CONTROL VARIABLES

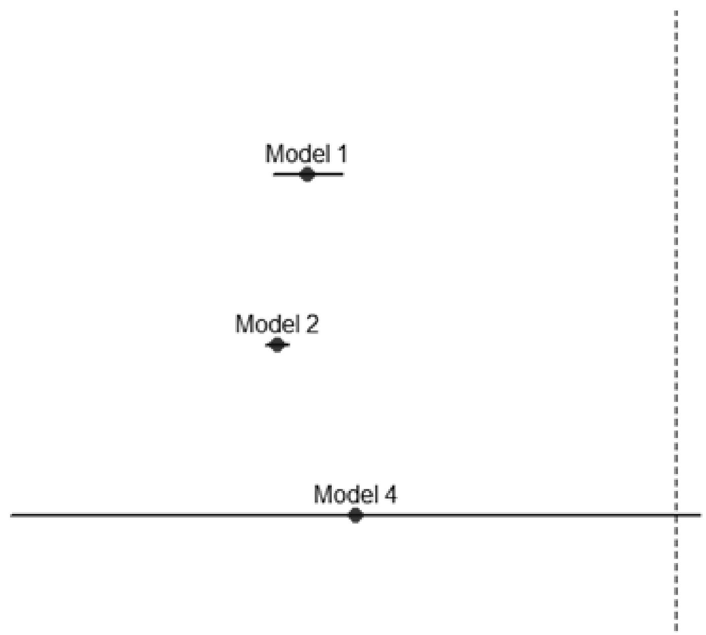
Fundamentals of

PROGRAM EVALUATION

JESSE LECY

	Dependent Variable: Test Scores				
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
Classroom Size	-4.22***	-3.91***		-2.67	-2.22***
	(0.18)	(0.03)		(1.63)	(0.23)
Teacher Quality		55.01***	55.03***		55.01***
		(0.25)	(0.26)		(0.25)
Socio-Economic Status			40.94***	16.34	17.77***
			(0.27)	(17.10)	(2.40)
Intercept	738.34***	456.70***	272.91***	665.29***	377.26***
	(4.88)	(1.48)	(1.39)	(76.57)	(10.82)

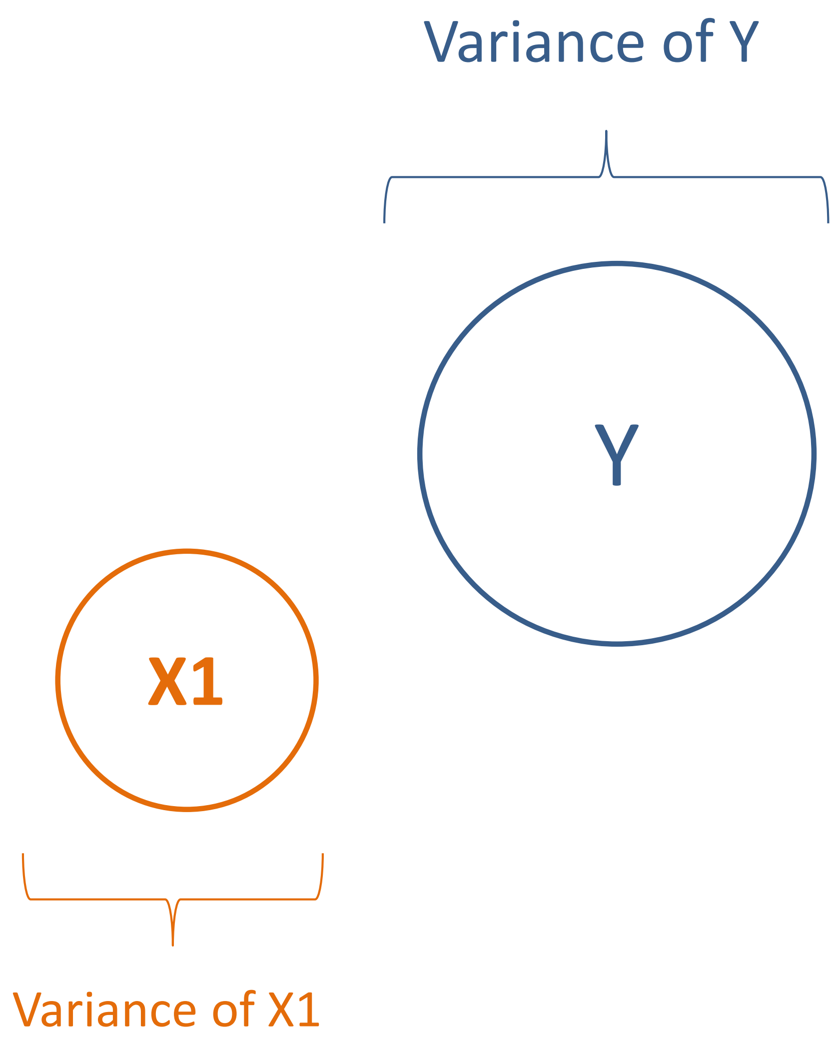
Why are **slopes** and **standard errors** changing when we add **"control" variables**?



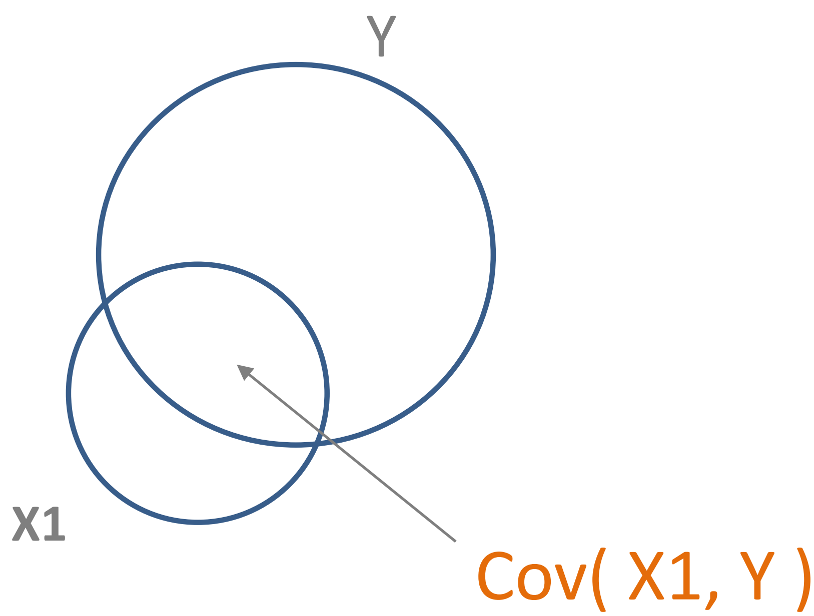
Visual representations of multiple regression models to allow for reasoning regarding model specification and fit.

BALLENTINE VENN DIAGRAMS

BALLENTINE VENN DIAGRAM

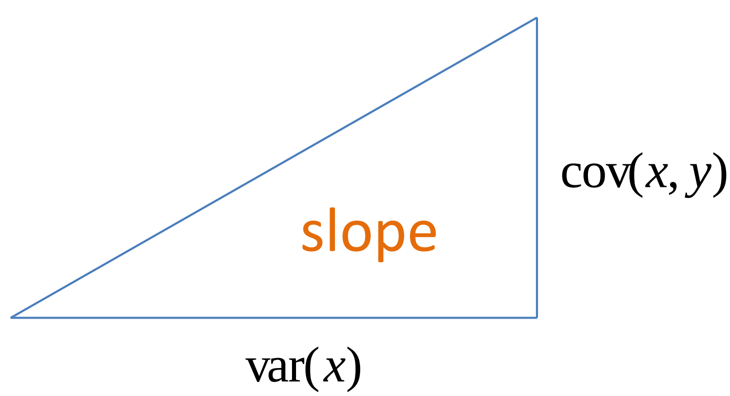


BALLENTINE VENN DIAGRAM



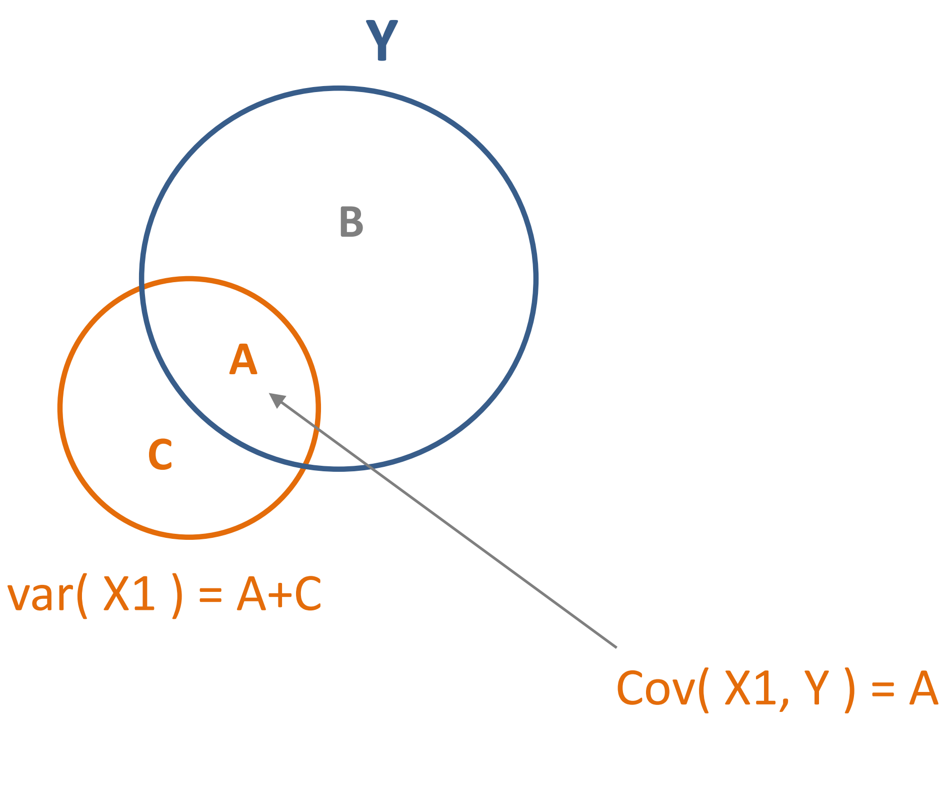
SLOPE

$$b_1 = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$



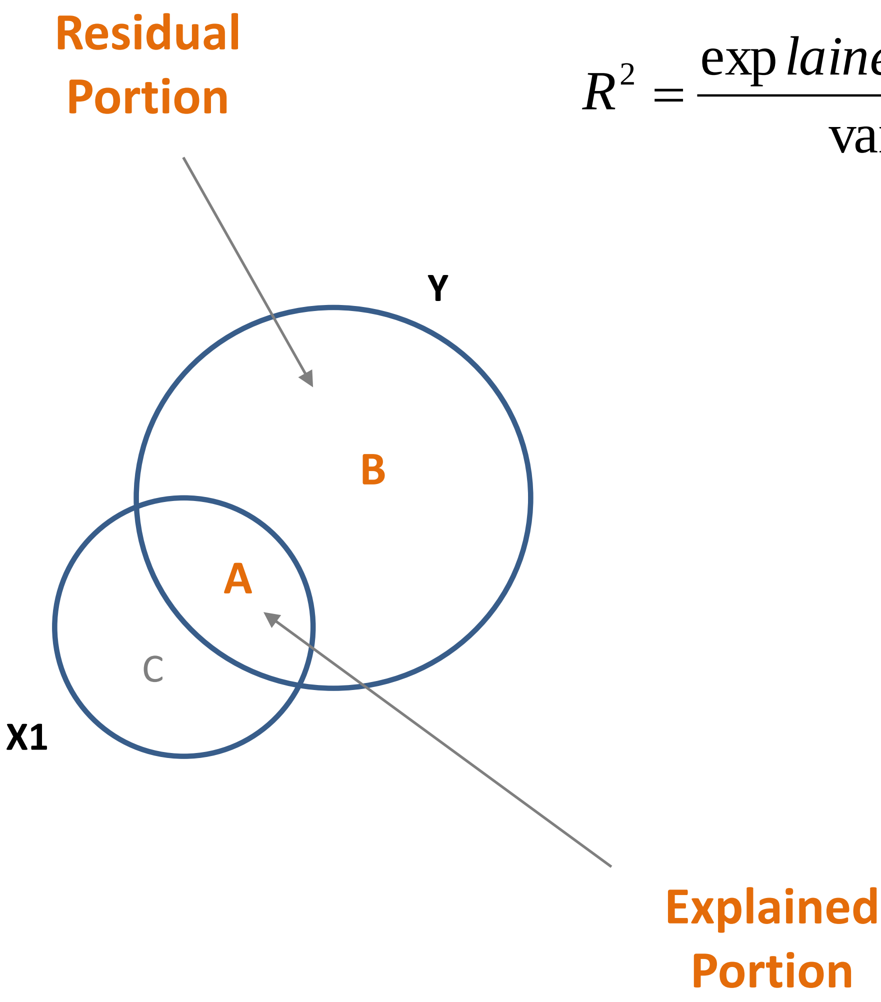
SLOPE

$$b_1 = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \frac{A}{A + C}$$

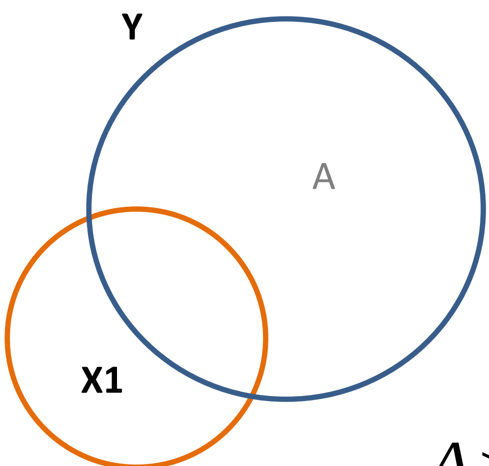


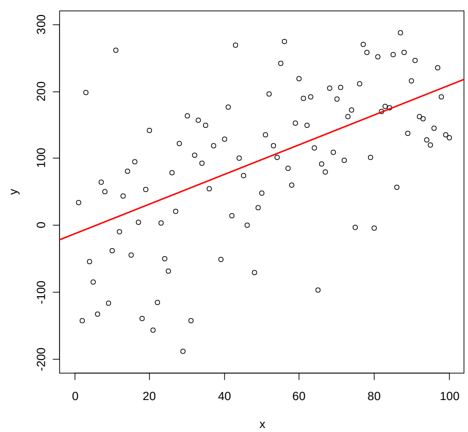
THE RESIDUAL AND R²

$$R^2 = \frac{\text{explained var}(y)}{\text{var}(y)} \approx \frac{A}{A+B}$$



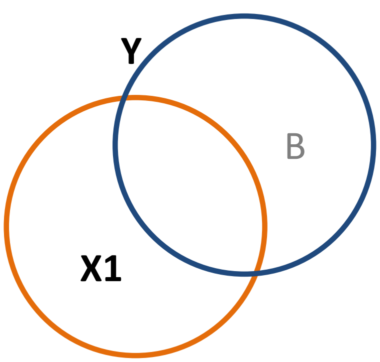
R-SQUARED AND REGRESSION RESIDUAL

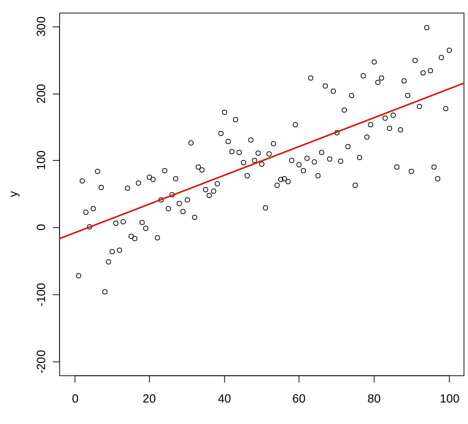




$$var(y) = 14508$$
$$var(x) = 841$$
$$R^2 = 0.29$$
$$Residual = 102.3$$

$$A > B$$

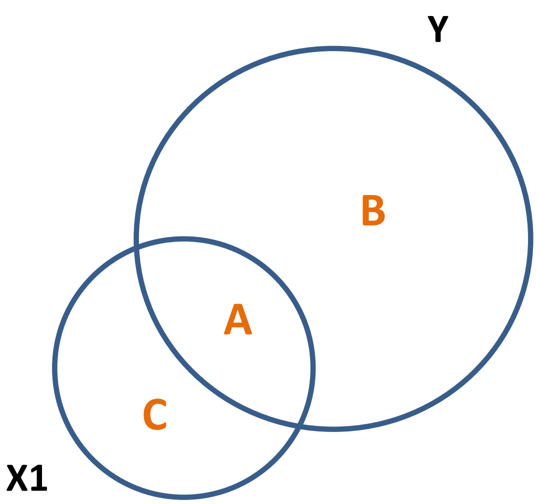




$$var(y) = 6291$$
$$var(x) = 841$$
$$R^2 = 0.61$$
$$Residual = 49.49$$

The variance of X1 and cov(X1,Y) are the same in these two cases. The var(Y) is larger in the top case. Although the “explained” portion is the same in both models, there is more variance to explain on top.

COEFFICIENT STANDARD ERROR



There are three ways to make the standard error smaller, and thus improve the confidence intervals around b_1 :

(1) Increase sample size
(2) Explain more variance of Y
(3) Increase variance of X

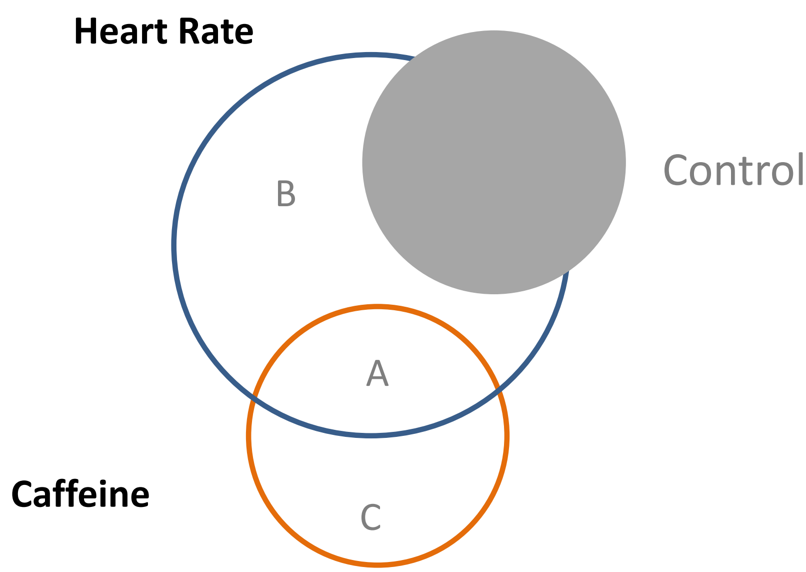
$$SE_{b1} \approx \frac{\text{residual}}{\text{sample size} \cdot \text{variance X1}} \approx \frac{B}{n \cdot (A+C)}$$

EXPLAIN MORE Y

There are three ways to make the standard error smaller, and thus improve the confidence intervals around b_1 :

$$\textbf{SE} \approx \frac{B}{n \cdot (A+C)}$$

(1) Increase sample size
(2) Explain more variance of Y
(3) Increase variance of X



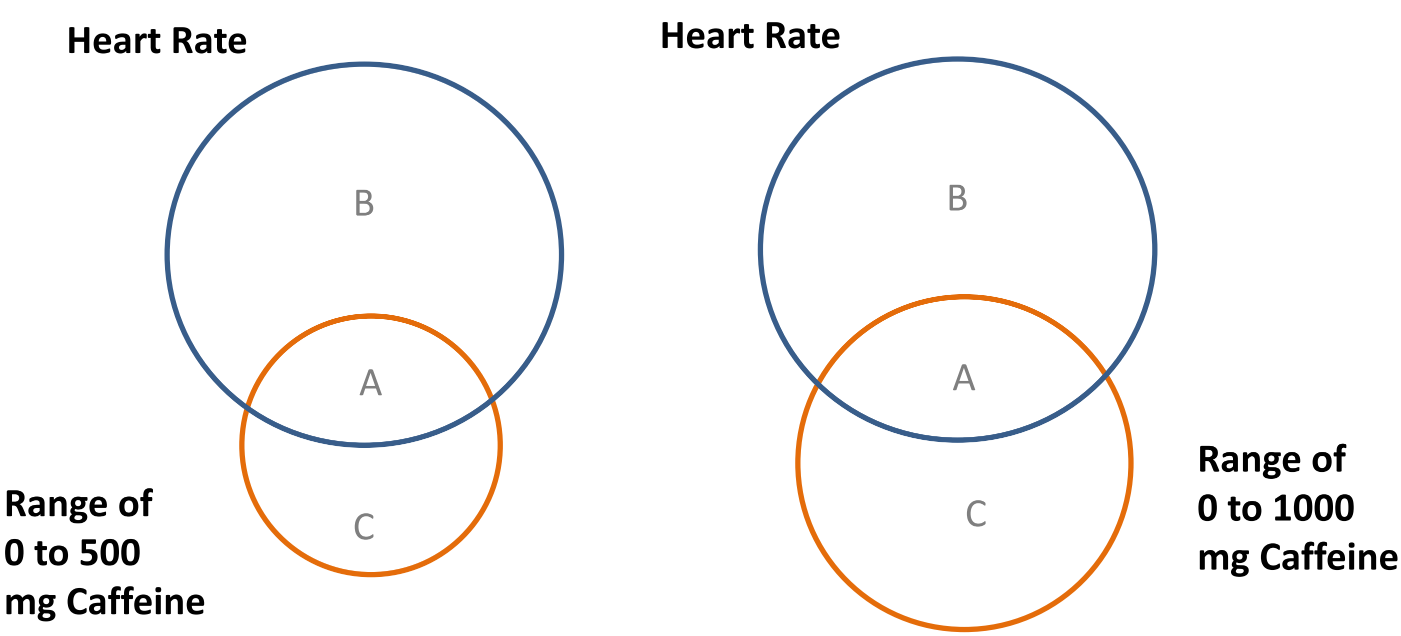
Adding a control variable can explain some of Y, thus leading to smaller residuals.

INCREASE VAR(X)

There are three ways to make the standard error smaller, and thus improve the confidence intervals around b_1 :

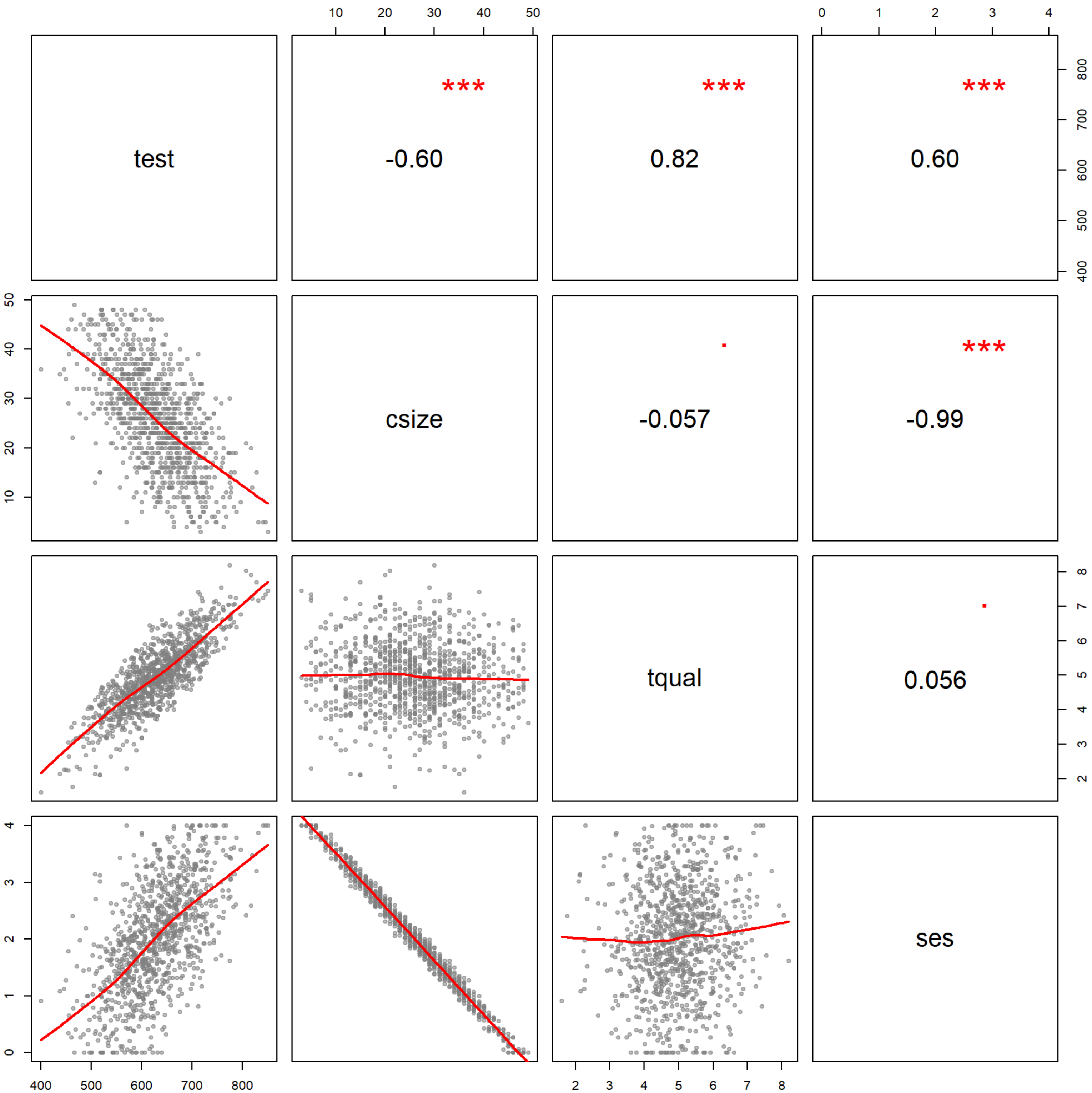
$$SE_{b1} \approx \frac{B}{n \cdot (A+C)}$$

(1) Increase sample size
(2) Explain more variance of Y
(3) Increase variance of X

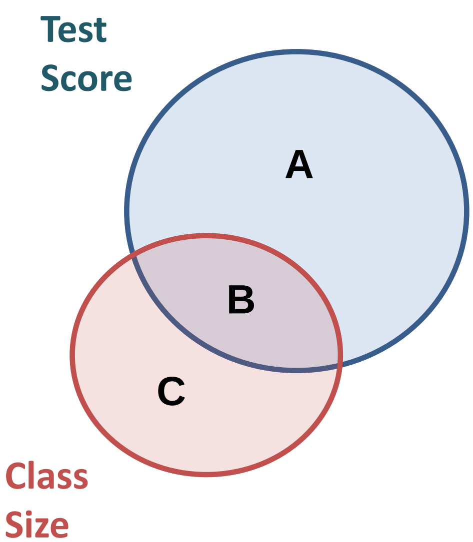


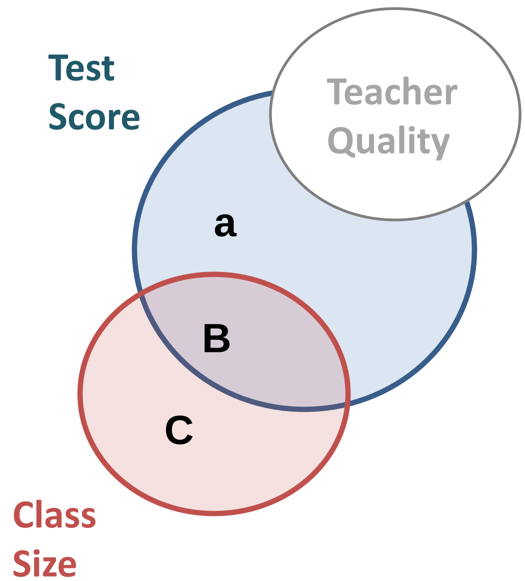
In an experiment assigning the treatment levels over a range of 0 to 1000mg increases the variance of X compared to a study that uses 0 to 500mg.

TWO TYPES OF CONTROL VARIABLES:


test
csize
tqual
ses
-0.60
0.82
0.60
-0.057
-0.99
0.056
***
***
***
***
10 20 30 40 50
0 1 2 3 4
400 500 600 700 800
2 3 4 5 6 7 8

FIRST TYPE: UNCORRELATED WITH THE POLICY VARIABLE





$$slope: \frac{B}{B+C} \rightarrow \frac{B}{B+C}$$

Slope does not change.

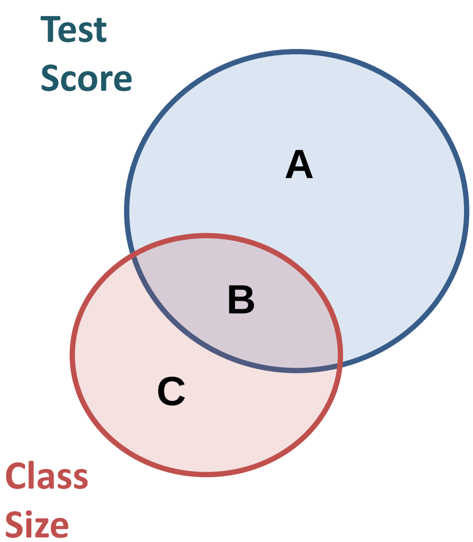
$$SE_{b1}: \frac{A}{B+C} \rightarrow \frac{a}{B+C}$$

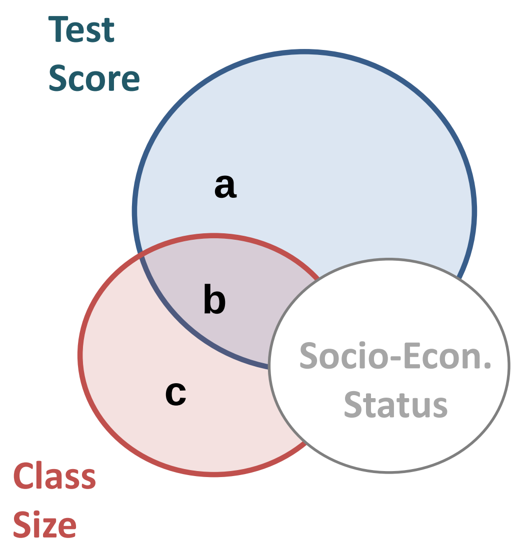
Standard error becomes smaller.

	Dependent Variable: Test Scores				
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
Classroom Size	-4.22***	-3.91***		-2.67	-2.22***
	(0.18)	(0.03)		(1.63)	(0.23)
Teacher Quality		55.01***	55.03***		55.01***
		(0.25)	(0.26)		(0.25)
Socio-Economic Status			40.94***	16.34	17.77***
			(0.27)	(17.10)	(2.40)
Intercept	738.34***	456.70***	272.91***	665.29***	377.26***
	(4.88)	(1.48)	(1.39)	(76.57)	(10.82)

The **slope** is approximately the same, but the **standard error** is six times smaller.

SECOND TYPE: CORRELATED WITH THE POLICY VARIABLE





$$slope: \frac{B}{B+C} \rightarrow \frac{b}{b+c}$$

$$SE_{b1}: \frac{A}{B+C} \rightarrow \frac{a}{b+c}$$

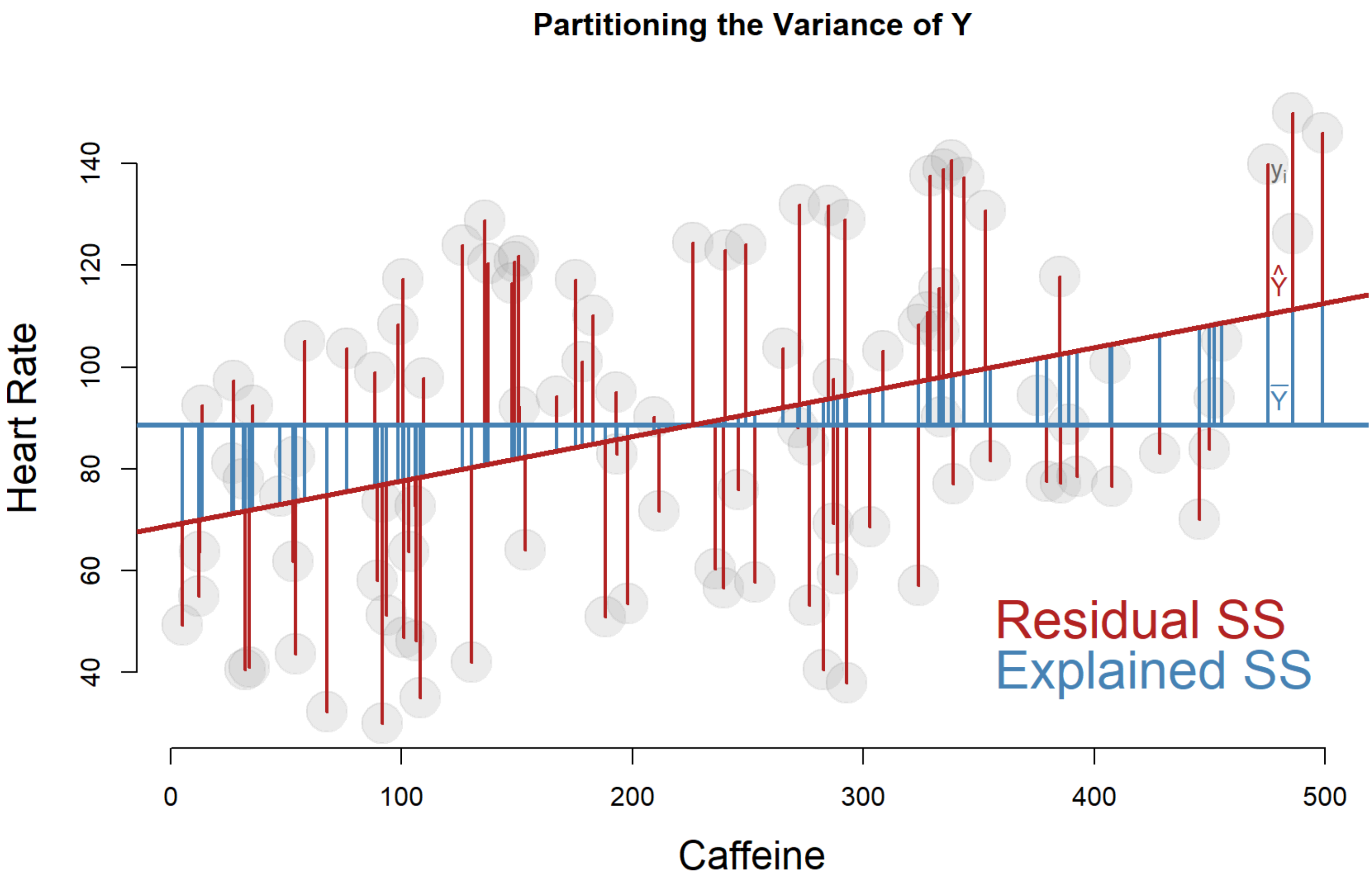
Slope changes (can increase or decrease depending on size of b relative to c).

The standard error typically becomes larger (again depends on ratio).

	Dependent Variable: Test Scores				
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
Classroom Size	-4.22***	-3.91***		-2.67	-2.22***
	(0.18)	(0.03)		(1.63)	(0.23)
Teacher Quality		55.01***	55.03***		55.01***
		(0.25)	(0.26)		(0.25)
Socio-Economic Status			40.94***	16.34	17.77***
			(0.27)	(17.10)	(2.40)
Intercept	738.34***	456.70***	272.91***	665.29***	377.26***
	(4.88)	(1.48)	(1.39)	(76.57)	(10.82)

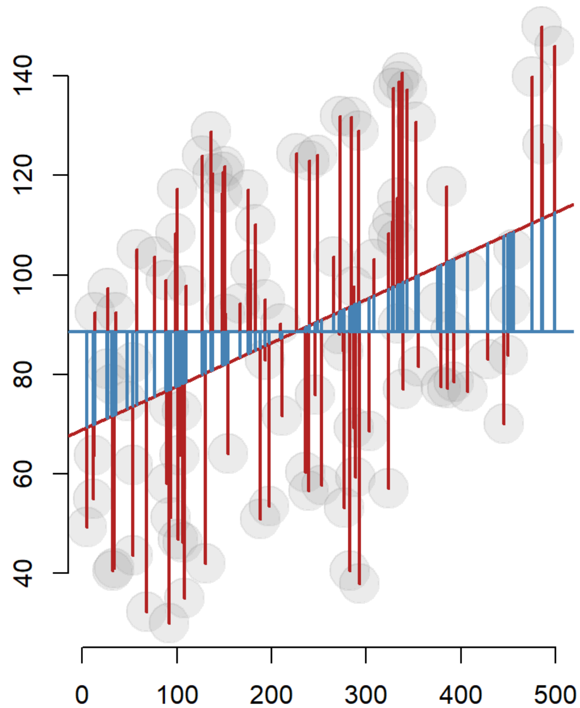
The **slope** is smaller (close to zero), and the **standard error** is almost ten times as large.

COFFEE STUDY EXAMPLE

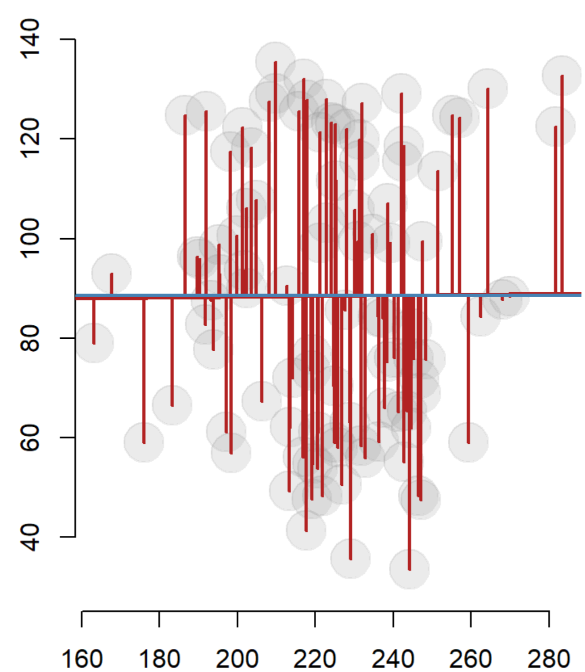


Reconsider the caffeine study as an observational study on coffee consumption. Now caffeine is not assigned, but level of consumption is a choice by individuals. We can add controls.

$$Y = b_0 + b_1 X + e$$

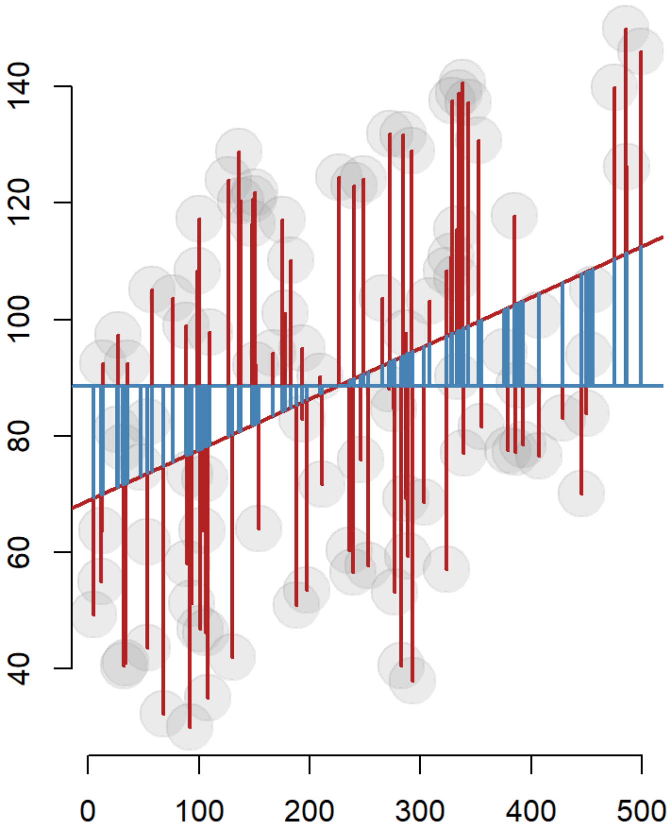


$$Y = \beta_0 + \beta_1 X + \beta_2 X2 + \varepsilon$$

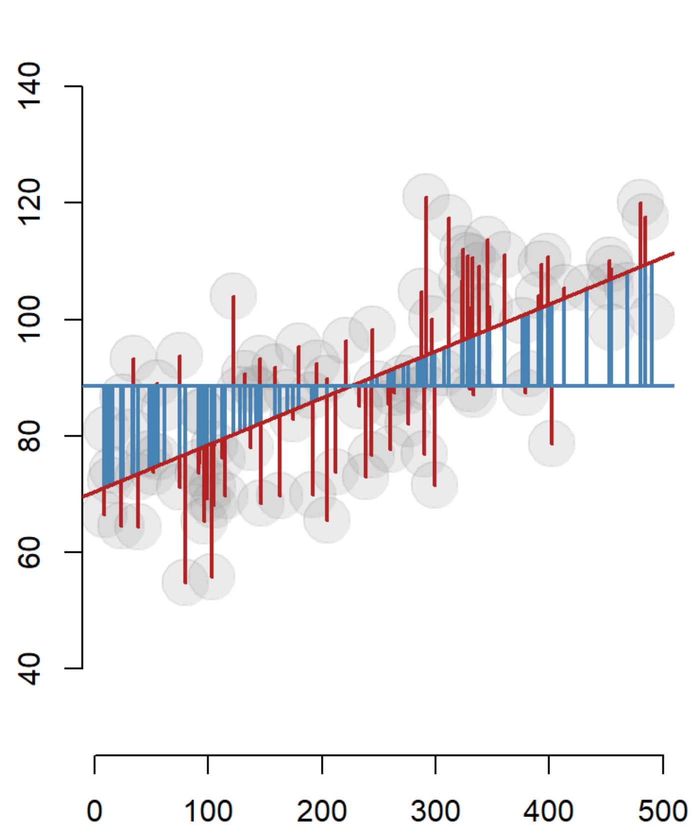


If the new control variable X2 is highly-correlated with our policy variable, we primarily remove the EXPLAINED variance from the model.

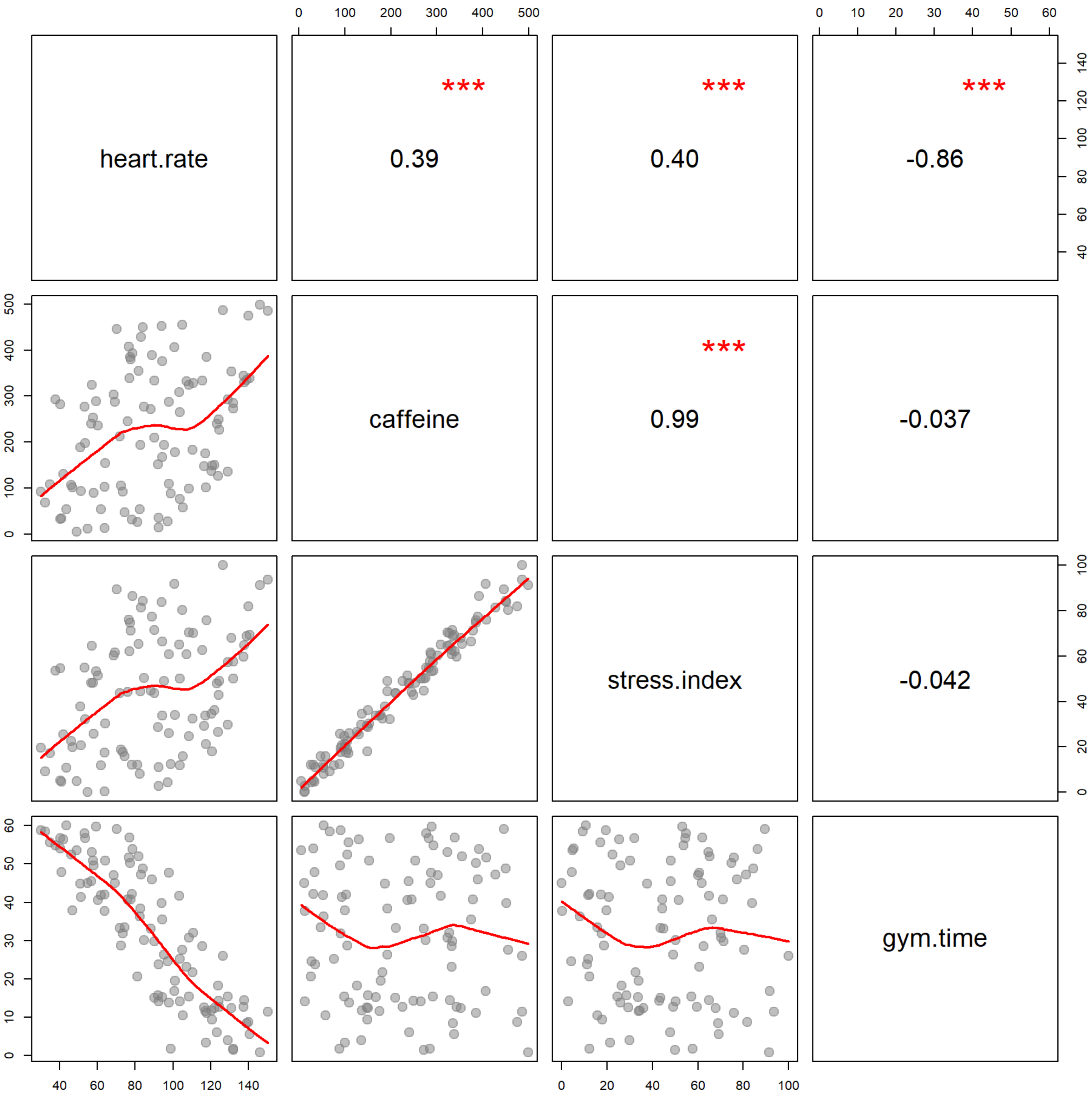
$$Y = b_0 + b_1 X + e$$



$$Y = \beta_0 + \beta_1 X + \beta_2 X2 + \varepsilon$$



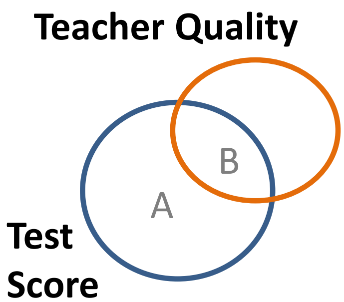
If the new control variable X2 is NOT correlated with our policy variable and correlated with the outcome, we primarily remove the RESIDUALS from the model.

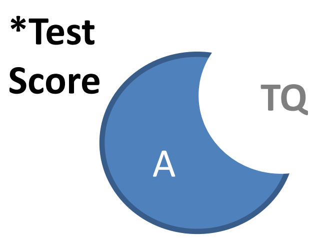

heart.rate
caffeine
stress.index
gym.time
0.39
0.40
-0.86
0.99
-0.037
-0.042
***
***
***
***
0
100
200
300
400
500
0
10
20
30
40
50
60
40
60
80
100
120
140
0
20
40
60
80
100

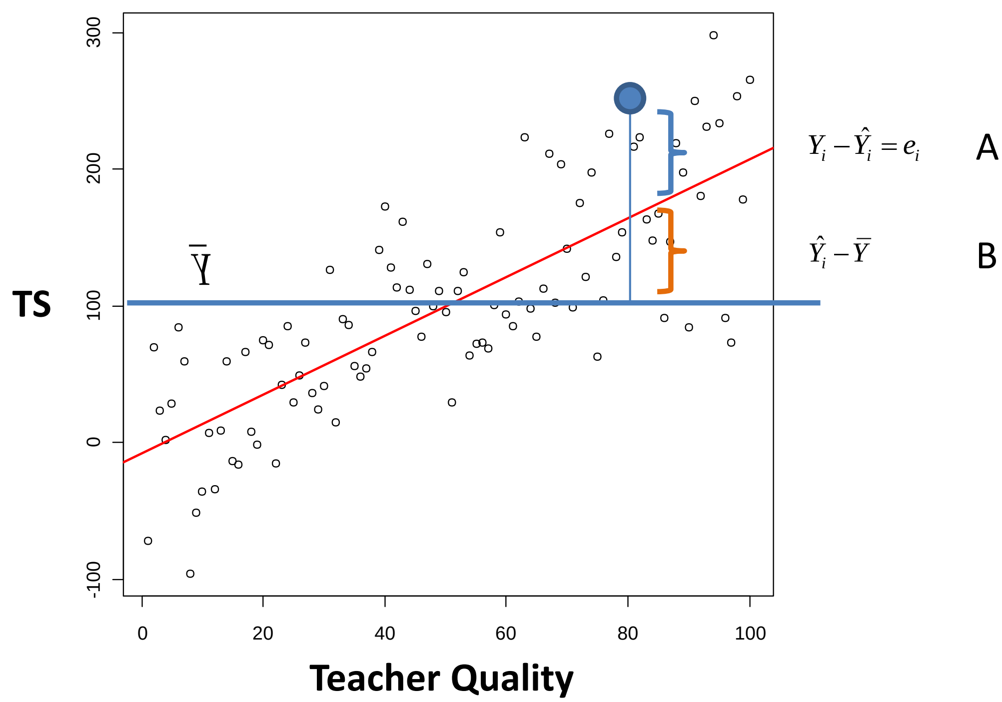
	Dependent variable:				
	heart.rate				
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
Caffeine	0.087***		0.009	0.080***	0.037
	(0.021)		(0.121)	(0.008)	(0.047)
Stress Index		0.460***	0.414		0.228
		(0.108)	(0.631)		(0.246)
Time Spent at Gym				-1.441***	-1.440***
				(0.062)	(0.062)
Intercept	68.953***	68.251***	68.267***	116.461***	116.022***
	(5.454)	(5.535)	(5.568)	(2.942)	(2.982)
Observations	100	100	100	100	100
R^2	0.153	0.157	0.157	0.872	0.873
Note:					$p<0.1$; $p<0.05$; $p<0.01$

CLASS SIZE EXAMPLE

PARTITIONING THE VARIANCE OF Y

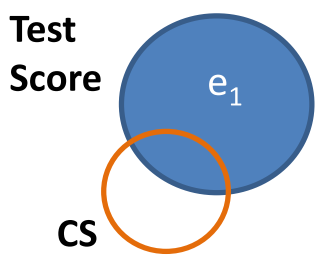


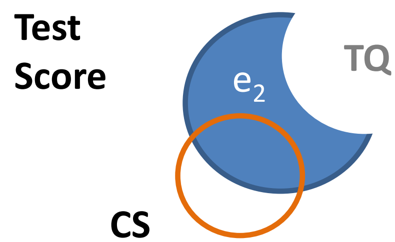


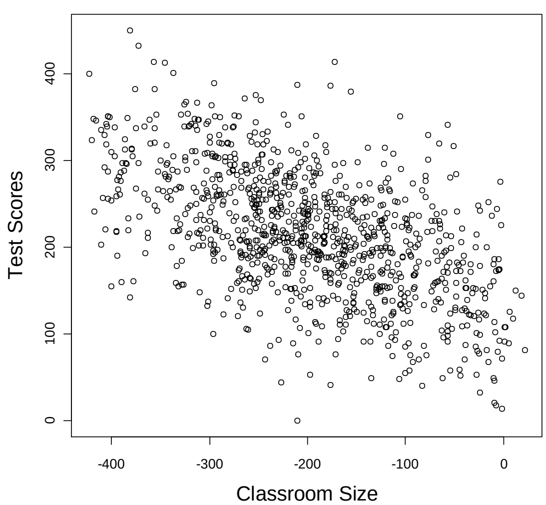


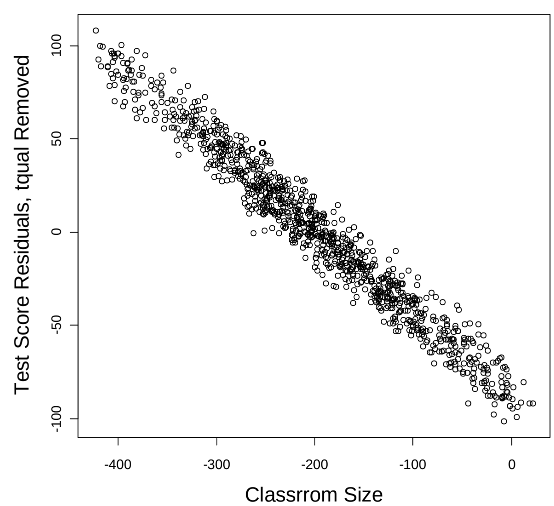
Control variables target either the **Explained SS** or the **Residual SS**.

EFFECTS OF THE CONTROL VARIABLE



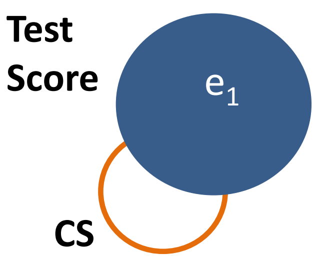




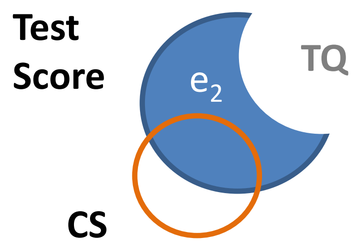


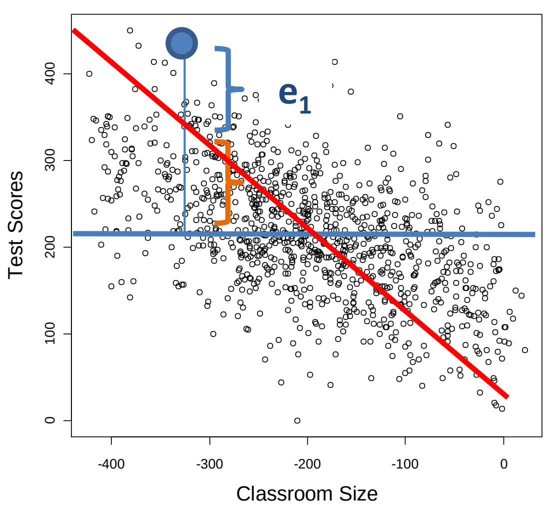
Uncorrelated control variables target the Residual SS, thus removing unexplained variance from the model and improving the relationship between the policy variable and Y.

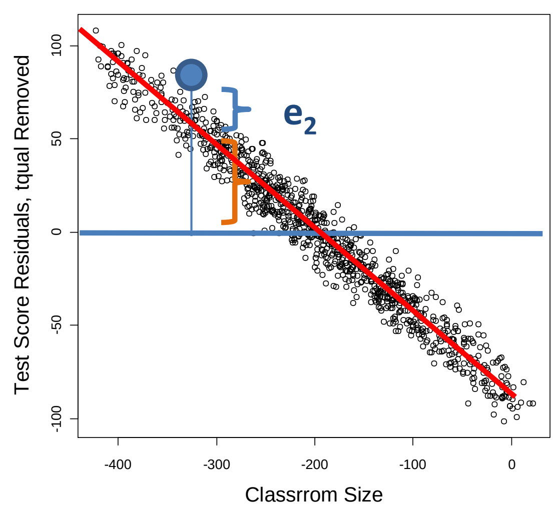
EFFECTS OF THE CONTROL VARIABLE



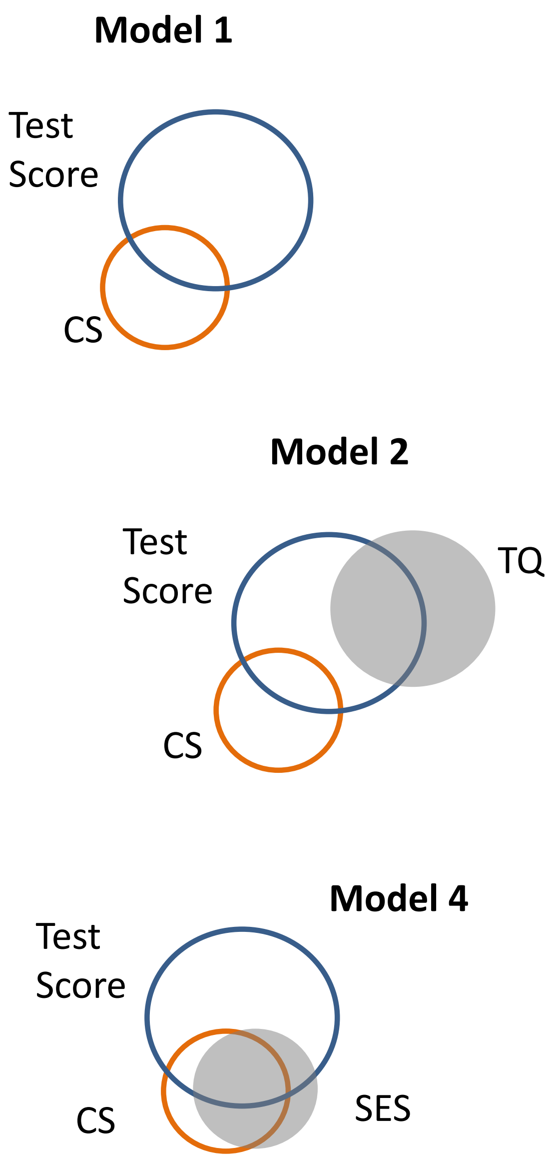
$e_1 > e_2$

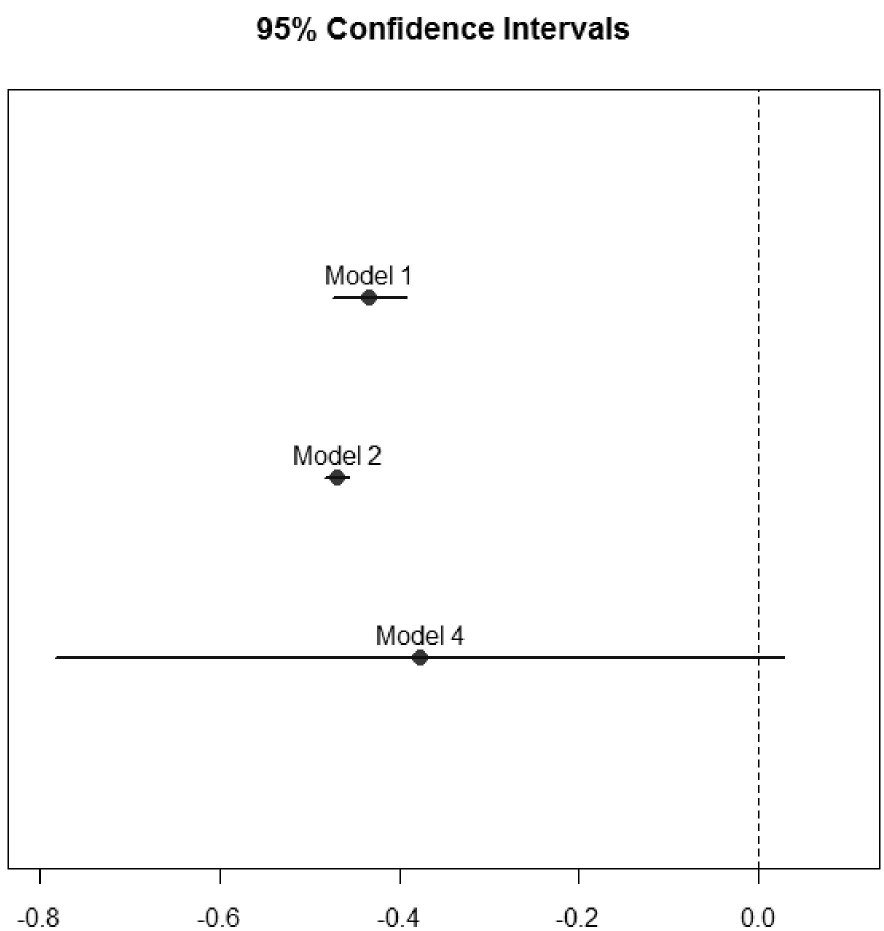




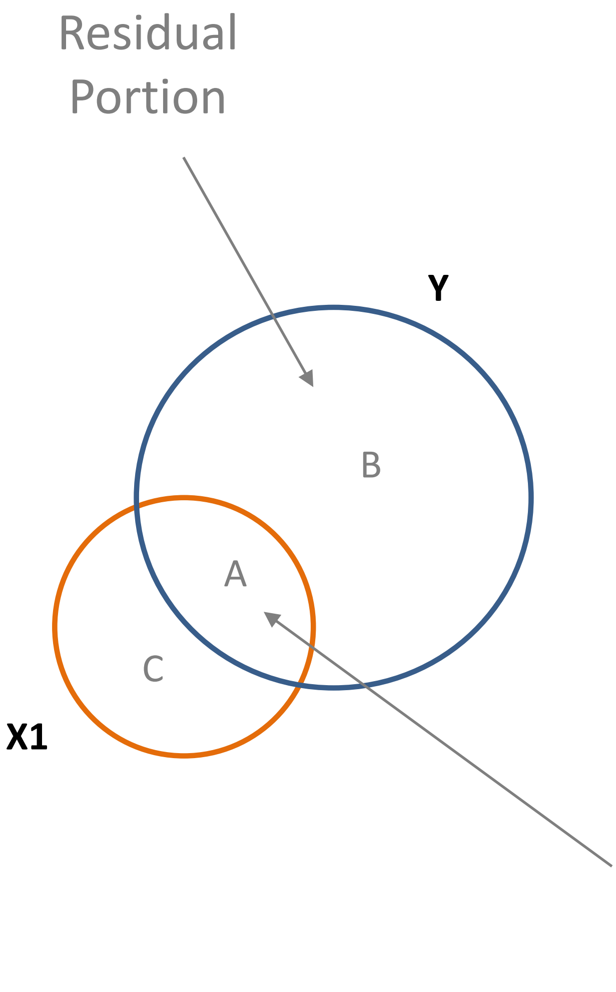


MODEL 2 HAS SMALLER STANDARD ERRORS





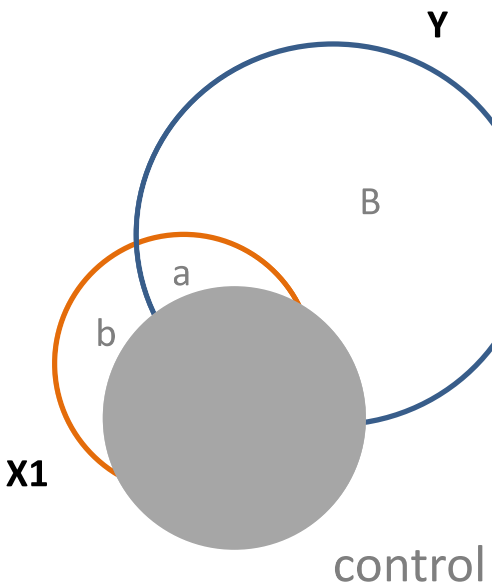
THE RESIDUAL AND R^2



$$R^2 = \frac{\text{explained var}(y)}{\text{var}(y)} \approx \frac{A}{A+B}$$

ε *is proportional to* B

THE OTHER TYPE OF CONTROL TARGETS THE EXPLAINED SS



WHAT HAPPENS WHEN WE REMOVE THE EXPLAINED SS FROM THE MODEL?