psyc3010 lecture 7

correlation and regression
multiple regression

last week: ancova and blocking
next week: standard vs hierarchical regression

last week → this week

- last week we looked at some methodological and statistical ways to improve power
- one of those strategies (ANCOVA) involved predicting what a group mean on the DV would have been if all groups were equal on a control variable – it achieved this via regression
- today we review bivariate correlation & regression, and introduce *multiple regression*

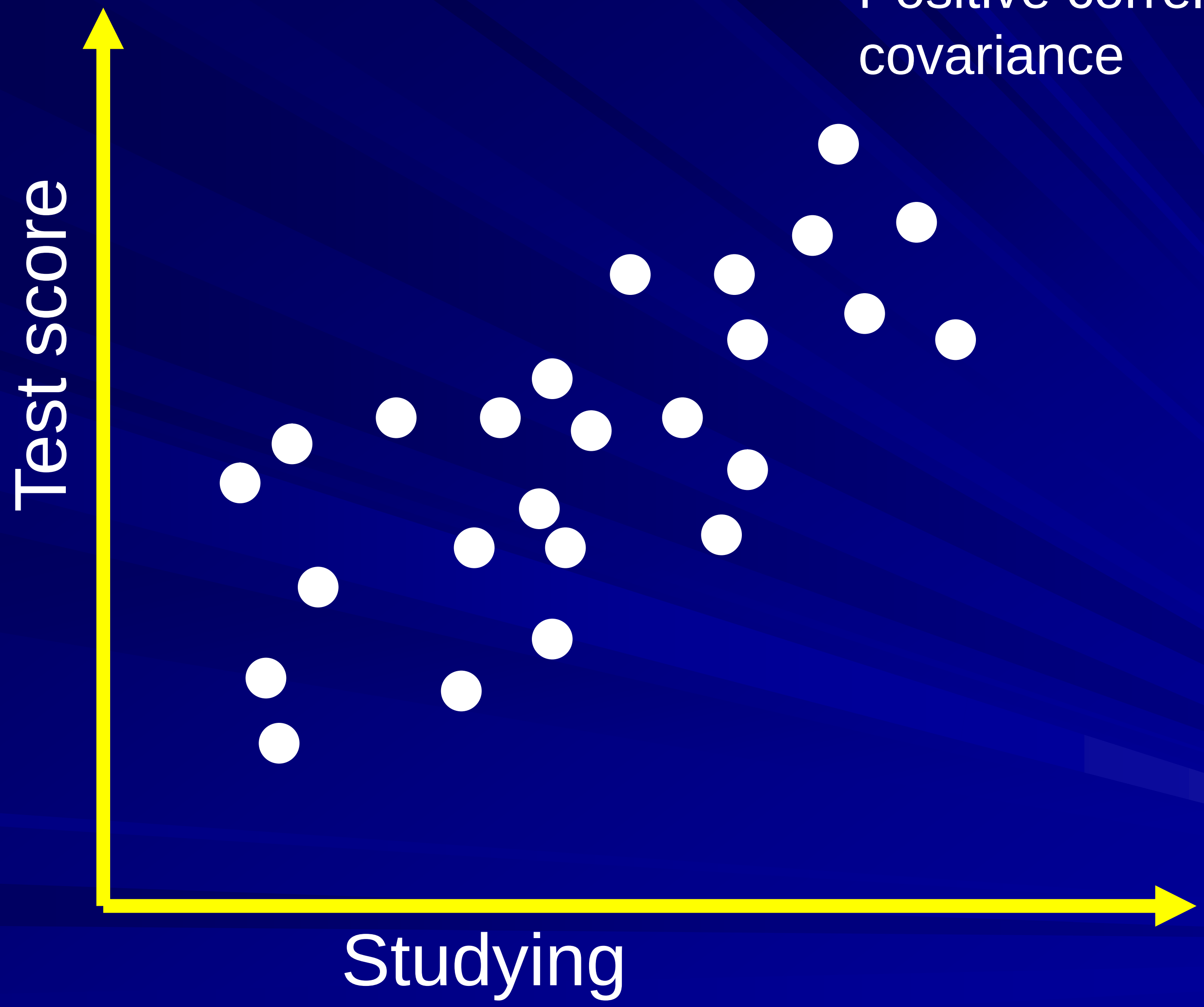
Experimental vs Correlational Research

- Experiments:
  - Determine causation through manipulation of IVs in controlled setting, assessing effect on DV
  - Impossible or unethical to manipulate factors such as personality, brain damage, long-term stress
- Correlational research:
  - Measure variables (predictors) and correlate with outcome / dependent variable (criterion)
  - Use bivariate regression (1 predictor) or multiple regression (>1 predictor)
- ANOVA (group diffs) $\neq$ Experiment (random assignment).
  - Random assignment is the basis for inferring causation
  - Sometimes use ANOVA to look at factors like gender that are observed – in which case cannot conclude causality
  - Sometimes analyse experiments using ‘group’ coded [categorical] variables in regression (e.g., so can do groovy mediation / moderation – see later in course)
  - Adding to the confusion, in regression (correlational analysis) you get an *F* statistic which tests whether the model is significant
  - Point: Don’t confuse the statistical issue (ANOVA vs regression) with the design issue (Experimental vs Correlational Research)

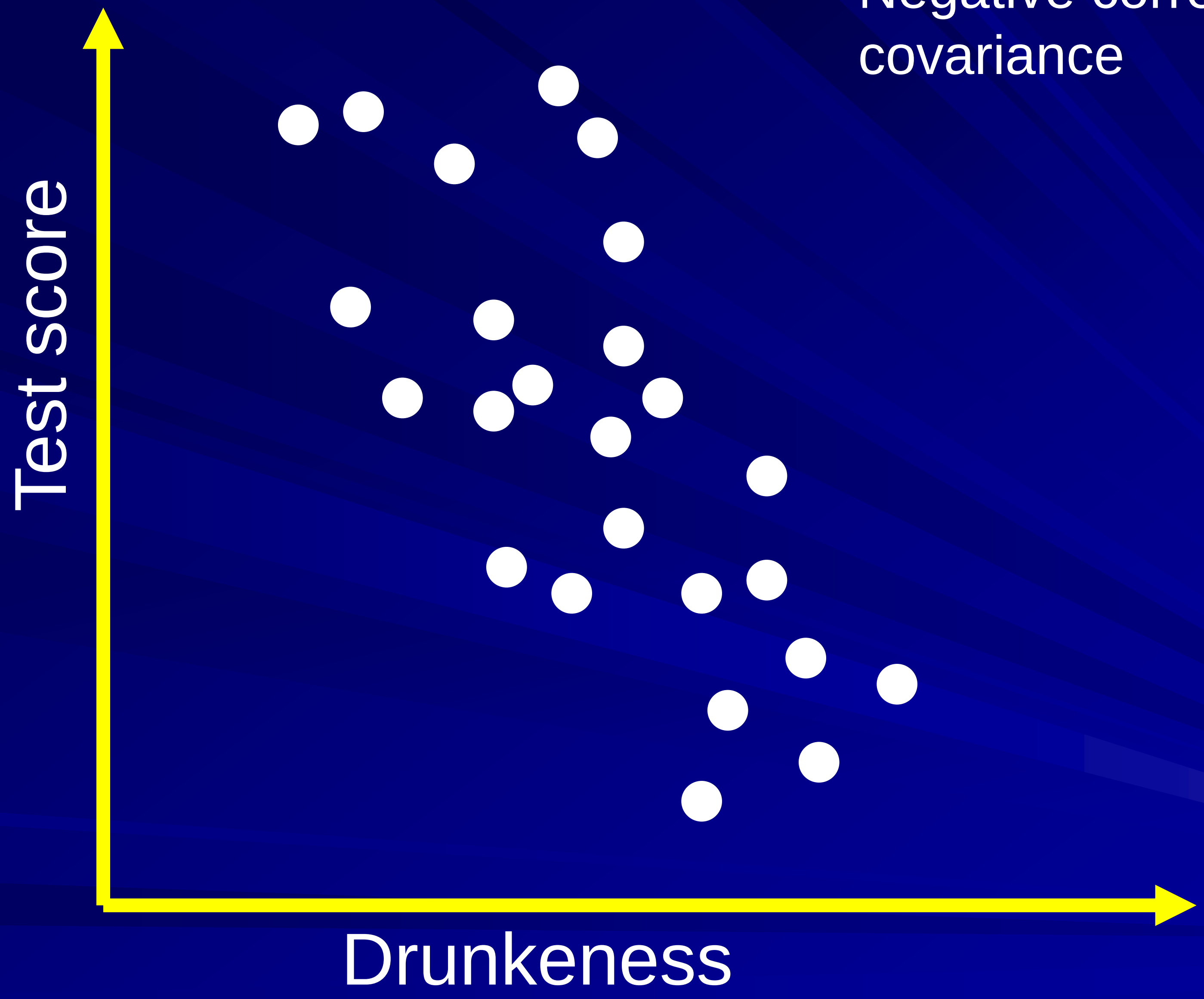
Measuring association

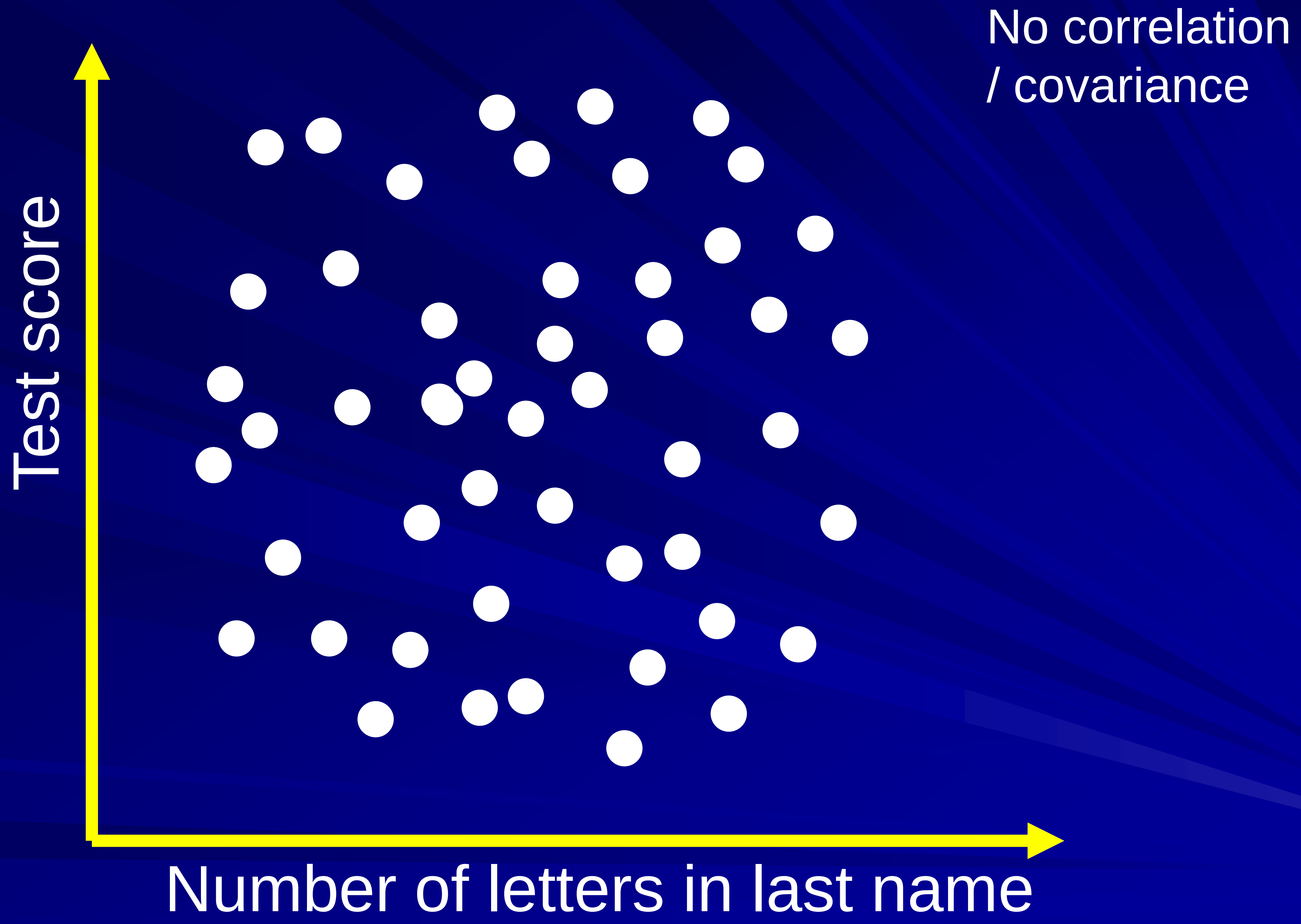
- is there a relationship between the number of social events attended per month (X) and life satisfaction, measured on 9-point scale (Y)
  - do scores on X and Y ***covary***?
  - is there a ***correlation*** between them?
  - scatterplot:
    - slope indicates general direction (+ or -)
    - width of ellipse indicates magnitude

Positive correlation / covariance



Negative correlation / covariance




No correlation
/ covariance
Test score
Number of letters in last name

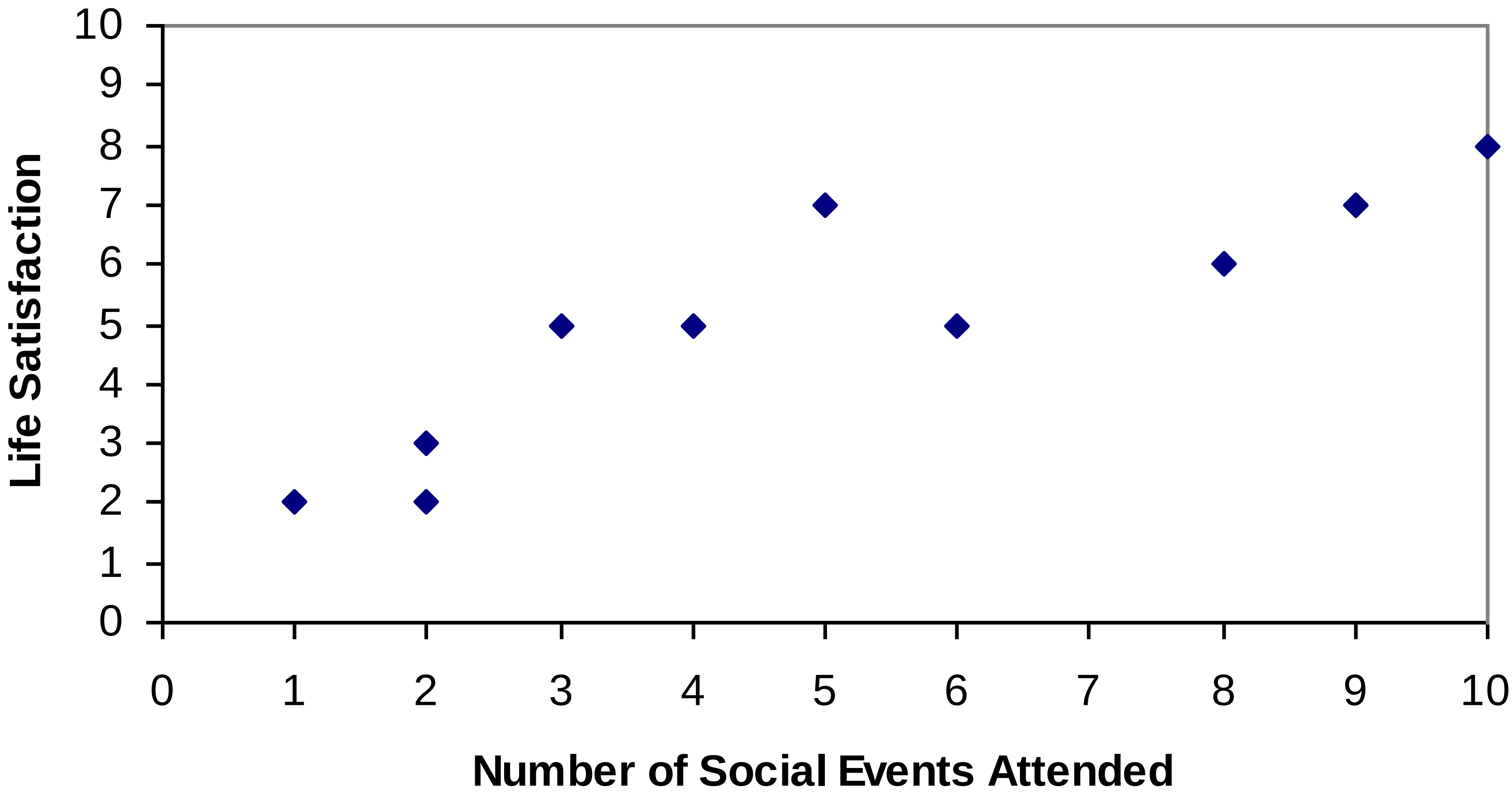
association between number of social events attended (*X*) and life satisfaction (*Y*)

participant	social events attended	life satisfaction
1	1	2
2	2	3
3	2	2
4	3	5
5	4	5
6	5	7
7	6	5
8	8	6
9	9	7
10	10	8

Notice the changing convention re labelling of the DV and IVs:

- ANOVA: IVs usually A, B, C and DV = X
- Regression / Correlation: IVs usually X, Z, W (or X1, X2, X3) and DV = Y

association between number of social events attended (*X*) and life satisfaction (*Y*)



quantifying the relationship

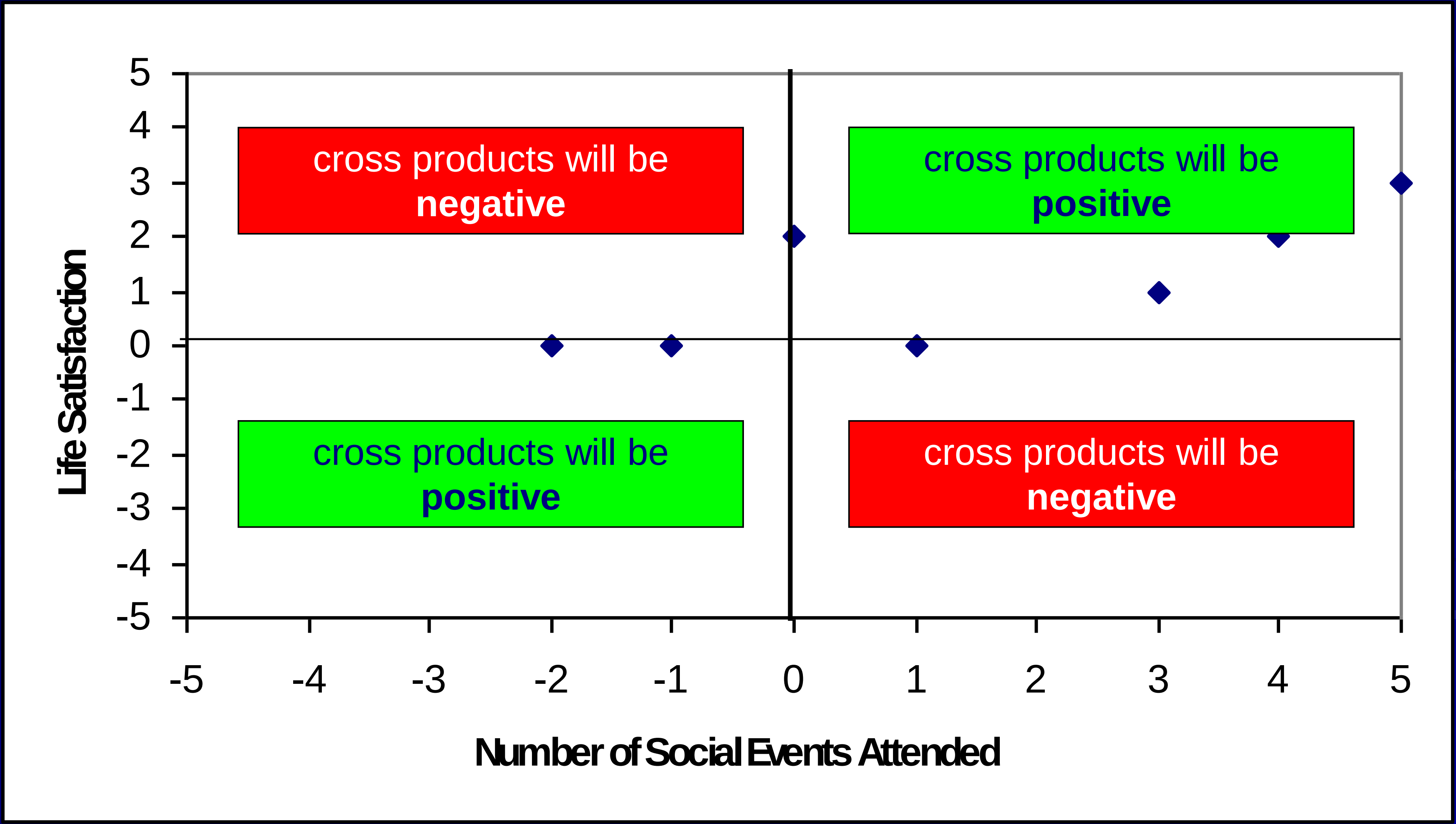
***covariance*:** average cross-product of the deviation scores (as seen in Lecture 6)

$$\mathrm{cov}_{XY} \text{ or } S_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1}$$

$$= 53/9 = 5.89$$

positive values → positive relationship
negative values → negative relationship

association between number of social events attended (*X*) and life satisfaction (*Y*)
(deviation scores)



quantifying the relationship

- unfortunately covariance is ***scale dependent*** – we can only tell whether a covariance of 5.89 is strong/weak if we know the scale of our variables
  - e.g., 1-9 rating scale
    with cov 5.89 (53/9)
    vs. 3-27 rating scale
    with cov 17.69 (159/9) if triple all scores
  - smiley faces ?
  - direct measure (e.g., brain volume) ?

$$\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N-1}$$

quantifying the relationship

- ***correlation***: standardised covariance – i.e., covariance relative to standard deviations of *X* and *Y*:

$$r_{XY} = \frac{\text{cov}_{XY}}{S_X S_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y}) / N - 1}{S_x S_y}$$

$$S_y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N - 1}} = \sqrt{\frac{40}{9}} = 2.11$$

$$S_x = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}} = \sqrt{\frac{90}{9}} = 3.16$$

quantifying the relationship

- ***correlation*:** average cross-product of the standard scores of two variables

$$r_{XY} = \frac{\text{cov}_{XY}}{S_X S_Y} = \frac{\sum Z_x Z_y}{N-1}$$

$= 5.89/(3.16 \times 2.11)$

$= .8833$ (and <u>same</u> .88 with 3-27 scale for Y)

positive values → positive relationship (max = 1)

negative values → negative relationship (min = -1)

- ***r*** is comparable across studies & scales (vs covariance) !
- When we say **correlation**, we usually mean r. But there are other correlation statistics. To be precise, r is called a **Pearson correlation** or **zero-order correlation**.

interpreting r in terms of variance

remember, PSYC3010 is all about variance!

- r^2 - ***the coefficient of determination:*** **proportion of variance in one variable that is explained by the variance in another**

$$r^2 = .7802$$

– 78% of the variance in life satisfaction is explained by # of social events attended per month

-And therefore, $1 - r^2$ = error or residual variance in data (22% of the variance in life satisfaction is not explained by social events)

testing r for significance

- *is r large enough to conclude that there is a non-zero correlation in the population?*

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{.8833\sqrt{8}}{\sqrt{1-.7802}} = \underline{5.3195}$$

df = N-2 =8

$t_{\alpha=.05}(8) = 2.306$

$$\sqrt{\frac{r^2(N-2)}{(1-r^2)}}$$

Effect variance accounted for * # of observations used to arrive at statistic, divided by error variance - Like ANOVA!

– the relationship between life satisfaction and # of social events attended per month is significant, r = .88, $t(8)$=5.32, p<.05.

r as a population estimate - r_{adj}

- r is a sample statistic and is biased to sample (like eta-squared in our 'estimates of effect size' lecture).
- can calculate rho, ρ , the unbiased estimate of the population correlation coefficient – estimated by r_{adj}

$$r_{adj} = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{N - 2}} = \sqrt{1 - \frac{(1 - .88^2)(9)}{8}} = \underline{.8676}$$

  – r^2_{adj}= .7527 (vs 78% for r^2)
- r_{adj} is always smaller than r (more conservative) – like ω^2
- The difference between the two becomes greater as sample size decreases

correlation and prediction - regression

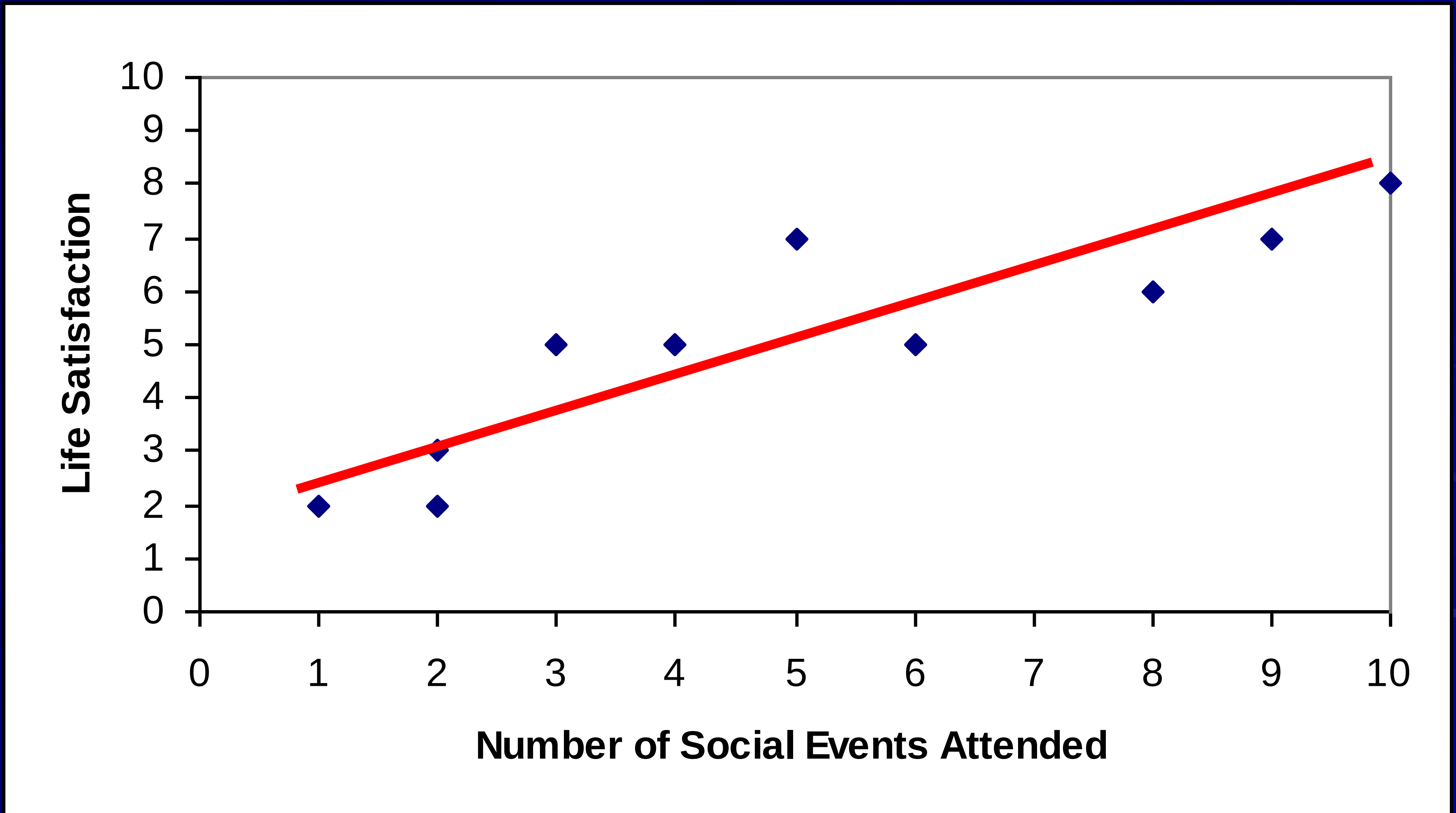
- estimating a score on one variable (*Y, criterion*) on the basis of scores on another variable (*X, predictor*)

  →note, prediction is implied in a conceptual sense, not a literal sense. Cannot conclude definitively re causality because no random assignment – in correlational designs, we infer causality based on theory. If theory is wrong, DV causes IV or both are caused by 3rd variable!

  →regression of Y on X (X is IV) ≠ regression of X on Y (X is DV) – phrase is "regress DV on IV(s)"

  →objective is to find the best fitting line on the scatter plot. This represents the best *linear model* of the data.

predicting life satisfaction (Y) from number of social events attended (*X*)



correlation and prediction - regression

bivariate regression equation:

$$\hat{Y} = bX + a$$

$\hat{Y}$ = predicted value of Y
b = slope of regression line (change in Y associated with a 1- unit change in X)
X = value of predictor
a = intercept (value of Y when X = 0)

correlation and prediction - regression

bivariate regression equation:

$$\hat{Y} = bX + a$$

For a 1-unit change in X, expect Y to change +.59 units

$$b = \frac{\mathrm{cov}_{XY}}{s_X^{\,2}} = 5.89/3.16^2 = \underline{.5898}$$

$$= r\,\frac{s_Y}{s_X} = .8833 \times 2.11/3.16 = \underline{.5898}$$

correlation and prediction - regression

bivariate regression equation:

$$\hat{Y} = bX + a$$

Intercept or constant
- Where X = 0, Y = 2.05

a $= \overline{Y} - b\overline{X}$

$= 5 - .5898(5)$

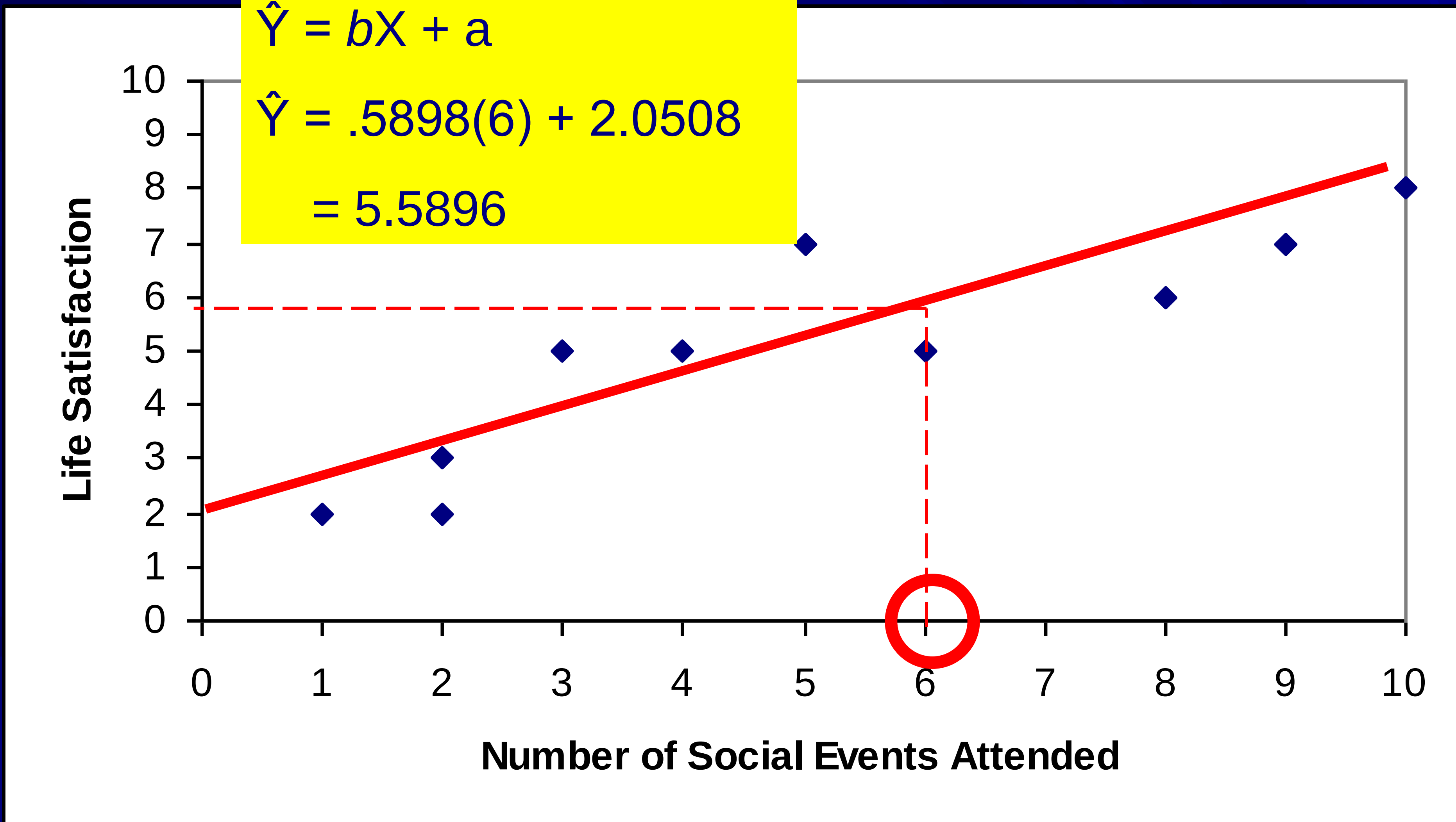
$= \underline{2.0508}$

association between number of social events attended (X) and life satisfaction (Y)

$\hat{Y} = bX + a$

$\hat{Y} = .5898(6) + 2.0508$

$= 5.5896$



the regression slope

b is important because it describes the best line of fit to the data – line of fit that achieves the least squares criterion

- Formula looks like r in correlation:

$$b = \frac{COV_{XY}}{S_X^2} \qquad r = \frac{COV_{XY}}{S_X S_Y}$$

- so if the data were standardised…
  - $S_X = S_Y$
  - ***and*** **S_X and S_Y = 1**

- $COV_{XY} = r_{XY} = b$

- and *b* would become a standardised regression coefficient, β (beta)

-β indicates Z score change in Y predicted from a 1 SD increase in X

- *How many SDs change in Y would you expect from 1 SD change in X?*

correlation and prediction - regression

(unstandardised) bivariate regression equation:

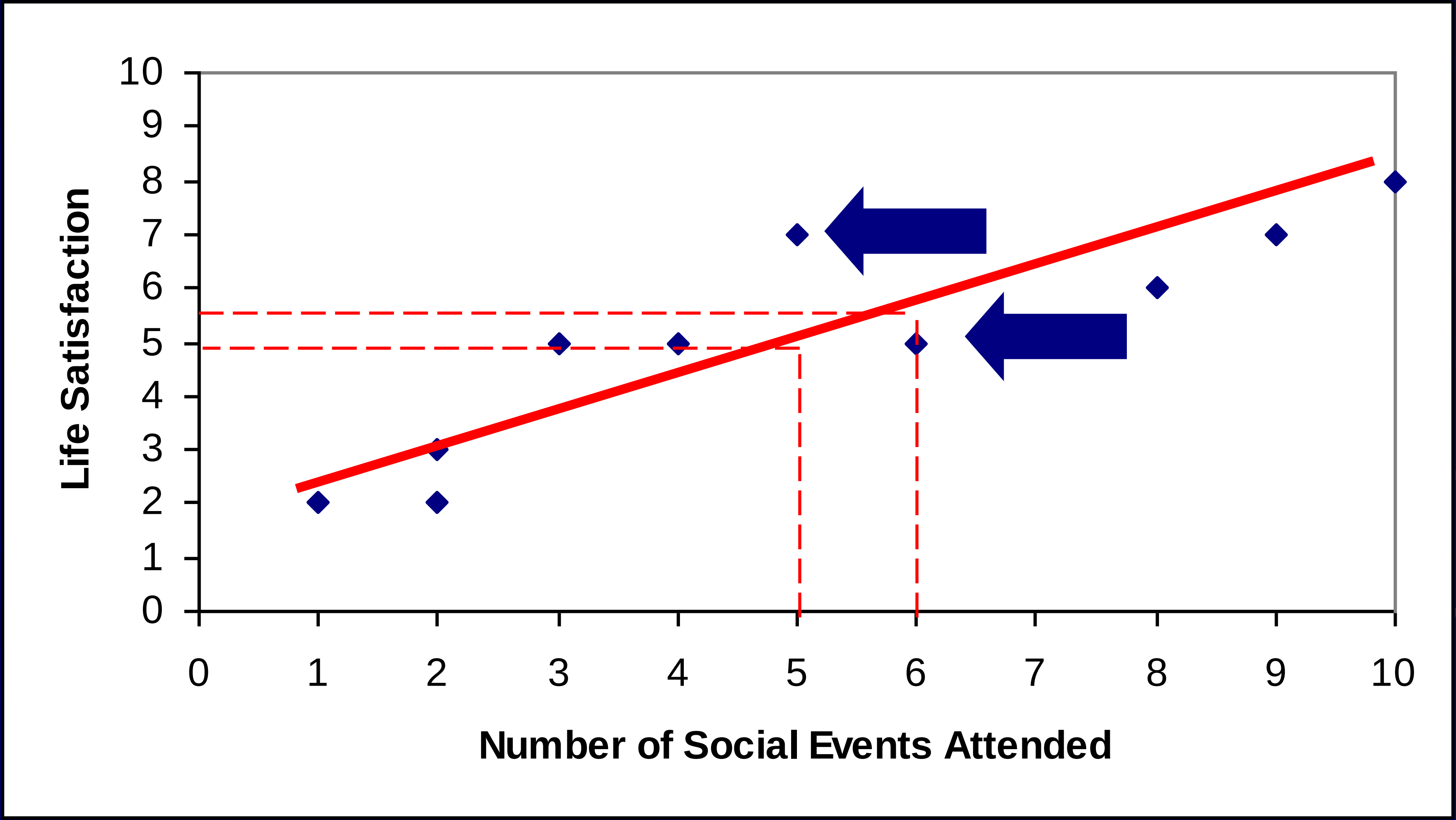
$$\hat{Y} = bX + a$$

Standardised bivariate regression equation:

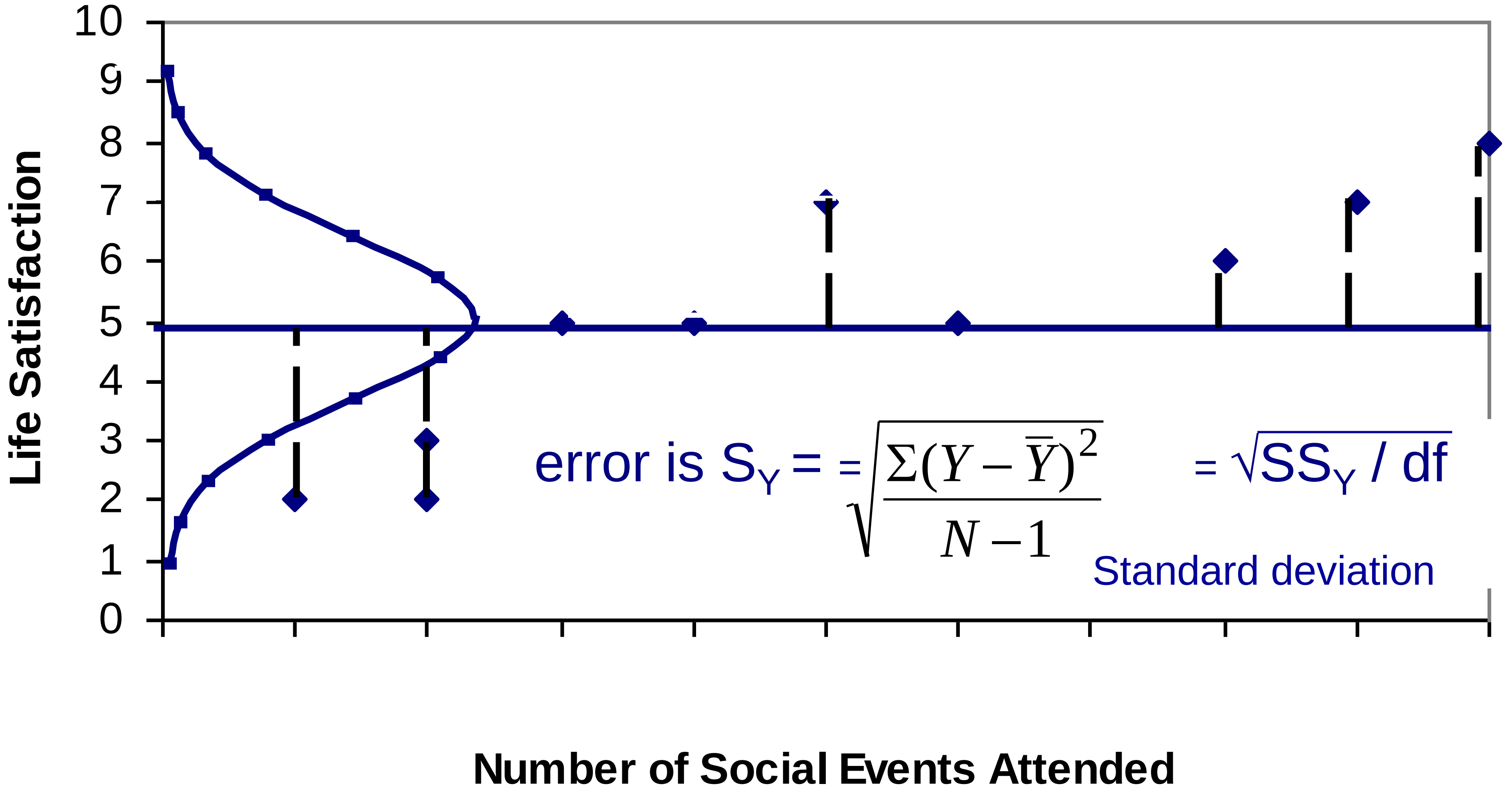
$$\hat{Z}_Y = \boldsymbol{beta}Z_X = r_{XY}Z_X = .88Z_X$$

– For a 1 SD change in X, expect .88SD increase in Y

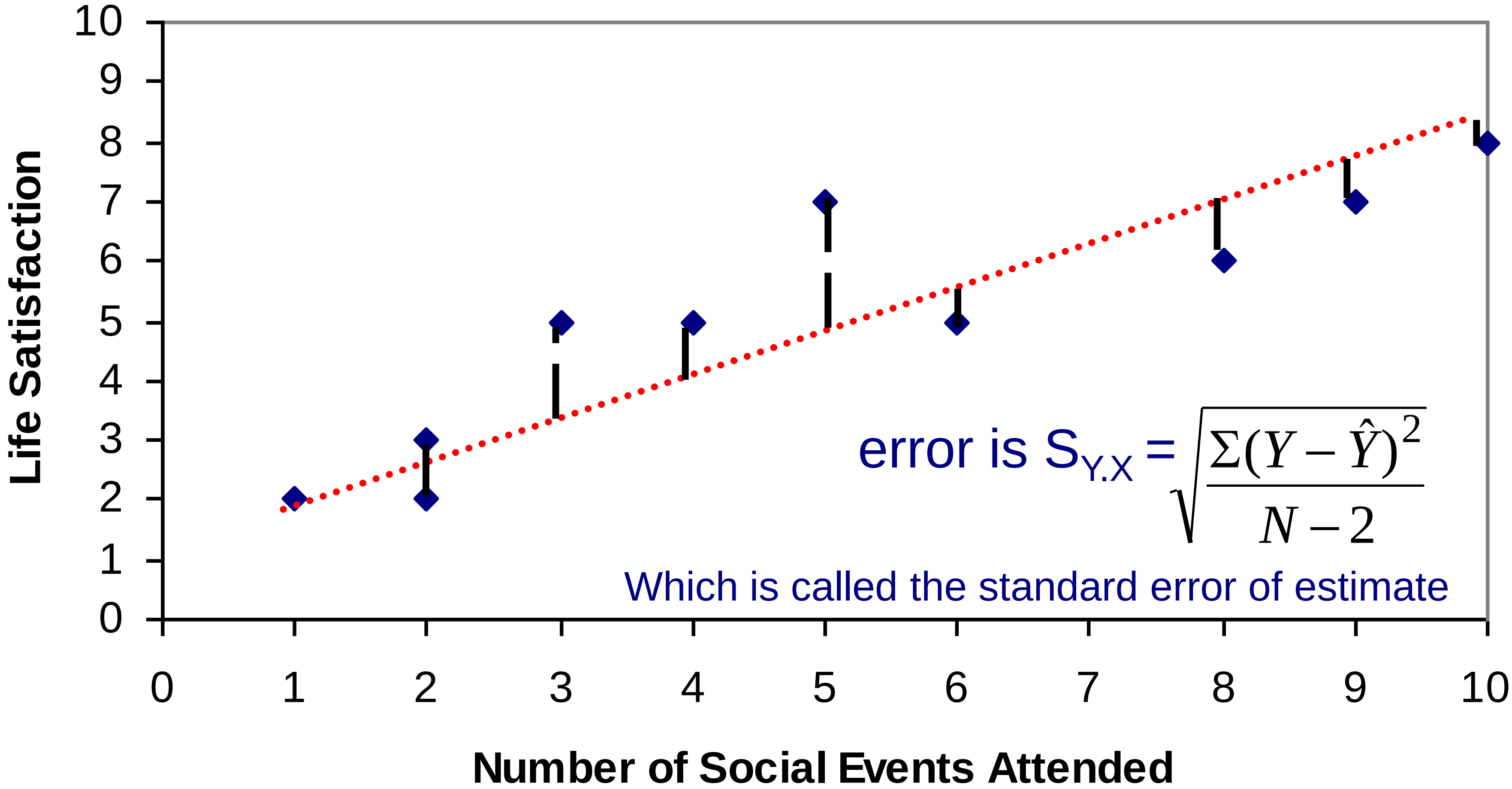
errors of prediction



X unknown – best predictor of Y is $\overline{Y}$



X known – best predictor of Y is Ŷ (a conditional value – depends on X)



the standard error of the estimate

$S_{Y.X}$ reflects the amount of variability around the regression slope (Ŷ, a conditional value), and is an important statistic in correlation and regression

- $S_{Y.X}$ = $$\sqrt{\frac{\sum (Y-\hat{Y})^2}{N-2}} = \sqrt{\frac{SS_{error}}{df}} = S_Y\sqrt{1-r^2}$$

- the regression line is fitted according to the ***least squares criterion:***
  – such that $\Sigma(Y-\hat{Y})^2$ is a minimum
  – i.e., such that that ***errors of prediction*** are a minimum

$e_i = Y_i - \hat{Y}_i$ = *errors of prediction*

E.g. see Howell 6th ed pp. 245-7

association between number of social events attended (*X*) and life satisfaction (*Y*)

participant	social events attended	life satisfaction
1	1	2
2	2	3
3	2	2
4	3	5
5	4	5
6	5	7
7	6	5
8	8	6
9	9	7
10	10	8

e.g., for this data, b = .59 and a = 2.05.

$\hat{Y} = bX + a$

$\hat{Y}_1 = .59(1) + 2.05$

$= 2.64$

$Y_1 = 2$

$e_1 = 2 – 2.64 = -.64$

Goal: minimize $\sum e_i^2$

significance of the regression slope

b and β, like r, can be tested for significance *(if assumptions are met):*

$$t = \frac{(b)(s_X)(\sqrt{N-1})}{S_{Y.X}} \qquad df = N - 2$$

$$= \frac{(.5898)(3.16)(\sqrt{9})}{.9892}$$

$$= \underline{5.6523}$$

Significant regression coefficient = a slope that significantly differs from zero (+ or -). Null hypothesis is that b=0, i.e. there is no systematic change in Y when X increases by a unit.

same formula for β (if *b* is significant, β will be also)

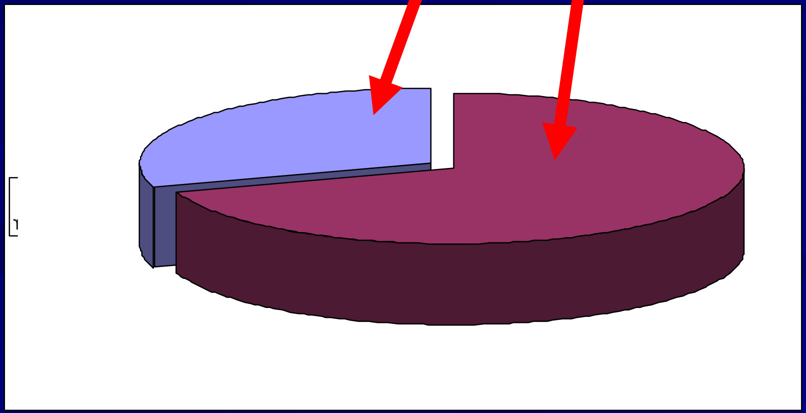
partitioning the variance

quite similar to anova (its all about variance!)

in anova,

SS Between groups, SS residual

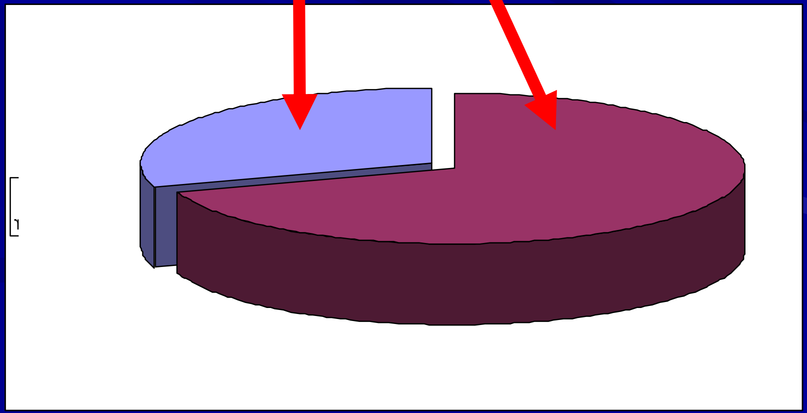
$$X_i = \mu + \boxed{\tau_j} + \boxed{e_{ij}}$$



in regression,

SS predicted, SS residual

$$Y_i = \boxed{bX_i + c} + \boxed{e_i}$$



$$SS_Y = SS_{regression} + SS_{residual}$$

where

$$SS_Y = \Sigma(Y-\bar{Y})^2$$

$$SS_{regression} = \Sigma(\hat{Y}-\bar{Y})^2$$

$$SS_{residual} = \Sigma(Y-\hat{Y})^2$$

$Df_Y = N - 1$

$Df_{regression} = p$ [# predictors]
i.e. 1, for bivariate

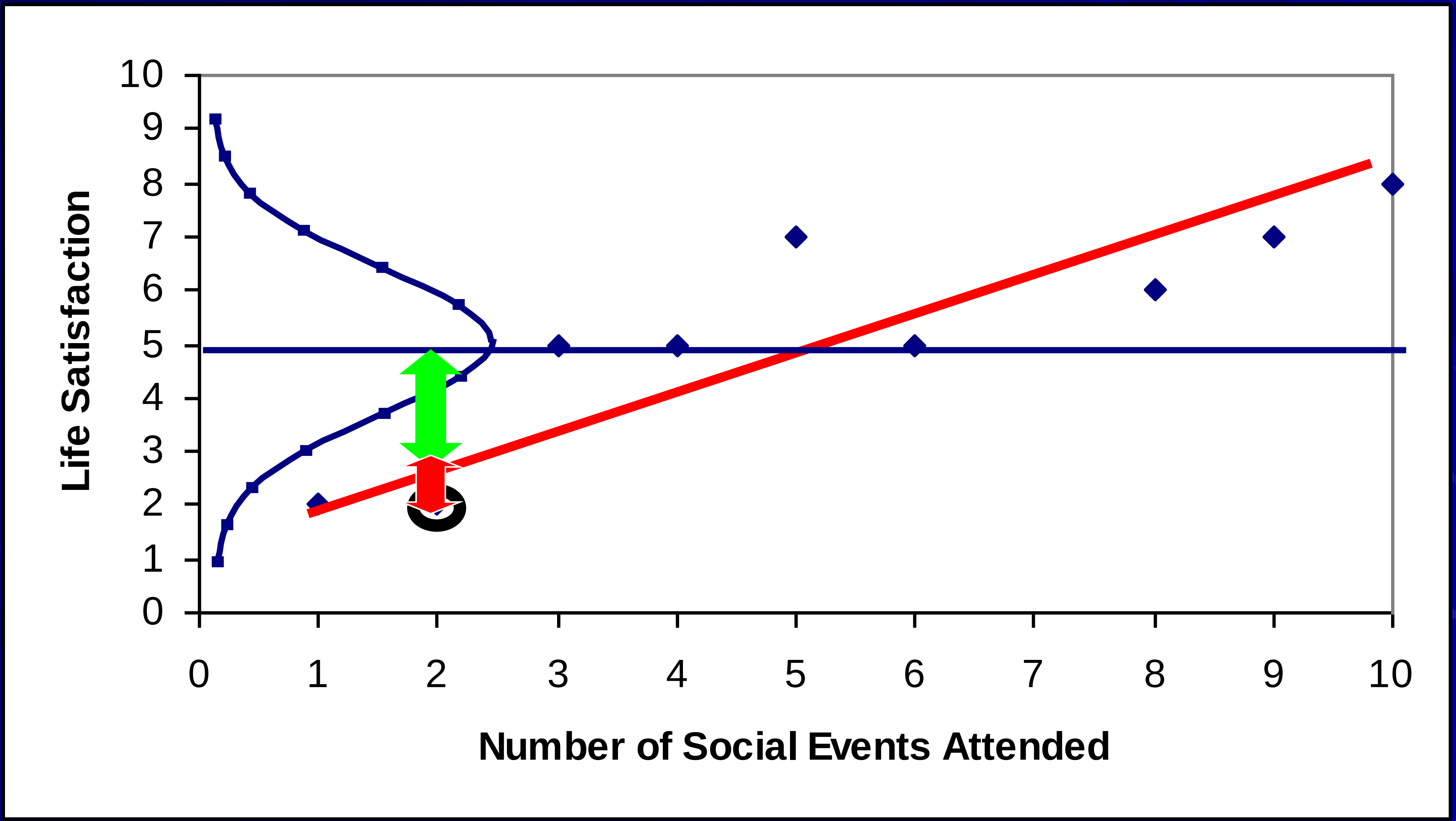
$Df_{residual} = N – p - 1$

$$F\,(1, N – 2) = \frac{\text{MS regression}}{\text{MS residual}}$$

→ Tests hypothesis that the model accounts for significant variance in the DV. Null hypothesis is that R2 = 0.

$$SS_Y = SS_{regression} + SS_{residual}$$

$$\text{(also, } df_{total} = df_{regression} + df_{residual}\text{)}$$



the standard error of the estimate

SY.X = $S_Y\sqrt{1-r^2} = 2.11\sqrt{1-.7802^2} = .9892$

- 68% of individuals will score within + or - .9892 units of the predicted score (Yhat).

Bigger r_{xy} -> smaller $S_{y.x}$ – i.e., a high correlation between X and Y reduces the standard error of estimate and enhances the accuracy of the prediction.

Remember r^2 is overly liberal (inflated) with small samples - we also find therefore that $S_{y.x}$ is underestimated for small samples

bivariate regression → multiple regression

correlation and bivariate regression

→ single predictor

multiple regression

→ variation is a function of *multiple* predictors usually acting simultaneously – therefore achieve better prediction

→ Multiple correlation: relation between criterion Y and a set of predictors

→ Multiple regression: scores on criterion Y are predicted using > 1 predictor

bivariate regression → multiple regression

multiple regression

→ 2 major steps/issues

- ***strength of relationship*** between criterion and set of predictors: multiple R, R^2
- ***importance of individual predictors***: b, β, sr, pr

→ predictors are usually correlated so their contribution overlaps – this has implications for <u>both</u> steps

What you test for:

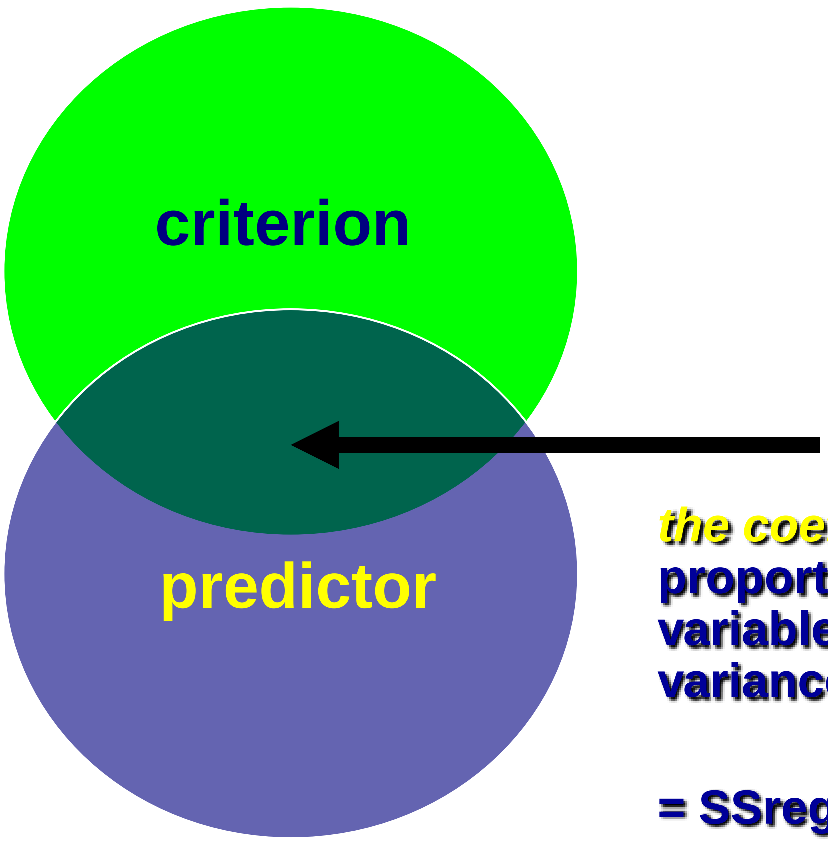
Bivariate regression:

- Does the predictor account for significant variance in the DV?
- Two tests with same significance value:
  - *F* test for Model R^2 = squared *t*-test of β for IV

Multiple regression:

- Do the predictors jointly account for significant variance in the DV?
  - F test of Model R^2
- For each IV: does it uniquely account for variance in the DV?
  - t-test of β for each IV
- Model R^2 can be sig even if individual β(s) are not, or vice versa

bivariate regression



$r^2 = R^2$

the coefficient of determination:
proportion of variance in one variable that is explained by the variance in another

$= SSregression / SS_Y$
$= (SS_Y – SS_{error}) / SS_Y$

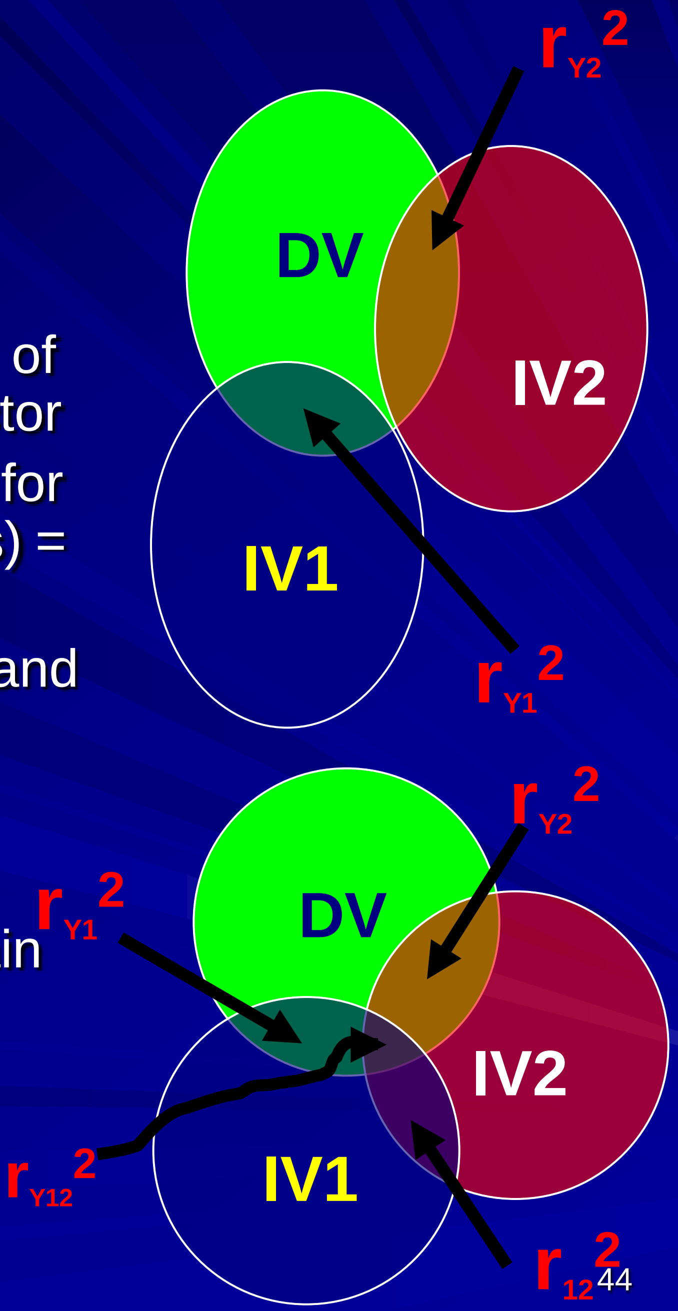
While 1 – r2 = error variance = SSerror / SSY

Multiple Regression

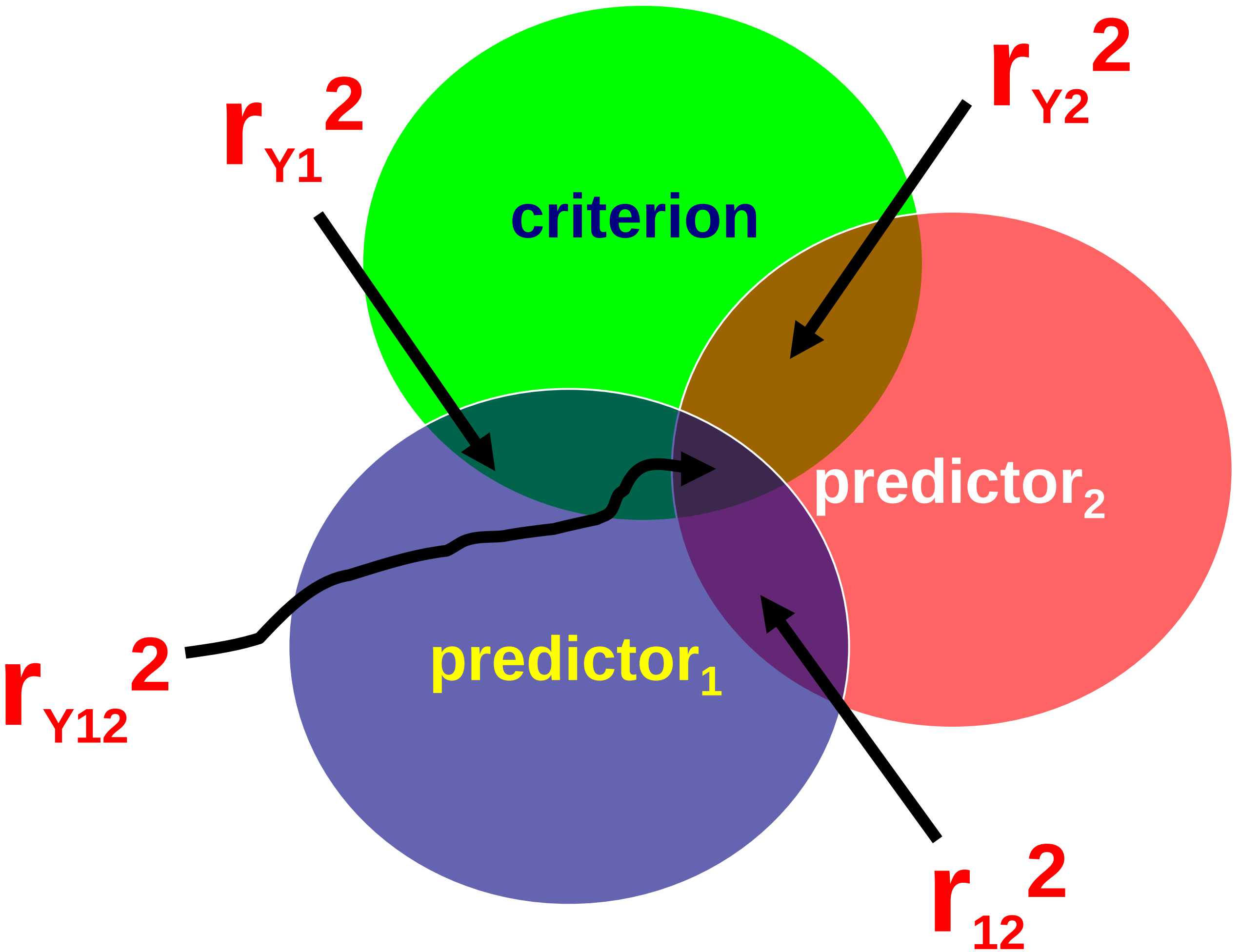
- when predictors are uncorrelated
  - can unambiguously identify proportion of variance accounted for by each predictor
  - R^2 (i.e., the variance in DV accounted for by linear model including all predictors) = $r_{Y1}^2 + r_{Y2}^2$
  - predictor importance indicated by r_{Y1}^2 and r_{Y2}^2 respectively

….but predictors are rarely uncorrelated

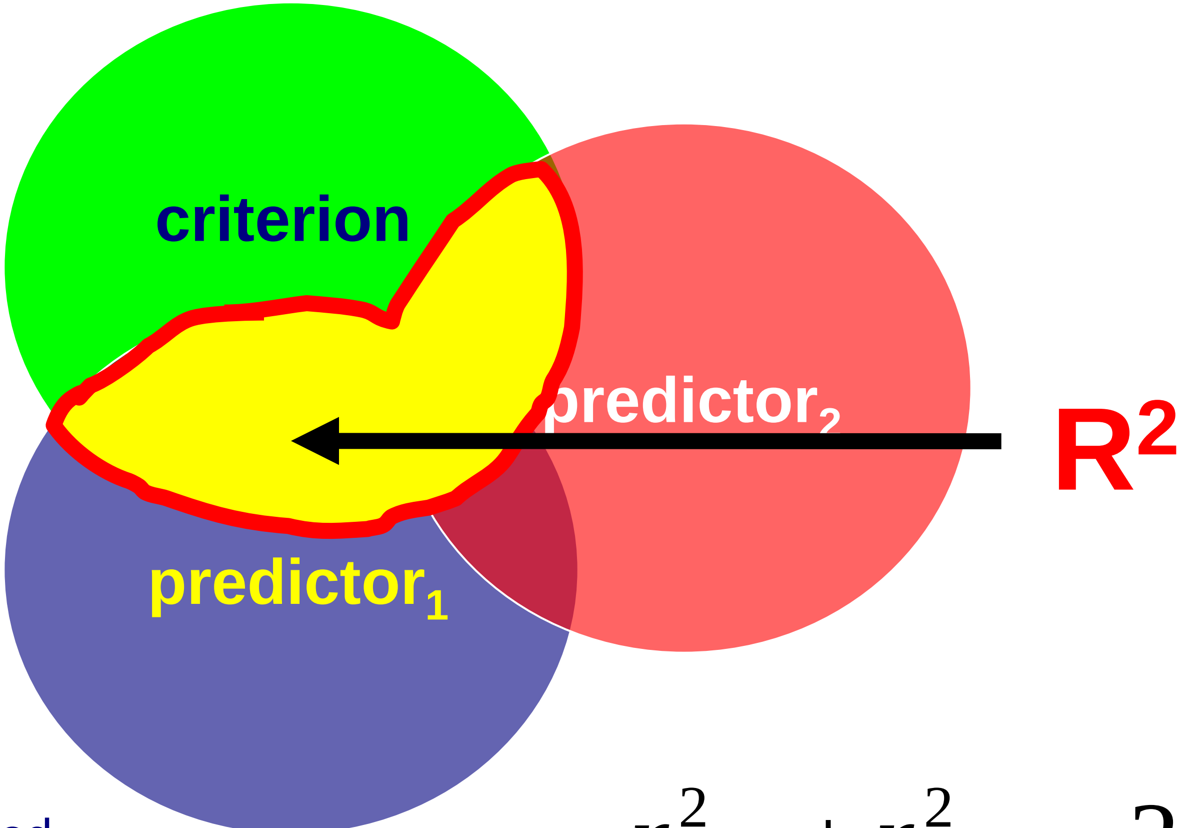
- so R^2 generally < $r_{Y1}^2 + r_{Y2}^2$
- predictor importance difficult to ascertain



multiple regression



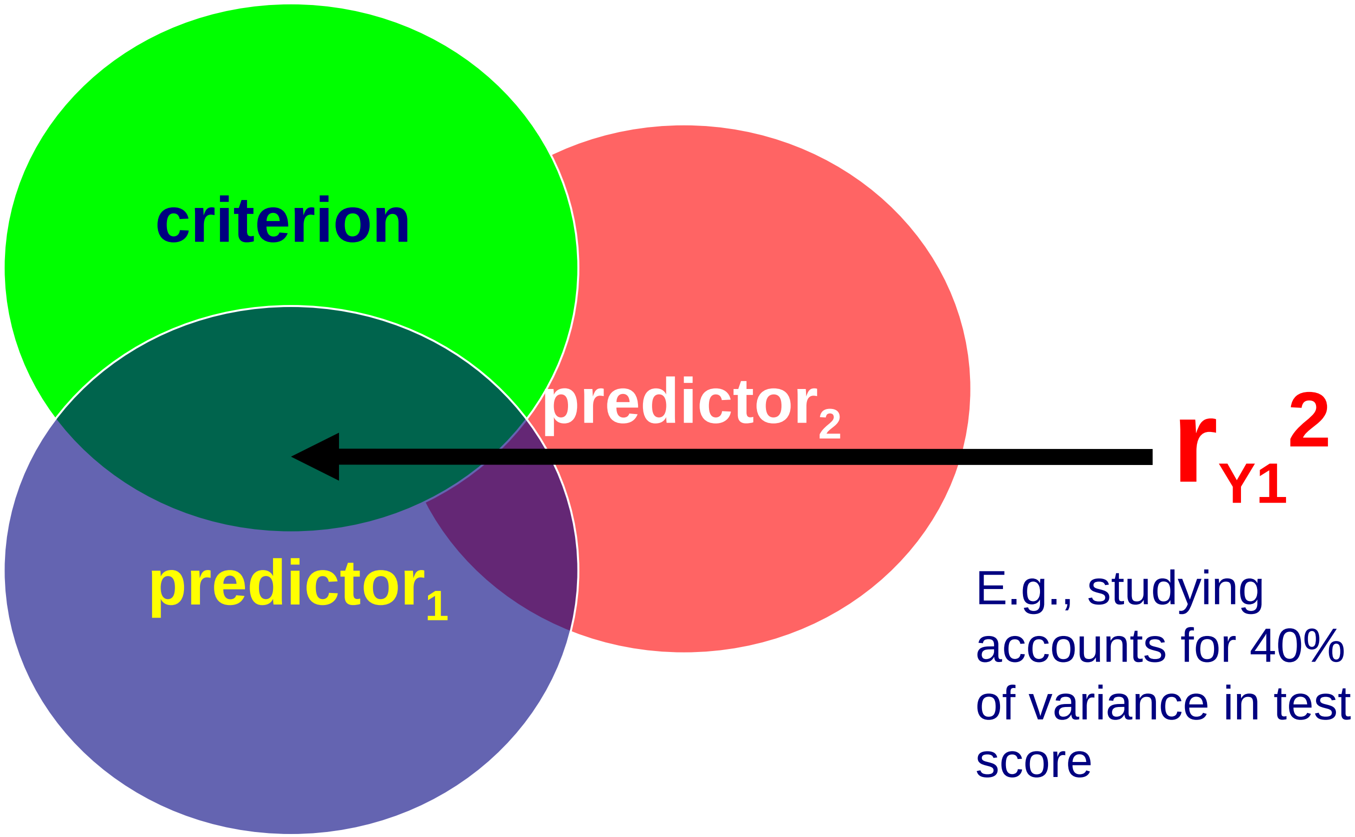
multiple regression



You do not need to know the formula – only understand the point about how interpreting the zero-order correlations can be misleading

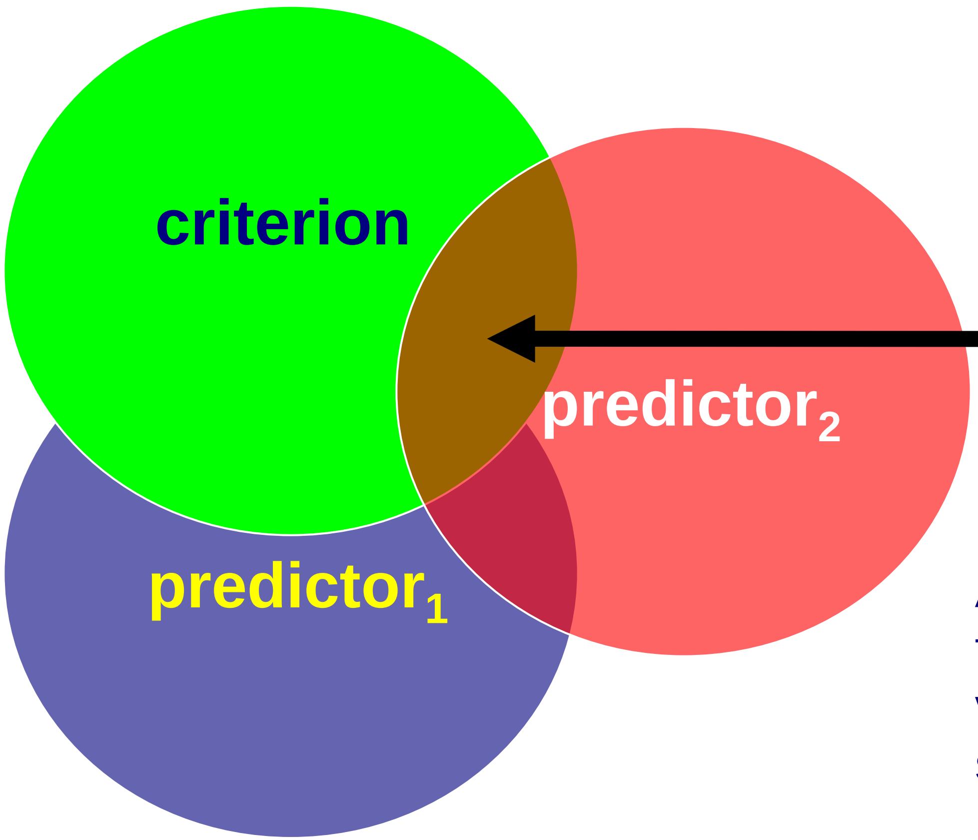
$$R^2_{Y.12} = \frac{r^2_{Y1} + r^2_{Y2} - 2r_{Y1}r_{Y2}r_{12}}{1 - r^2_{12}}$$

contribution of each predictor in terms of r



E.g., studying accounts for 40% of variance in test score

contribution of each predictor in terms of r

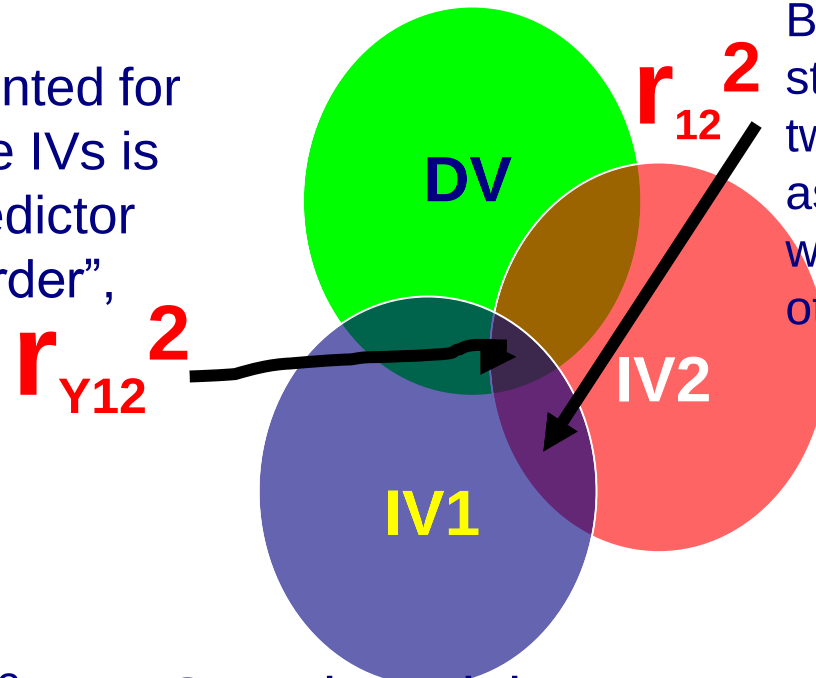


$r_{Y1}{}^2$

And IQ accounts for 30% of variance in test score

multiple regression

The variance in the DV accounted for by the **shared variance** of the IVs is double counted if have >1 predictor and focus on regular (“zero-order”, Pearson) correlations.

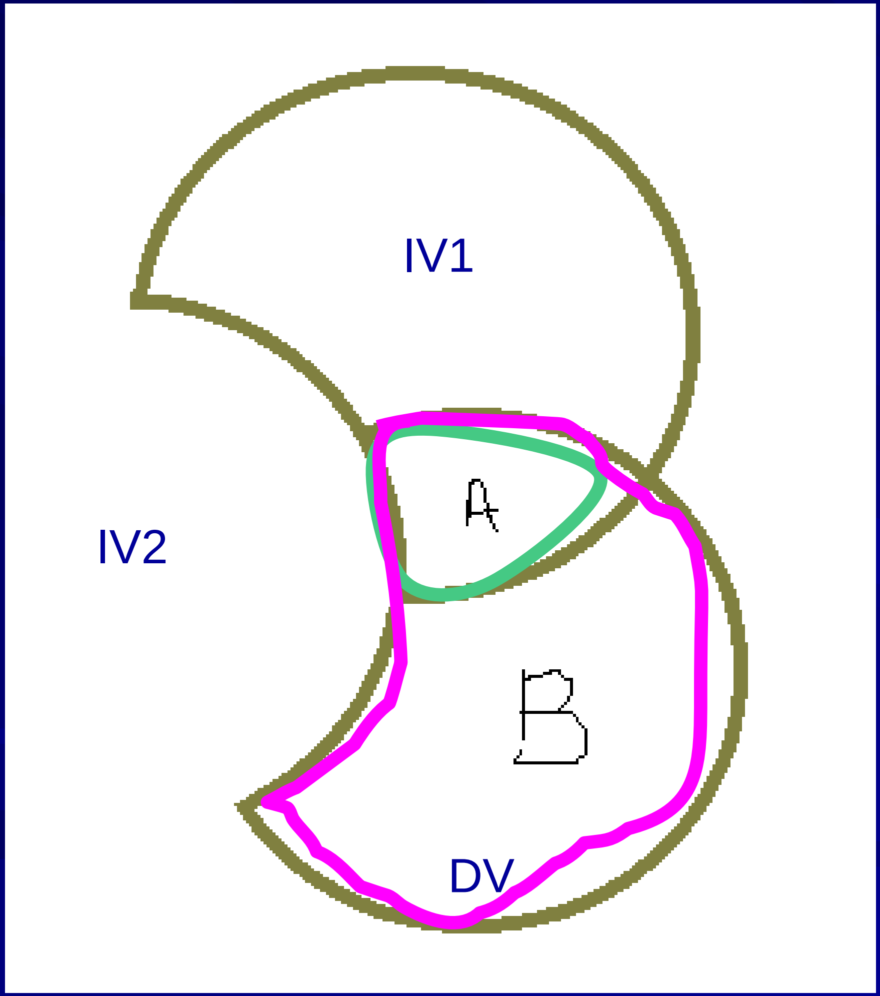


But IQ and studying, the two IVs, are associated with each other

Implications:

- (1) $R^2_{Y.12} < r^2_{Y1} + r^2_{Y2}$ -- IQ and studying account for < 70% of variance in test score. R2 measures the ***non-redundant*** variance in DV accounted for from combo of variables.
- (2) Need to think about correlations between each IV and the DV adjusted to control for the effects of other IVs – partial and semi-partial correlations.

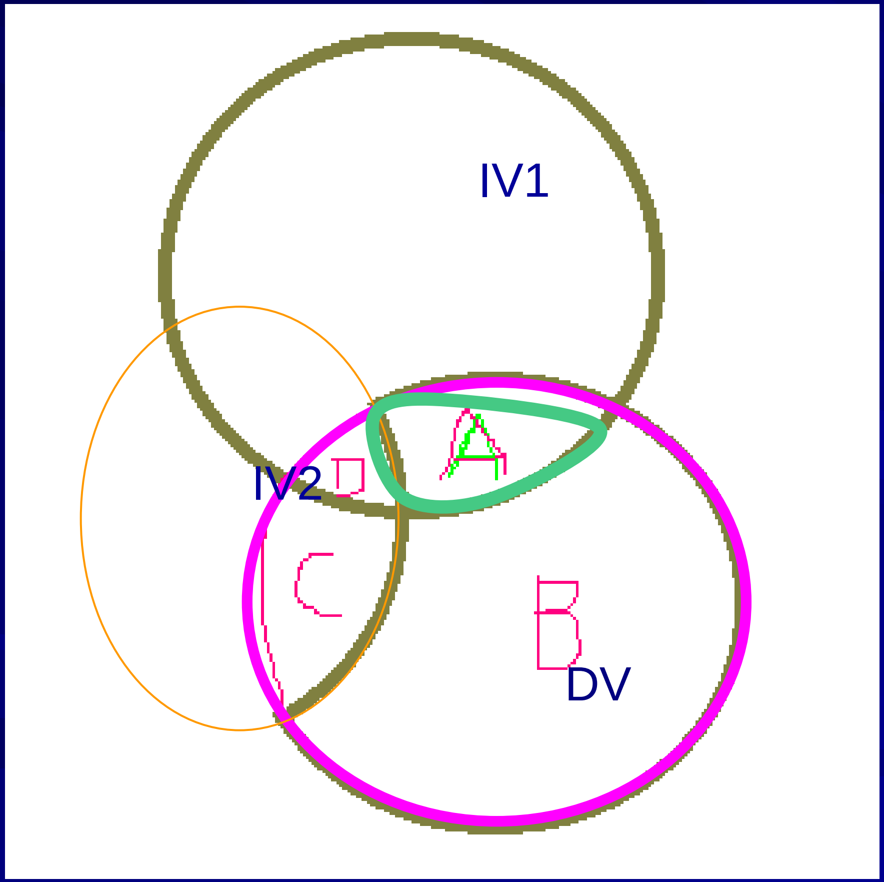
the partial correlation



examines the relationship between predictor 1 and the criterion, with the variance shared with predictor 2 partialled out of **BOTH**

pr^2 = the proportion of residual variance in the criterion uniquely accounted for by predictor 1 [A/(A+B)]

the semi-partial correlation



examines the relationship between predictor 1 and the criterion, after partialling out of predictor 1 the variance shared between predictor 1 and 2

spr² = the proportion of total variance in the criterion **UNIQUELY** accounted for by predictor 1 [A/(A+B+C+D)]

Very useful!

Difference between structure of ANOVA tests and MR tests

<u>ANOVA:</u>

1. No test of overall model
2. Tests main effect of each IV (differences in marginal means across levels of IV, regardless of other variables' effects – other variables assumed to be un-correlated [equal n, random assignment])
3. Tests all interactions automatically
4. Report Fs and effect sizes for each IV and interaction, plus relevant follow-ups

<u>Multiple Regression:</u>

1. Tests overall model automatically
2. Tests unique effect of each IV (i.e., covariation of residual DV scores with IV once all other IVs' effects are controlled (partialled out))
3. Does not test for interactions (unless you ask it to – moderated multiple regression, which we cover in 2 weeks)
4. Report Model R^2 with F test, plus each IVs' βs with t-tests, plus relevant follow-ups

Next week in class:

- Hierarchical vs standard regression

In the tutes:

- This week: Correlational designs, SPSS
- Next week: Multiple regression, SPSS

readings :

- Howell Ch 15
- Field Chapter 5