Learning Objectives

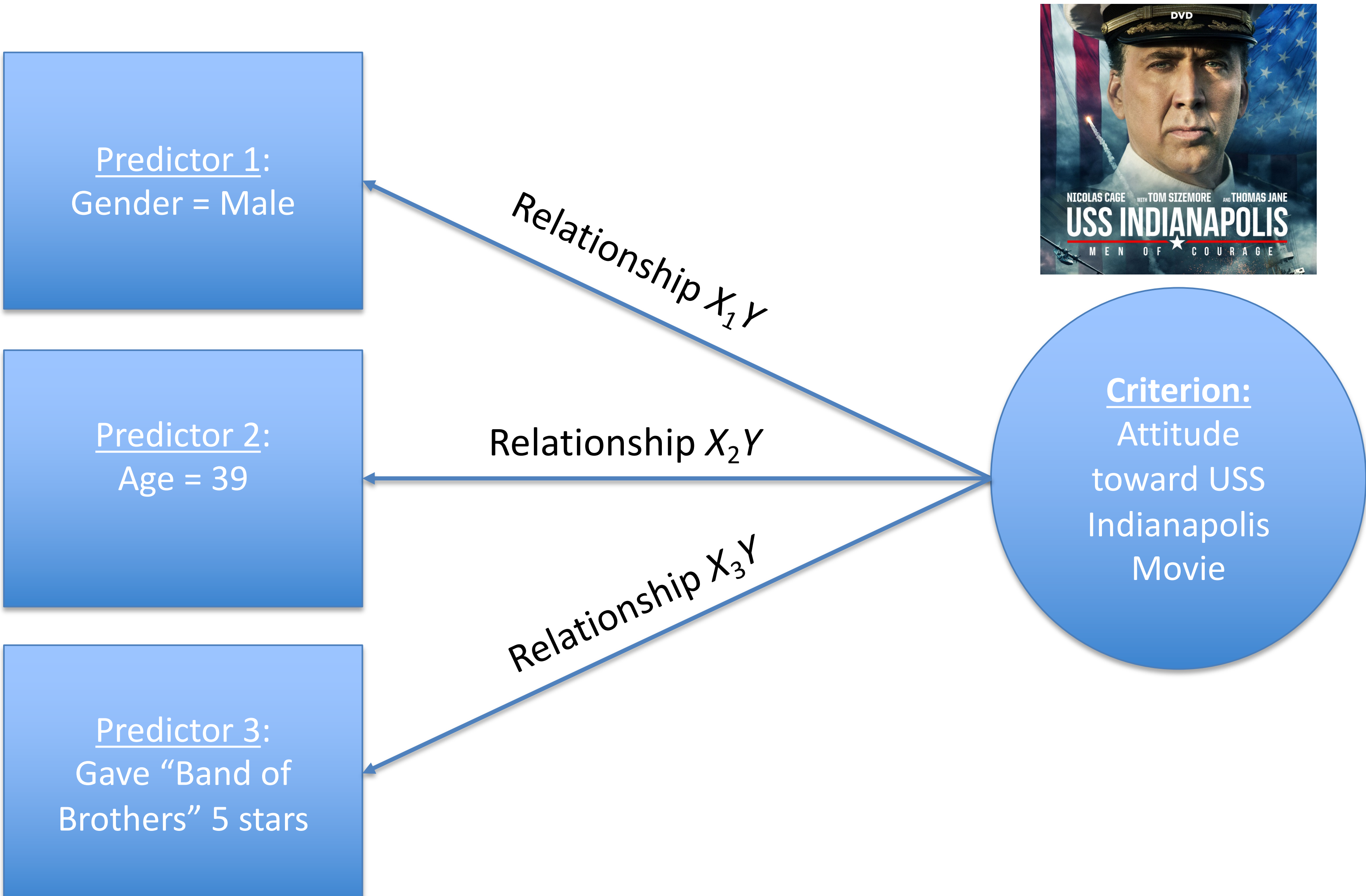
- **Describe** *linear regression* and fit linear models to observed data
- **Build** equations for simple linear regression and multiple regression
- **Calculate** and **interpret** *standard error of the estimate* ($s_{y|x}$)
- **Contrast** r, r^2, R, and R^2

Regression is making predictions

- How does *Youtube* decide what video to show next?
- How does *Netflix* know how well I'll like a movie I've never seen?
- What is machine learning?

Answer: Regression, regression, regression!

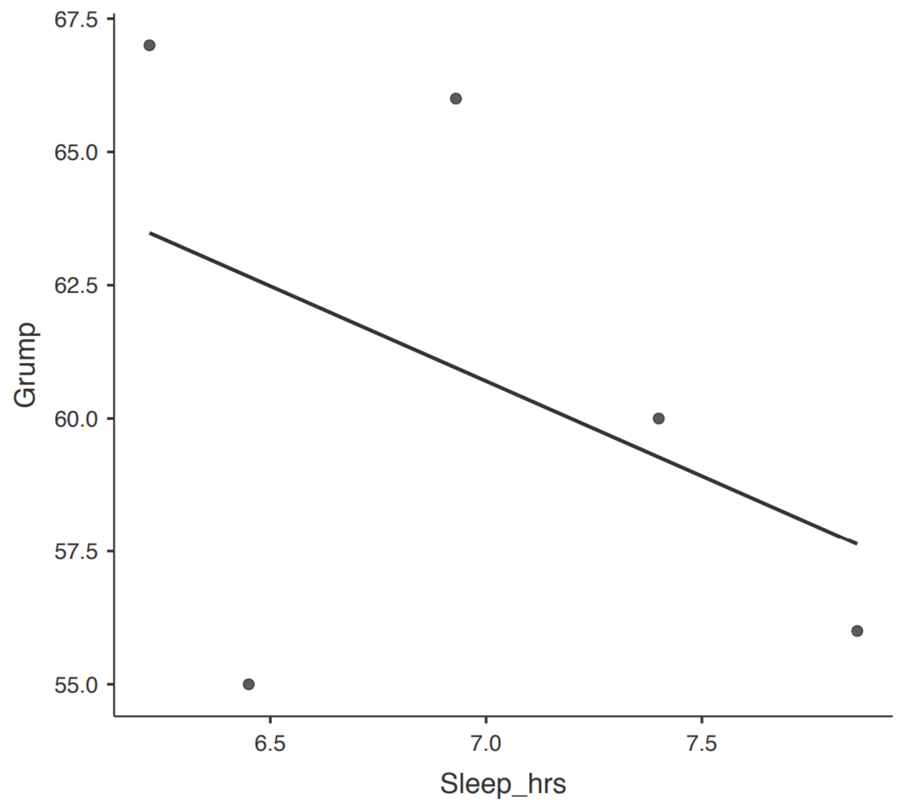
Path diagram of multiple regression



How grumpy is Dr. Dan?

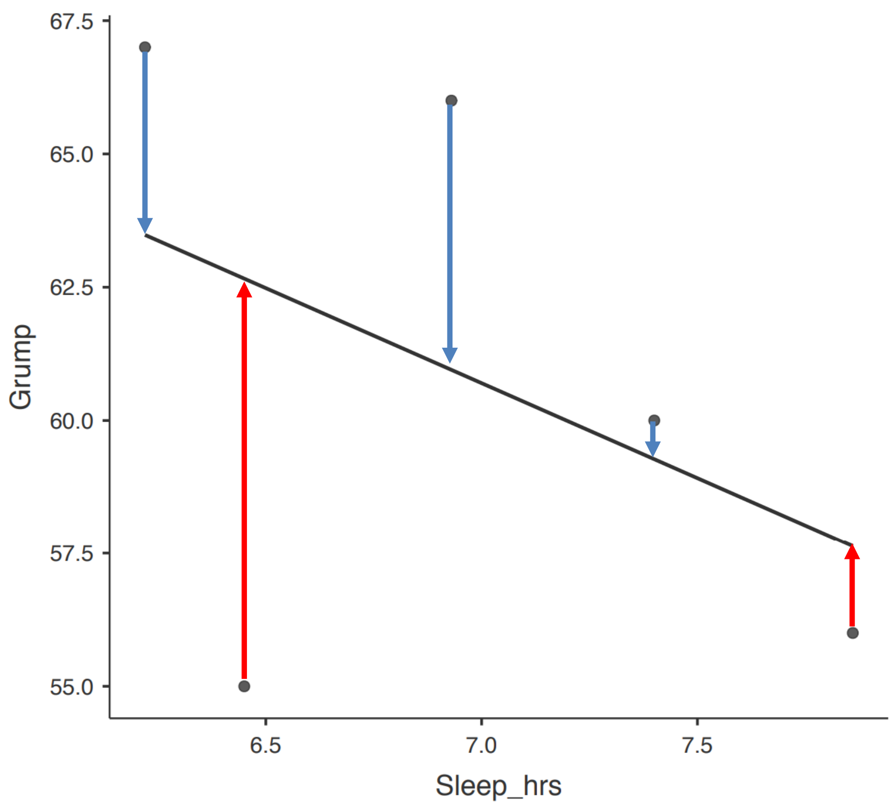
Simple regression

- <u>Criterion</u> *(Y):* Self-reported grumpiness on 0-100 scale
- <u>Predictor</u> (*X*): Self-reported hours of sleep



- *Line of best fit* is a model that minimizes prediction errors for *Y* ('criterion' or 'predicted' variable)
  - Balances the magnitude of positive and negative errors
    - What model is this similar to?
  - Minimizes $\Sigma(Y - Y')^2$: **Least squares regression line**

$$\sum \;\; ^2 + \;\; ^2 + \;\; ^2 + \;\; ^2 + \;\; ^2$$



Line of Best Fit = Regression Line

Formula for linear regression line:

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

Y' = Criterion variable (′ =predicted; Grumpiness)

X = Predictor variable (hrs of sleep)

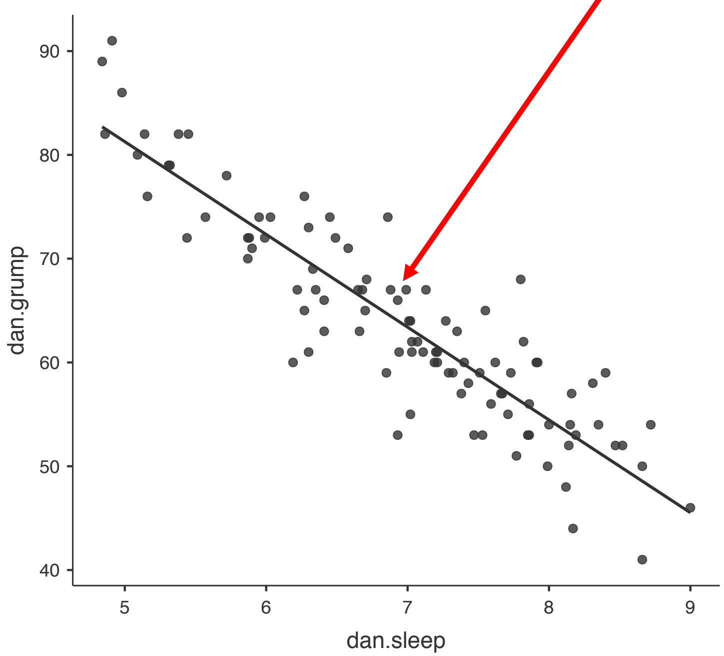
b_Y = Slope of regression line

a_Y = intercept (when sleep = 0, how 😡 is Dan?)

Line of Best Fit = Regression Line

$$Y' = b_Y X + a_Y$$

$$Y' = -8.94(X) + 125.96$$



- the regression line?

Calculating Slope, b_Y

Formula 1: Simple, when you know r, s_Y, and s_X

$$b_Y = r\frac{s_Y}{s_X}$$

Formula 2: When you have only raw data

$$b_Y = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{SS_X}$$

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

Note: N = paired X & Y scores

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

Night	Sleep (*X*)	Grump (*Y*)	X^2	Y^2	*XY*
9	7.40	60	54.76	3600	444
24	7.86	56	61.78	3136	440.16
28	6.93	66	48.025	4356	457.38
60	6.22	67	38.688	4489	416.74
99	6.45	55	41.602	3025	354.75
$N = 5$	$\Sigma X = 34.86$	$\Sigma Y = 304$	$\Sigma X^2 =$ 244.855	$\Sigma Y^2 =$ 18606	$\Sigma XY =$ 2113.03

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x}$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

Night	Sleep (*X*)	Grump (*Y*)	X^2	Y^2	*XY*
9	7.40	60	54.76	3600	444
24	7.86	56	61.78	3136	440.16
28	6.93	66	48.025	4356	457.38
60	6.22	67	38.688	4489	416.74
99	6.45	55	41.602	3025	354.75
$N = 5$	$\Sigma X = 34.86$	$\Sigma Y = 304$	$\Sigma X^2 = 244.855$	$\Sigma Y^2 = 18606$	$\Sigma XY = 2113.03$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$
$\Sigma X = 34.86$
$\Sigma Y = 304$
$(\Sigma X)^2 = 1215.22$
$\Sigma X^2 = 244.855$
$\Sigma XY = 2113.03$
$SS_X = ???$

Step 1: Calculate SS_X

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$SS_X = ???$

$$SS_X = 244.855 - \frac{1215.22}{5}$$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$\boldsymbol{SS_X = 1.811}$

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x}$$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$SS_X = 1.811$

$$b_Y = \frac{2113.03 - \frac{(34.86)(304)}{5}}{1.811}$$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$SS_X = 1.811$

$$b_Y = \frac{2113.03 - \frac{\mathbf{10597.44}}{5}}{1.811}$$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$SS_X = 1.811$

$$b_Y = \frac{2113.03 - \mathbf{2119.49}}{1.811}$$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$SS_X = 1.811$

$$b_Y = \frac{\mathbf{-6.46}}{1.811}$$

Calculate b_Y from raw data

$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_x} \qquad SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\Sigma Y = 304$

$(\Sigma X)^2 = 1215.22$

$\Sigma X^2 = 244.855$

$\Sigma XY = 2113.03$

$SS_X = 1.811$

$b_Y = -3.57$

Every extra hour of sleep, we predict 3.57 less grumpy units

Calculate a_Y from raw data

$$a_Y = \bar{Y} - b_Y\bar{X}$$

$$\bar{Y} = \frac{\Sigma Y}{N}; \quad \bar{X} = \frac{\Sigma X}{N}$$

$N = 5$

$\Sigma X = 34.86$

$\bar{X} = 6.97$

$\Sigma Y = 304$

$\bar{Y} = 60.8$

$SS_X = 1.811$

$b_Y = -3.57$

$$a_Y = 60.8 - -3.57(6.97)$$

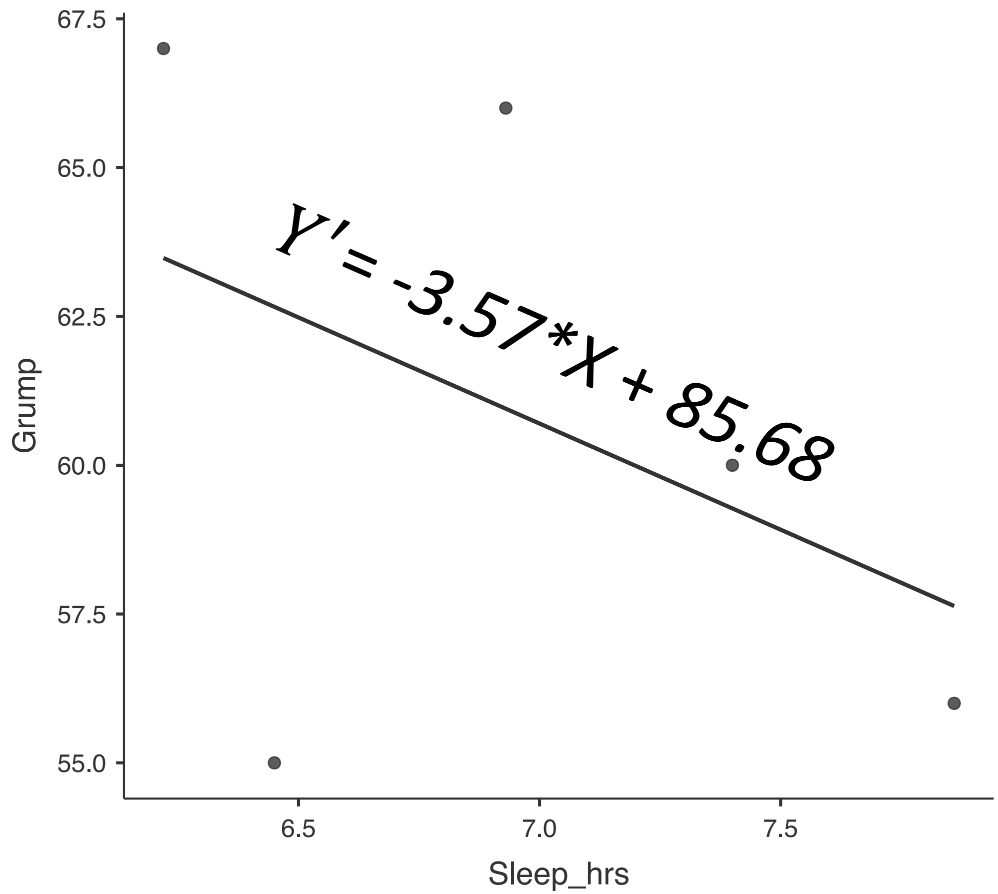
$$a_Y = 85.68$$

When Dan gets no sleep (X=0), we predict he'll be 85.68 units grumpy!!

$$Y' = b_Y X + a_Y$$

b_Y = -3.57

a_Y = 85.68



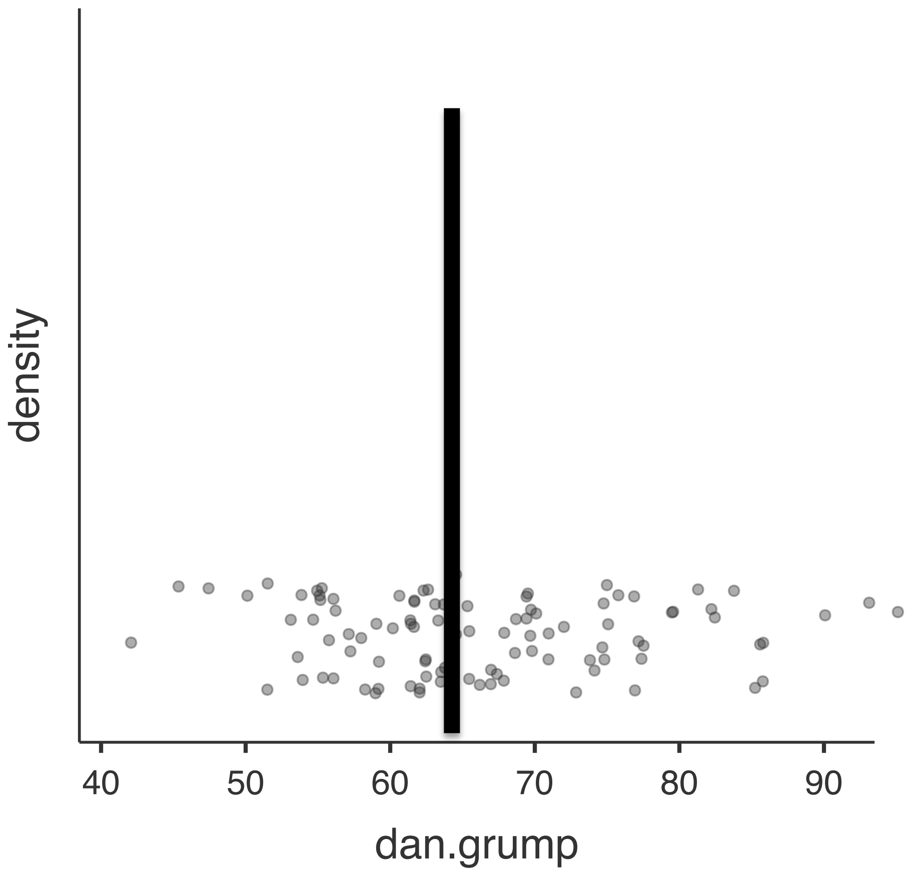
Required for Linear Regression

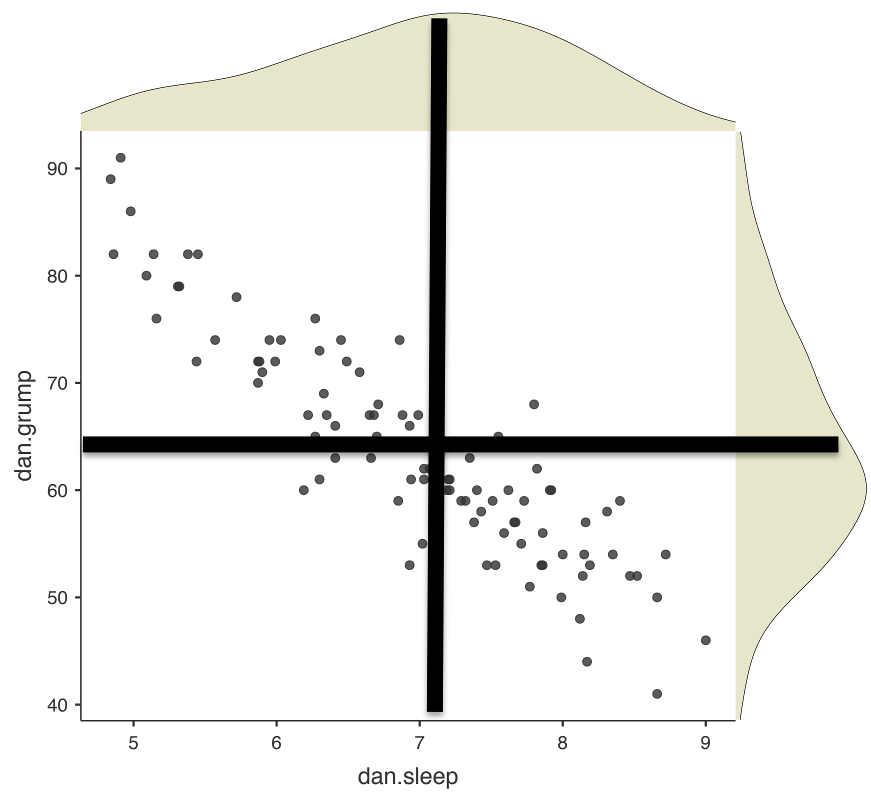
Must have:

- Linear relationship
- Sample relationship is representative (of linear model)
- Prediction is within the range of original variables
  - We would have low confidence in our predictions for very <u>low</u> and very <u>high</u> amounts of sleep

Two Models, lots of similarity

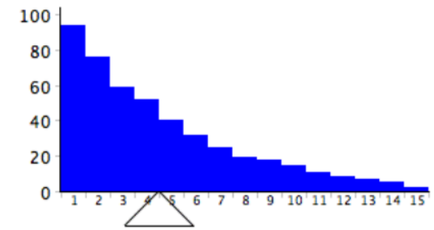
- Mean & correlation/linear regression

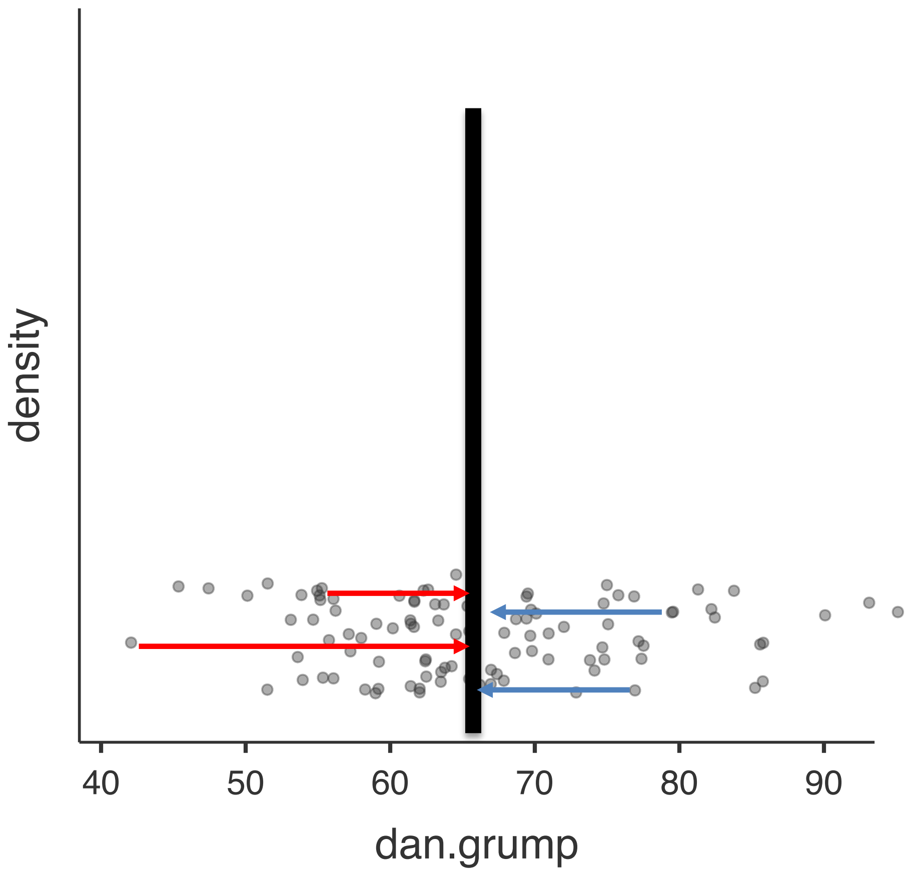


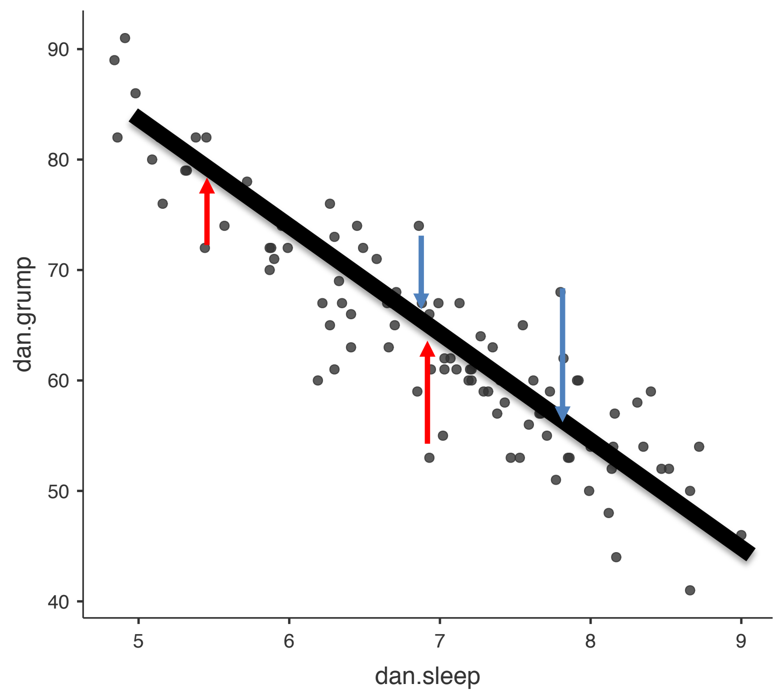


Two Models, lots of similarity

- <u>Best Prediction</u>: Mean/Line of Best Fit
  - Both are fulcrums, balancing the data

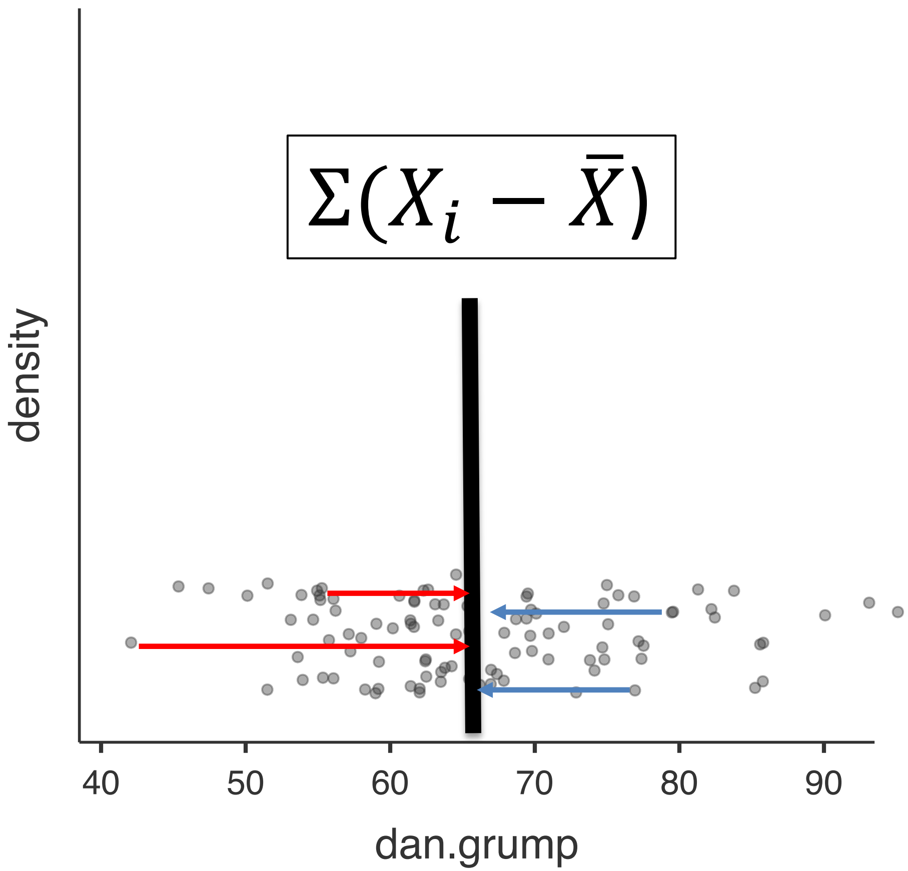


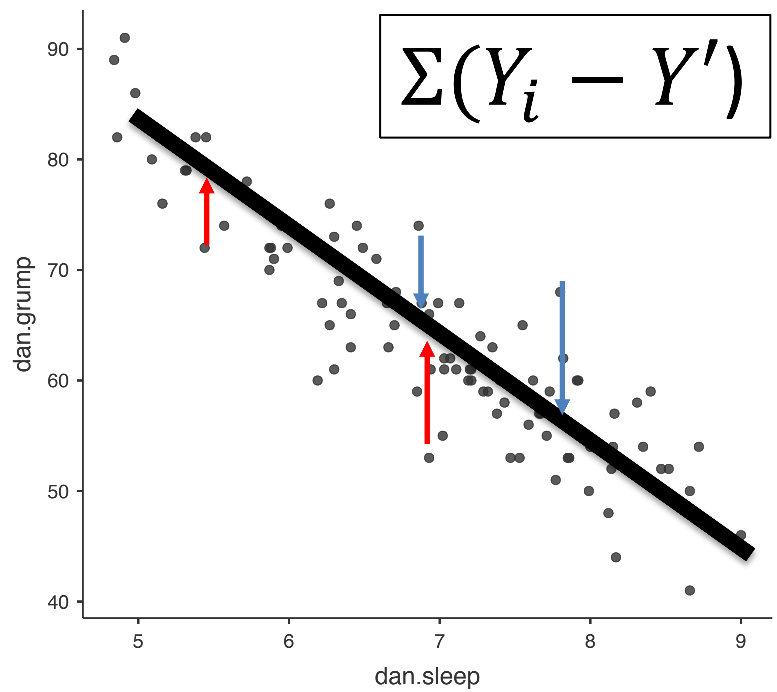




Two Models, lots of similarity

- How much error in our model:
  - *Deviation scores & Prediction errors*

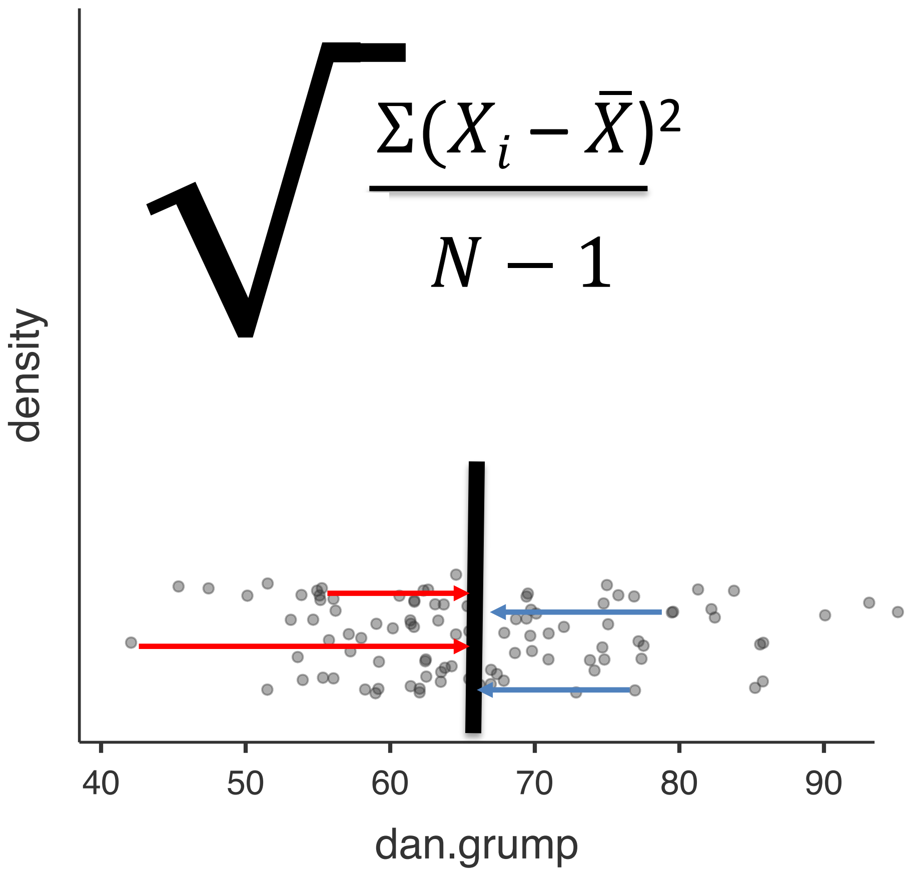




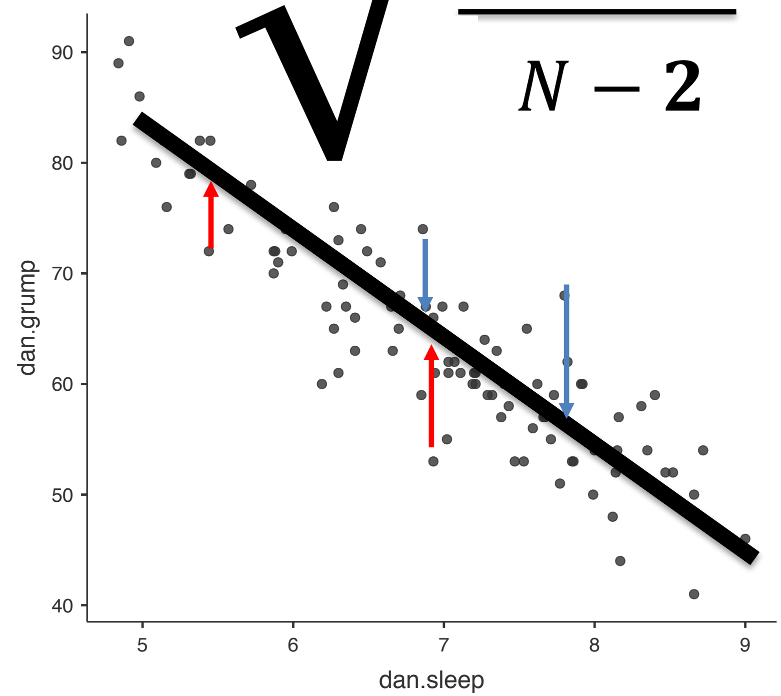
Two Models, lots of similarity

- <u>Average (mean) error</u>:
  - *s & Standard Error of the Estimate* ($s_{y|x}$)

$$\sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N-1}}$$

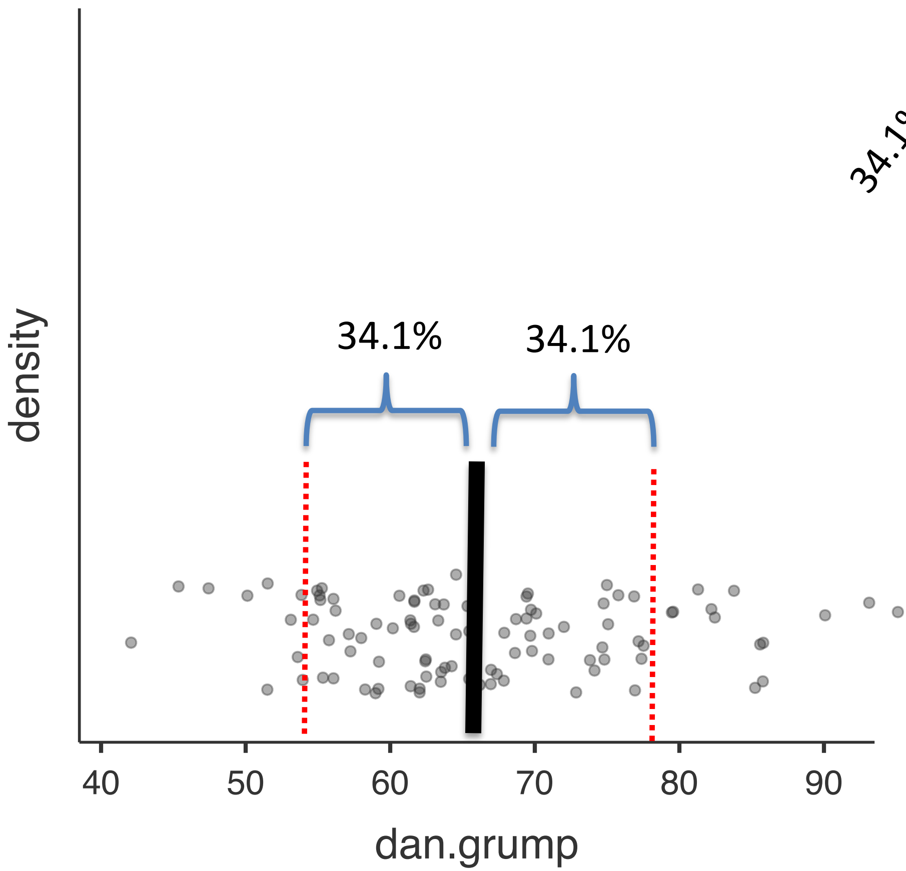


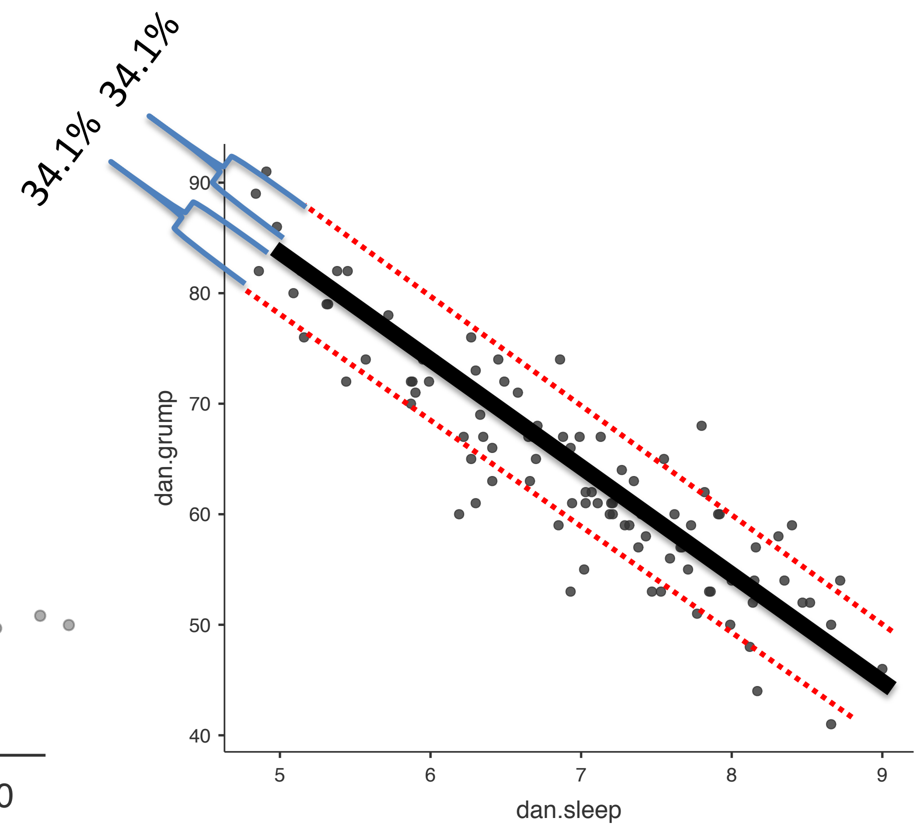
$$\sqrt{\frac{\boxed{\Sigma(Y_i - Y')^2}}{N-\mathbf{2}}}$$



Two Models, lots of similarity

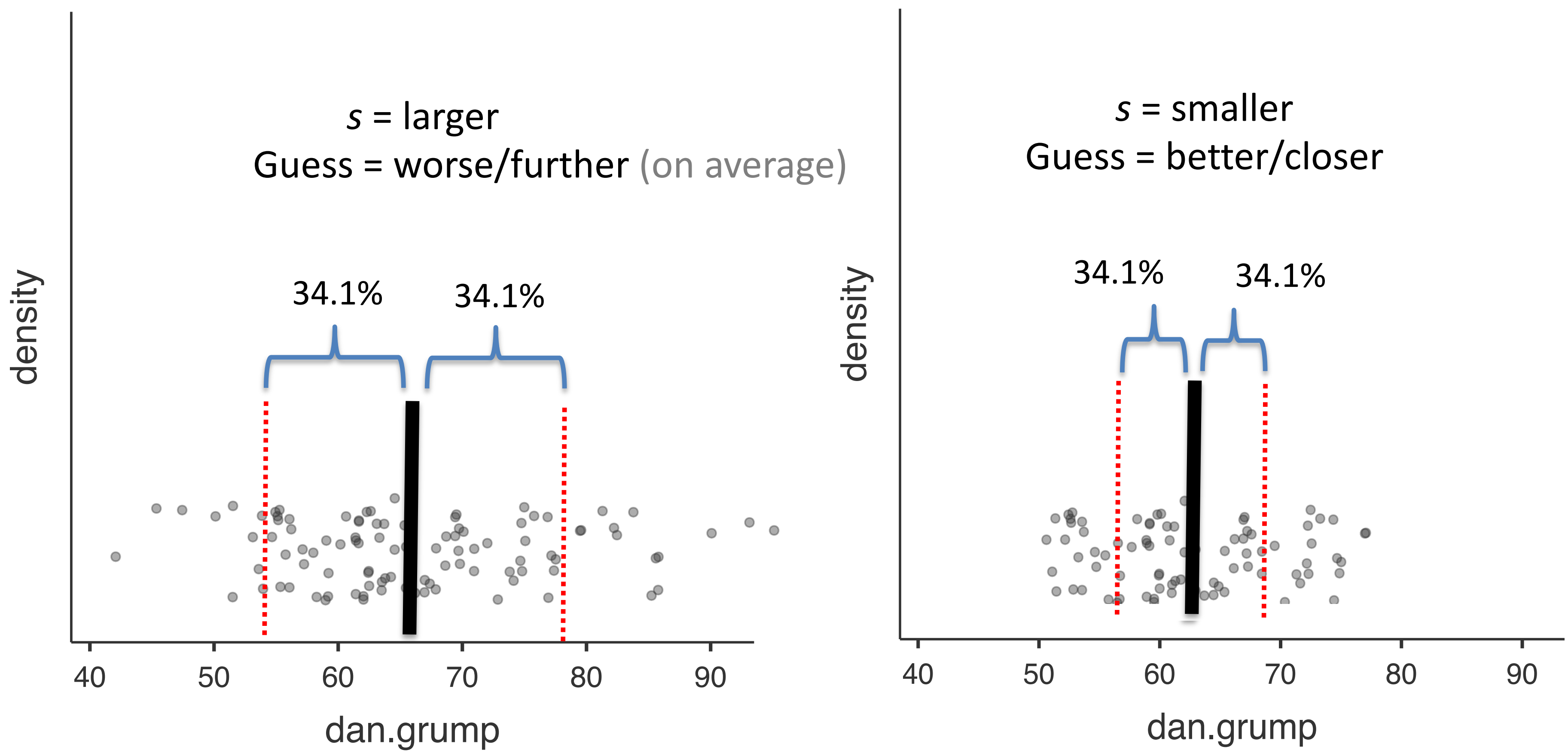
- <u>Distribution of scores</u>:**
  - Given s_x & $s_{y|x}$





Another way to think about variability

- Standard deviation: Average 'miss' when using $\bar{X}$ to predict a score



Another way to think about variability

- <u>Standard error of the estimate</u>: Average 'miss' when using the regression line to predict a score