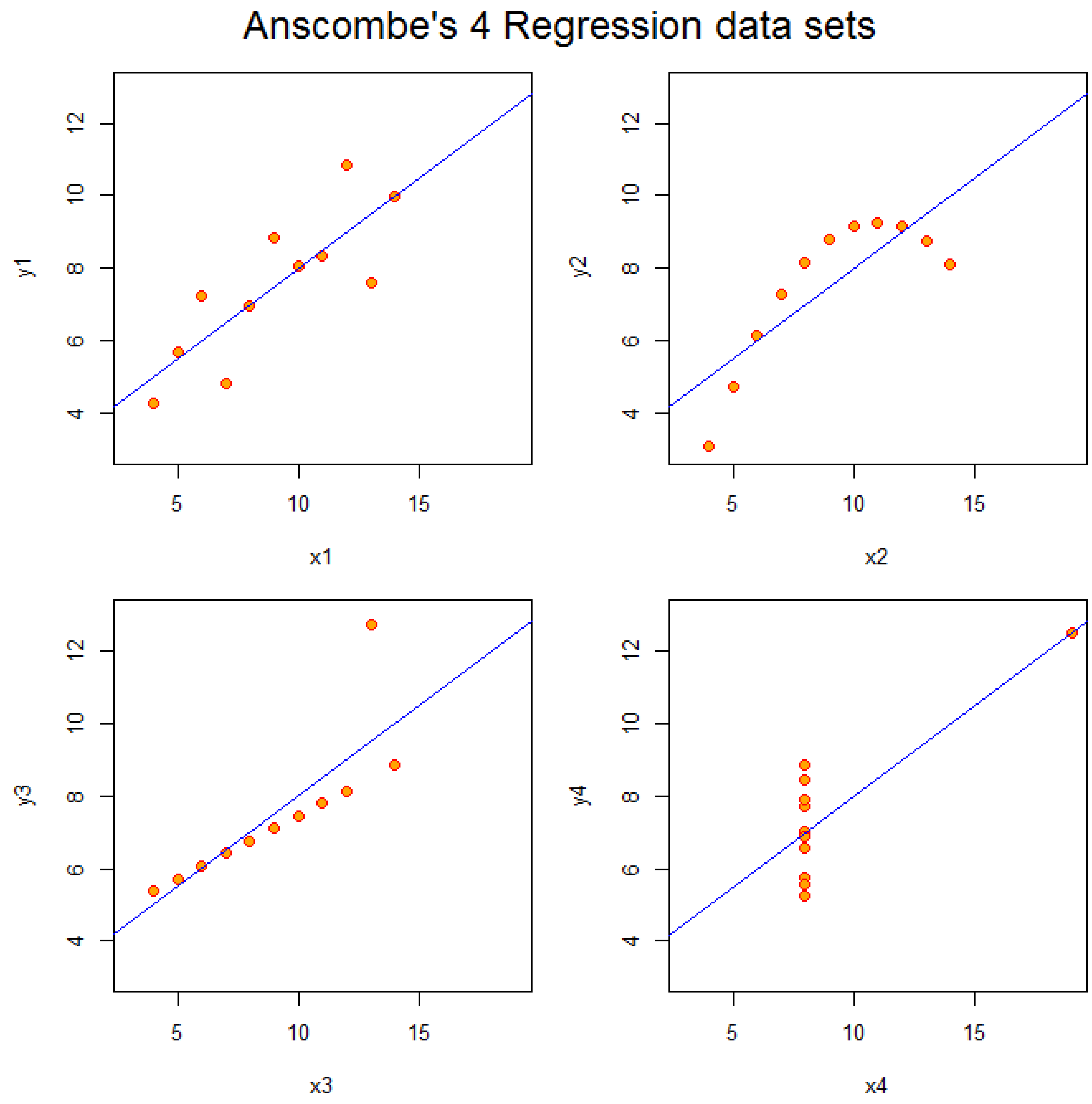
BIAS FROM SPECIFICATION OR MEASUREMENT

ANSCOMBE'S QUARTET
A LESSON IN MODEL FIT

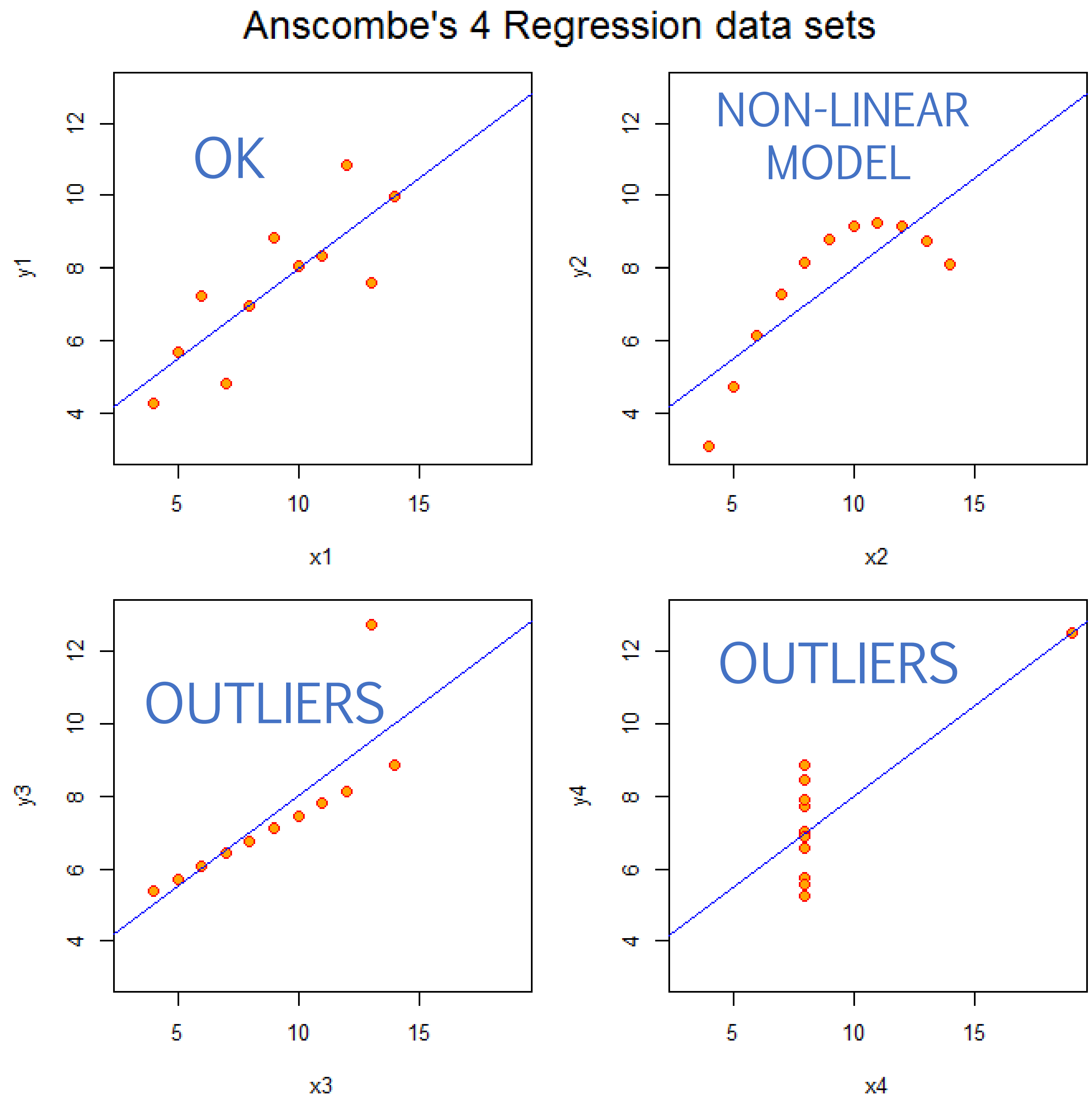
Anscombe's Quartet

	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Variance	11	4.12	11	4.12	11	4.12	11	4.12
Correlation	0.816		0.816		0.816		0.816	
Regression	y = 3 + 0.5x		y = 3 + 0.5x		y = 3 + 0.5x		y = 3 + 0.5x	

Four datasets that produce IDENTICAL descriptive stats, correlations, and regression models



BUT THEY ARE VERY DIFFERENT RELATIONSHIPS!



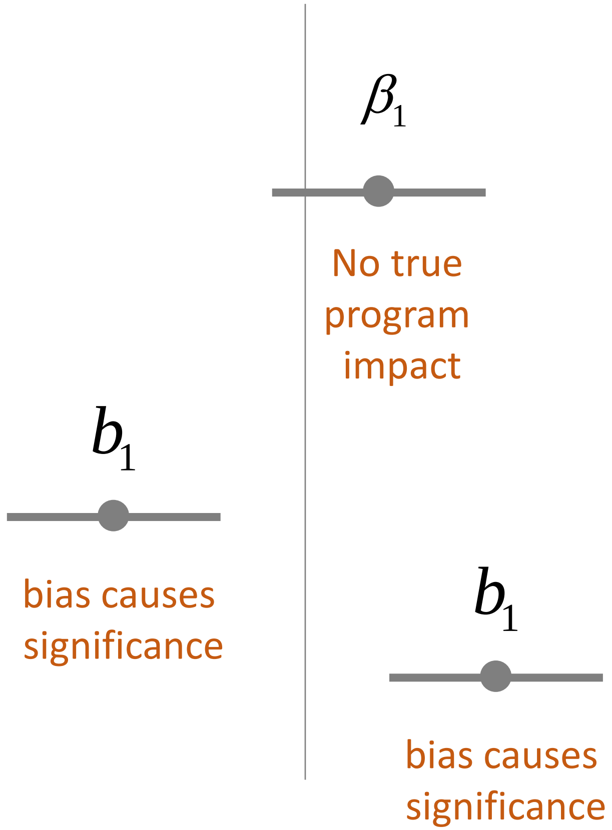
Anscombe's Quartet is often cited because (a) whoever created this example is a genius, and (b) it is a vivid demonstration of causes and consequences of **SPECIFICATION BIAS.**

We will consider what happens to slopes when outliers are present, or we use a linear specification when the relationship is non-linear.

CLASSES OF INFERENTIAL FAILURE

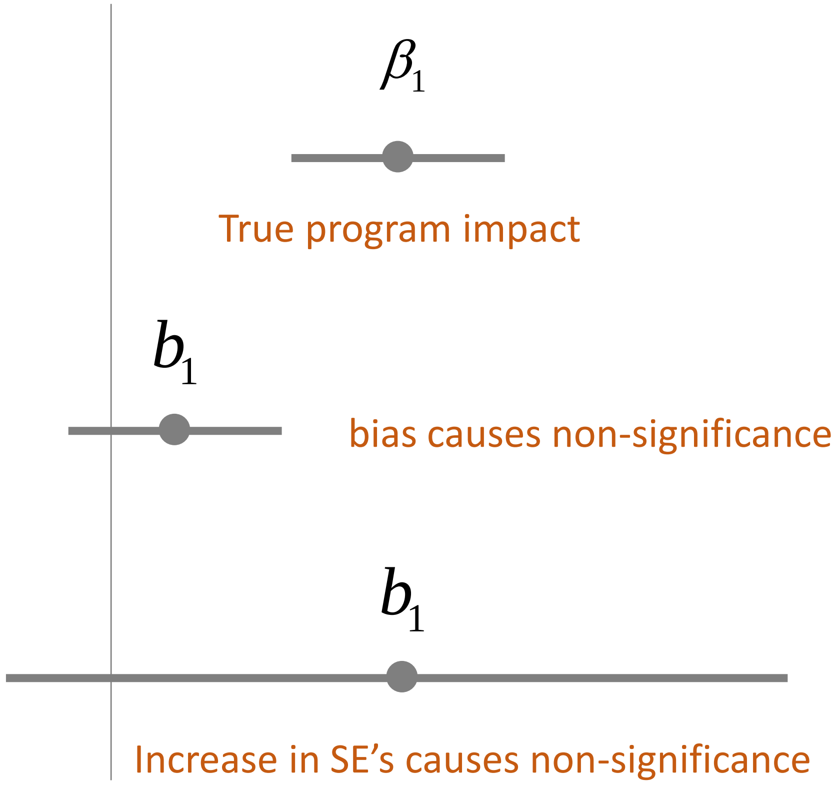
TYPE I AND TYPE II ERRORS

TYPE I ERROR
FALSE POSITIVE
CLAIMING PROGRAM HAS IMPACT WHEN IT DOESN'T



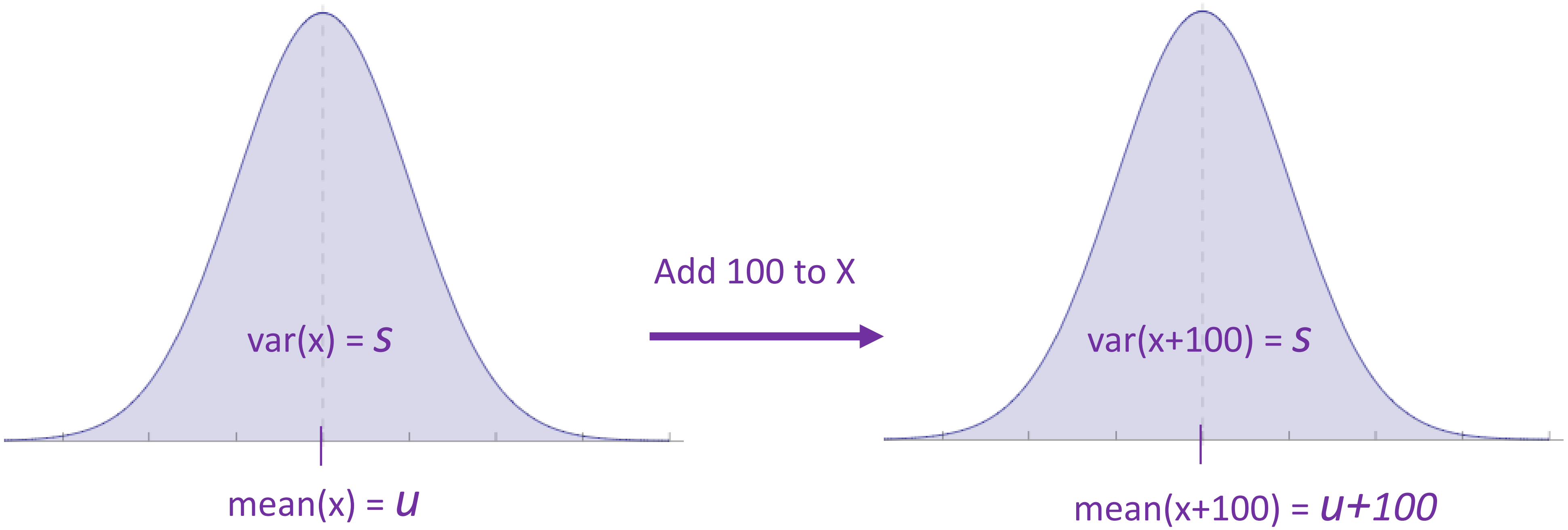
Type I errors are typically caused by OVB

TYPE II ERROR
FALSE NEGATIVE
FAILING TO IDENTIFY TRUE PROGRAM IMPACT



Type II Errors can be caused by bias or inflated standard errors

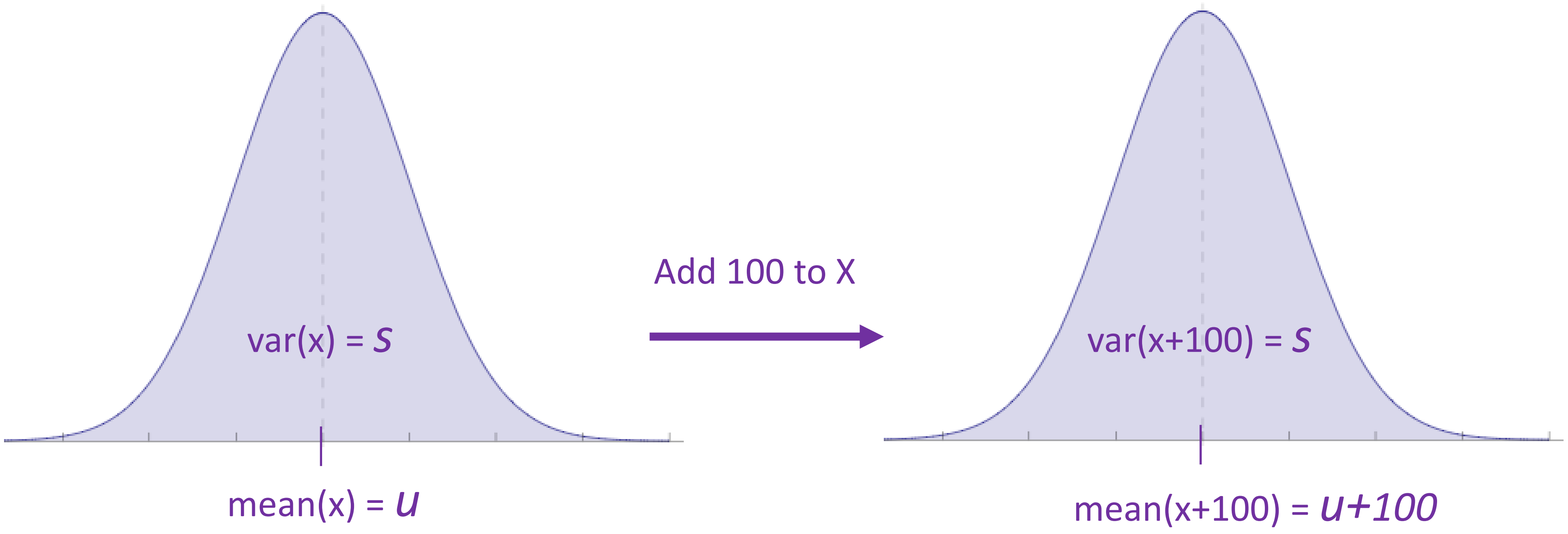
IMPLICATIONS OF MEASUREMENT ERROR

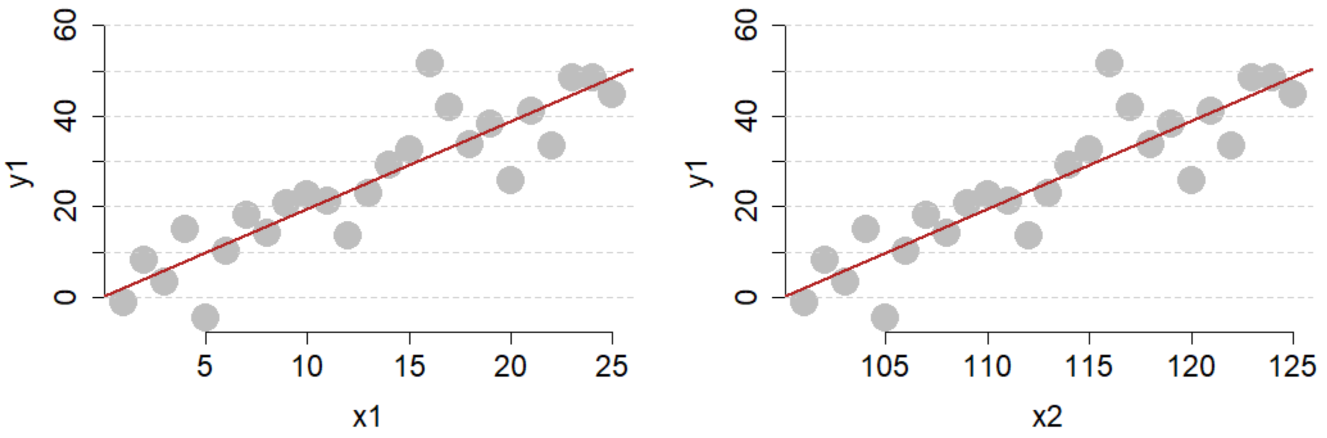


"Linear Transformations"

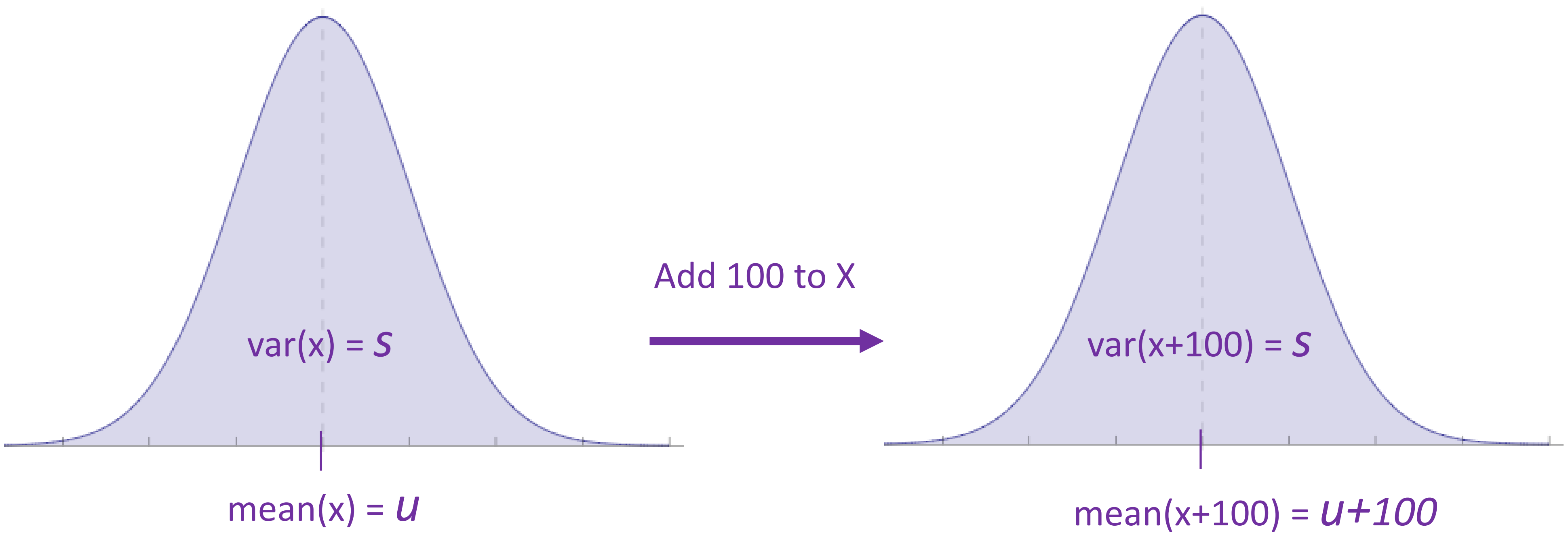
X2 = X1 + 100

Variance of X is unchanged





After linear transformations: Regressions are unaffected

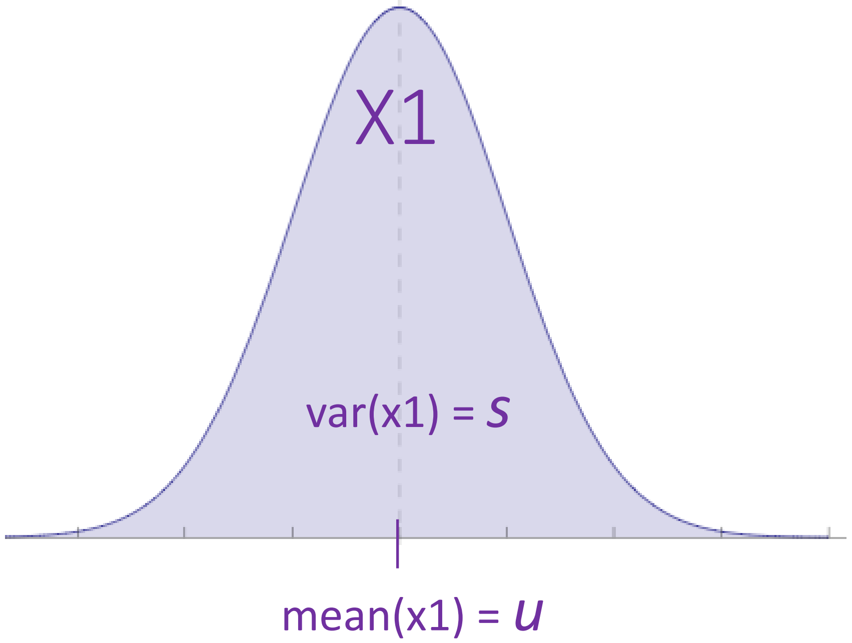


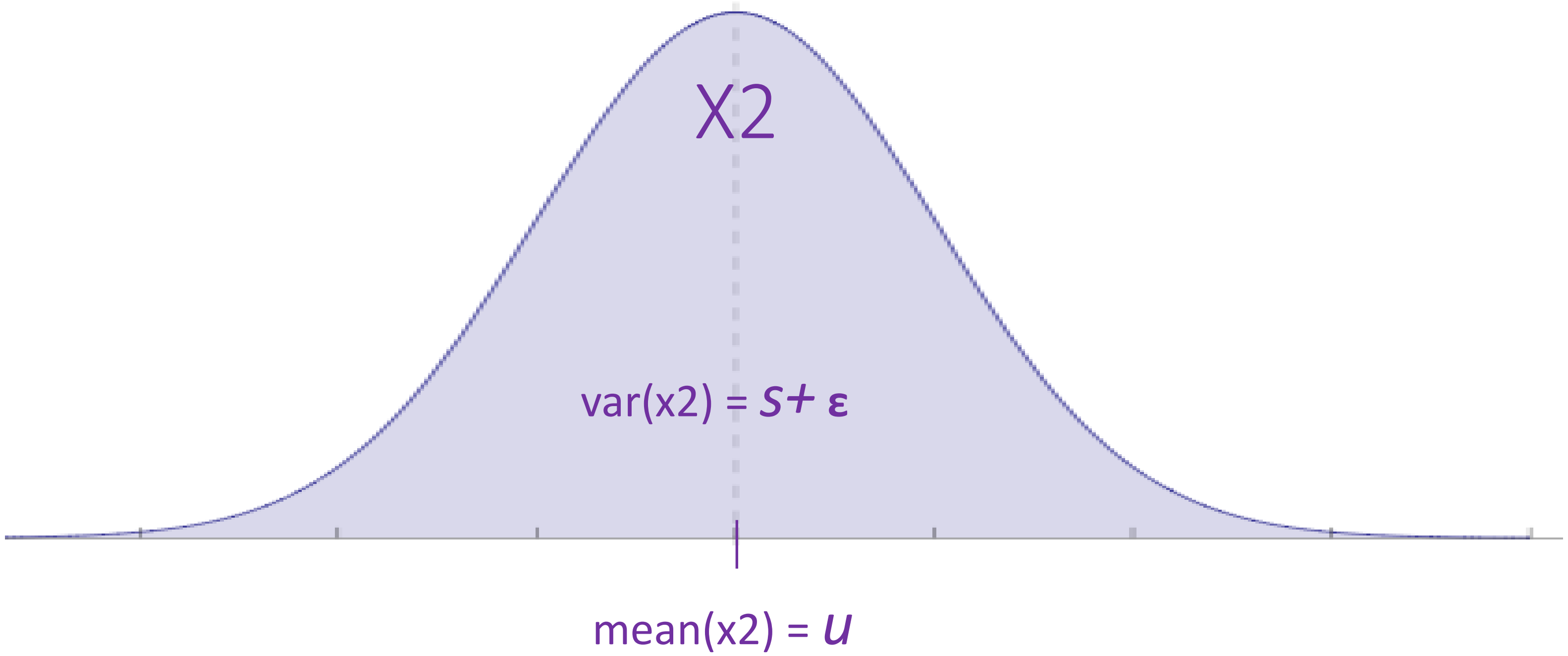
“Linear Transformations”

X2 = X1 + **100**

Must add the **same constant** to every value of X

Just moves the distribution to right or left





Measurement Error

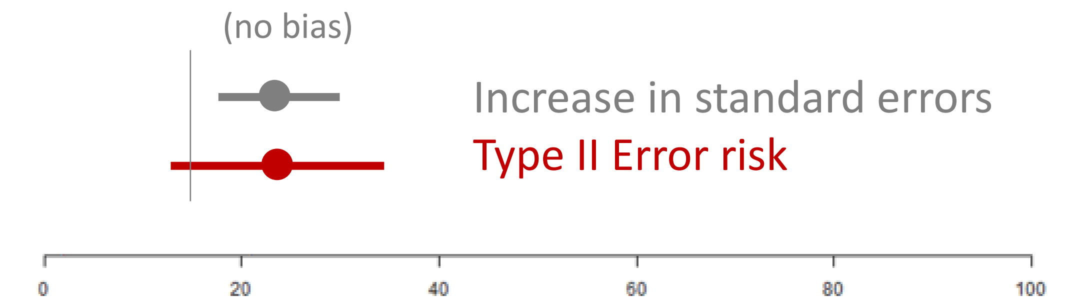
$$X2 = X1 + \boldsymbol{\varepsilon}$$

Add random error to every X.

Random means each X is equally likely to be over-measured as under-measured.

X2 has the same mean as X1, but more variance

ADDING MEASUREMENT ERROR TO THE DV



ADDING MEASUREMENT ERROR TO THE INDEPENDENT VARIABLE: “ATTENUATION BIAS”

β_1

True program impact

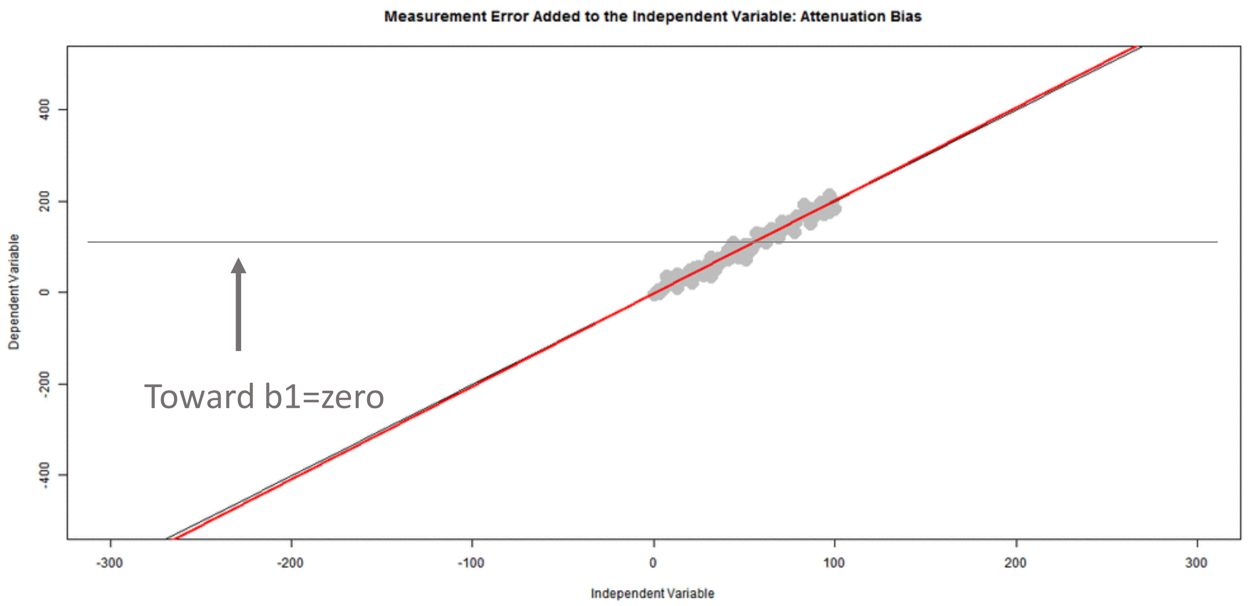
slope with measurement error

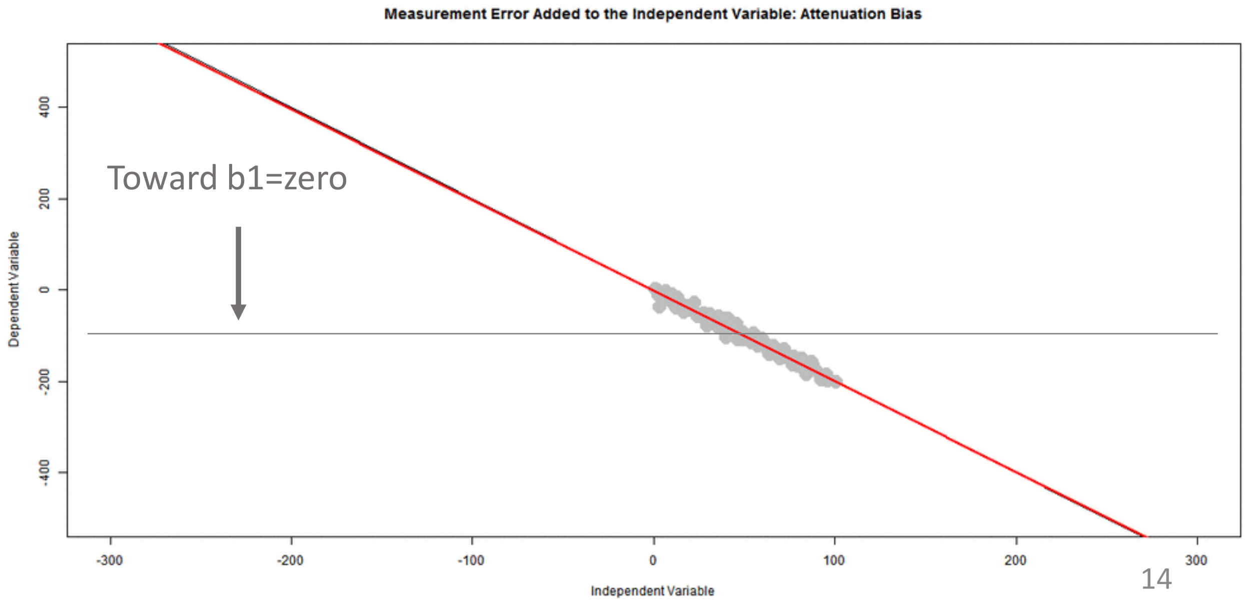
b_1

Smaller slope
Smaller SE

$$b_1 \downarrow = \frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)\ \uparrow}$$

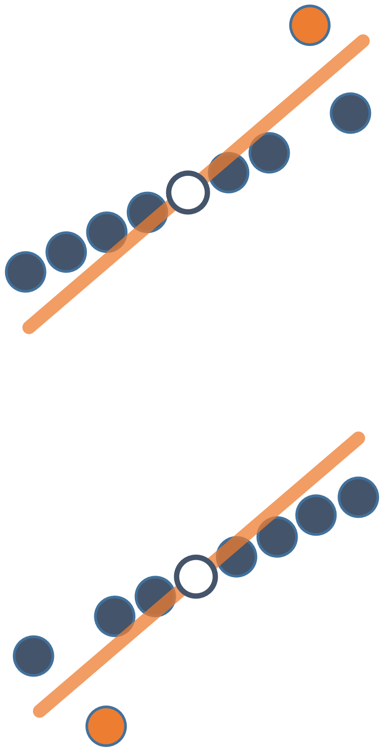
$$SE_{b1} \downarrow = \frac{residual}{samplesize \cdot \mathrm{var}(x_1) \uparrow}$$



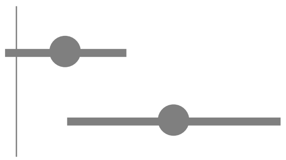


OUTLIERS

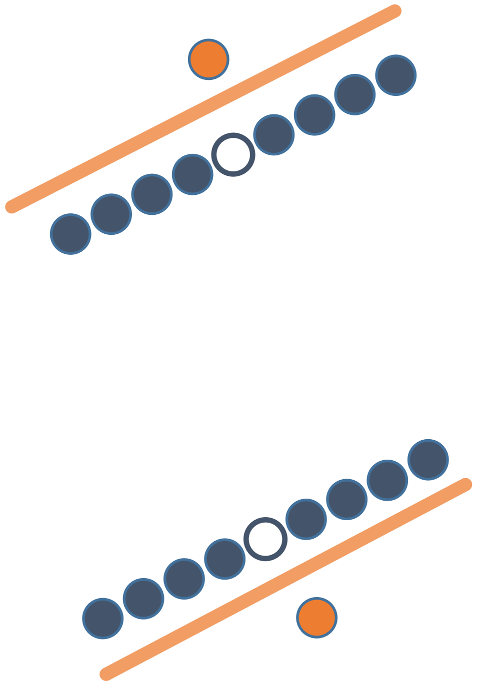
SLOPES TOO LARGE
SE LARGER



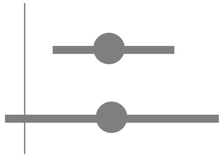
Extreme of X:
Risk of bias in slope ↑
Risk of false positive



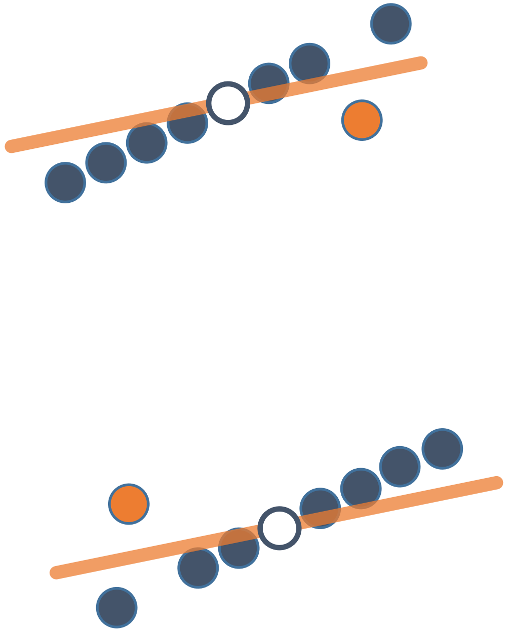
SLOPES OK
SE LARGER



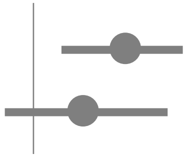
Middle of X:
Don't bias slope
Increased risk of false negative

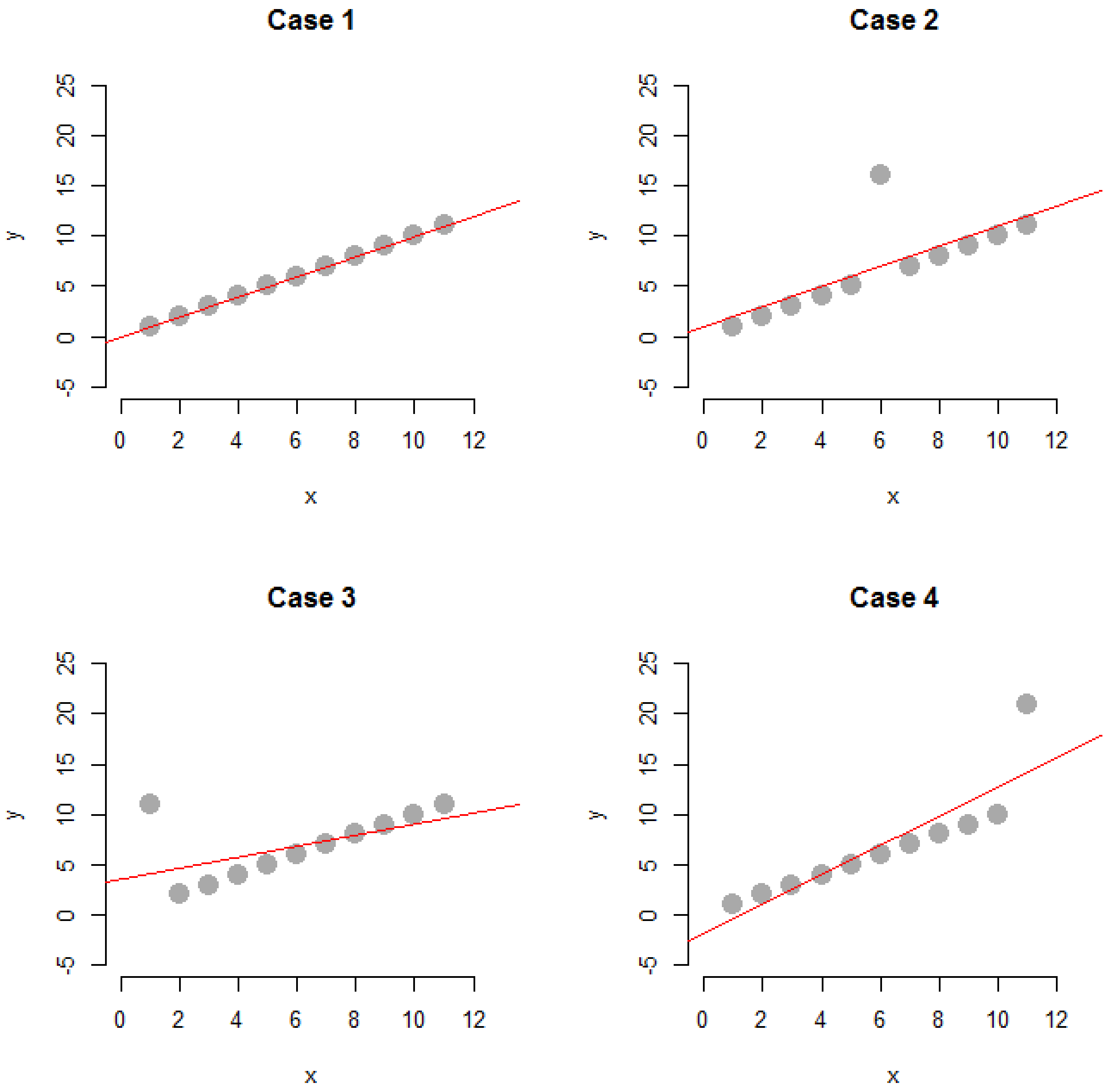


SLOPES TOO SMALL
SE LARGER



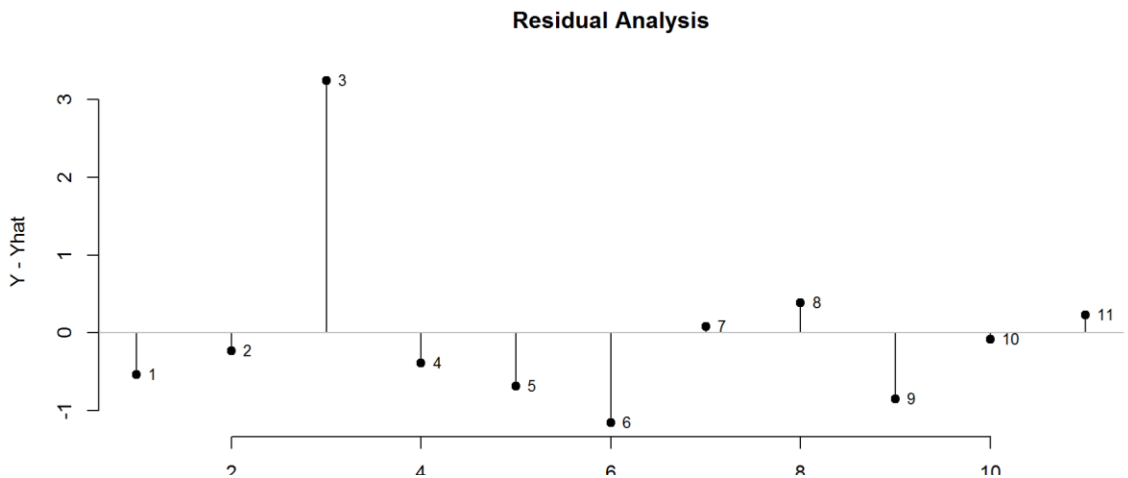
Extreme of X:
Risk of bias in slope ↓
Increased risk of false negative

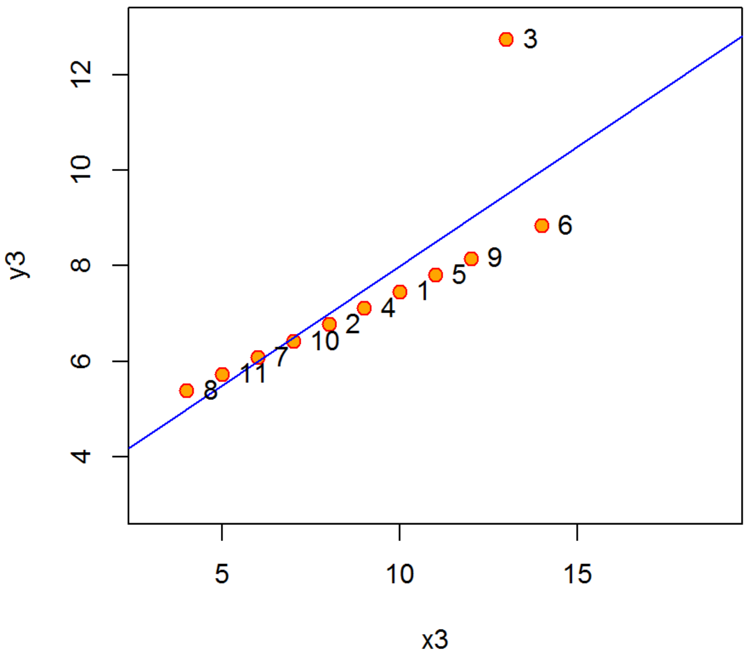


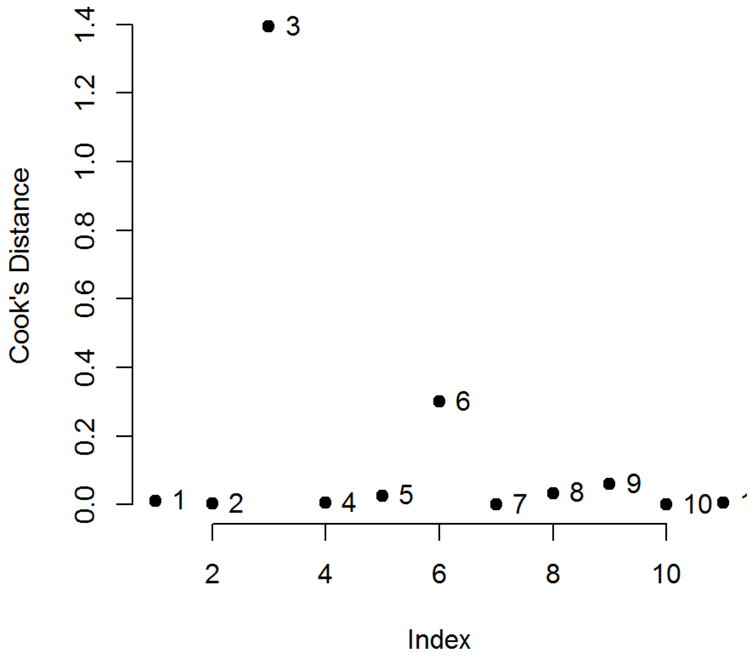


	Dependent variable:			
	y			
	(1)	(2)	(3)	(4)
x	1.00***	1.00***	0.55*	1.45***
	(0.00)	(0.30)	(0.26)	(0.26)
Constant	0.00***	0.91	3.64*	-1.82
	(0.00)	(2.06)	(1.78)	(1.78)

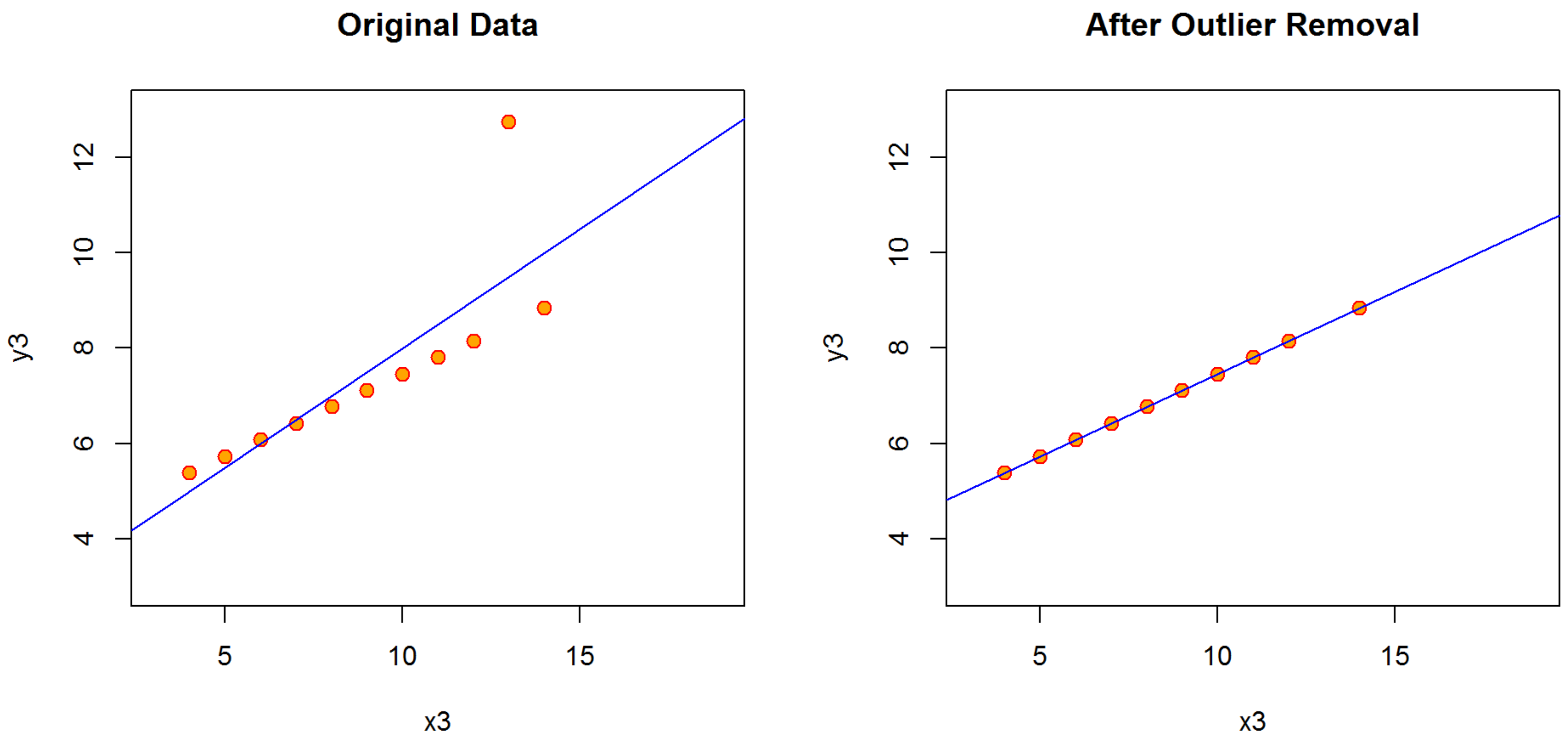
IDENTIFYING OUTLIERS USING RESIDUALS AND COOK'S DISTANCE



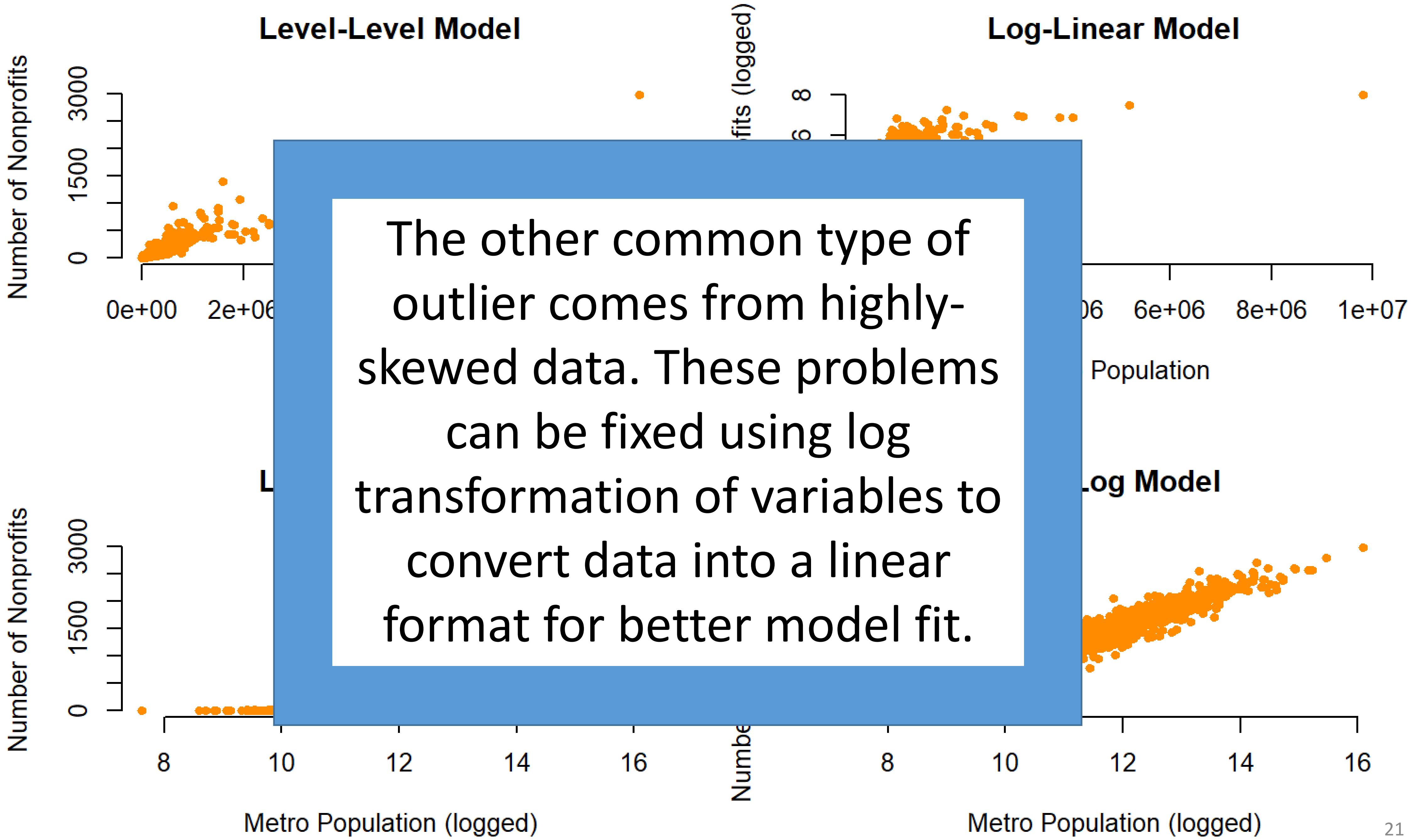




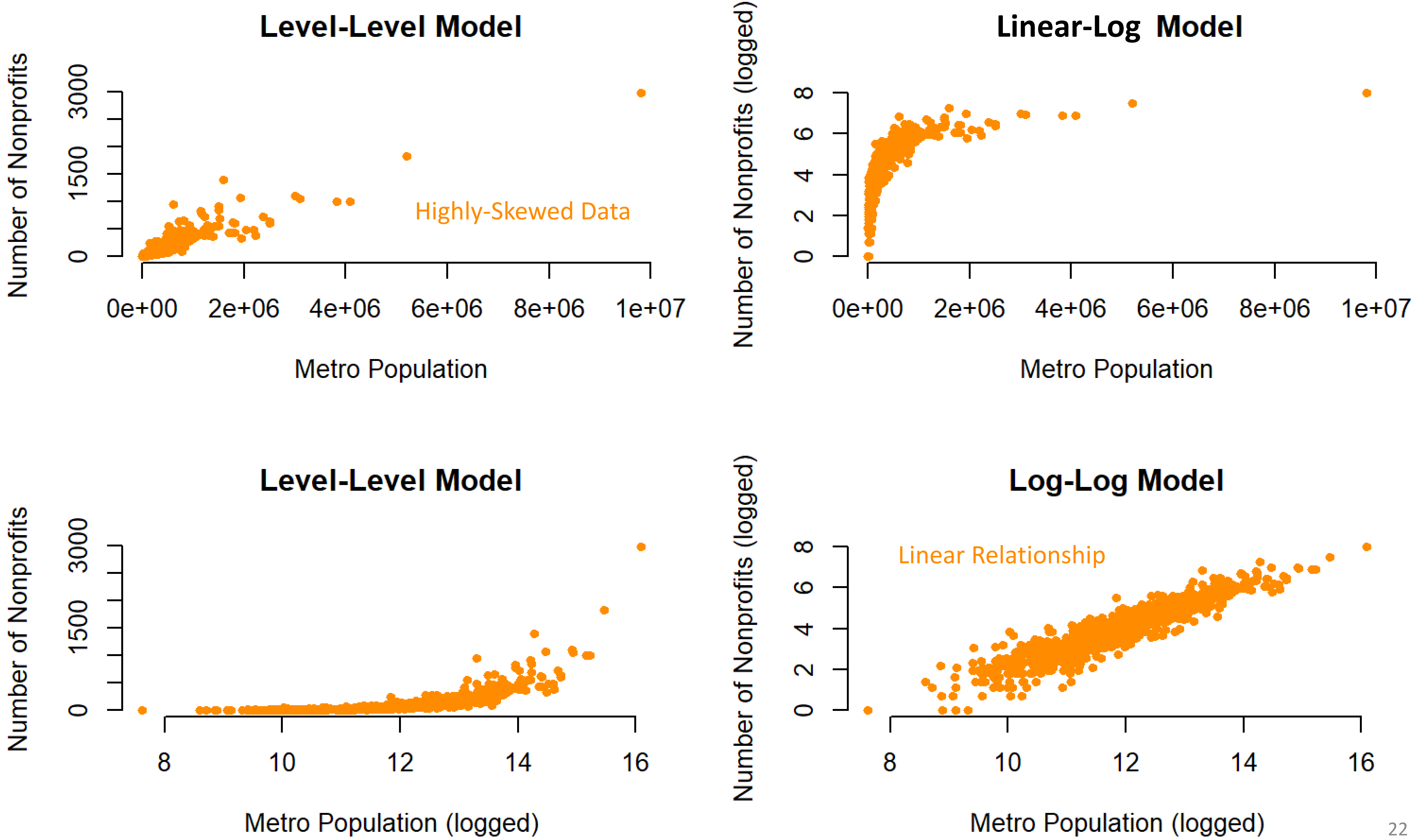
	Dependent variable:	
	y3	
	(1)	(2)
x3	0.50***	0.35***
	(0.12)	(0.0003)
Constant	3.00**	4.01***
	(1.12)	(0.003)
Observations	11	10
R^2	0.67	1.00
Adjusted R^2	0.63	1.00
Note:	*p<0.1; **p<0.05;** p<0.01*	



LOGGED
REGRESSION MODELS



The other common type of outlier comes from highly-skewed data. These problems can be fixed using log transformation of variables to convert data into a linear format for better model fit.


Level-Level Model
Number of Nonprofits
0
1500
3000
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Metro Population
Highly-Skewed Data
Linear-Log Model
Number of Nonprofits (logged)
0
2
4
6
8
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Metro Population
Level-Level Model
Number of Nonprofits
0
1500
3000
8
10
12
14
16
Metro Population (logged)
Log-Log Model
Number of Nonprofits (logged)
0
2
4
6
8
8
10
12
14
16
Metro Population (logged)
Linear Relationship

NON-LINEAR RELATIONSHIPS
QUADRATIC MODELS

Linear: $Y = b0 + b1(X_1) + e$

Quadratic: $Y = b0 + b1(X_1) + b2(X_1)^2 + e$