

# Simple Regression

- correlation vs. prediction research
- $y' = bx + a$
- prediction and relationship strength
- interpreting regression formulas
  - quantitative vs. binary predictor variables
  - raw score vs. standardized formulas
- selecting the correct regression model
- regression as linear transformation (how it works!)
- “equally predictive” vs. “equivalent predictors”
- process of a prediction study

## Correlation Studies vs. Prediction Studies

### Correlation research (95%)

- purpose is to identify the direction and strength of linear relationship between two quantitative variables
- usually theoretical hypothesis-testing interests

### Prediction research (5%)

- purpose is to take advantage of linear relationships between quantitative variables to create (linear) models to predict values of hard-to-obtain variables from values of available variables
- use the predicted values to make decisions about people (admissions, treatment availability, etc.)

Also, to fully understand important things about the correlation models requires a good understanding of the regression model upon which prediction is based...



## Linear regression for prediction...

- linear regression “assumes” there is a linear relationship between the variables involved
  - “if two variables aren’t linearly related, then you can’t use one as the basis for a linear prediction of the other”
  - “a significant correlation is the minimum requirement to perform a linear regression”
- sometimes even a small correlation can lead to useful prediction (if it is not a Type I error)
- must have a “meaningful criterion” in order to obtain a useful prediction formula

Predictors, predicted criterion, criterion and residuals

Here are two formulas that contain “all you need to know”

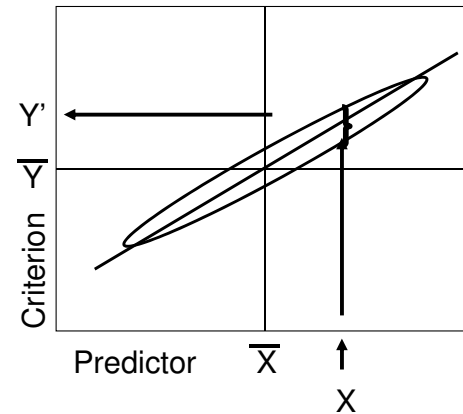
$$y' = bx + a \quad \text{residual} = y - y'$$

- x the predictor -- variable related to criterion that you will use to make an estimate of criterion value for each participant
- y' the predicted criterion value -- “best guess” of each participant’s y value, based on their x value --that part of the criterion that is related to (predicted from) the predictor
- y the criterion -- variable you want to use to make decisions, but “can’t get” for each participant (time, cost, ethics)
- residual difference between criterion and predicted criterion values -- the part of the criterion not related to the predictor -- the stronger the correlation the smaller the residual (on average)

Let’s take a look at the relationship between the strength of the linear relationship and the accuracy of linear prediction.

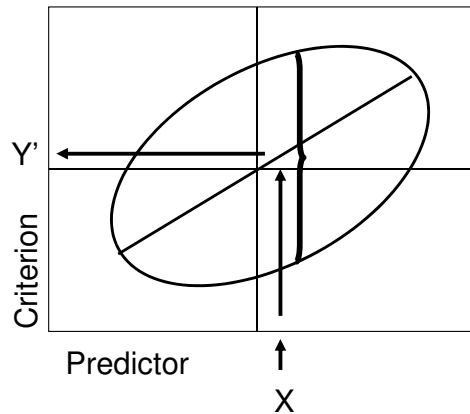
- graphic equivalent of  $y' = bx + a$
- for a given value of X
- project up to the regression line
- project over to the y' value

When the linear relationship is very strong, there is a narrow range of Y values for any X value, and so the Y' “guess” will be close



Notice that everybody with the same X score will have the same predicted Y score.

There won't be much error, though, because there isn't much variability among the Y scores for any given X score.



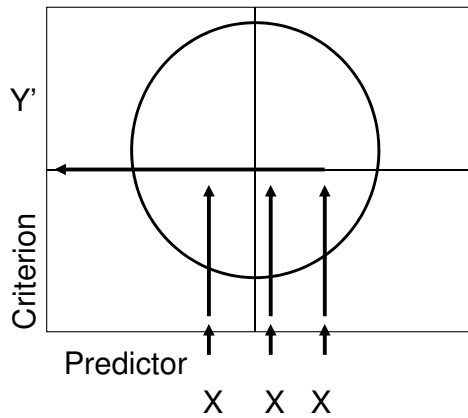
However, when the linear relationship is very weak, there is a wide range of Y values for any X value, and so the Y' “guess” will be less accurate, on the average.

There is still some utility to the linear regression, because larger values of X still “tend to” go with larger values of Y.

Notice that everybody with the same X score will have the same predicted Y score.

Now there will be more error, because there is a lot of variability among the Y scores for any given X score.

So the linear regression might supply useful information, even if it isn't very precise -- depending upon what is “useful”?



When there is no linear relationship, everybody has the same predicted Y score – the mean of Y.

This is known as “univariate prediction” – when we don’t have a working predictor, our best guess for each individual is that they will have the group mean.

Some key ideas we have seen are:

- everyone with a given “X” value will have the same predicted “Y” value
- if there is no (statistically significant & reliable) linear relationship, then there is no basis for linear prediction (bivariate prediction)
- the stronger the linear relationship, the more accurate will be the linear prediction (on the average)



## Simple regression

$$y' = bx + a \quad \text{raw score form}$$

b -- raw score regression slope or coefficient

a -- regression constant or y-intercept

For a quantitative predictor

a = the expected value of y if x = 0

b = the expected direction and amount of change in y for a 1-unit change in x

For a binary x with 0-1 coding

a = the mean of y for the group coded “0”

b = the direction and amount of difference in the mean of y between the group coded “0” and the group coded “1”

Let’s practice -- quantitative predictor ...

#1  $\text{discom}' = (2.5 * \text{stress}) + 23$

apply the formula -- patient has stress score of 10  $\text{discom}' = 48$

interpret “b” -- for each 1-unit increase in stress, discomfort is expected to increase by 2.5

interpret “a” -- if a person has a stress score of “0”, their expected discomfort score is 23

#2  $\text{job errors} = (-6 * \text{interview score}) + 95$

apply the formula -- applicant has interview score of 10, expected number of job errors is 35

interpret “b” -- for each 1-unit increase in intscore, errors are expected to decrease by 6

interpret “a” -- if a person has a interview score of “0”, their expected number of job errors is 95

Let's practice -- binary predictor ...

$$\#1 \text{ discomfort}' = (7.5 * \text{tx group}) + 15.0 \quad \text{code: Tx}=1 \text{ Cx}=0$$

interpret "b" -- the Tx group has mean 7.5 more than Cx

interpret "a" -- mean of the Cx group (code=0) is 15

so ... mean of Tx group is 22.5

$$\#2 \text{ job errors} = (-2.0 * \text{job}) + 8 \quad \text{code: mgr}=1 \text{ sales}=0$$

the mean # job errors of the sales group is 8

the mean difference # job errors between the groups is -2

the mean # of job errors of the mgr group is 6

$$\text{Linear regression as linear transformations: } y' = bX + a$$

this formula is made up of two linear transformations --

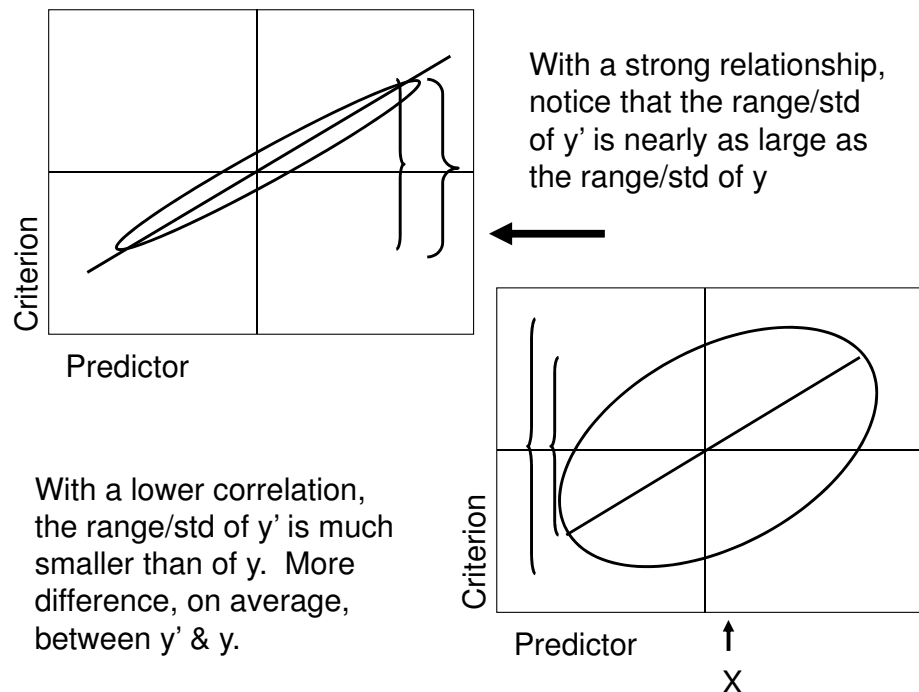
$bX$  = a multiplicative transformation that will change the standard deviation and mean of  $X$

$+a$  = an additive transformation which will further change the mean of  $X$

A good  $y'$  will be a "mimic" of  $y$  -- each person having a value of  $y'$  as close as possible to their actual  $y$  value.

This is accomplished by "transforming"  $X$  into  $Y$  with the mean and standard deviation of  $y'$  as close as possible to the mean and standard deviation of  $Y$

- First, the value of  $b$  is chosen to get the standard deviation of  $y'$  as close as possible to  $y$  -- this works better or poorer depending upon the strength of the  $x,y$  linear relationship (see next slide).
- Then, the value of  $a$  is chosen to get the mean of  $y'$  to match the mean of  $Y$  -- this always works exactly --  $\text{mean } y' = \text{mean } Y$ .



Selecting the proper regression model (predictor & criterion)

For any correlation between two variables (e.g., SAT and GPA) there are two possible regression formulas

- depending upon which is the Criterion & which is the Predictor

crit	regression	predictor
GPA'	=	b(SAT) + a
SAT'	=	b(GPA) + a

(Note: the b and a values are NOT interchangeable between the two models)

The criterion is the variable that “we want a value for but can’t have” (because “hasn’t happened yet”, cost or ethics).

The predictor is the variable that “we have a value for”.

standard score regression  $Z_y' = \beta Z_x$

Basically, this is regression done with both Y & X as Z-scores.

for a quantitative predictor  $\beta$  tells.. if  $Z_x$  changes by 1 unit (1 std), but how much of a unit (std) is  $Z_y'$  expected to change?

for a binary predictor,  $\beta$  tells size/direction of group mean difference on criterion variable in Z-units (useful?)

Why no “a”

The mean of  $Z_x = 0$ . So, the mean of  $\beta Z_x = 0$ , which mimics the mean of  $Z_y' = 0$  (without any correction).

Which regression model to use, raw or standardized?

- depends upon the predictor data you have ...
  - Have raw predictor scores → use the raw score model
  - Have standardized scores → use the standardized model

## GRE & SAT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.525 <sup>a</sup>	.275	.272	91.098

a. Predictors: (Constant), SAT

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	75.039	1	75.039	90.421	.000 <sup>a</sup>
	Residual	197.513	238	.830		
	Total	272.552	239			

a. Predictors: (Constant), SAT  
b. Dependent Variable: 1st year gpa (on 1-6 scale)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.472	.371		-1.272	.205
	SAT	.006	.001	.525	9.509	.000

a. Dependent Variable: 1st year gpa (on 1-6 scale)

Descriptive Statistics

	N	Mean	Std. Deviation
SAT	240	574.3200	92.34360
1st year gpa (on 1-6 scale)	240	3.0128	1.06789
Valid N (listwise)	240		

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.525 <sup>a</sup>	.275	.272	78.77534

a. Predictors: (Constant), 1st year gpa (on 1-6 scale)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56112.460	1	56112.460	90.421	.000 <sup>a</sup>
	Residual	1476921.910	238	6205.554		
	Total	2038034.370	239			

a. Predictors: (Constant), 1st year gpa (on 1-6 scale)  
b. Dependent Variable: SAT

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	437.618	15.249		28.698	.000
	1st year gpa (on 1-6 scale)	45.373	4.772	.525	9.509	.000

a. Dependent Variable: SAT

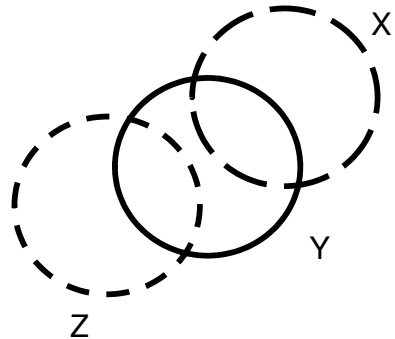
- b is influenced by **both** the correlation and the relative scales of the predictor and criterion
- $\beta$  is not influenced by relative scale – next page!



## “equally predictive predictors” vs. “equivalent predictors”

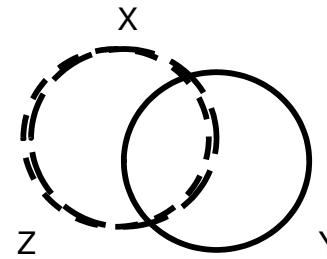
When we compare predictors we must consider both:

- “how much” of Y each predictor predicts &
- “what part” of Y each predictor predicts



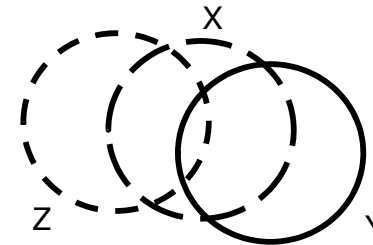
X & Z are “equally predictive”

But they are not “equivalent predictors” because they predict “different parts” of Y



X & Z are “equally predictive”

X & Z are also “equivalent predictors” because they predict “the same parts” of Y (as best I can draw)



X & Z are not “equally predictive”

“everything Z can do X can do better”

## Predictors: “equally predictive” vs. “equivalent predictors”

If two predictors are correlated  $\pm 1.00$  with each other...

- they will be equally correlated with any criterion
- they will produce the same  $y$ 's and residuals as each other

If two predictors are correlated  $< \pm 1.00$  with each other ...

- they may be differentially correlated with any criterion
- they may produce different  $y$ 's and residuals
- which participants give higher/lower  $y$ ' values may differ
- equally predictive predictors might not predict the same  $y$ ' values for all or any individuals !!!!

So...

... a person's predicted criterion score ( $y'$ ) may differ depending upon which predictor is used



How we describe our correlation/regression findings & results...

Originally, regression was developed to analyze data from randomized experiments...

- the predictor was a randomly assigned, manipulated quantitative variable (eg, amount of fertilizer/acre)
- the criterion was a quantitative variable (eg yield/acre)
- all other variables were constants (rain fall) or control variables (soil chemistry)

The correlation analysis tests whether or not “more is better,” but more precise information was wanted!!!

- What yield can be expected for a given amount of fertilizer?
- How much is yield improved by using more fertilizer?

Both answered by the regression model  $\rightarrow y' = bx + a$

Since this is a true experiment with good experimental controls, we are likely to get very useful answers to both questions!!

HOWEVER...

Most of the time we won't have data from well-controlled true experiments, so we are limited to making associative interpretations of our correlations and our regression models.

However, we “really want to understand cause” and, so, we have to be really careful about the language we use to represent our “associations” !!

Identifying “just a correlate” is not nearly as cool as identifying “a cause”!!

So watch out for others' and your own “becauseing” of associations!

Important to remember... Conducting a Prediction Study...

This is a **2-step process**

Step 1 -- using the “Modeling Sample” which has values for both the predictor and criterion.

- Determine that there is a significant linear relationship between the predictor and the criterion.
- If there is an appreciable and significant correlation, then build the regression model (find the values of b and a)

Step 2 -- using the “Application Sample” which has values for only the predictor.

- Apply the regression model, obtaining a  $y'$  value for each member of the sample