Bivariate & Multivariate Regression

- correlation vs. prediction research
- prediction and relationship strength
- interpreting regression formulas
- process of a prediction study
- Multivariate research & multiple regression
- Advantages of multiple regression
- Interpreting multiple regression weights
- Inspecting & describing multiple regression models

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...

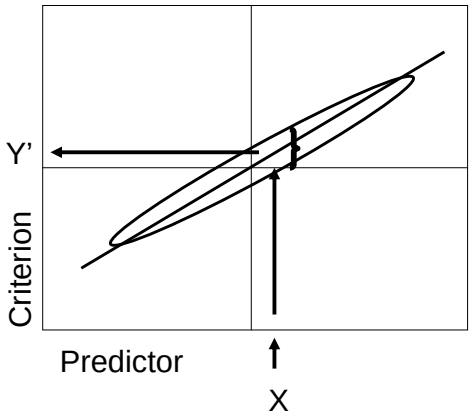
- "prediction" is using "what you know" to estimate "what you wish you knew"
  - "what you know" is called the predictor variable
  - "what you wish you knew" is called the criterion variable
- "what you wish you knew" must be estimated/predicted because it "hasn't happened yet" and you don't want to wait

For each, which is the criterion and which is the predictor?

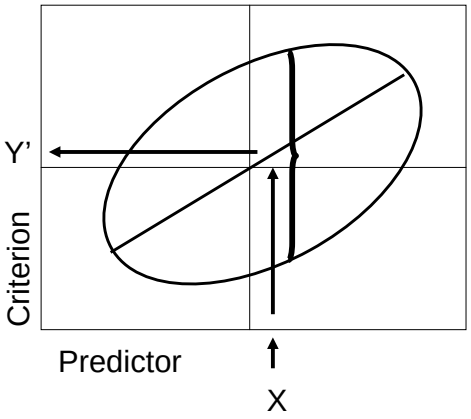
- score on the Final Exam & score on Exam 1
- undergraduate GPA & graduate GPA
- length of treatment & recidivism
- performance & practice

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 to the predicted Y value



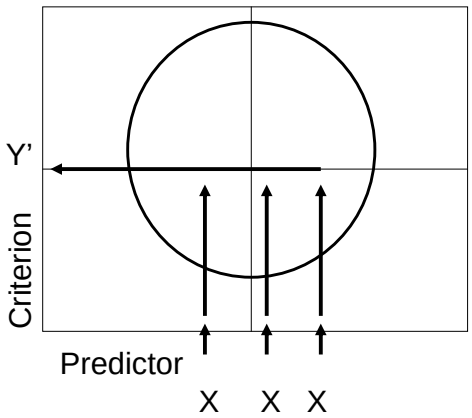
When the linear relationship is very strong, there is a narrow range of Y values for any X value, and so the Y' "guess" will be close



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"

$y' = bx + a$ $\quad\quad$ $residual = y - y'$

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

For a quantitative predictor

a = expected value of y if x = 0

b = direction and extent of expected change in the criterion for a 1-unit increase in the predictor

For a binary x with 0-1 coding

a = the mean of y for the group with the code value = 0

b = the y mean 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' = _____

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

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

#2 job errors = (-6 * interview score) + 95

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

interpret "b" -- for each 1-unit increase in interview score, errors are expected to _____ by _____

interpret "a" -- if a person has a interview score of "0", their expected number of job errors is _____

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 _____ _____ than Cx

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

so … mean of Tx group is _____

#2 job errors = (-2.0 * job) + 38 code: mgr=1 sales=0

the mean # of job errors of the sales group is _____

the mean # job errors of the management group is _____

if we measured another group of salespersons, what would be the expected # of job errors? _____

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

Advantages of Multiple Regression

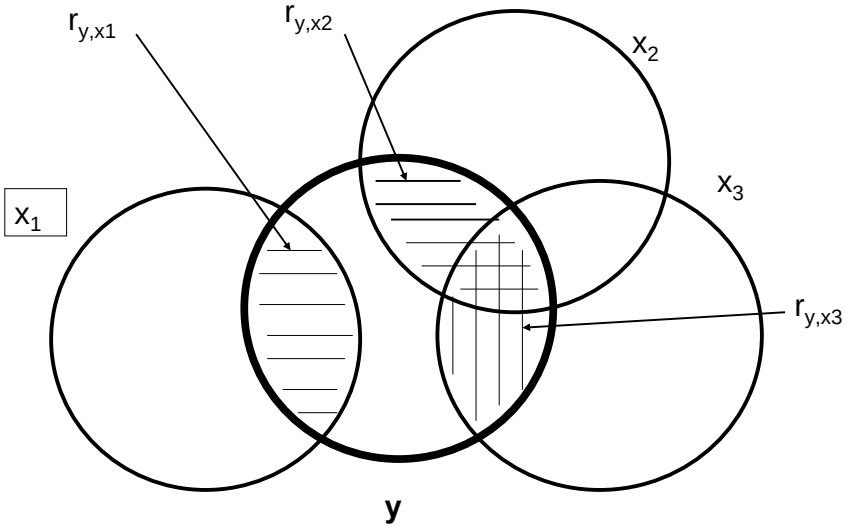
Practical issues …

- better prediction from multiple predictors

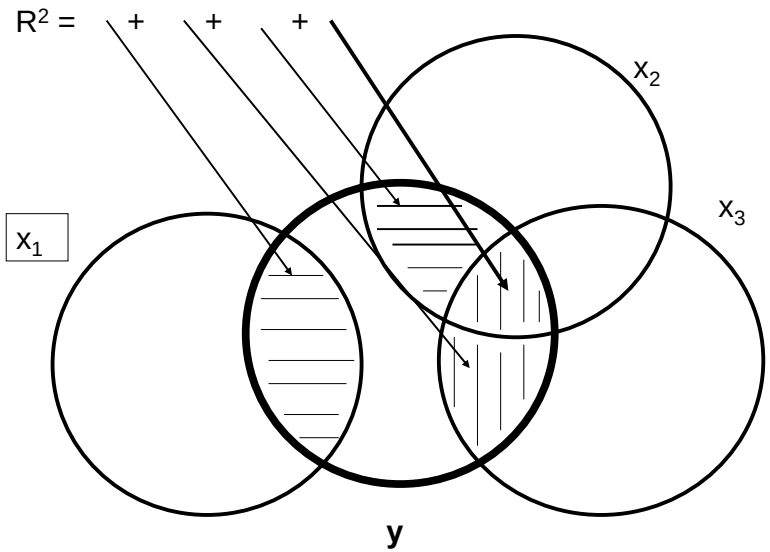
Theoretical issues …

- 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
- can examine "unique relationships" between individual predictors within a model and the criterion

Venn diagrams representing r, b and R^2



Remember R^2 is the total variance shared between the model (all of the predictors) and the criterion



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

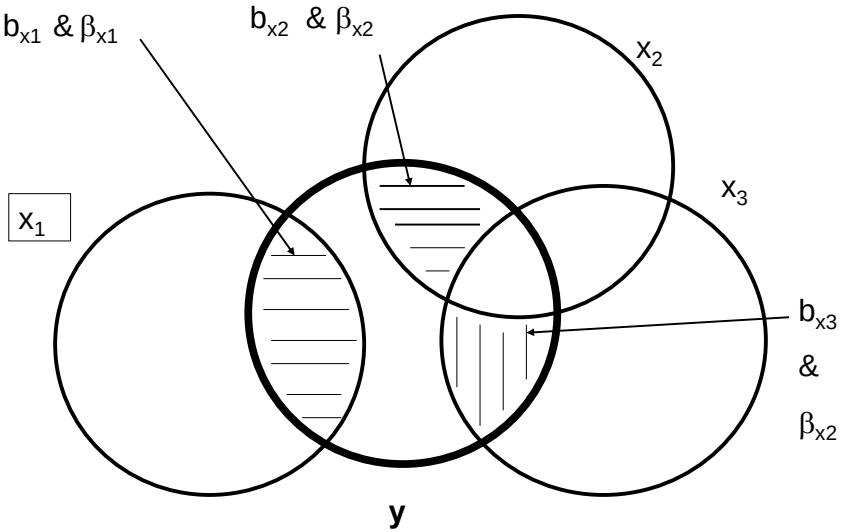
each b

- 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

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



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' = _____
- interpret "b" for stress -- for each 1-unit increase in stress, depression is expected to _____ by _____, when holding all other variables constant
- interpret "b" for support -- for each 1-unit increase in support, depression is expected to _____ by _____, 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 _____ _____ 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 _____

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

β carries the same information as b, but is scaled so that they are more comparable – allowing us to think about the "relative importance" of the different predictors to the model

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!!!).

Inspecting & describing results of a multiple regression formula

1. Does the model work?

   F-test (ANOVA) of H0: $R^2 = 0$ (R=0)

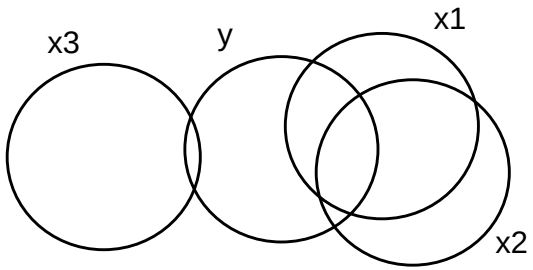
2. How well does the model work?
   - R^2 is an "effect size estimate" telling the proportion of variance of the criterion variable that is accounted for by the model

3. Which variables contribute to the model ??
   - t-test of H0: b = 0 for each variable

4. Which variables "contribute most" to the model ??
   - Compare the **β** weights of the variables (never b weights)
   - This must be done carefully – only trust large differences (e.g. at least .1 - .15 different)

Important Stuff !!! There are two different reasons that a predictor might not be contributing to a multiple regression model...

- the variable isn't correlated with the criterion
- the variable is correlated with the criterion, but is collinear with one or more other predictors, and so, has no independent contribution to the multiple regression model



X1 has a substantial r with the criterion and has a substantial b

x2 has a substantial r with the criterion but has a small b because it is collinear with x1

x3 has neither a substantial r nor substantial b