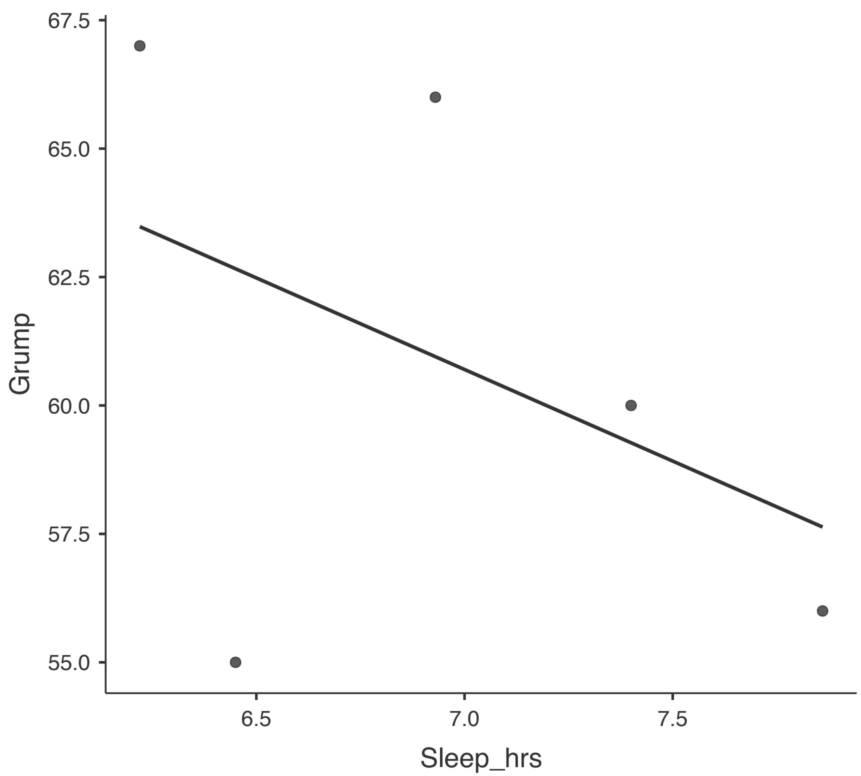
Learning Objectives

- Describe *linear regression* and construct (fit) linear models to data
- **Build** equations for simple linear regression and multiple regression
- **Calculate** and **interpret** *standard error of the estimate*
- Contrast r, r^2, R, and R^2

Calculate $s_{Y|X}$ from raw data

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

Night	Sleep (*X*)	Grump (*Y*)
9	7.40	60
24	7.86	56
28	6.93	66
60	6.22	67
99	6.45	55
N = 5	ΣX = 34.86	ΣY = 304



Calculate Standard Error of the Est.

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{[\sum XY - \frac{(\sum X)(\sum Y)}{N}]^2}{SS_X}}{N-2}}$$

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

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma Y)^2 = 92416$

$\Sigma Y^2 = 18606$

$\Sigma XY = 2113.03$

$SS_X = 1.811$

$SS_Y = 122.8$

$$s_{Y|X} = \sqrt{\frac{122.8 - \frac{[2113.03 - \frac{(34.86)(304)}{5}]^2}{1.811}}{5-2}}$$

Calculate $s_{Y|X}$

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$(\Sigma Y)^2 = 92416$
$\Sigma X^2 = 244.855$
$\Sigma Y^2 = 18606$
$\Sigma XY = 2113.03$
$SS_X = 1.811$
$SS_Y = 122.8$

$$s_{Y|X} = \sqrt{\frac{122.8 - \frac{[2113.03 \ - \mathbf{2119.488}]^2}{1.811}}{\mathbf{3}}}$$

Calculate $s_{Y|X}$

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$(\Sigma Y)^2 = 92416$
$\Sigma X^2 = 244.855$
$\Sigma Y^2 = 18606$
$\Sigma XY = 2113.03$
$SS_X = 1.811$
$SS_Y = 122.8$

$$s_{Y|X} = \sqrt{\frac{122.8 - \frac{[\mathbf{-6.458}]^2}{1.811}}{3}}$$

Calculate $s_{Y|X}$

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$(\Sigma Y)^2 = 92416$
$\Sigma X^2 = 244.855$
$\Sigma Y^2 = 18606$
$\Sigma XY = 2113.03$
$SS_X = 1.811$
$SS_Y = 122.8$

$$s_{Y/X} = \sqrt{\frac{122.8 - \frac{\mathbf{41.706}}{1.811}}{3}}$$

Calculate $s_{Y|X}$

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

N = 5
ΣX= 34.86
ΣY= 304
$(\Sigma X)^2$= 1215.22
$(\Sigma Y)^2$= 92416
ΣX^2= 244.855
ΣY^2= 18606
ΣXY= 2113.03
SS_X = 1.811
SS_Y = 122.8

$$s_{Y/X} = \sqrt{\frac{122.8 - \mathbf{23.03}}{3}}$$

Calculate $s_{Y|X}$

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$(\Sigma Y)^2 = 92416$
$\Sigma X^2 = 244.855$
$\Sigma Y^2 = 18606$
$\Sigma XY = 2113.03$
$SS_X = 1.811$
$SS_Y = 122.8$

$$s_{Y/X} = \sqrt{\frac{\mathbf{99.77}}{3}}$$

Calculate $s_{Y|X}$

$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$(\Sigma Y)^2 = 92416$
$\Sigma X^2 = 244.855$
$\Sigma Y^2 = 18606$
$\Sigma XY = 2113.03$
$SS_X = 1.811$
$SS_Y = 122.8$

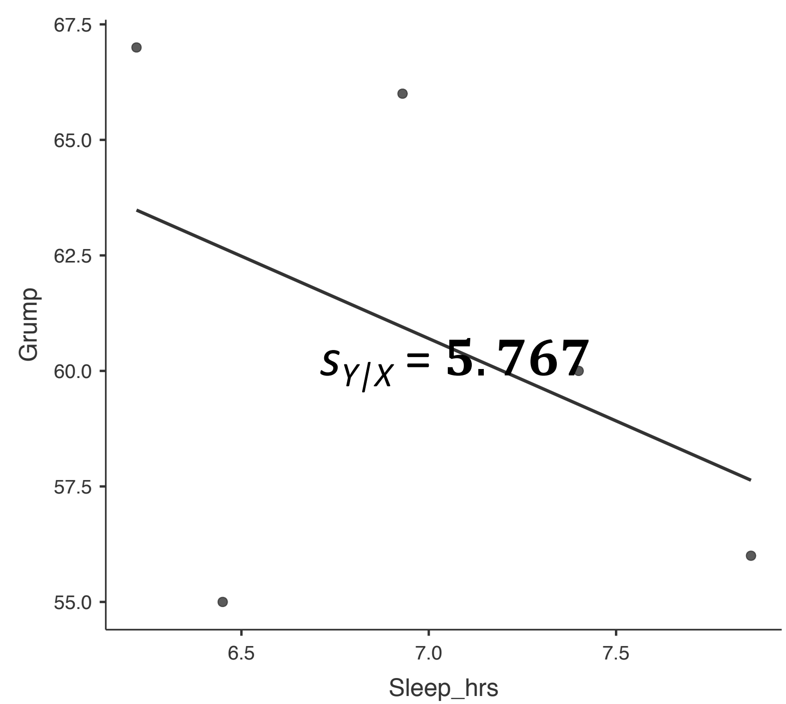
$$s_{Y/X} = \sqrt{\mathbf{33.257}}$$

Calculate $s_{Y|X}$

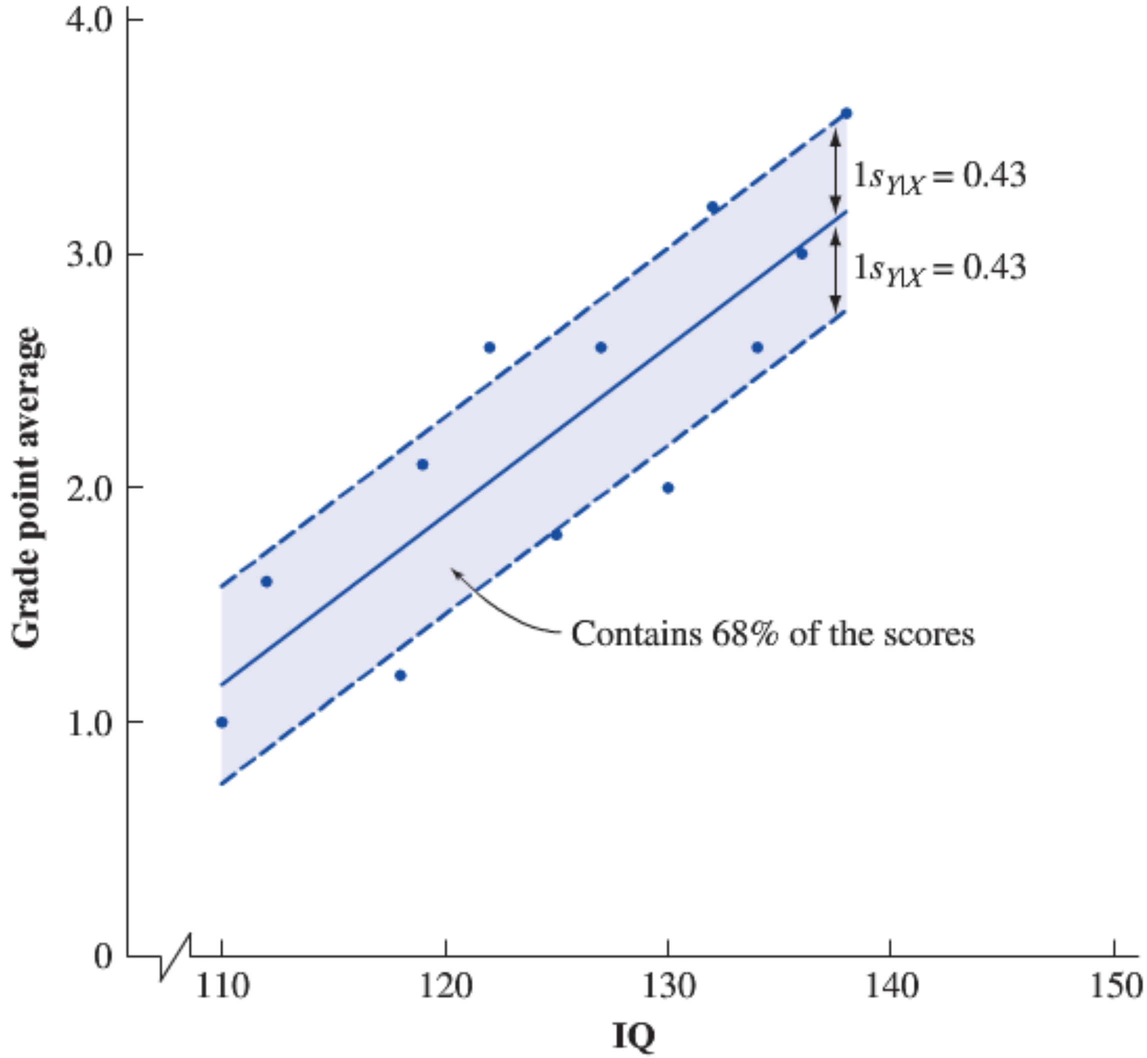
$$s_{Y|X} = \sqrt{\frac{SS_Y - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{SS_X}}{N-2}}$$

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

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$(\Sigma Y)^2 = 92416$
$\Sigma X^2 = 244.855$
$\Sigma Y^2 = 18606$
$\Sigma XY = 2113.03$
$SS_X = 1.811$
$SS_Y = 122.8$

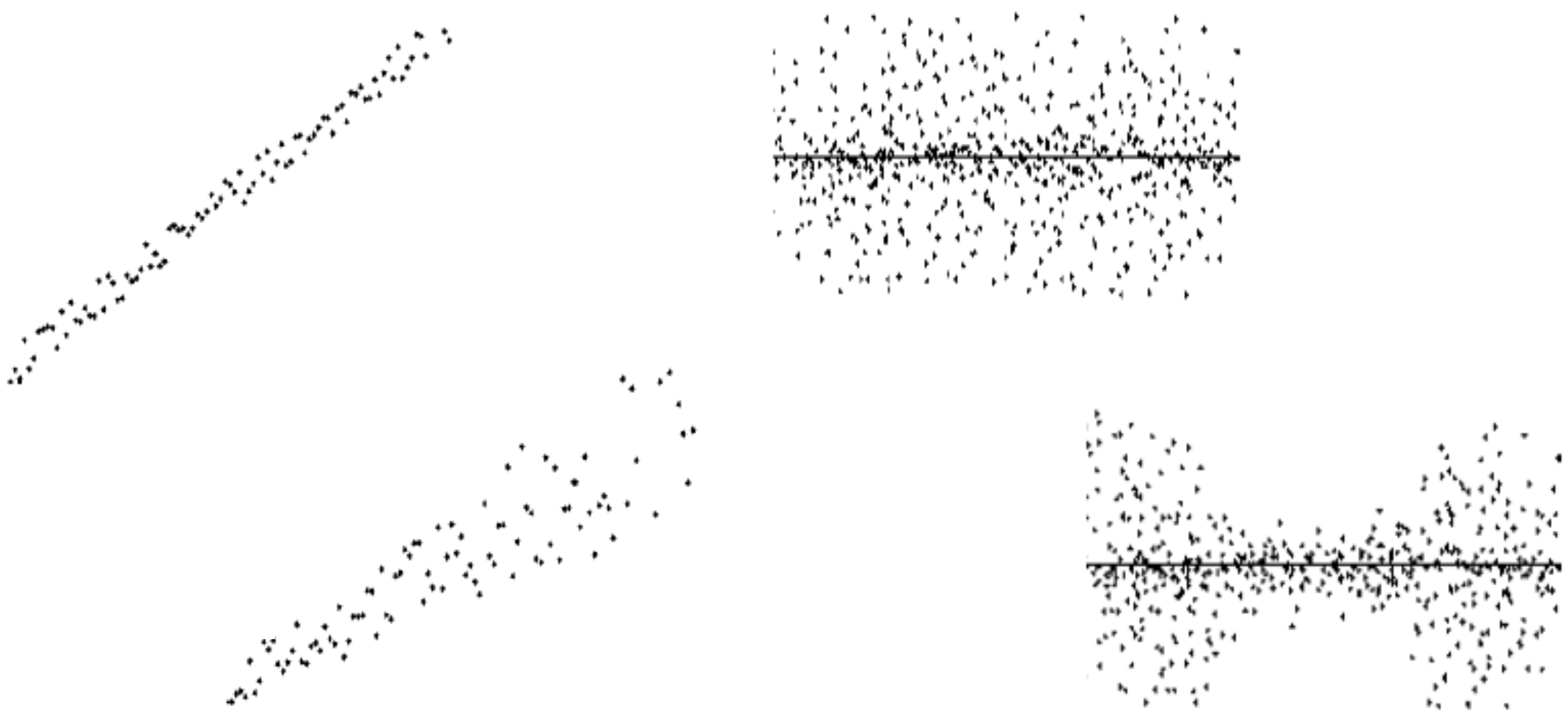


Homoscedasticity = consistent error (or same scatter)



Homoscedasticity = consistent error (or same scatter)

- Beware! Standard errors are only meaningful if the variability in Y is constant over values of X



Multiple Regression

- Regression model contains 2 or more predictors
  - Still have only 1 criterion

$$\boxed{Y' = b_Y X + a_Y}$$

$$\boxed{Y' = b_1 X_1 + b_2 X_2 + a_Y}$$

- Quantitatively, our prediction will *always* improve…how would we know?
  - Standard error of the estimate ($s_{y|x}$) decreases, or…
  - <u>Multiple</u> coefficient of determination (R^2) increases

Example: Predicting Happiness

- **Two predictors**: Income & Optimism

$$Y'_{\text{happy}} = b_{\$}X_{\$} + b_{opt.}X_{opt.} + a_{\text{Y}}$$

Correlation matrix			
	Optimism	Income	Happiness
Optimism	1	0.23	$\boxed{0.56}$
Income	0.23	1	$\boxed{0.48}$
Happiness	0.56	0.48	1

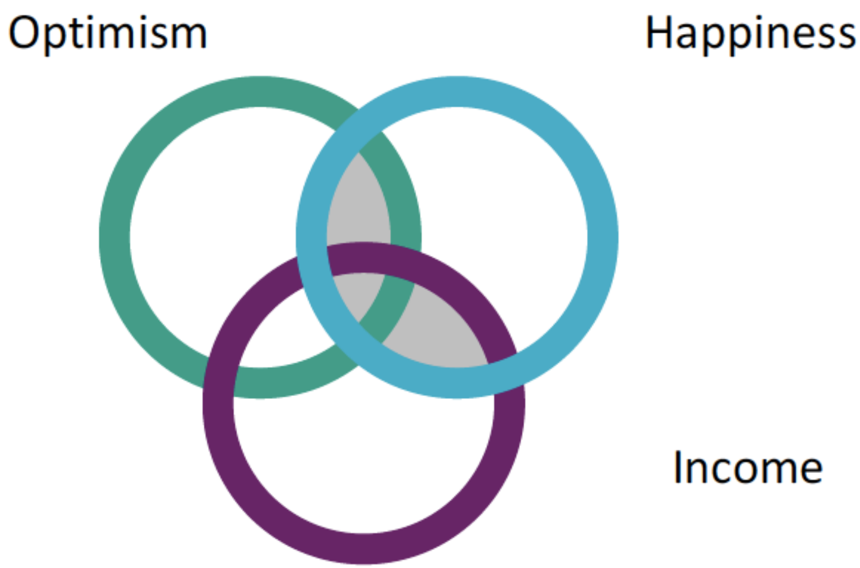
$r^2 = .31$

$r^2 = .23$

Trick Question: What is R^2 ?

R^2 in Multiple Regression

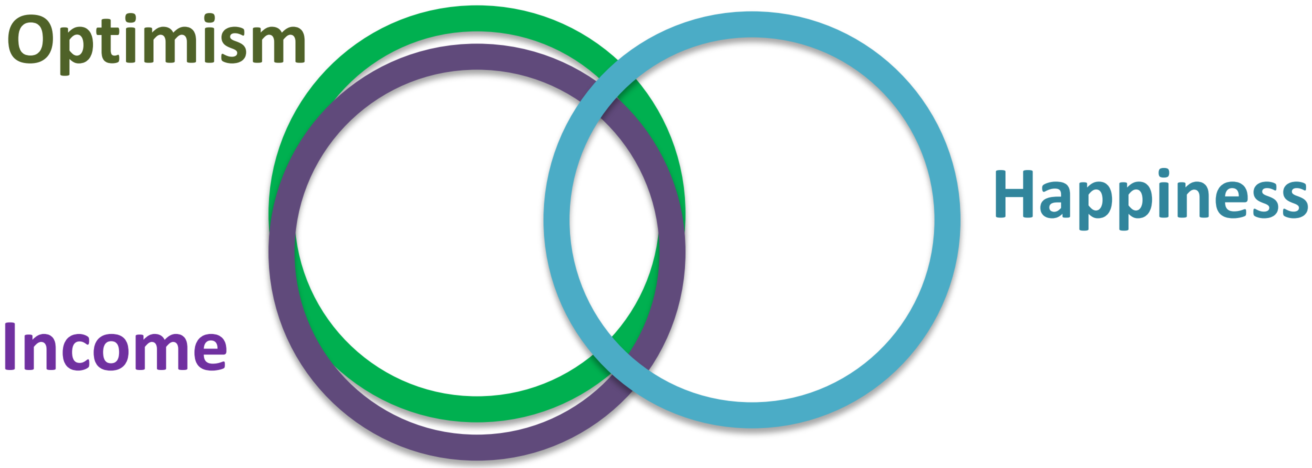
- We cannot simply add up the r^2 values



- **Goal of Prediction:** shade as much of the blue circle as possible
  - **Challenge:** Predictors might be explaining the same variability

R^2 in Multiple Regression

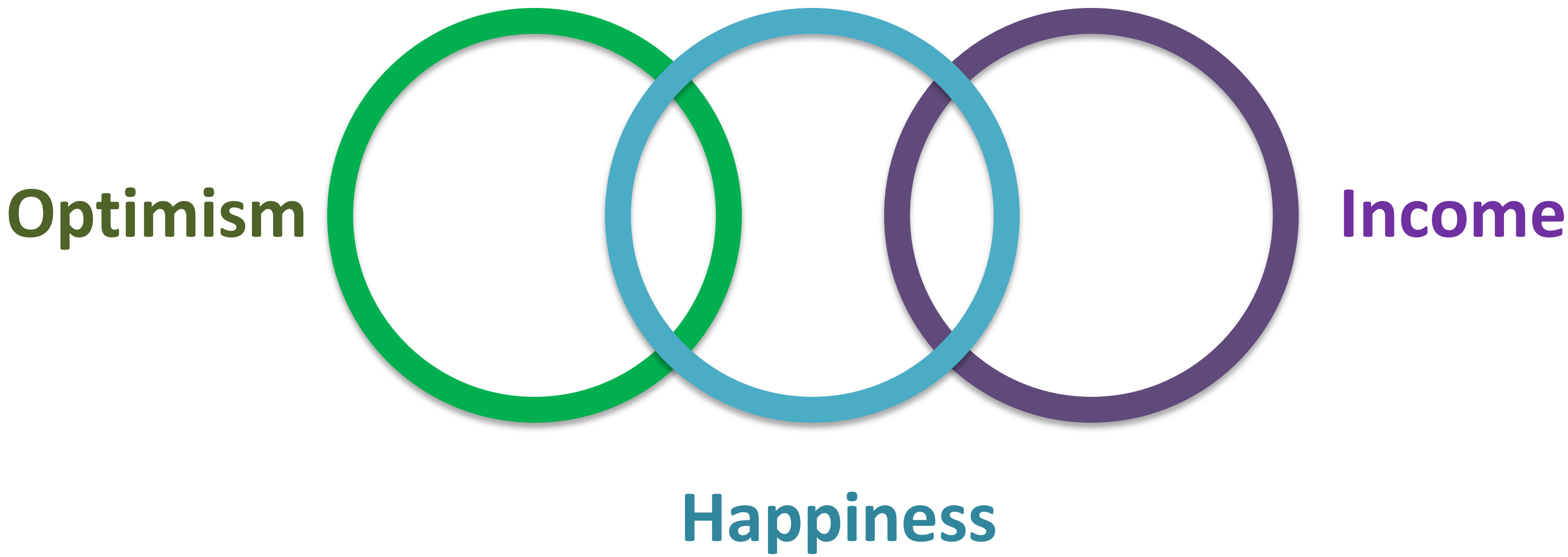
- **Worst case**: Fully redundant predictors



- **Goal:** shade as much of the blue circle as possible
  - **Challenge:** Predictors explaining exact same variability

R^2 in Multiple Regression

- **Best case**: Orthogonal predictors



- **Goal:** shade as much of the blue circle as possible
  – Predictors explain unique variability

Symbology Old & New

r

r_s or ρ

r_{pb}

ϕ

β = *standardized* slope coefficient

z_x, z_y

b = *unstandardized* slope coefficient

X_i, Y_i

r^2 = *variability explained by predictor variable*

R^2 = *variability explained by regression model*

R^2 Formula: Multiple Regression

Coefficient of Determination:
$Y \sim X_1$

$r^2 = Y \sim X_2$

Get rid of parts we double-counted

$$R^2 = \frac{r^2_{YX_1} + r^2_{YX_2} - 2(r_{YX_1} \times r_{YX_2} \times r_{X_1X_2})}{1 - r^2_{X_1X_2}}$$

Residual variability of
$X_1 \sim X_2$

Meehl's 6th Law of Psychology

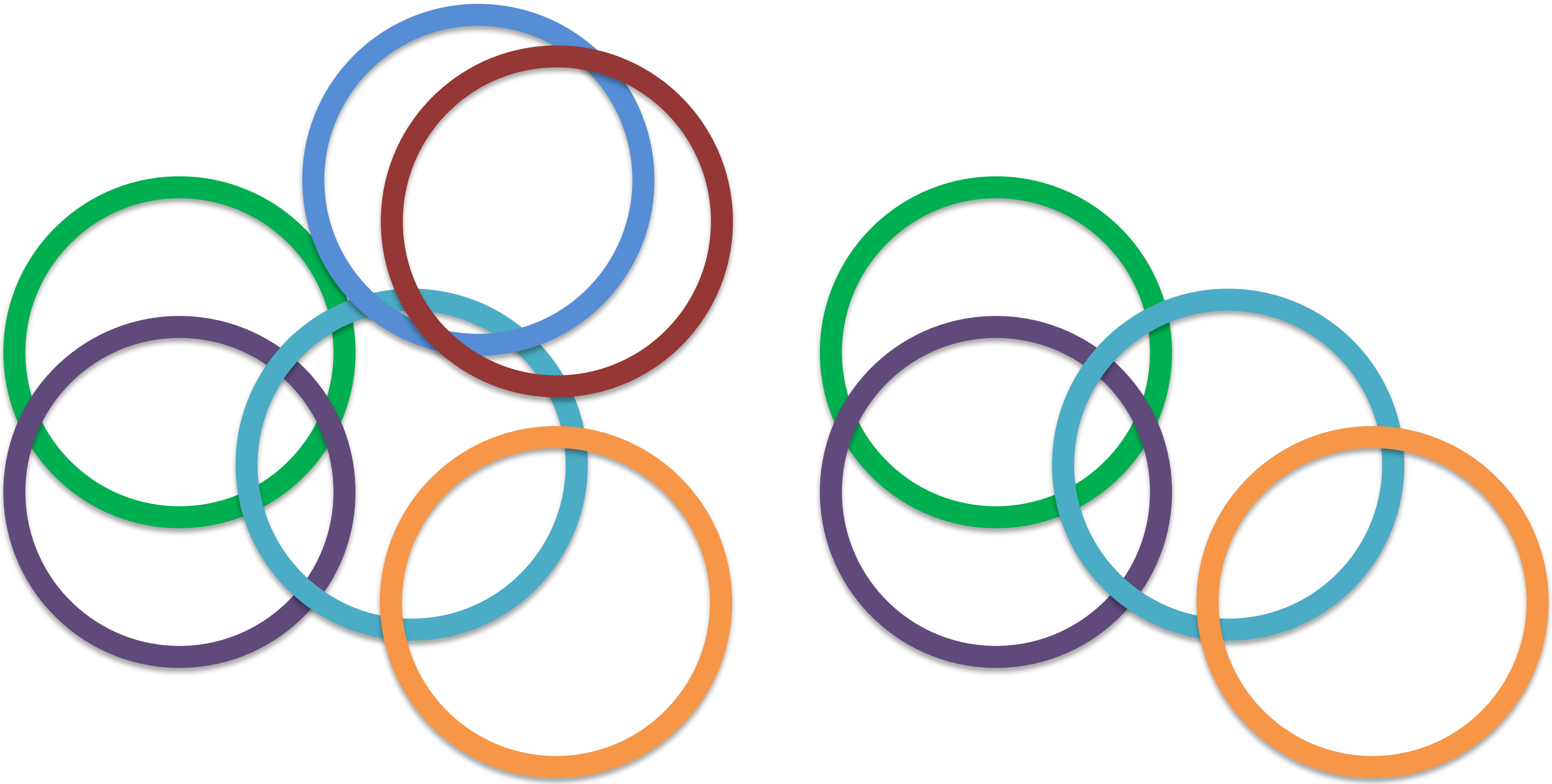
"Damnit; everything correlates with everything else!"

$R^2_{\text{adj.}}$

- If everything correlates with everything else…
  - Then, every predictor correlates w/criterion to some extent
  - Some of these correlations will be ***spurious*** ☹
    - "Crud" is a meaningless correlation
    - Results from *capitalization on chance*

- Adding predictors *always* increases R^2, even crud predictors
  - $R^2_{\text{adj.}}$ is a penalized R^2
    - Each new predictor could be crud, so let's be cautious

“Everything correlates with everything else”



Another approach to Meehl’s law: **Try** to trash all crud predictors

Categorical Predictors in Regression:

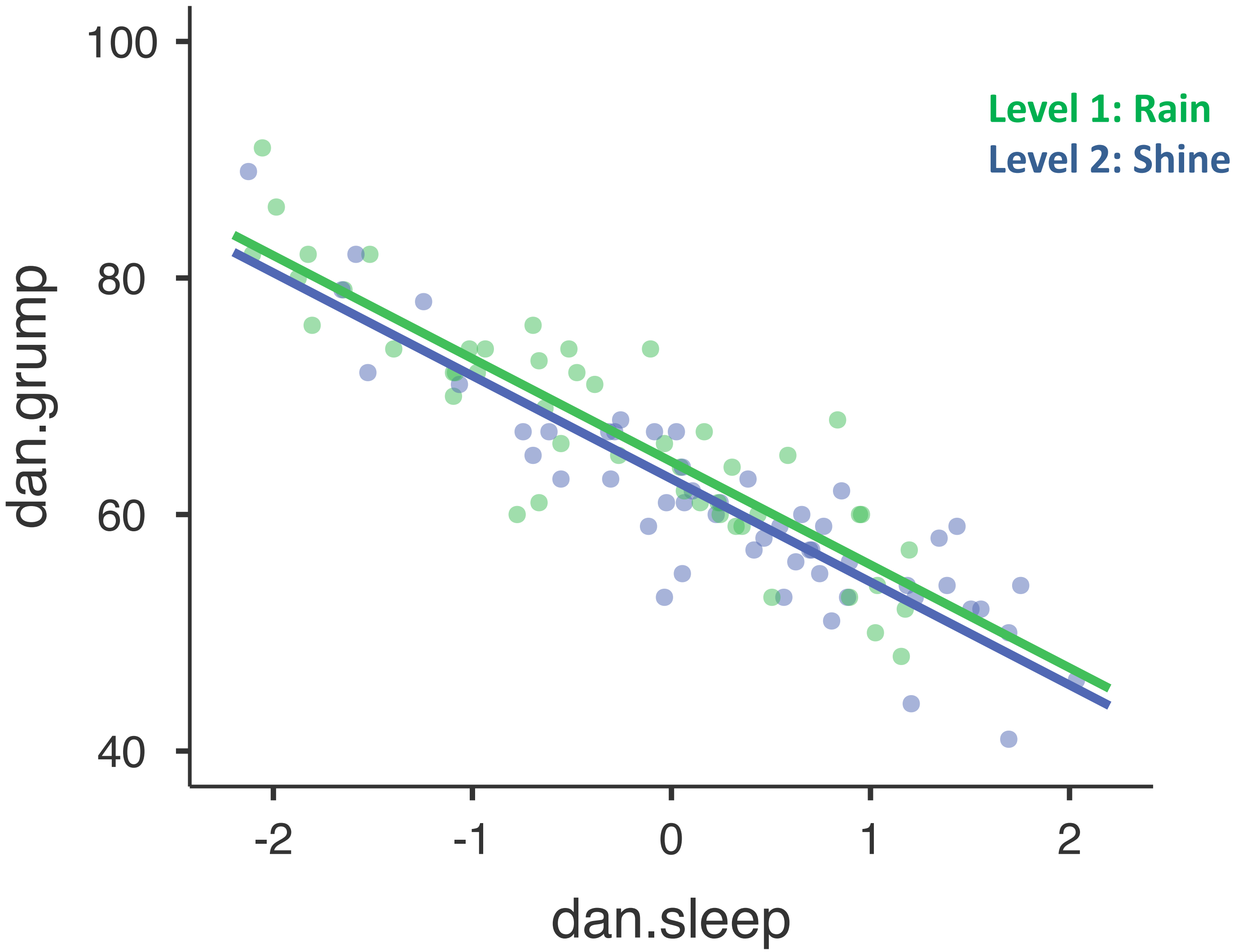
Sleep, Weather, and Grumpiness

Sleep, Weather & Grumpiness

$$Y' = b_1 X_1 + b_2 X_2 + a$$

- Criterion Y: How grumpy is Dr. Dan?
- Predictor X_1: Hours of sleep
- Predictor X_2: $Rain_{L1}$ or $Shine_{L2}$
  - b_1 = -8.715
  - b_2 = -1.455
  - a = 124.442
- If rain level of X_2 is coded as ‘1’ and shine as ‘2’, then good weather predicts 1.455 fewer grumpy units

Figure. Regression lines predicting grumpiness by continuous predictor sleep (standardized) and by categorical predictor (rain vs. shine).



Coding Categorical Predictors

$$Y' = -8.715 * X_1 + -1.455 * X_2 + 124.442$$

- Criterion Y: How grumpy is Dr. Dan?
- Predictor X_1: Hours of sleep
- Predictor X_2: Rain_{L1} or Shine_{L2}
  - $b_1 = -8.715$
  - $b_2 = -1.455$
  - $a = 124.442$
- What if Dan sleeps 5 hours and it's raining?

Coding Categorical Predictors

$$Y' = b_1X_1 + b_2X_2 + a$$

- Criterion Y: How grumpy is Dr. Dan?
- Predictor X_1: Hours of sleep
- Predictor X_2: $\mathbf{Rain_{L1}}$, $\mathbf{Clouds_{L2}}$, or $\mathbf{Shine_{L3}}$
  - b_1 = -8.968
  - b_2 = ???
    - Estimate 1 and Estimate 2???
  - a = 126.178

Coding Categorical Predictors

- Criterion Y: How grumpy is Dr. Dan?
- Predictor X_1: Hours of sleep
- Predictor X_2: $\mathbf{Rain_{L1}}$, $\mathbf{Clouds_{L0}}$, or $\mathbf{Shine_{L0}}$
  - ***Dummy variable*** (for Rain)

  - $b_1 = -8.968$
  - $b_2 = 0.350$
  - $a = 125.908$

- What is the effect of rain on grumpiness?

Coding Categorical Predictors

- Coding of categorical predictors changes interpretation of *b*

- ***Dummy coding*** is most common and simplest method

- Many other coding methods exist depending on hypotheses