Bivariate & Multiple Regression

- correlation vs. prediction research
- prediction and relationship strength
- interpreting regression formulas
- selecting the correct regression model
- regression as linear transformation (how it works!)
- process of a prediction study
- Advantages of multiple regression
- Parts of a multiple regression model & interpretation
- Raw score vs. Standardized models
- Differences between r, b_{biv}, b_{mult} & β_{mult}

Correlation Studies and 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.)

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

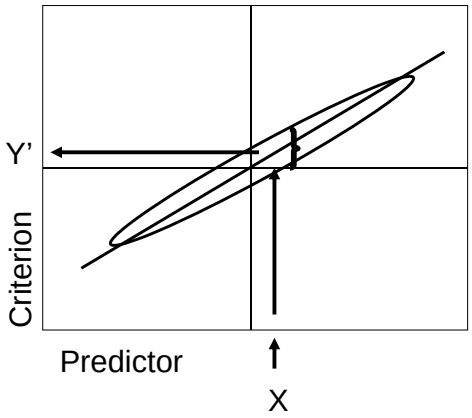
Linear regression for prediction...

- 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

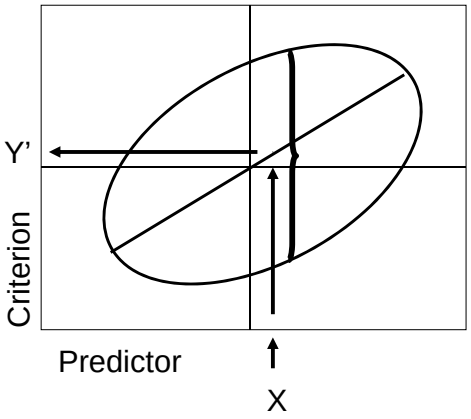
Story time
- Hotel Manager prediction story (.2 = $)
- "Which criterion" sales prediction story

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

- for a given value of X
- draw up to the regression line
- draw over the predicted value of Y



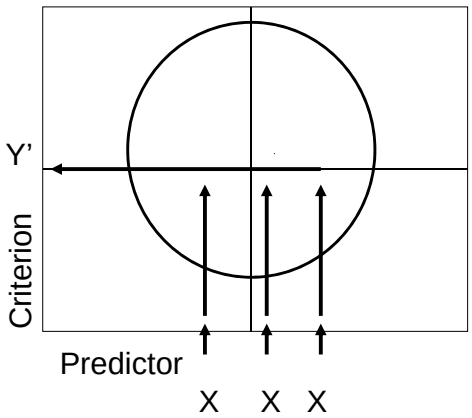
When the linear relationship is very strong, there is a narrow range of Y values for any X value, and so the Y' "guess" will be close



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.

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



However, when there is no linear relationship,not only is there a very wide range of Y values for any X value, but all X values lead to the SAME Y value estimate (the mean of Y)

Some key ideas 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
- the stronger the linear relationship, the more accurate will be the linear prediction (on the average)

Predictors, predicted criterion, criterion and residuals
Here are two formulas that contain "all you need to know"

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

y the criterion -- variable you want to use to make decisions, but "can't get" for each participant (time, cost, ethics)

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

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)

Simple regression

$$y' = bx + a$$ raw score form

a -- regression constant or y-intercept

- for a quantitative predictor = the expected value of y if x = 0
- for a binary x with 0-1 coding = the mean of y for the group with the code value = 0

b -- raw score regression slope or coefficient

- for a quantitative predictor = the expected change (direction and amount) in the criterion for a 1-unit change in the predictor
- for a binary x with 0-1 coding = the mean y difference between the two coded groups

Let's practice -- quantitative predictor ...

#1 depression' = (2.5 * stress) + 23

apply the formula -- patient has stress score of 10 dep' = 48

interpret "b" -- for each 1-unit increase in stress, depression is expected to increase by 2.5

interpret "a" -- if a person has a stress score of "0", their expected depression score is 23

#2 job errors = (-6 * 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 depression'=(7.5 * tx group) +15.0 code: Tx=1 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 job errors = (-2.0 * job) + 8 code: mgr=1 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

Selecting the proper regression model (predictor & criterion)

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

-- depending upon which is the Criterion and Predictor

criterion		predictor
GRE'	=	b(GPA) + a
GPA'	=	b(GRE) + 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".

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.

Then, the value of a is chosen to get the mean of y' to match the mean of Y -- this always works exactly -- mean y' = mean Y.

Let's consider models for predicting GRE and GPA

Each GRE scale has mean = 500 and std = 100

GPA usually has a mean near 3.2 and std near 1.0

say we want to predict GRE from GPA GRE' = b(GPA) + a

- we will need a very large b-value -- to transform GPA with a std of 1 into GRE' with a std of 100

but, say we want to predict GPA from GRE GPA' = b(GRE) +a

- we will need a very small b-value -- to transform GRE with a std of 100 into GPA' with a std of 1

Obviously we can't use these formulas interchangeably -- we have to properly determine which variable is the criterion and which is the predictor and obtain and use the proper formula!!!

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

Tell the Pepperidge Farm story !

Advantages of Multiple Regression

Practical issues …

- better prediction from multiple predictors
- can "avoid" picking/depending on a single predictor
- can "avoid" non-optimal combinations of predictors (e.g., total scores)

Theoretical issues …

- even when we know in our hearts that the design will not support causal interpretation of the results, we have thoughts and theories of the causal relationships between the predictors and the criterion -- and these thoughts are about multi-causal relationships
- multiple regression models allow the examination of more sophisticated research hypotheses than is possible using simple correlations
- gives a "link" among the various correlation and ANOVA models

raw score regression $y' = b_1x_1 + b_2x_2 + b_3x_3 + a$

each b

- represents the unique and independent contribution of that predictor to the model
- for a quantitative predictor tells the expected direction and amount of change in the criterion for a 1-unit change in that predictor, while holding the value of all the other predictors constant
- for a binary predictor (with unit coding -- 0,1 or 1,2, etc.), tells direction and amount of group mean difference on the criterion variable, while holding the value of all the other predictors constant

a

- the expected value of the criterion if all predictors have a value of 0

Let's practice -- Tx (0 = control, 1 = treatment)

depression' = (2.0 * stress) - (1.5 * support) - (3.0 * Tx) + 35

- apply the formula patient has stress score of 10, support score of 4 and was in the treatment group dep' = 46
- interpret "b" for stress -- for each 1-unit increase in stress, depression is expected to increase by 2 , when holding all other variables constant
- interpret "b" for support -- for each 1-unit increase in support, depression is expected to decrease by 1.5 , when holding all other variables constant
- interpret "b" for tx – those in the Tx group are expected to have a mean depression score that is 3.0 lower than the control group, when holding all other variables constant
- interpret "a" -- if a person has a score of "0" on all predictors, their depression is expected to be 35

standard score regression $Z_y' = \beta Z_{x1} + \beta Z_{x2} + \beta Z_{x3}$

The most common reason to refer to standardized weights is when you (or the reader) is unfamiliar with the scale of the criterion. A second reason is to promote comparability of the relative contribution of the various predictors (but see the important caveat to this discussed below!!!).

It is important to discriminate among the information obtained from ...

bivariate r & bivariate regression model weights

r -- simple correlation

tells the direction and strength of the linear relationship between two variables (r = β for bivariate models)

b -- raw regression weight from a bivariate model

tells the expected change (direction and amount) in the criterion for a 1-unit change in the predictor

It is important to discriminate among the information obtained from ...

multivariate R & multivariate regression model weights

R^2 -- squared multiple correlation

tells how much of the Y variability is "accounted for,"

. "predicted from" or "caused by" the multiple regression model

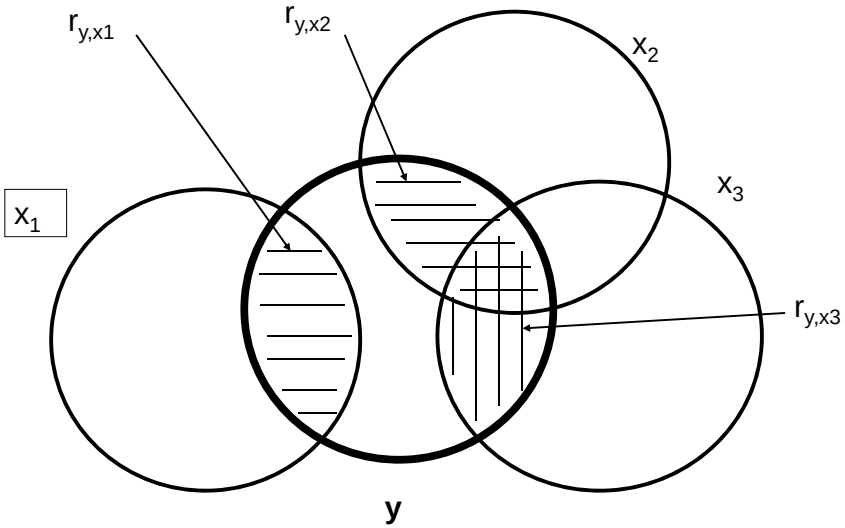
b_i -- raw regression weight from a multivariate model

tells the expected change (direction and amount) in the criterion for a 1-unit change in that predictor, holding the value of all the other predictors constant

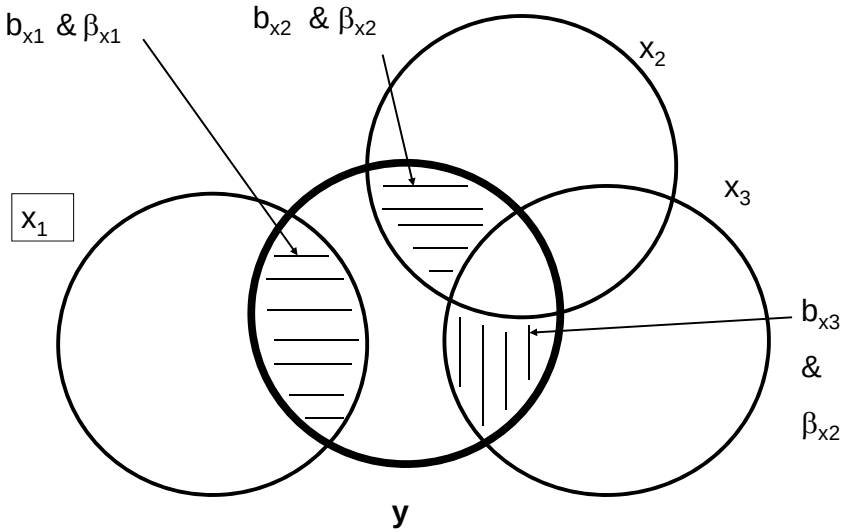
β_i -- standardized regression wt. from a multivariate model

tells the expected change (direction and amount) in the criterion in Z-score units for a 1-Z-score unit change in that predictor, holding the value of all the other predictors constant

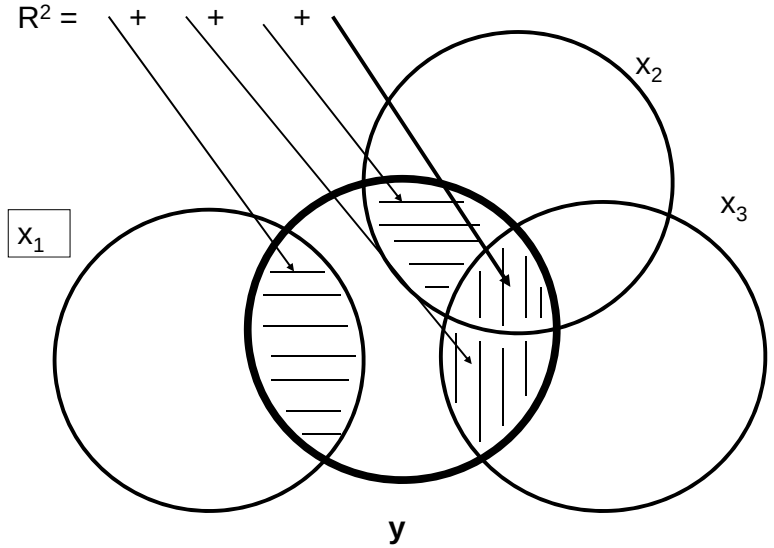
Venn diagrams representing r, b and R^2



Remember that the b of each predictor represents the part of that predictor shared with the criterion that is not shared with any other predictor -- the unique contribution of that predictor to the model



Remember R^2 is the total variance shared between the model (all of the predictors) and the criterion (not just the accumulation of the parts uniquely attributable to each predictor).



Model Specification & why it matters !!!

What we need to remember is that we will never, ever (even once) have a "properly specified" multiple regression model → one that includes all of & only the causal variables influencing the criterion !

What can we do about "misspecification" ?

• running larger models with every available predictor in them won't help – models with many predictors tend to get really messy

• our best hope is to base our regression models upon the existing literature & good theory and to apply programmatic researc

Proxy variables

Remember (again) we are not going to have experimental data!

The variables we have might be the actual causal variables influencing this criterion, or (more likely) they might only be correlates of those causal variables – proxy variables

Many of the "subject variables" that are very common in multivariate modeling are of this ilk…

- is it really "personality," "ethnicity", "age" that are driving the criterion – or is it all the differences in the experiences, opportunities, or other correlates of these variables?
- is it really the "number of practices" or the things that, in turn, produced the number of practices that were chosen?

Again, replication and convergence (trying alternative measure of the involved constructs) can help decide if our predictors are representing what we think the do!!

Proxy variables

In sense, proxy variables are a kind of "confounds" → because we are attributing an effect to one variable when it might be due to another.

We can take a similar effect to understanding proxys that we do to understanding confounds → we have to rule out specific alternative explanations !!!

An example $r_{\text{personality, performance}} = .4$ Is it really personality?

Motivation, amount of preparation & testing comfort are some variables that have and are all related to perf.

So, we run a multiple regression with all four as predictors.

If personality doesn't contribute, then it isn't personality but the other variables.

If personality contributes to that model, then we know that "personality" in the model is "the part of personality that isn't motivation, preparation or comfort".