variability partitioning

Dr. Mine Çetinkaya-Rundel
Duke University

variability partitioning

- So far: t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between *x* and *y*.
- Alternative: consider the variability in *y* explained by *x*, compared to the unexplained variability.
- Partitioning the variability in y to explained and unexplained variability requires analysis of variance (ANOVA).

anova output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

sum of squares

total variability in *y*: $SS_{Tot} = \sum (y - \bar{y})^2 = 6724.66$

unexplained variability in *y* (residuals): $SS_{Res} = \sum (y - \hat{y})^2 = \sum e_i^2 = 1493.53$

explained variability in *y*: $SS_{Reg} = 6724.66 - 1493.53 = 5231.13$

anova output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

degrees of freedom

total degrees of freedom: $df_{Tot} = 27 - 1 = 26$

regression degrees of freedom: $df_{Reg} = 1$ only 1 predictor

residual degrees of freedom: $df_{Res} = 26 - 1 = 25$

anova output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

mean squares

MS regression:

$$MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$$

MS residual:

$$MS_{Res} = \frac{SS_{Res}}{df_{Res}} = \frac{1493.53}{25} = 59.74$$

F statistic

ratio of explained to unexplained variability

$$F_{(1,25)} = \frac{MS_{Reg}}{MS_{Res}} = 87.56$$

anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

small p-value → reject H_0

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

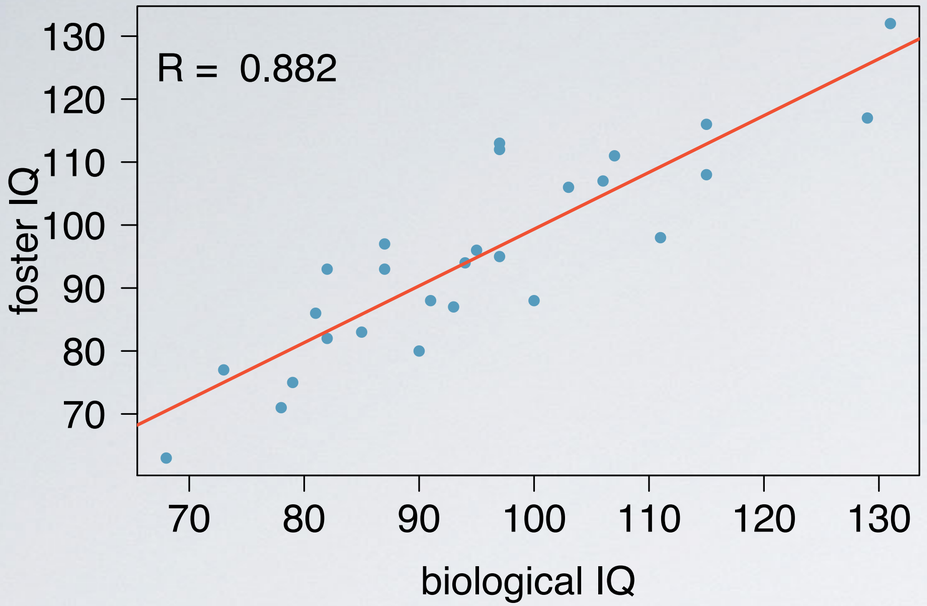
The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

revisiting R^2

- R^2 is the proportion of variability in *y* explained by the model:
  - large → linear relationship between *x* and *y* exists
  - small → evidence provided by the data may not be convincing
- Two ways to calculate R^2:

(1) using correlation: square of the correlation coefficient

(2) from the definition: proportion of explained to total variability



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

(1) $R^2 = \text{square of correlation coefficient} = 0.882^2 \approx 0.78$

(2) $R^2 = \dfrac{\text{explained variability}}{\text{total variability}} = \dfrac{SS_{Reg}}{SS_{Tot}} = \dfrac{5231.13}{6724.66} \approx 0.78$