Simple Regression

- correlation vs. prediction research
- 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!)
- process of a prediction study

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...

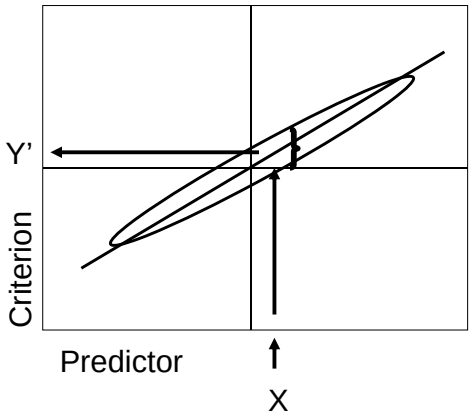
Linear regression for prediction...

- 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

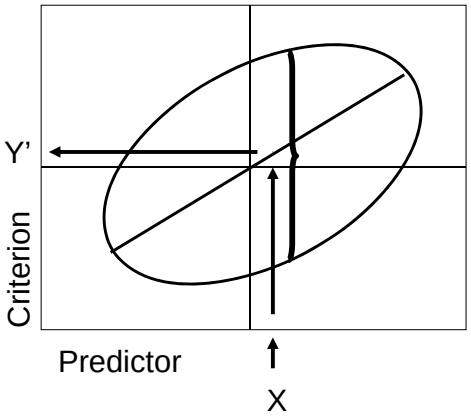
Story time
- Hotel Manager prediction story (.2 = $)
- “Which criterion” sales prediction story

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



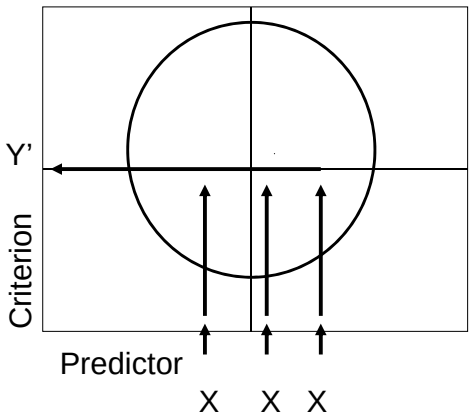
When the linear relationship is very strong, there is a narrow range of Y values for any X value, and so the Y' "guess" will be close



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 !