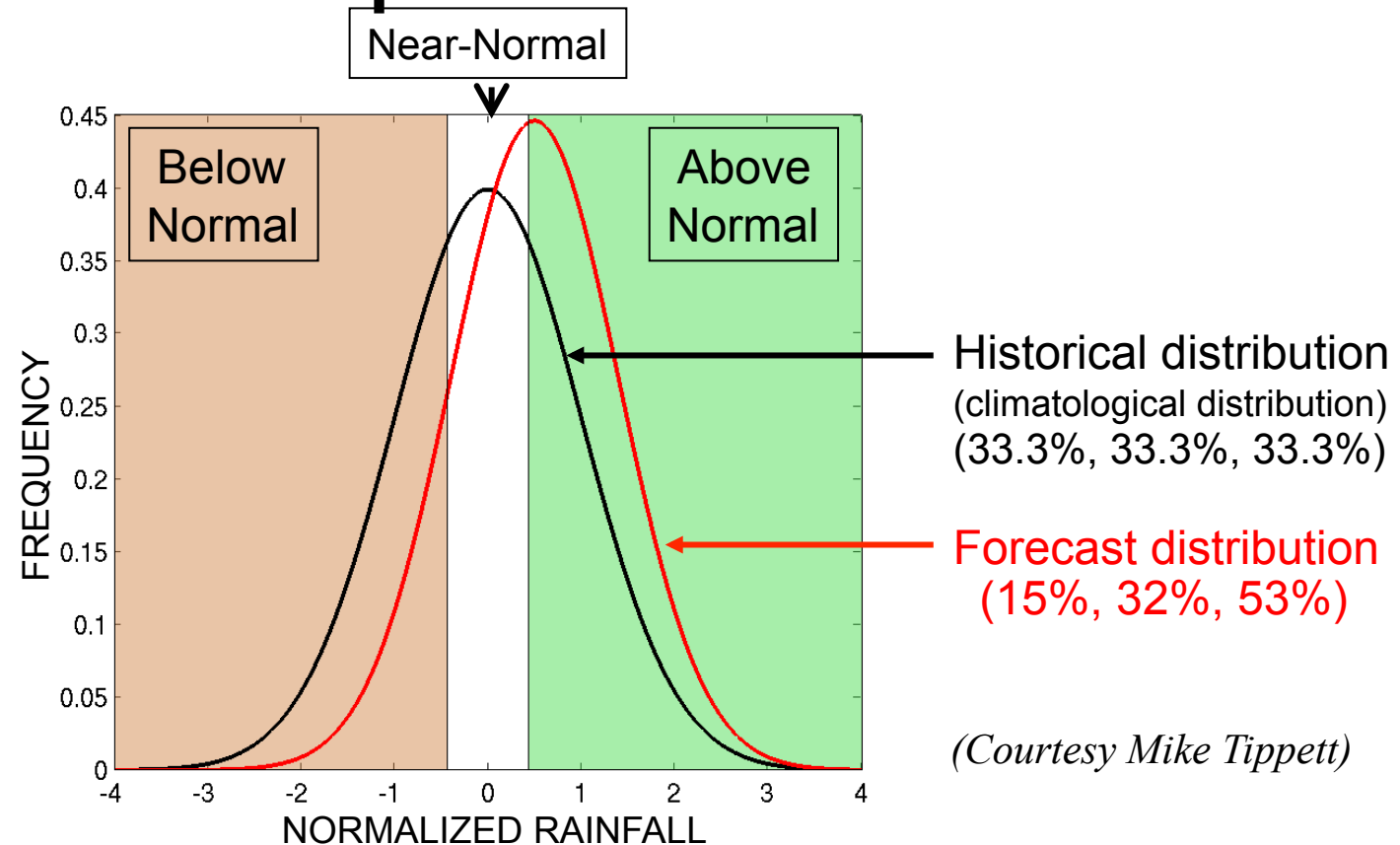
Forecast of precip =  $935 + 0.56(500)$   
=  $935 + 280 = 1215 \text{ mm}$

Standard error of estimate (standardized units) =  $\sqrt{1 - cor^2_{xy}} = .854$

Standard error of estimate (physical units) =  $SD_y (.854) = 427 \text{ mm}$

Since we do not know the sample size used to develop this regression model,  
We cannot compute the standard error of estimate for this forecast specifically.

# What probabilistic forecasts represent

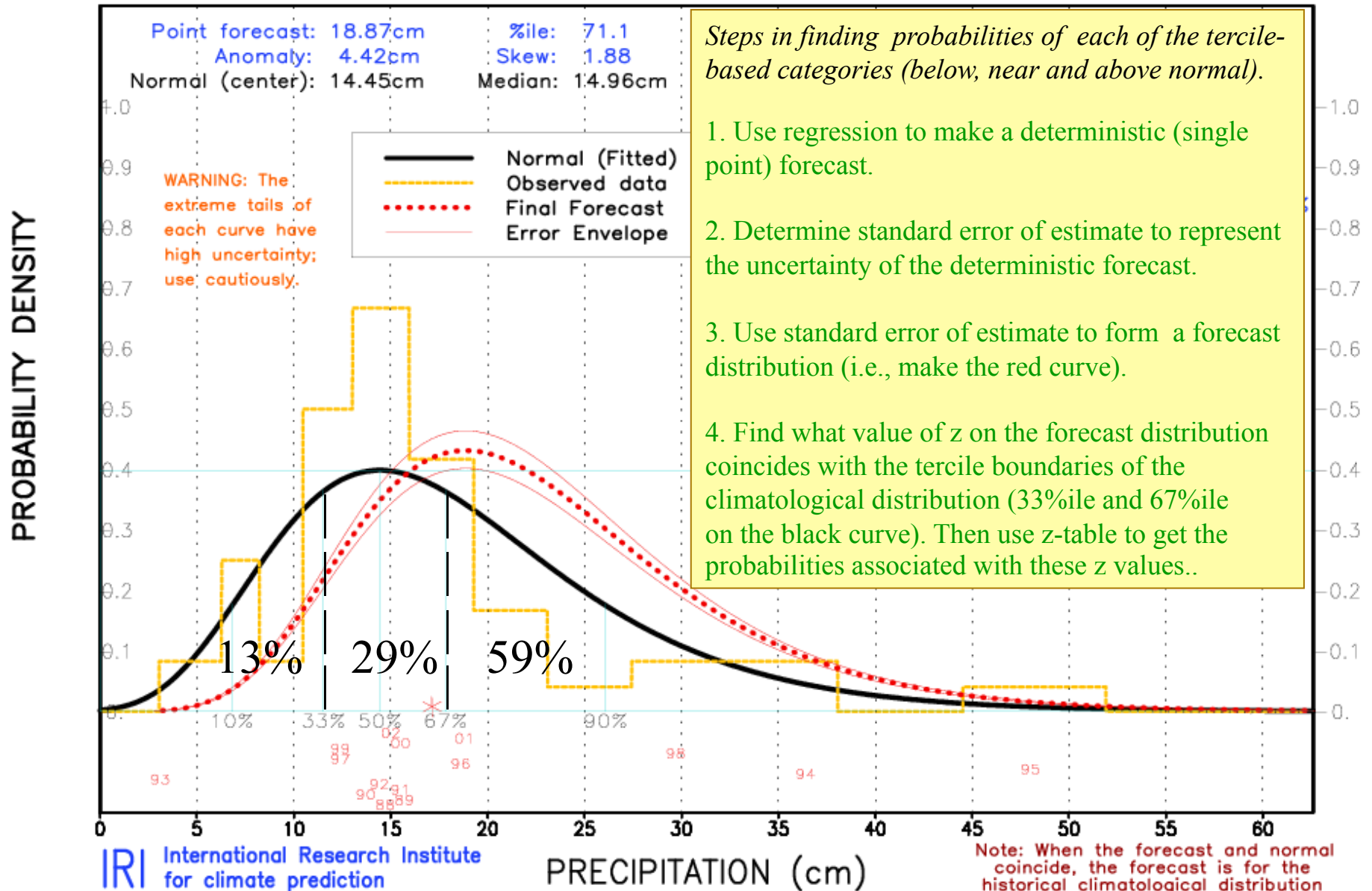


Historically, the probabilities of above and below are 0.33. Shifting the mean by one half standard deviation and reducing the variance by 20% changes the probability of below to 0.15 and of above to 0.53. Correlation skill would be 0.45, and predictor signal strength would be 1.11 SD units.

# A “strong” shift of odds in rainfall forecast for Kenya during El Nino

## 3-MONTH TOTAL PRECIPITATION MULTI-MODEL PROBABILITY FORECAST FOR OND 1997 2.5 MONTH LEAD OUTLOOK – MADE MID-JULY, 1997

Station 389                      WAJIR, Kenya                      1.45    40.30    152



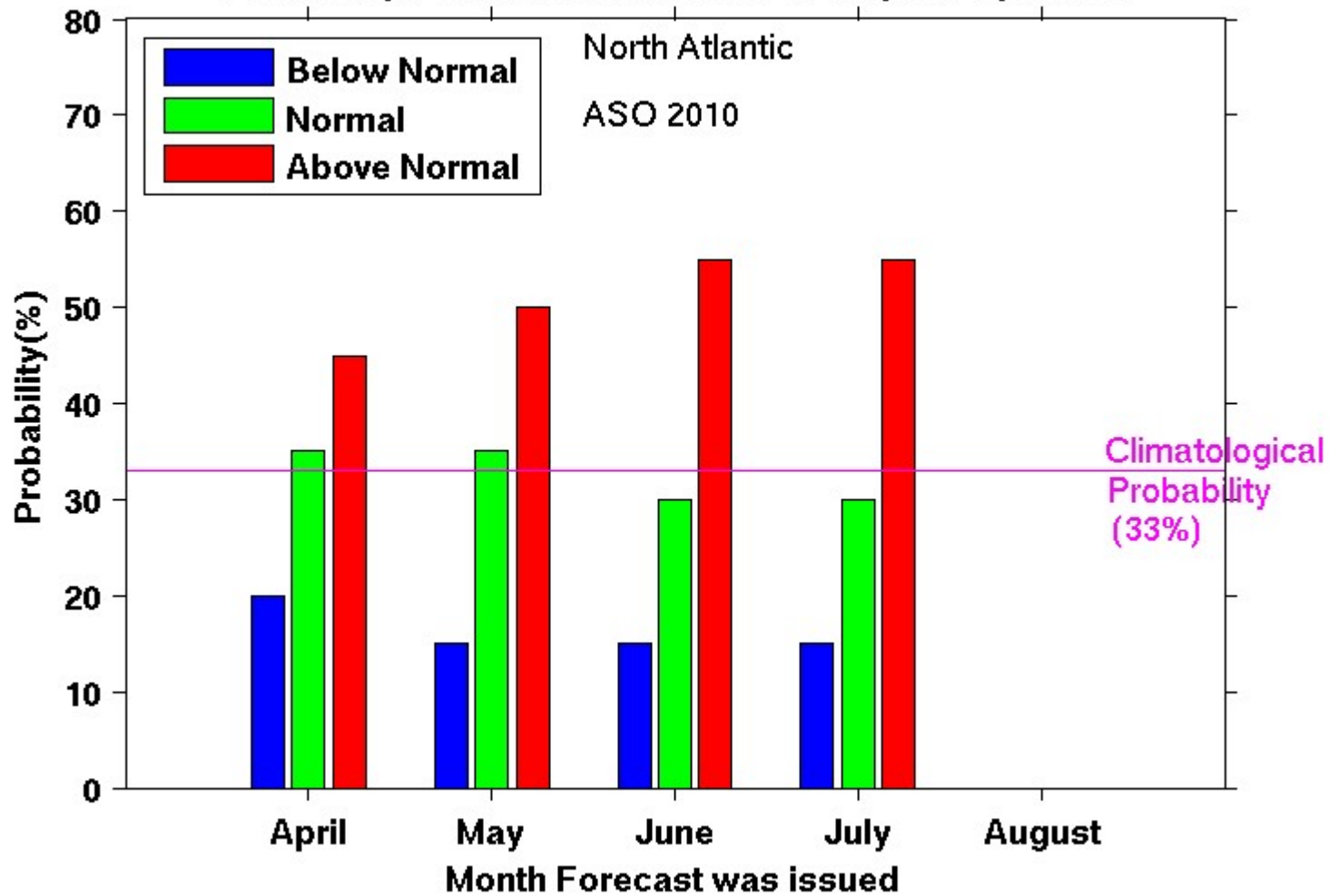
Tercile probabilities for various correlation skills and predictor signal strengths (in SDs). Assumes Gaussian probability distribution. Forecast (F) signal = (Predictor Signal) x (Correl Skill).

Correlation Skill	Predictor Signal=0.0	Predictor Signal +0.5	Predictor Signal +1.0	Predictor Signal +1.5	Predictor Signal +2.0
0.00	F signal 0.00 33 / 33 / 33	F signal 0.00 33 / 33 / 33	F signal 0.00 33 / 33 / 33	F signal 0.00 33 / 33 / 33	F signal 0.00 33 / 33 / 33
0.20	F signal 0.00 33 / 34 / 33	F signal 0.10 29 / 34 / 37	F signal 0.20 26 / 33 / 41	F signal 0.30 23 / 33 / 45	F signal 0.40 20 / 31 / 49
0.30	F signal 0.00 33 / 35 / 33	F signal 0.15 27 / 34 / 38	F signal 0.30 22 / 33 / 45	F signal 0.45 17 / 31 / 51	F signal 0.60 14 / 29 / 57
0.40	F signal 0.00 32 / 36 / 32	F signal 0.20 25 / 35 / 40	F signal 0.40 18 / 33 / 49	F signal 0.60 13 / 30 / 57	F signal 0.80 9 / 25 / 65
0.50	F signal 0.00 31 / 38 / 31	F signal 0.25 22 / 37 / 42	F signal 0.50 14 / 33 / 53	F signal 0.75 9 / 27 / 64	F signal 1.00 5 / 21 / 74
0.60	F signal 0.00 30 / 41 / 30	F signal 0.30 18 / 38 / 44	F signal 0.60 10 / 32 / 58	F signal 0.90 5 / 23 / 72	F signal 1.20 2 / 15 / 83
0.70	F signal 0.00 27 / 45 / 27	F signal 0.35 13 / 41 / 46	F signal 0.70 6 / 30 / 65	F signal 1.05 2 / 17 / 81	F signal 1.40 1 / 8 / 91
0.80	F signal 0.00 24 / 53 / 24	F signal 0.40 8 / 44 / 48	F signal 0.80 2 / 25 / 73	F signal 1.20 0* / 10 / 90	F signal 1.60 0** / 3 / 97

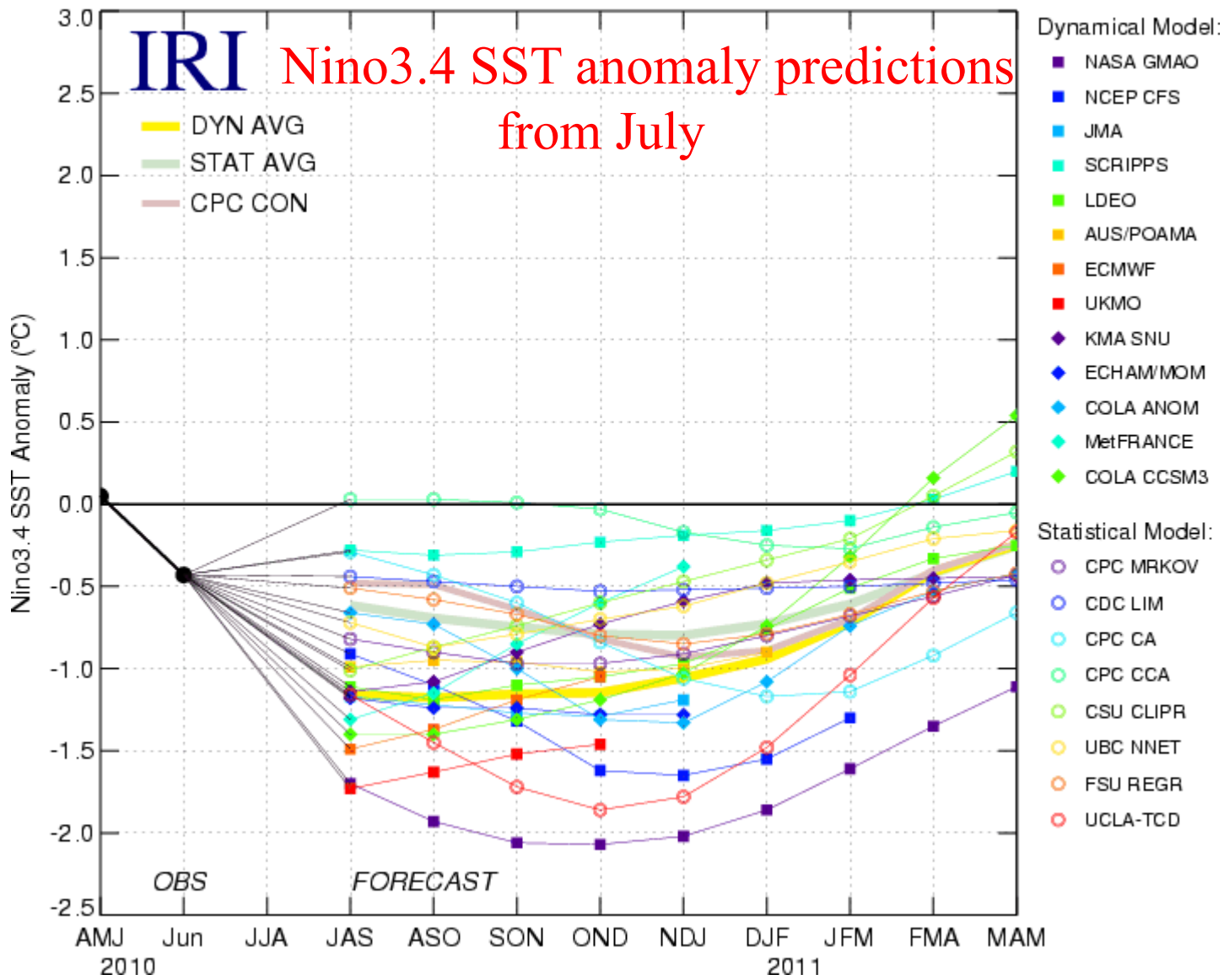
\*0.3

\*\*0.04

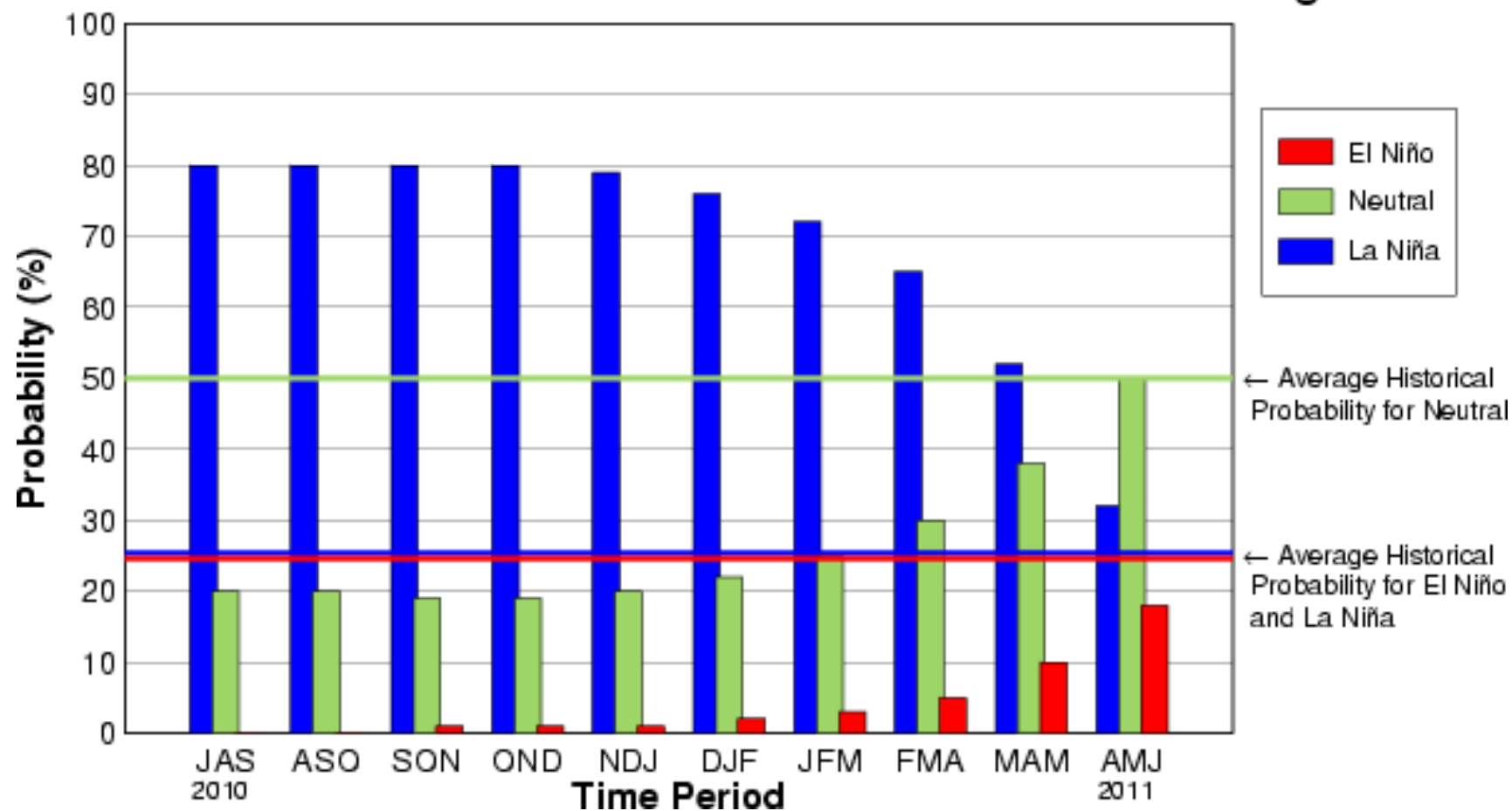
### Probability Forecasts for Number of Tropical Cyclones



### Model Predictions of ENSO from Jul 2010

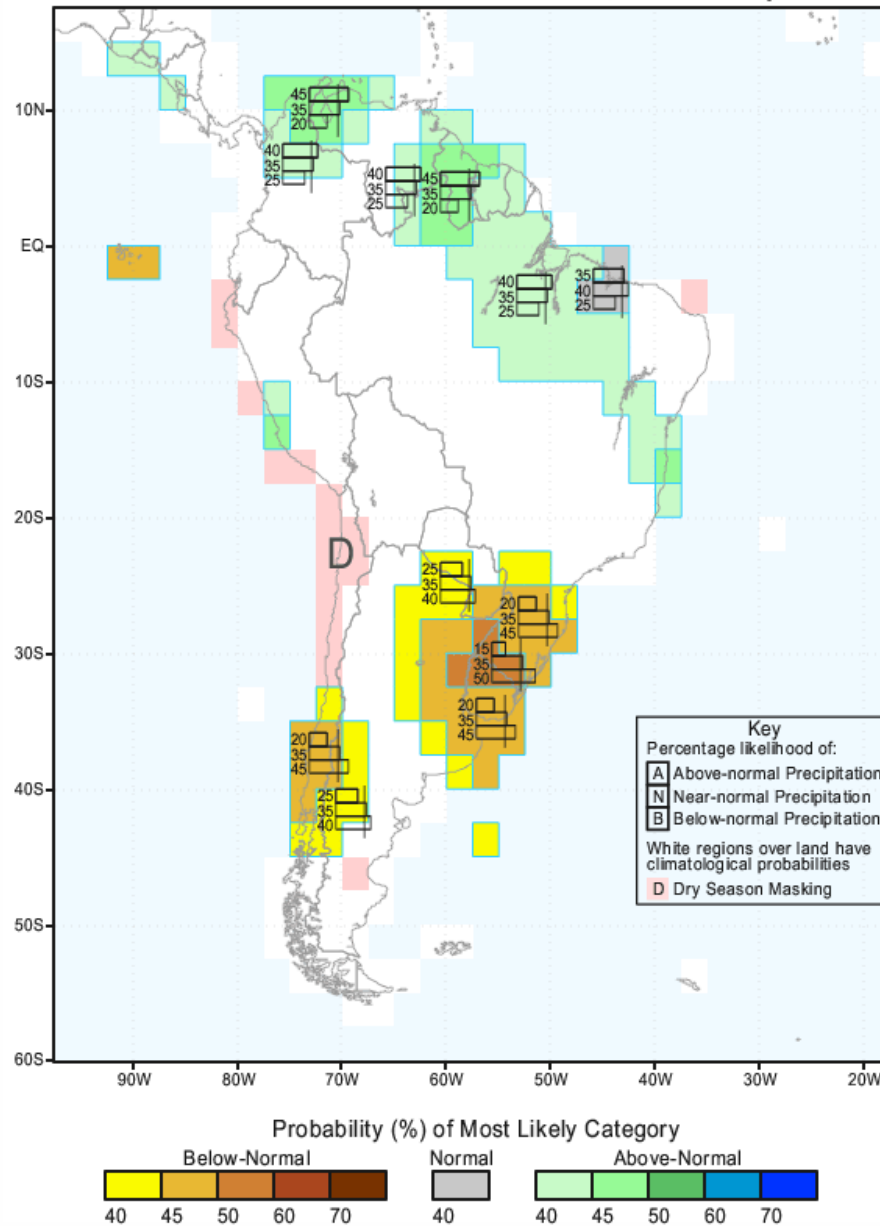


## IRI Probabilistic ENSO Forecast for NINO3.4 Region





# IRI Multi-Model Probability Forecast for Precipitation for October-November-December 2010, Issued July 2010



Raw scores  
regression formula:  $\hat{y} = \bar{y} + (cor_{xy}) \frac{SD_y}{SD_x} (x - \bar{x})$

where  $\hat{y}$  is predicted value of y from regression from x

$$slope = b = (cor_{xy}) \frac{SD_y}{SD_x}$$

Slope can be nonzero only if correlation is nonzero.

Therefore, testing if the slope is significantly different from zero should give the same result (the probability that it is different from zero by chance alone) as testing if the correlation is significantly different from zero.

# Hypothesis test for a correlation value

How can we reject the (null) hypothesis that a correlation value comes from a population having zero correlation?

Standard error of correlation coefficient **with respect to zero correlation** (approximate; slightly too strict for  $n < 10$ ):

$$StError_{(zerocorrel)} = \frac{1}{\sqrt{n-1}} \quad (\text{can also be called } \sigma_{0-cor})$$

See if your sample correlation falls outside of plus or minus 1.96 (for 2-tailed, 5% level) times the above standard error. **If not, it could have come from population with zero correlation.** For 1-sided test (if sign of the correlation is known or expected in advance of seeing the resulting experimental correlation), see if your sample correlation is greater in magnitude than 1.65 times the above standard error. For correlation, 1-sided tests are common.

## Example of a hypothesis test for a correlation

Suppose we test the correlation between malaria incidence following the November – March rainy season in Botswana, and the amount of rainfall during that rainy season. We know, before investigating the correlation, that more rainfall (except for extreme flooding conditions) creates a more favorable environment for the vector and thus greater risk for malaria.

Suppose for 10 years of data for rainfall during Nov – March and malaria during March – May, we get a correlation of 0.64. Is this statistically significant in terms of the hypothesis that the true population correlation is zero? That is, could the 0.64 have come about just by chance, due to natural sampling variations, and not due to a physical association between rainfall amount and malaria? Since the slate is wiped clean for the rainfall – malaria relationship with each new year, we can use 10 as the degrees of freedom. (This might not be true if the cases were not independent, such as for adjacent seasons that have nonzero lag correlation in both rainfall and malaria.)

## Example of a hypothesis test for a correlation

Sample size for rainfall and subsequent malaria incidence:  $n = 10$

Correlation between rainfall and malaria incidence:  $0.64$

$$\sigma_{0-cor} = StError_{(zerocorrel)} = \frac{1}{\sqrt{n-1}} = \frac{1}{\sqrt{9}} = \frac{1}{3} = 0.333$$

We set up a 1-sided z test for the correlation of 0.64. It is 1-sided because we have physical reason to expect a positive correlation rather than a negative correlation.

← Numerator shows correlation difference between sample outcome and population having zero correlation

$$z = \frac{SampleCor - 0.00}{StandardError_{zerocorrel}} = \frac{0.64}{0.333} = 1.92$$

Looking at a z table, the chance of equaling or exceeding  $z = 1.92$  is 0.0274. Significance at the 5% level is therefore achieved.

$$\sigma_{0-cor} = StError_{(zerocorrel)} = \frac{1}{\sqrt{n-1}}$$

but  $StError_{(non-zerocorrel)} < \frac{1}{\sqrt{n-1}}$

As mentioned earlier.....

Confidence intervals for a nonzero correlation are smaller than those for zero correlation, and are asymmetric such that the interval toward lower absolute values is larger.

Significance tests against populations with nonzero correlation require the Fisher r-to-Z transformation, whose tables are available in many statistics books.

Temporal degrees of freedom (number of independent time samples) can be less than the number of cases, due to autocorrelation in the data.

To assess the effective degrees of freedom (from Livezey and Chen, 1983, Mon. Wea. Rev.), the time between independent samples is estimated:

$$\text{Integral time} = 1 + 2 \sum_{lag=1}^n (\text{autocor}_{x(lag)}) (\text{autocor}_{y(lag)})$$

Then the effective degrees of freedom is **Total period / integral time**

For example, if there are 20 years of data and the integral time is 1.4 years, then there are  $20/1.4 =$  about 14 degrees of freedom.

Monte Carlo techniques can also be used to estimate temporal degrees of freedom and also spatial degrees of freedom.

Standard error of the slope  $b$  (depends on  $b$  itself, and on the correlation and on sample size  $n$ ):

$$StError(b) = \frac{b}{cor_{xy}} \frac{\sqrt{1 - cor_{xy}^2}}{\sqrt{n - 2}}$$

See if confidence interval around your sample slope, reaching about double (for 2-tailed, 5% level) the  $StError(b)$  on either side of your sample slope, contains zero slope. If so, could have come from population with zero slope (retain null hypoth).

Again, a significance test on the slope should agree with a significance test on the correlation itself.



# Multiple Linear Regression

uses 2 or more predictors

General form:  $z_y = b_1 z_{x_1} + b_2 z_{x_2} + b_3 z_{x_3} + \dots + b_n z_{x_n}$

Let us take simplest multiple regression case -- **two predictors**:

$$z_y = b_1 z_{x_1} + b_2 z_{x_2}$$

Here, the b's are **not simply**  $COR_{x_1,y}$  and  $COR_{x_2,y}$ , unless  $x_1$  and  $x_2$  have zero correlation with one another. Any correlation between  $x_1$  and  $x_2$  makes determining the b's less simple. The b's are related to the **partial correlation**, in which the value of the other predictor(s) is held constant. Holding other predictors constant eliminates the part of the correlation due to the other predictors and not just to the predictor at hand.

Notation: partial correlation of y with  $x_1$ , with  $x_2$  held constant, is written  $COR_{y,x_1.x_2}$

$$z_y = b_1 z_{x_1} + b_2 z_{x_2}$$

For 2 (or any n) predictors, there are 2 (or any n) equations in 2 (or any n) unknowns to be solved simultaneously.

When  $n > 3$  or so, determinant operations are necessary.

For case of 2 predictors, and using z values (variables standardized by subtracting their mean and then dividing by the standard deviation) for simplicity, the solution can be done by hand. The two equations to be solved simultaneously are:

$$\begin{array}{rcl} b_{1.2} & + b_{2.1}(\text{cor}_{x_1,x_2}) & = \text{cor}_{y,x_1} \\ b_{1.2}(\text{cor}_{x_1,x_2}) & + b_{2.1} & = \text{cor}_{y,x_2} \end{array}$$

Goal: to find the two coefficients,  $b_{1.2}$  and  $b_{2.1}$  (called simply  $b_1$  and  $b_2$  in the equation at the top)

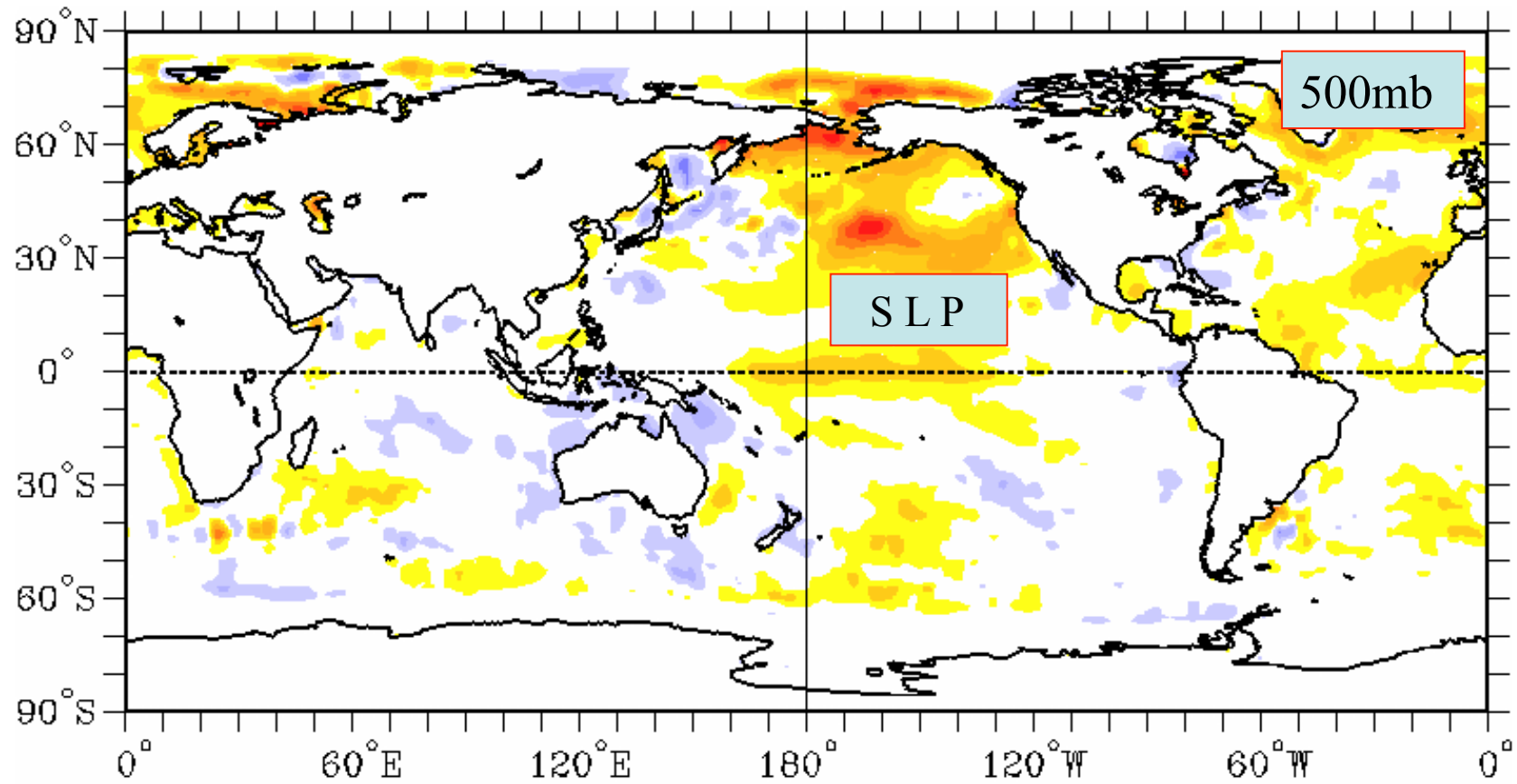
## Example of a multiple regression problem with two predictors

The number of **Atlantic hurricanes** between June and November is slightly predictable 6 months in advance (in early December) using several precursor atmospheric and oceanic variables. Two variables used are:

(1) 500 mb geopotential height in November in the polar north Atlantic (67.5N-85°N latitude, 10E-50°W longitude)

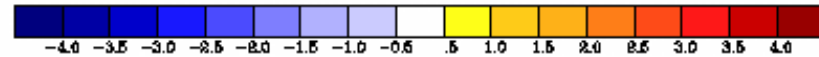
(2) Sea level pressure in November in the North tropical Pacific (7.5N-22.5°N latitude, 125-175°W longitude).

# Location of two long-lead Atlantic hurricane predictor regions



SST ANOM 9/ 5/04-10/ 2/04

Base Period: 1982-96



Physical reasoning behind the two predictors:

(1) **500 millibar geopotential height in November in the polar north Atlantic.** High heights are associated with a negative North Atlantic Oscillation (NAO) pattern, tending to associate with a stronger thermohaline circulation, and also tending to be followed by weaker upper atmospheric westerlies and weaker low-level trade winds in the tropical Atlantic the following hurricane season. All of these favor hurricane activity.

(2) **sea level pressure in November in the North tropical Pacific.** High pressure in this region in winter tends to be followed by La Nina conditions in the coming summer and fall, which favors easterly Atlantic wind anomalies aloft, and hurricane activity.

First step: Find “regular” correlations among all the variables  
( $x_1, x_2, y$ ):  $\text{cor}_{x_1,y}$   $\text{cor}_{x_2,y}$   $\text{cor}_{x_1,x_2}$

$X_1$ : Polar north Atlantic 500 mb height

$X_2$ : North tropical Pacific sea level pressure

$$COR_{Atlantic500mb,hurricanes} = 0.20 (x_1,y)$$

$$COR_{PacificSLP,hurricanes} = 0.40 (x_2,y)$$

$$COR_{Atlantic500mb,PacificSLP} = 0.30 (x_1,x_2) \leftarrow \rightarrow \text{one predictor vs the other}$$

Simultaneous equations to be solved:

$$\begin{aligned} b_{1.2} + (0.30)b_{2.1} &= 0.20 \\ (0.30)b_{1.2} + b_{2.1} &= 0.40 \end{aligned}$$

Want to get one of the predictors to cancel out.

Solution: Multiply 1<sup>st</sup> equation by 3.333, then subtract second equation from first equation:

$$(3.033)b_{1.2} + 0 = 0.267 \quad \text{Dividing by 3.033:}$$

$$b_{1.2} = 0.088. \text{ Then use this in either equation to find } b_{2.1} = 0.374.$$

Then regression equation is  $Z_y = (0.088)z_{x1} + (0.374)z_{x2}$

More detail on solving the two simultaneous equations:

$$\begin{array}{rcl} b_{1.2} & + (0.30)b_{2.1} & = 0.20 \\ (0.30)b_{1.2} & + b_{2.1} & = 0.40 \end{array}$$

Solution: Multiply 1<sup>st</sup> equation by 3.333:

$$(3.333)b_{1.2} + (3.333)(0.30)b_{2.1} = (3.333)(0.20)$$

then the 1<sup>st</sup> equation becomes: Want to get one of the predictors to cancel out. Do it with  $b_{2.1}$

$$3.333 b_{1.2} + (1.0)b_{2.1} = 0.667$$

Now subtract second equation from the new first equation:

$$(3.333 - 0.3)b_{1.2} + (1 - 1)b_{2.1} = 0.667 - 0.40$$

Doing the subtraction yields:

$$3.033b_{1.2} = 0.267$$

Then divide both sides by 3.033:  $b_{1.2} = 0.267 / 3.033 = 0.088$ .

Use this value of  $b_{1.2}$  in either equation, and get  $b_{2.1} = 0.374$ .

Then the regression equation is  $Z_y = (0.088)z_{x1} + (0.374)z_{x2}$

Multiple correlation coefficient =  $R$  = correlation between predicted  $y$  and actual  $y$  using multiple regression.

$$R = \sqrt{b_{1.2} \text{cor}_{x_1 y} + b_{2.1} \text{cor}_{x_2 y}}$$

In example above,  $R = \sqrt{(.088)(.20) + (.373)(.40)} = 0.408$

Note this is only very slightly better than using the second predictor alone in simple regression. This is not surprising, since the first predictor's total correlation with  $y$  is only 0.2, and it is correlated 0.3 with the second predictor, so that the second predictor already accounts for some of what the first predictor has to offer. A decision would probably be made concerning whether it is worth the effort to include the first predictor for such a small gain. **Note: the multiple correlation can never decrease when more predictors are added.**



Multiple R is usually inflated somewhat compared with the true relationship, since additional predictors fit the accidental variations found in the data sample.

Adjustment (decrease) of R for the existence of multiple predictors gives a less biased estimate of R:

$$\text{Adjusted R} = \sqrt{\frac{R^2(n-1) - k}{n - k - 1}} \quad \begin{array}{l} n = \text{sample size} \\ k = \text{number of predictors} \end{array}$$

Sampling variability of a **simple (x, y)** correlation coefficient around zero when population correlation is zero is approximately

$$\sigma_{0-cor} = StError_{(zerocorrel)} = \frac{1}{\sqrt{n-1}}$$

In multiple regression the same approximate relationship holds except that **n must be further decreased**, depending on the number of predictors additional to the first one.

If the number of predictors (x's) is denoted by k, then the sampling variability of R around zero, when there is no true relationship with any of the predictors, is given by

$$\sigma_{0-cor} = StError_{(zerocorrel)} = \frac{1}{\sqrt{n-k}}$$

It becomes easier to get a given multiple correlation by chance as the number of predictors increases.

# Hypothesis test for a multiple correlation value

How can we reject the (null) hypothesis that a multiple correlation value comes from a population having zero correlation?

Standard error of correlation coefficient **with respect to zero correlation** (approximate; slightly too strict for  $n < 10$ ):

$$StError_{(zerocorrel)} = \frac{1}{\sqrt{n - k}} \quad (\text{also called } \sigma_{0-cor})$$

See if your sample correlation (R) equals or exceeds 1.96 (for 2-sided, 5% level) times the above standard error, or 1.65 (for 1-sided, 5% level) times it. **If not, it could have come from a population with zero correlation, with a probability of >5%.**

For multiple correlations, a 1-sided test can be used only when the signs of the correlations between each individual predictor and the predictand (y) are anticipated before the experiment, and when the results confirm those expected correlation signs. (Note: R is always positive.)

## Example of a hypothesis test for a multiple correlation

As a follow-up to the hypothesis test of the positive **rainfall vs. malaria correlation in Botswana** presented in the section on simple regression, suppose we now use both rainfall and temperature as predictors of malaria incidence. We expect greater rainfall to result in greater malaria incidence, but also expect higher temperature to increase incidence, so we use both as predictors in multiple regression.

Suppose for **10 years** of data for rainfall during Nov – March and malaria during the following March – May, using a correlation of **0.64** for rainfall vs. malaria, **0.46** for temperature vs. malaria, and **0.35** for rainfall vs. temperature, we get a multiple correlation of **0.69**. Is this statistically significant in terms of the null hypothesis that the true population multiple correlation is zero? (Could the 0.69 have come about just by chance, due to natural sampling variations among  $x_1$ ,  $x_2$ , and  $y$ , and not due to a physical association involving the combined predictive effects rainfall and temperature, and malaria?)

## Example of a hypothesis test for a multiple correlation

Sample size for rainfall, temperature, and malaria incidence:  $n = 10$

Multiple correlation between (rain, temp) and malaria incidence:  $0.69$

here  $k = 2$

$$\sigma_{0-cor} = StError_{(zerocorrel)} = \frac{1}{\sqrt{n-k}} = \frac{1}{\sqrt{8}} = \frac{1}{2.83} = 0.354$$

We do a 1-sided z test for the 0.69 correlation. 1-sided is justified, given that the correlations between malaria and both climate variables are both positive, as expected on basis of malaria knowledge.

$$z = \frac{SampleCor - 0.00}{StandardError_{zerocorrel}} = \frac{0.69}{0.354} = 1.95$$

Numerator shows correlation difference between sample outcome and population having zero correlation

Looking at the z table, the chance of equaling or exceeding 1.95 is  $0.5 - 0.4744 = 0.0256$ . Significance at the 5% level is achieved.

**Partial Correlation** is correlation between  $y$  and  $x_1$ , where a variable  $x_2$  is not allowed to vary. Example: in an elementary school, reading ability ( $y$ ) is well correlated with the child's weight ( $x_1$ ). But both  $y$  and  $x_1$  are really caused by something else: the child's age (call  $x_2$ ). What would the correlation be between weight and reading ability if the age were held constant? (Would it drop down to zero?)

$$r_{y,x_1.x_2} = \frac{r_{y,x_1} - (r_{y,x_2})(r_{x_1,x_2})}{\sqrt{(1 - r_{y,x_2}^2)(1 - r_{x_1,x_2}^2)}}$$

$$b_1 = r_{y,x_1.x_2} \frac{StErrorEst_{y,x_2}}{StErrorEst_{x_1,x_2}}$$

A similar set of equations exists for  $b_2$  (second predictor).

Suppose the three correlations in a school study are:

$$\text{reading vs. weight : } r_{y,x1} = 0.66$$

$$\text{reading vs. age: } r_{y,x2} = 0.82$$

$$\text{weight vs. age: } r_{x1,x2} = 0.80$$

The two partial correlations come out to be:

$$r_{y,x1.x2} = 0.012$$

$$r_{y,x2.x1} = 0.648$$

Finally, the two regression weights, for standardized variables, turn out to be:

$$b_1 = 0.011$$

$$b_2 = 0.811$$

$$R = 0.820$$

Body weight is seen to be a minor factor compared with age, as its regression weight is near zero.

Suppose a group of people observes an **increase in global temperature** but does not believe it is due to greenhouse gas increases. Instead, they believe that the warming is due to the simple passage of time, as stipulated by their religious doctrine.

To try to judge whether global warming can be attributed more to increases in greenhouse gas concentrations or to the march of time, we do a 2-predictor multiple regression:

$x_1$  = CO<sub>2</sub> concentration (annual average)

$x_2$  = the year number

$y$  = global mean temperature (annual average)



The correlations among  $x_1$ ,  $x_2$  and  $y$ :

CO<sub>2</sub> vs. global temperature: **0.89** ( $x_1$ ,  $y$ )

year vs. global temperature: **0.85** ( $x_2$ ,  $y$ )

CO<sub>2</sub> vs. year: **0.96** ( $x_1, x_2$ )

The two partial correlations come out to be:

$$r_{y, x_1 \cdot x_2} = 0.502$$

$$r_{y, x_2 \cdot x_1} = -0.034$$

Finally, the two regression weights, for standardized variables, turn out to be:

$$b_1 = 0.944$$

$$b_2 = -0.056$$

$$R = 0.890$$

CO<sub>2</sub> concentration is seen to be the dominant predictor.

Suppose the CO<sub>2</sub> vs. year correlation is even higher:

CO<sub>2</sub> vs. global temperature: **0.89** (x<sub>1</sub>, y)

year vs. global temperature: **0.85** (x<sub>2</sub>, y)

**CO<sub>2</sub> vs. year: 0.98** (x<sub>1</sub>, x<sub>2</sub>)

The two partial correlations then come out to be:

$$r_{y,x1.x2} = 0.544$$

$$r_{y,x2.x1} = -0.245$$

Finally, the two regression weights turn out to be:

$$b_1 = 1.439$$

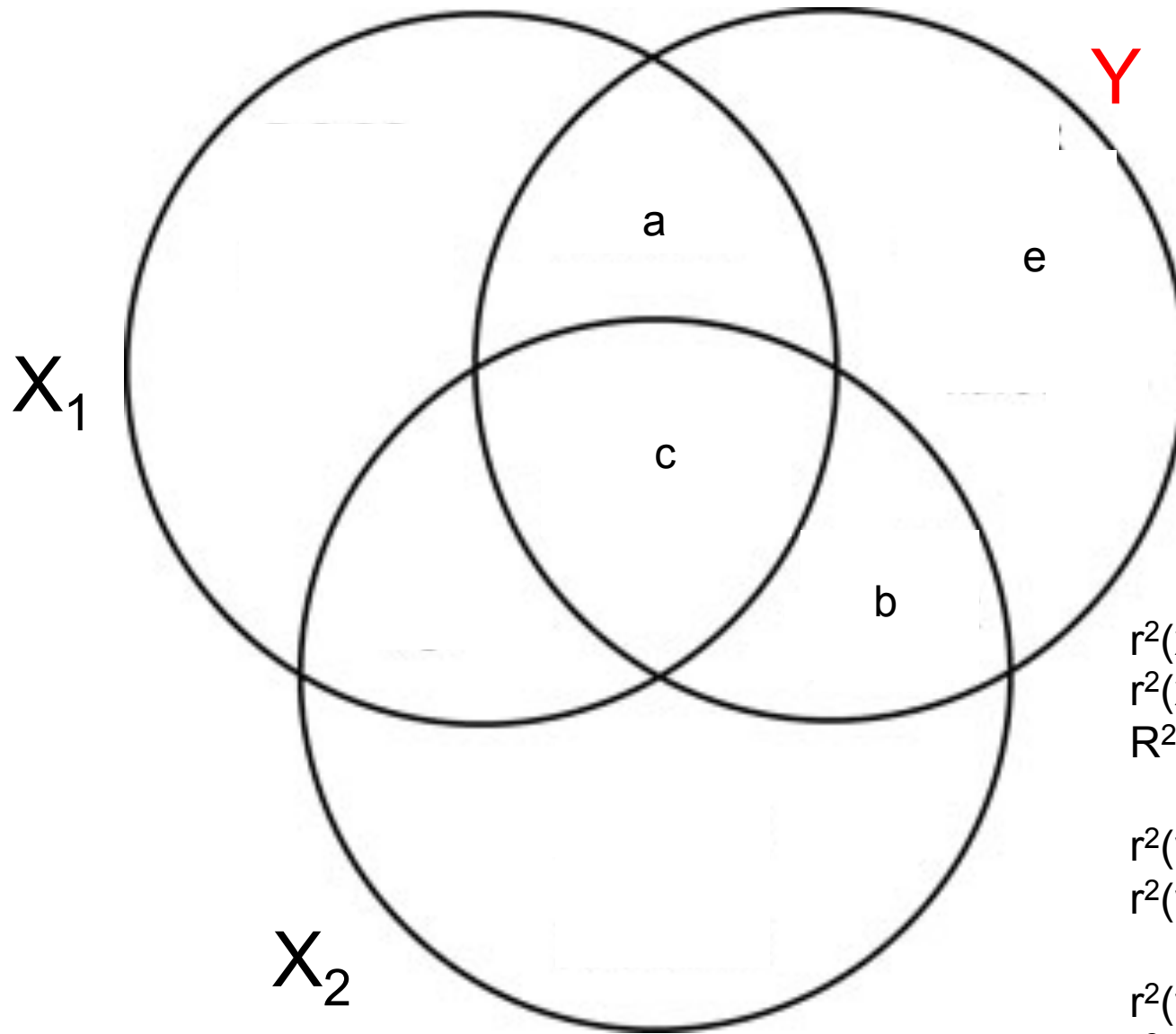
$$b_2 = -0.561$$

$$R = 0.897$$

When two predictors are correlated with one another and  $r(x_2, y) < [r(x_1, y) r(x_1, x_2)]$ , the weight for x<sub>2</sub> becomes signed opposite  $r(x_2, y)$ . (Here it becomes negative instead of positive.) In extreme cases the weights can take on very high magnitudes, and the regression can become unstable and even incalculable.

## Two-predictor multiple regression:

Some examples of behavior of regression weights when  $x_1$ ,  $x_2$  and  $y$  are all standardized to equalize their units



Squared correlations are additive; correlations are not.

$r(y, x_1 \cdot x_2)$  is  $r(x_1, y)$  when  $x_2$  is held constant

Zero-order terms:

$$r^2(x_1, y) = a + c$$

$$r^2(x_2, y) = b + c$$

$$R^2(y, x_1 \& x_2) = a + b + c$$

Semipartials:

$$r^2(y, x_1 \cdot x_2) = a$$

$$r^2(y, x_2 \cdot x_1) = b$$

Partials:

$$r^2(y, x_1 \cdot x_2) = a / (a + e)$$

$$r^2(y, x_2 \cdot x_1) = b / (b + e)$$

In the following 2-predictor examples, colors are used as follows:

Black: **Independence** of predictors:  
Information provided by each is unique.

$$R_{mult} = \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$

$$r(x_1, x_2) = 0$$

Blue: **Partial redundancy** among predictors: Part, but not all, of what  $x_2$  offers is already provided by  $x_1$ . Both coeffs retain original sign.

$$R_{mult} < \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$

$$r(x_2, y) > r(x_1, y)r(x_1, x_2)$$

Green: **Maximum redundancy** among predictors:  $x_2$  adds nothing beyond what is provided by  $x_1$ , so  $x_2$  is useless and has coeff of zero.

$$R_{mult} = \sqrt{r(x_1, y)^2} = r(x_1, y)$$

$$r(x_2, y) = r(x_1, y)r(x_1, x_2)$$

Purple: Redundancy among predictors, but  $r(x_2, y)$  is low (or even zero), and  $x_2$  **beneficially suppresses** a part of  $x_1$  that is unrelated to  $y$ . Coeff of  $x_2$  becomes opposite sign of its simple correlation with  $y$ .

$$r(x_1, y) < R_{mult} < \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$

$$r(x_2, y) < r(x_1, y)r(x_1, x_2)$$

Red: One form of this condition is when the **redundancy is less than the beneficial suppression**, causing  $R$  to exceed that expected for independent predictors. A variation of this is when  $x_1$  and  $x_2$  have **negative redundancy**: e.g.  $r(x_1, y) > 0$ ,  $r(x_2, y) > 0$ ,  $r(x_1, x_2) < 0$

$$R_{mult} > \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$

Effect of the inter-predictor correlation on weights (w) and multiple correlation (R)

$r(x_1, y) = .50$ $w_1$	$r(x_2, y) = .50$ $w_2$	$r(x_1, x_2)$	$R_{\text{mult}}$
<b>.50</b>	<b>.50</b>	<b>.0</b>	<b>0.707</b>
.42	.42	.2	0.645
.36	.36	.4	0.598
.31	.31	.6	0.559
.28	.28	.8	0.527
.26	.26	.9	0.513
.256	.256	.95	0.506
.251	.251	.99	0.501

Independence

Increasing  
redundancy,  
decreasing  
benefit from  
using both  
predictors  
instead of one



Effect of the inter-predictor correlation on weights (w) and multiple correlation (R)

$r(x_1, y) = .50$ $w_1$	$r(x_2, y) = .50$ $w_2$	$r(x_1, x_2)$	$R_{mult}$	
.62	.62	-.2	<b>0.791</b>	Enhancement
.50	.50	.0	<b>0.707</b>	Independence
.42	.42	.2	<b>0.645</b>	Redundancy

$r(x_1, y) = .50$	$r(x_2, y) = -.50$	$r(x_1, x_2)$	$R_{mult}$	
.42	-.42	-.2	<b>0.645</b>	Redundancy
.50	-.50	.0	<b>0.707</b>	Independence
.62	-.62	.2	<b>0.791</b>	Enhancement

When  $r(x_1, x_2) = 0$ ,  $R_{mult} = \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$

Redundancy occurs when  $r(x_1, x_2)$  is of same sign as that of  $[r(x_1, y)] * [r(x_2, y)]$

Enhancement occurs when  $r(x_1, x_2)$  is of sign opposite that of  $[r(x_1, y)] * [r(x_2, y)]$

Effect of the inter-predictor correlation on weights (w) and multiple correlation (R)


$r(x_1, y) = .54$ $w_1$	$r(x_2, y) = .50$ $w_2$	$r(x_1, x_2)$	$R_{\text{mult}}$	
<b>.54</b>	<b>.50</b>	<b>.0</b>	<b>0.736</b>	Independence
.46	.41	.2	0.672	Increasing redundancy, $ry_2 > (ry_1)(r_{12})$  ↓
.41	.34	.4	0.623	
.38	.27	.6	0.583	
.39	.19	.8	0.552	
.47	.07	.9	0.541	
.54	.00	.926*	0.540	$ry_2 = (ry_1)(r_{12})$
.67	-.13	.95	0.542	$ry_2 < (ry_1)(r_{12})$

\*When  $R(x_1, x_2) = .926$ , all information about y provided by  $x_2$  is totally redundant with that carried by  $x_1$ .

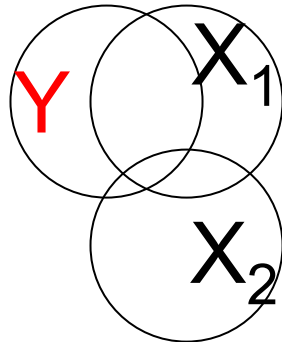
$ry_1$  is  $r(x_1, y)$   
 $ry_2$  is  $r(x_2, y)$   
 $r_{12}$  is  $r(x_1, x_2)$



Effect of the inter-predictor correlation on weights (w) and multiple correlation (R)

$r(x_1, y) = .70$ $w_1$	$r(x_2, y) = .20$ $w_2$	$r(x_1, x_2)$	$R_{\text{mult}}$	
<b>.70</b>	<b>.20</b>	<b>.0</b>	<b>0.728</b>	Independence
.69	.06	.2	0.703	redundancy, $ry_2 > (ry_1)(r_{12})$
70	.00	.286*	0.700	$ry_2 = (ry_1)(r_{12})$
.74	-.10	.4	0.705	$ry_2 < (ry_1)(r_{12})$
.80	-.20	.5	0.721	
.91	-.34	.6	0.752	

\*When  $r(x_1, x_2) = .286$ , all information about y provided by  $x_2$  is totally redundant with that carried by  $x_1$ .



$ry_1$  is  $r(x_1, y)$   
 $ry_2$  is  $r(x_2, y)$   
 $r_{12}$  is  $r(x_1, x_2)$

Effect of the inter-predictor correlation on weights (w) and multiple correlation (R)

$r(x_1, y) = .50$ $w_1$	$r(x_2, y) = .00$ $w_2$	$r(x_1, x_2)$	$R_{\text{mult}}$
.78	.47	-.6	0.625
.60	.24	-.4	0.646
.52	.10	-.2	0.510
.50	.00	.0	0.500
.52	-.10	.2	0.510
.60	-.24	.4	0.546
.78	-.47	.6	0.625

↑

Increasing beneficial suppression, enhancement

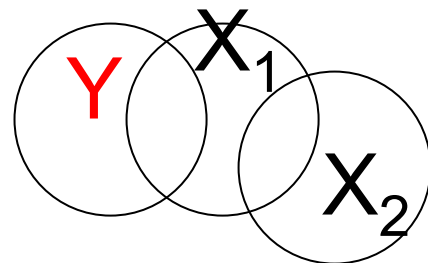
---

“Independence”,  
 $r_{y2} = (r_{y1})(r_{12})$

---

Increasing beneficial suppression, enhancement

↓



$X_2$  plays a role even though its correlation with  $y$  is zero.

$r_{y1}$  is  $r(x_1, y)$   
 $r_{y2}$  is  $r(x_2, y)$   
 $r_{12}$  is  $r(x_1, x_2)$

In the preceding 2-predictor examples, colors were used as follows:

Black: **Independence** of predictors:  
Information provided by each is unique.

$$R_{mult} = \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$
$$r(x_1, x_2) = 0$$

Blue: **Partial redundancy** among predictors: Part, but not all, of what  $x_2$  offers is already provided by  $x_1$ . Both coeffs retain original sign.

$$R_{mult} < \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$
$$r(x_2, y) > r(x_1, y)r(x_1, x_2)$$

Green: **Maximum redundancy** among predictors:  $x_2$  adds nothing beyond what is provided by  $x_1$ , so  $x_2$  is useless and has coeff of zero.

$$R_{mult} = \sqrt{r(x_1, y)^2} = r(x_1, y)$$
$$r(x_2, y) = r(x_1, y)r(x_1, x_2)$$

Purple: Redundancy among predictors, but  $r(x_2, y)$  is low (or even zero), and  $x_2$  **beneficially suppresses** a part of  $x_1$  that is unrelated to  $y$ . Coeff of  $x_2$  becomes opposite sign of its simple correlation with  $y$ .

$$r(x_1, y) < R_{mult} < \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$
$$r(x_2, y) < r(x_1, y)r(x_1, x_2)$$

Red: One form of this condition is when the **redundancy is less than the beneficial suppression**, causing  $R$  to exceed that expected for independent predictors. A variation of this is when  $x_1$  and  $x_2$  have **negative redundancy**: e.g.  $r(x_1, y) > 0$ ,  $r(x_2, y) > 0$ ,  $r(x_1, x_2) < 0$

$$R_{mult} > \sqrt{r(x_1, y)^2 + r(x_2, y)^2}$$

Various combinations of the above behaviors are likely to appear when there are 3 or more predictors.

Modelers *should not be insulted* when their model is assigned a negative weight in multiple regression!

When there are MANY predictors, and not very many time samples to develop the model, collinearity can become so severe that some of the weights have very high magnitudes. This is dangerous with respect to the addition of new cases.

**“Ridging” can ease this problem.**

**Ridging** is the addition of small amounts on the diagonal of the cross-correlation matrix; is like adding noise. It has the effect of reducing the cross-correlations among the models.

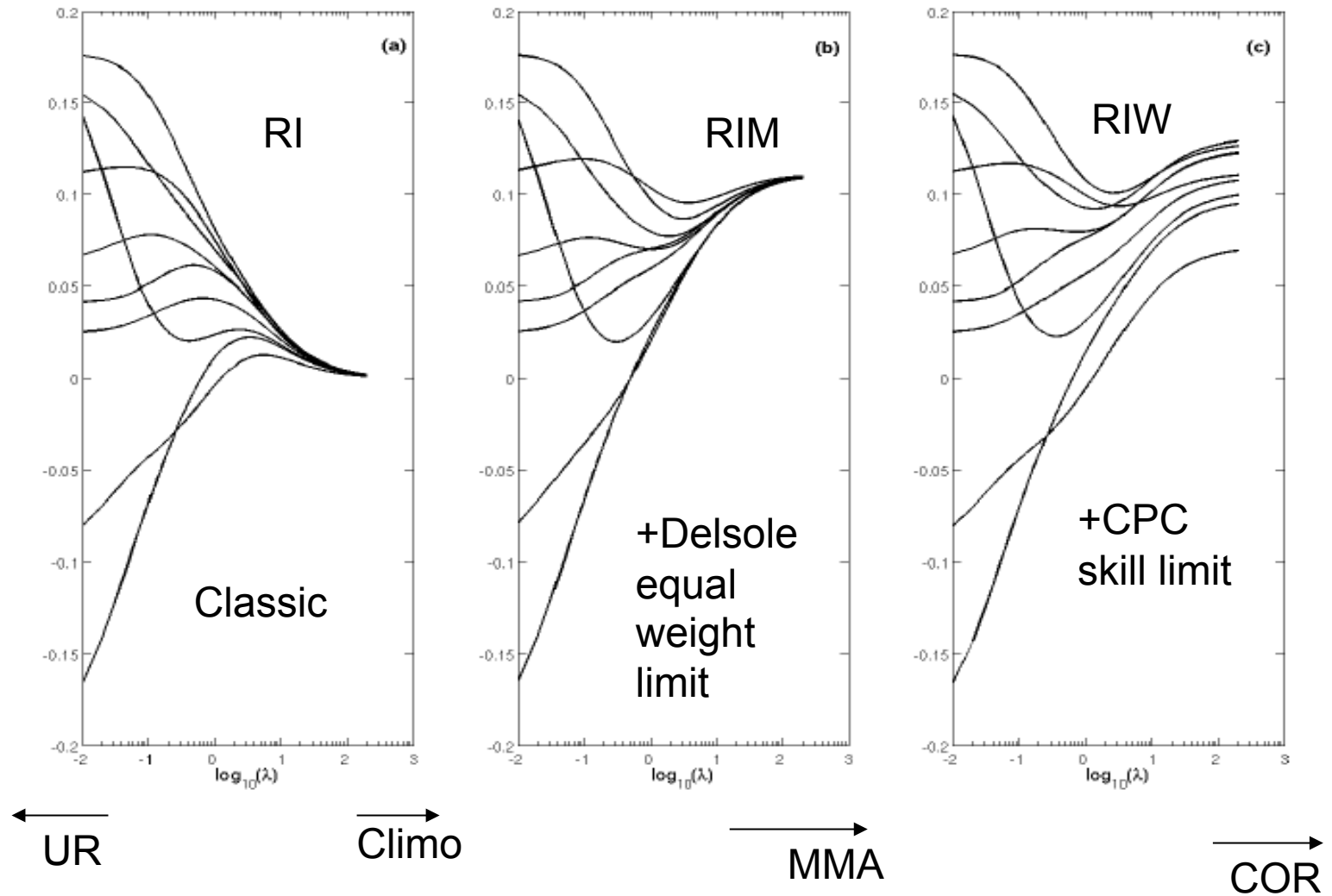


Figure 1.  weights for ensemble averages corresponding to the 9 models in the DEMETER+PLUS data as a function of ridging amount ( $\lambda$  in Log<sub>10</sub> scale) for (a) ridging, RID, (b) ridging with departure-from-equal-weight penalty, RIM, and (c) ridging with departure-from-skill-based-weights penalty, RIW.

## Another Example – Sahel Drying Trend

Suppose 50 years of climate data suggest that the drying of the Sahel in northern Africa in July to September may be related to **warming in the tropical Atlantic and Indian oceans ( $x_1$ )** as well as local **changes in land use in the Sahel itself ( $x_2$ )**.  $x_1$  is expressed as SST, and  $x_2$  is expressed as percentage vegetation decrease (expressed as a negative percentage) from the vegetation found at the beginning of the 50 year period. Both factors appear related to the downward trend in rainfall (in opposite sense), and the two predictors are negatively correlated. Suppose the correlations come out as follows:

$$\text{Cor}(y, x_1) = -0.52 \quad \text{Cor}(y, x_2) = 0.37 \quad \text{Cor}(x_1, x_2) = -0.50$$

What would be the multiple regression equation in “unit-free” standard deviation ( $z$ ) units?

$$\text{Cor}(x_1, y) = -0.52 \quad \text{Cor}(x_2, y) = 0.37 \quad \text{Cor}(x_1, x_2) = -0.50$$

First we set up the two equations to be solved simultaneously

$$\begin{aligned} b_{1.2} &+ b_{2.1}(\text{cor}_{x_1, x_2}) &= \text{cor}_{y, x_1} \\ b_{1.2}(\text{cor}_{x_1, x_2}) &+ b_{2.1} &= \text{cor}_{y, x_2} \end{aligned}$$

$$\begin{aligned} b_{1.2} &+ (-0.50)b_{2.1} &= -0.52 \\ (-0.50)b_{1.2} &+ b_{2.1} &= 0.37 \end{aligned}$$

Want to eliminate (or cancel)  $b_{1.2}$  or  $b_{2.1}$ . To eliminate  $b_{2.1}$ , multiply first equation by 2 and add second one to it:

$$1.5 b_{1.2} = -0.67 \quad \text{and} \quad b_{1.2} = -0.447 \quad \text{and} \quad b_{2.1} = 0.147$$

Regression equation is  $Z_y = -0.447 z_{x_1} + 0.147 z_{x_2}$

Regression equation is  $Z_y = -0.447z_{x_1} + 0.147z_{x_2}$

If want to express the above equation in physical units, then must know the means and standard deviations of  $y$ ,  $x_1$  and  $x_2$  and make substitutions to replace the  $z$ 's.

$$y = \bar{y} + z_y SD_y$$

$$z_y = (y - \bar{y}) / SD_y$$

$$x_1 = \bar{x}_1 + z_{x_1} SD_{x_1}$$

$$z_{x_1} = (x_1 - \bar{x}_1) / SD_{x_1}$$

$$x_2 = \bar{x}_2 + z_{x_2} SD_{x_2}$$

$$z_{x_2} = (x_2 - \bar{x}_2) / SD_{x_2}$$

When substitute and simplify results,  $y$ ,  $x_1$  and  $x_2$  terms will appear instead of  $z$  terms. There generally will also be a constant term that is not found in the  $z$  expression because the original variables probably do not have means of 0 the way  $z$ 's always do.



The means and the standard deviations of the three data sets are

y: Jul-Aug-Sep Sahel rainfall (mm): mean 230 mm, SD 88 mm

x<sub>1</sub>: Tropical Atlantic/Indian ocean SST: mean 28.3 C, SD 1.7 C

x<sub>2</sub>: Deforestation (percent of initial): mean 34%, SD 22%

$$Z_y = -0.447z_{x_1} + 0.147z_{x_2}$$

$$\frac{(y - \bar{y})}{SD_y} = -0.447 \frac{(x_1 - \bar{x}_1)}{SD_{x_1}} + 0.147 \frac{(x_2 - \bar{x}_2)}{SD_{x_2}}$$

$$\frac{(y - 230)}{88} = -0.447 \frac{(x_1 - 28.3)}{1.7} + 0.147 \frac{(x_2 - 34)}{22}$$

After term collection and algebraic simplification, final form will be:

$$y = \underset{b_1}{\text{coeff}} x_1 + \underset{b_2}{\text{coeff}} x_2 + \text{constant} \quad (\text{here, } b_1 < 0, b_2 > 0)$$

We now compute the multiple correlation R, and the standard error of estimate for the multiple regression.

Using the two individual correlations and the b terms:

$$\text{Cor}(x_1, y) = -0.52 \quad \text{Cor}(x_2, y) = 0.37 \quad \text{Cor}(x_1, x_2) = -0.50$$

$$\text{Regression equation is } Z_y = -0.447 z_{x1} + 0.147 z_{x2}$$

$$R = \sqrt{b_{1.2} \text{cor}_{x_1 y} + b_{2.1} \text{cor}_{x_2 y}}$$

$$R = \sqrt{(-.447)(-.52) + (.147)(.37)} = 0.535$$

The deforestation factor helps the prediction accuracy only a bit. If there were weaker negative correlation between the two predictors, then the second predictor would be more valuable.

$$\text{Standard Error of Estimate} = \sqrt{1 - R^2_{y,(x_1 x_2)}} = 0.845$$

$$\text{In physical units, it is } (88 \text{ mm}) (0.845) = 74.3 \text{ mm}$$

**Multiple regression with three or more predictors** is just an extension of two-predictor multiple regression.

Matrix math takes over (very tedious to do by hand) but the ideas are entirely the same; is quick on computer.

Partials and semipartials become with respect to *all* of the predictors other than the given one.

The shrinkage formula Adjusted R = 
$$\sqrt{\frac{R^2(n-1) - k}{n - k - 1}}$$

can result in huge shrinkage when number of predictors (k) is more than two-thirds of the sample size number (n), and very high R is required to show good prediction skill.

# 3-predictor multiple regression formulas

---

- Regression equation

$$\text{Standardized units form: } z_y = b_{1z}z_{x1} + b_{2z}z_{x2} + b_{3z}z_{x3}$$

$$\text{Physical units form: } y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

---

- Correlation coefficient

$$R = \sqrt{b_{1z}\text{cor}_{y,x1} + b_{2z}\text{cor}_{y,x2} + b_{3z}\text{cor}_{y,x3}}$$

---

- Adjusted correlation coefficient

$$R_{adjusted} = \sqrt{\frac{R^2(n-1) - k}{n-k-1}}$$

(n is sample size, k is number of predictors)

---

- Standard error of estimate

$$\text{In z unit: } \sqrt{1 - R^2_{adjusted}}$$

$$\text{In physical unit: } SD_y \sqrt{1 - R^2_{adjusted}}$$

---

- Statistical significance of a correlation

$$z = \frac{R_{adjusted}}{SE_{(zerocorrel)}}, \text{ where } SE_{(zerocorrel)} = \frac{1}{\sqrt{n-k}}$$

## Summary

Regression is a very useful prediction tool. Minimizing squared errors when several predictors are involved, using automation, is very powerful.

Regression in simplest form uses a single predictor.

Two-predictor regression adds more utility, can still be done by hand, and illustrates several awesome behaviors that will extend to cases of  $>2$  predictors.

Not only can many predictors be used, but many predictands also can (CCA, etc.) Use of EOFs in regression provides yet additional fuel for better understanding of some of the physical processes involved.

**WARNING:** Overfitting and artificial skill can fool us!