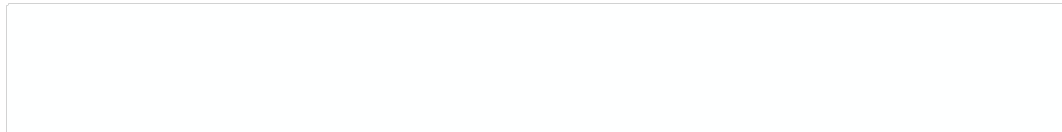


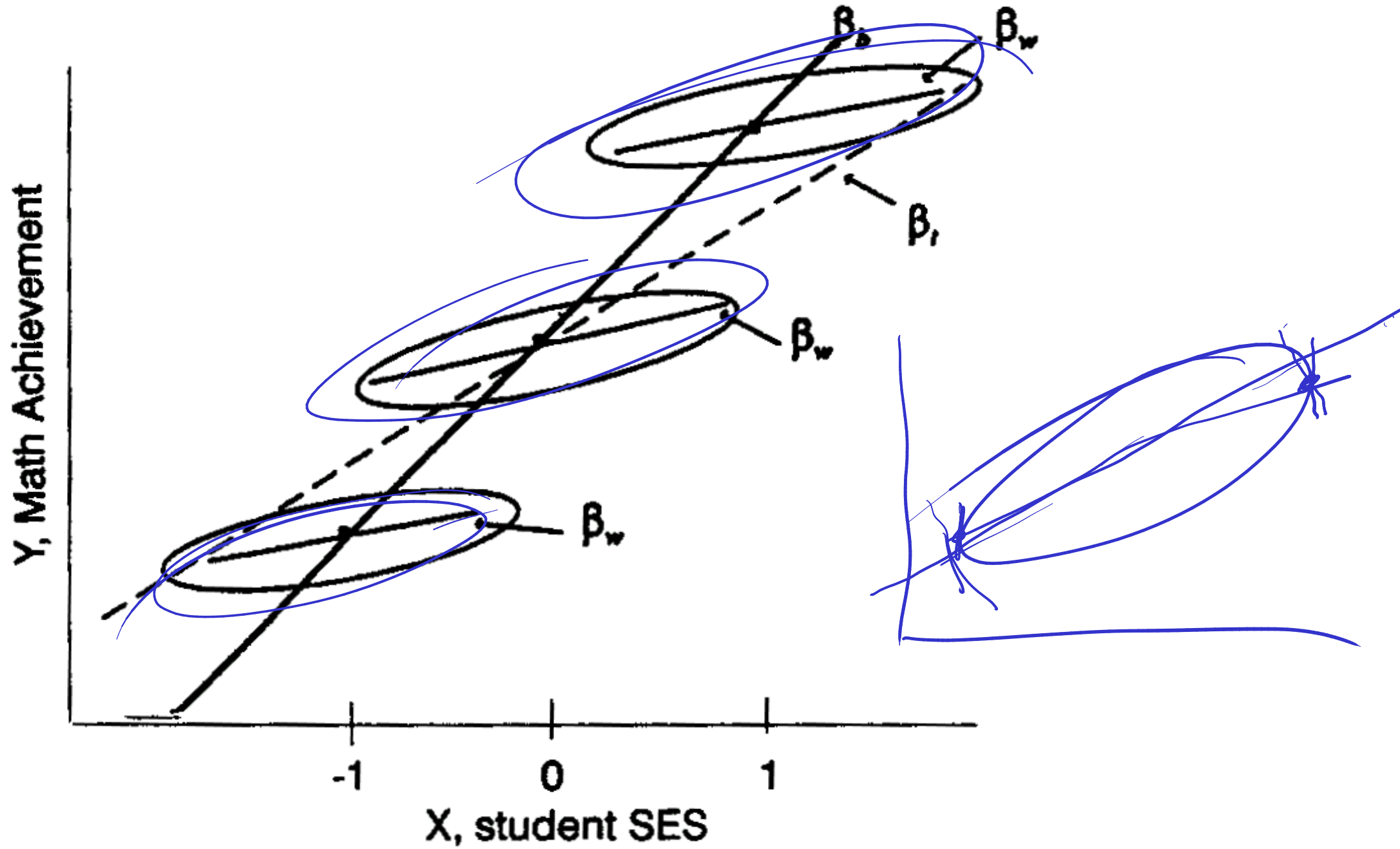
Statistical Reasoning with Ellipses: The Data Ellipse

Simple regression, the regression paradox, the anatomy of outliers: fit vs leverage, significance at a glance



Georges Monette

random@yorku.ca



Example: Origins of regression

Galton and Pearson studied the inheritance of height from father to son

To make things simple we use simulated data with adult heights in two generations where there is no drift:

Fathers: mean 68" sd 3"

Sons (adult) mean 68" sd 3"

Father's height Son's expected height

68

?

Father's height	Son's expected height
68	68

Father's height Son's expected height

$$\begin{array}{c} 68 \\ 71 = 68 + 3 \end{array}$$

$$\begin{array}{c} 68 \\ 71 \end{array}$$

Father's height Son's expected height

68

$$71 = 68 + 3$$

68

$$71 = 68 + 3$$

Father's height Son's expected height

68

$$71 = 68 + 3$$

$$74 = 68 + 6$$

$$65 = 68 - 3$$

$$62 = 68 - 6$$

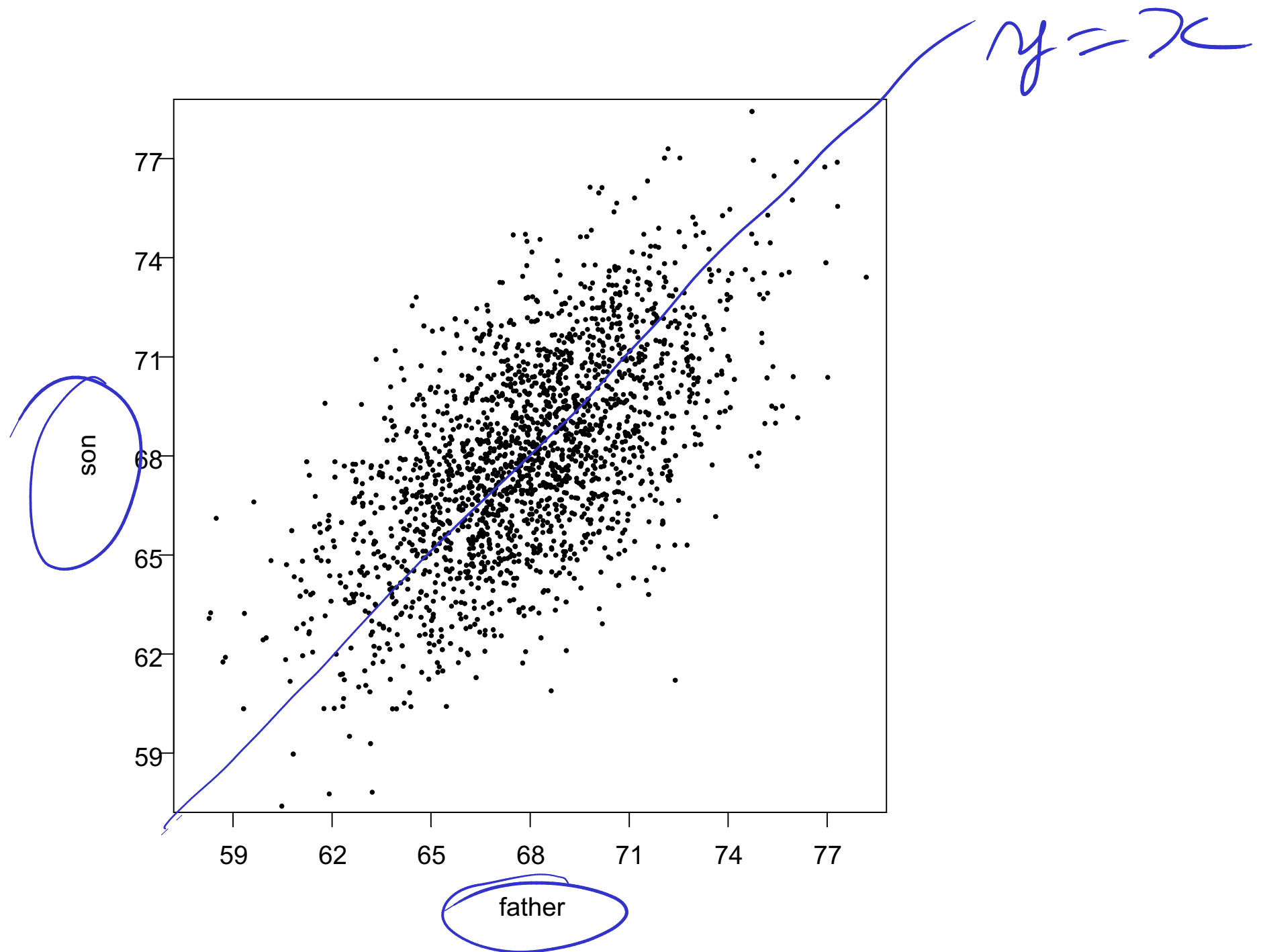
68

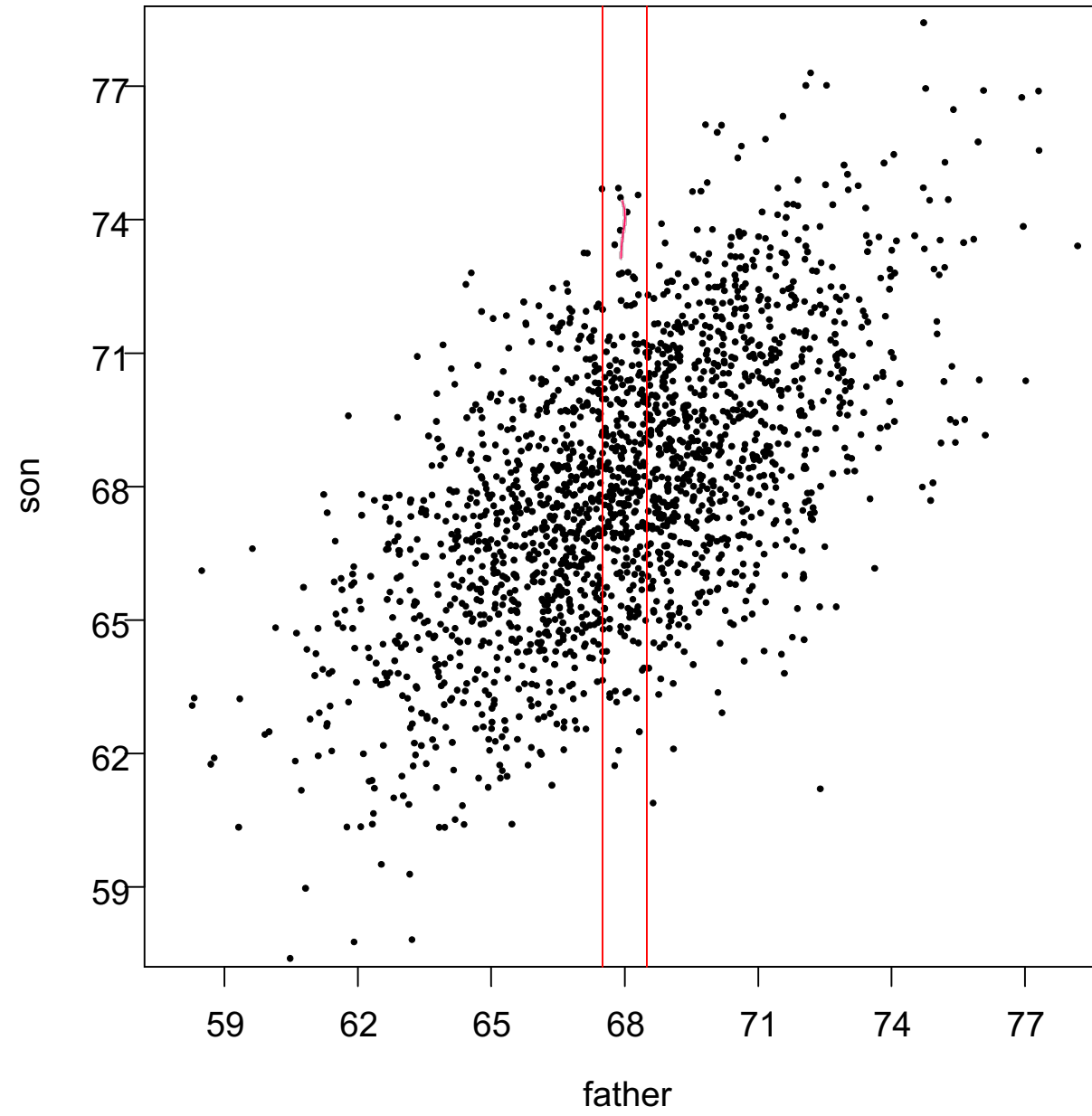
$$71 = 68 + 3$$

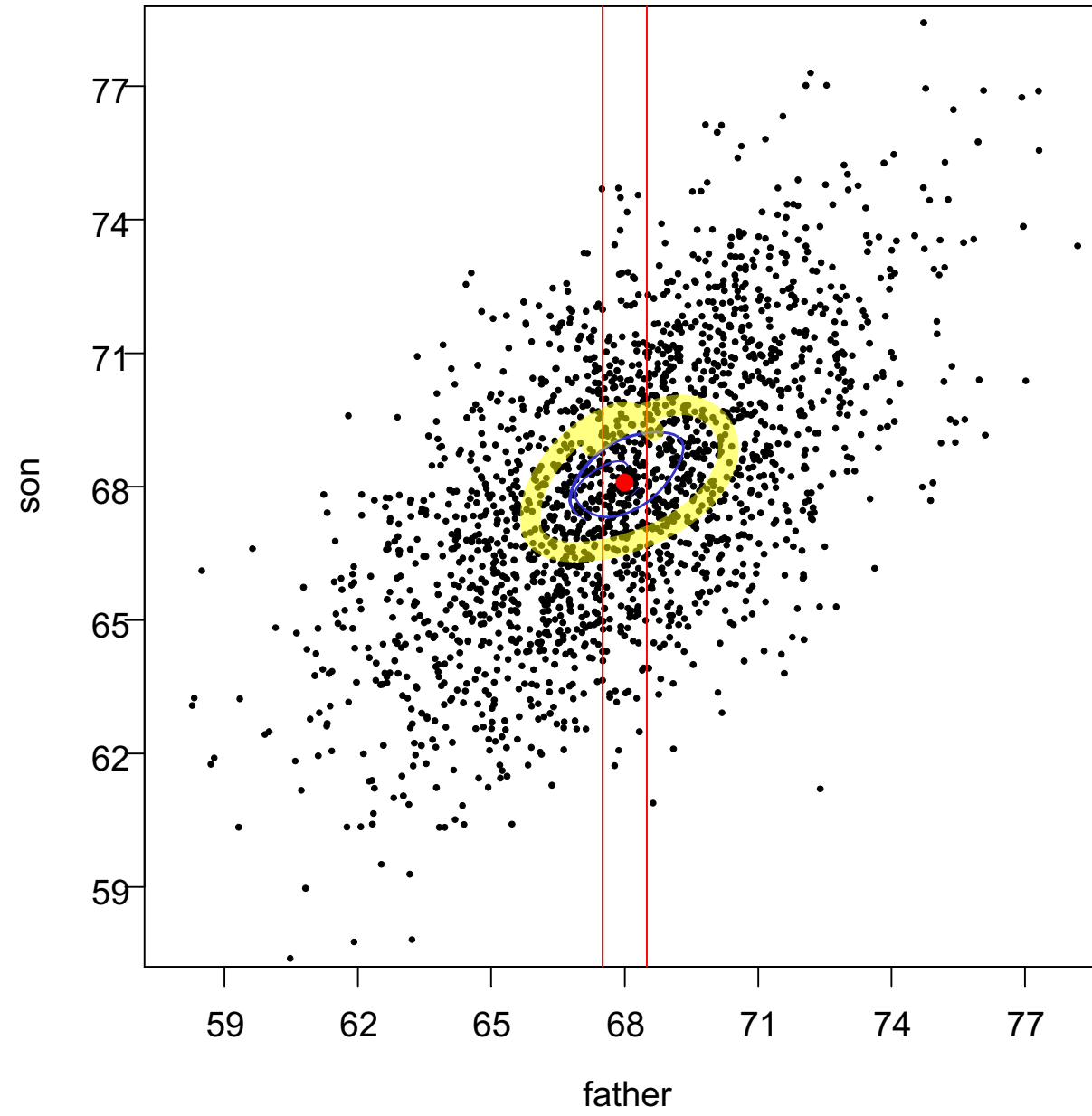
$$74 = 68 + 6$$

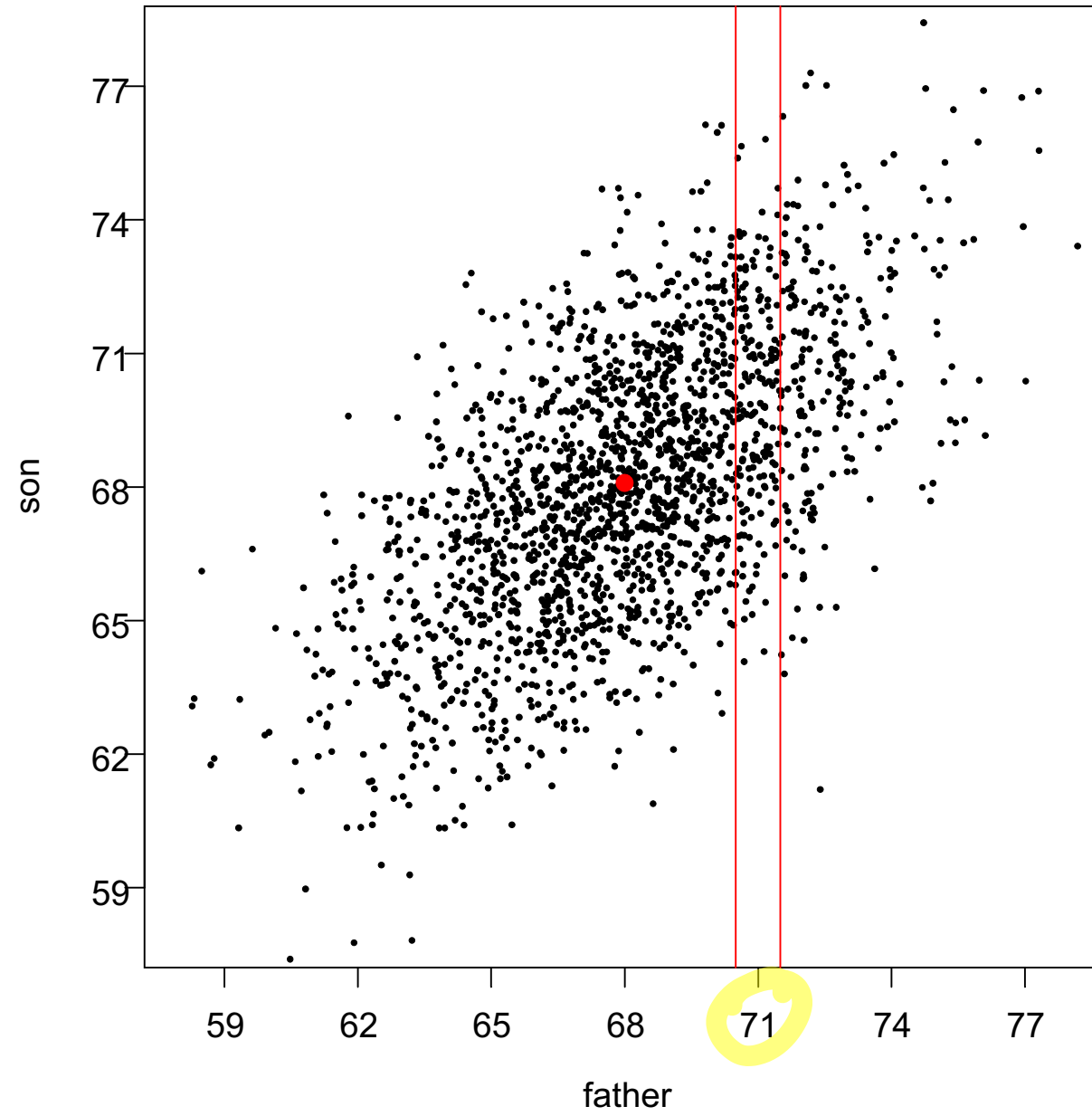
$$65 = 68 - 3$$

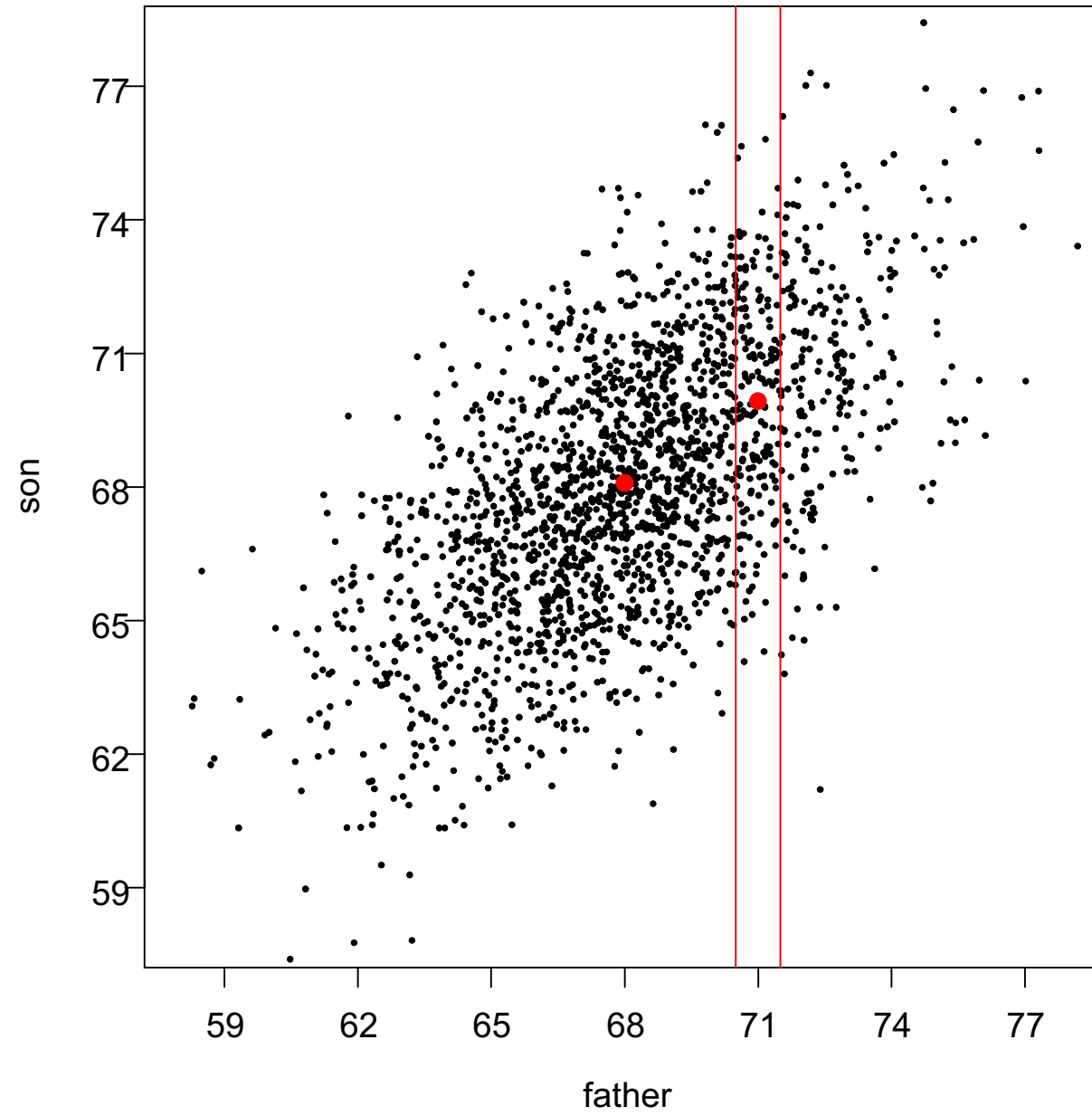
$$62 = 68 - 6$$

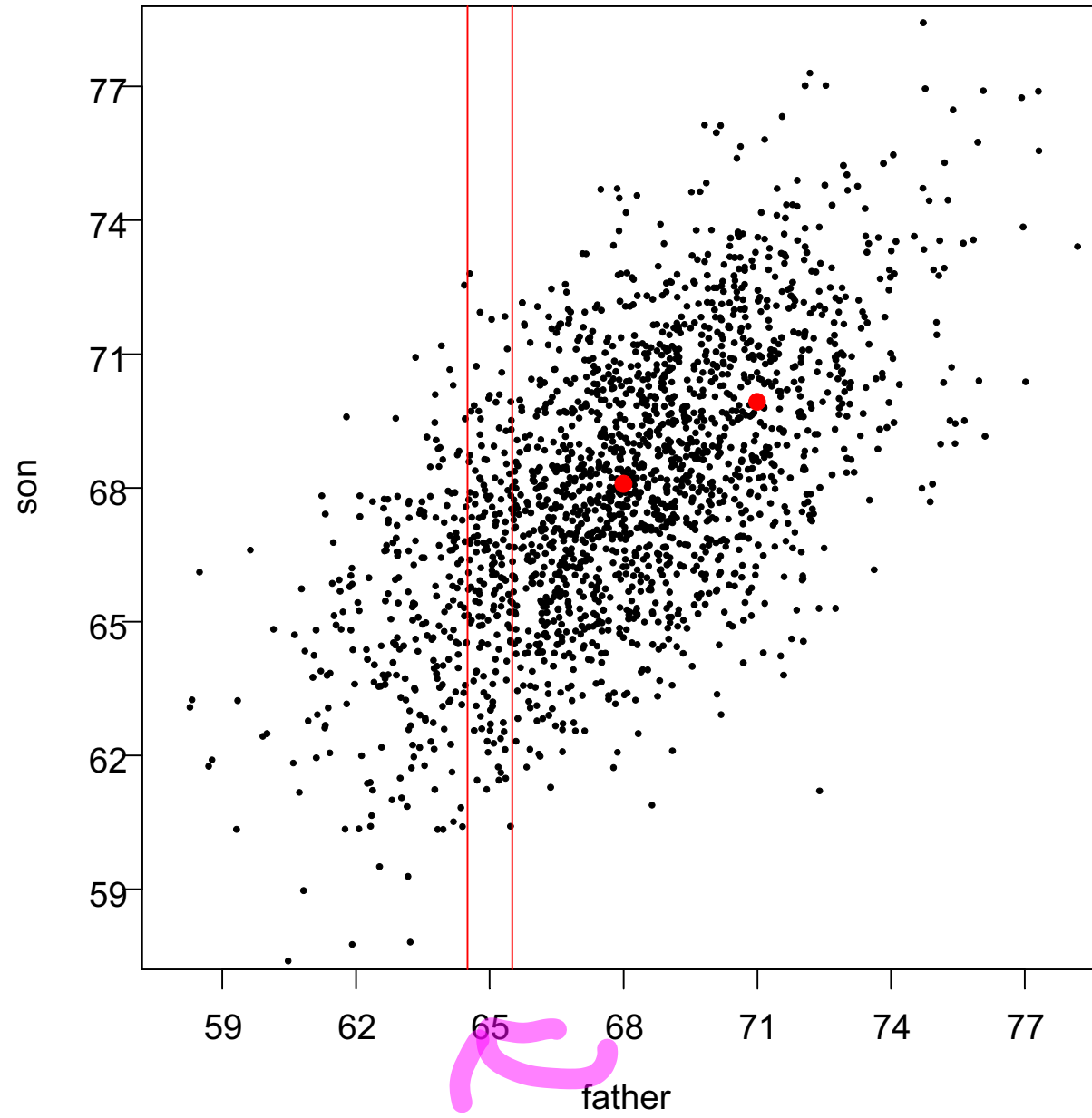


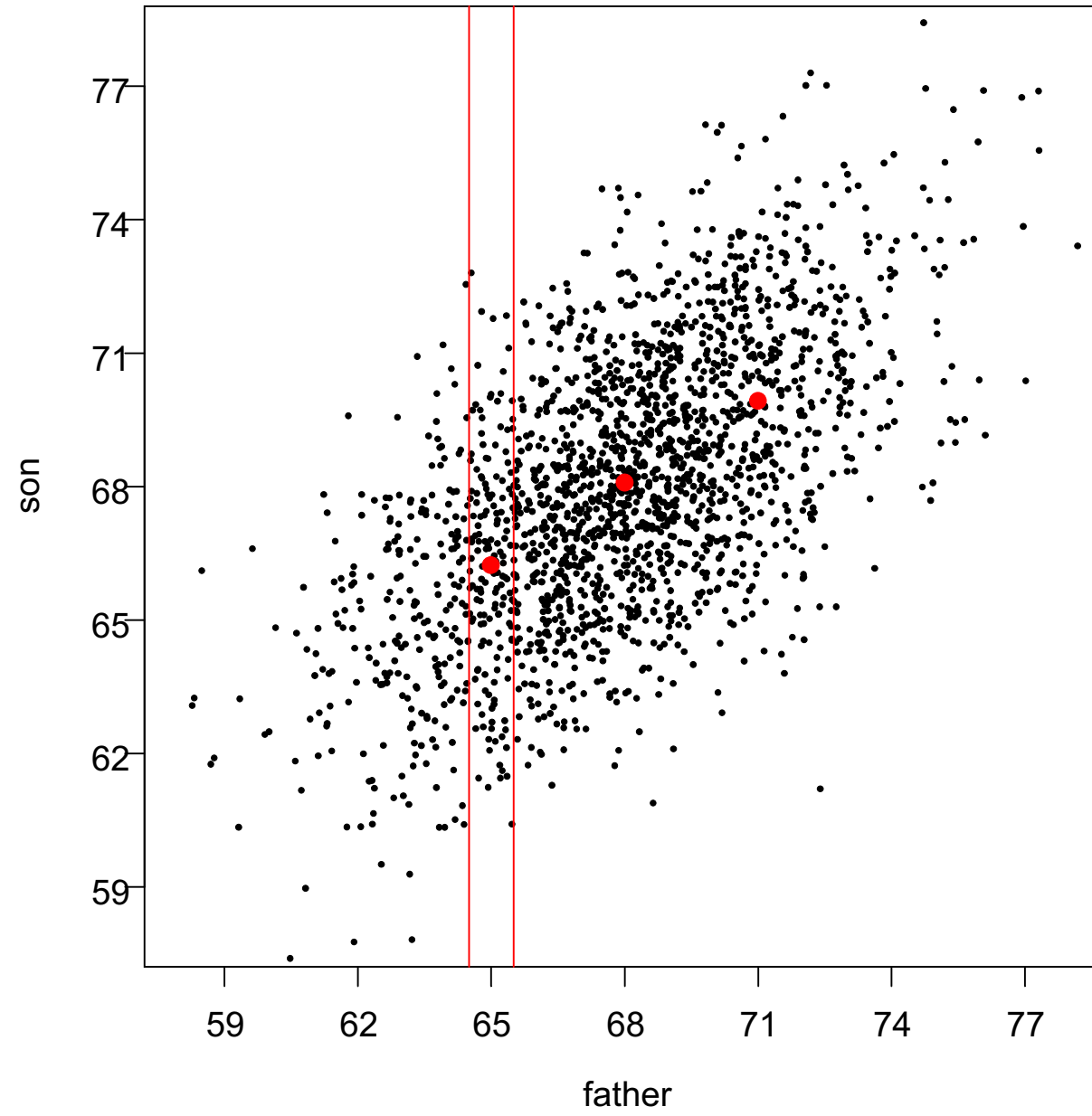


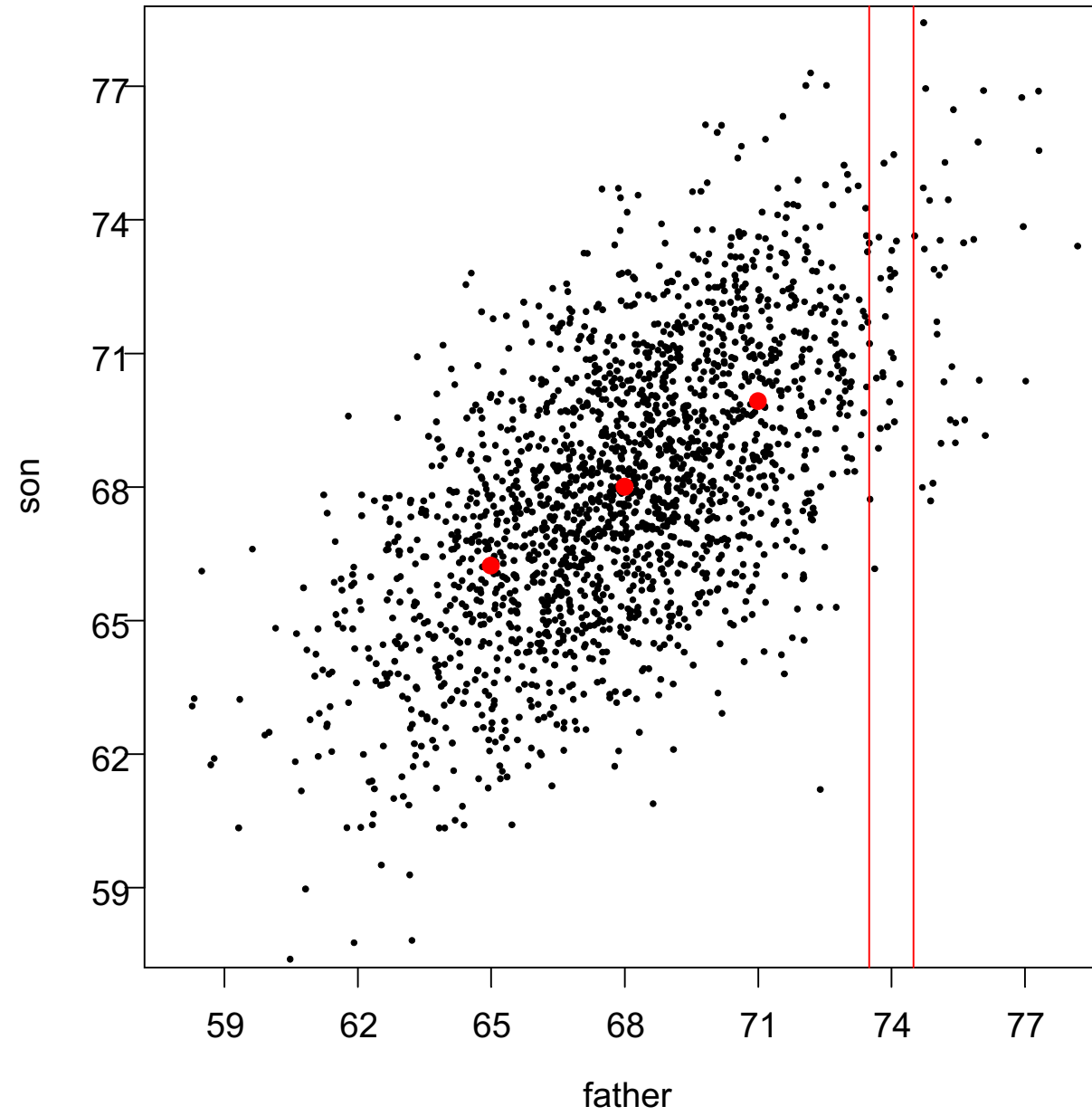


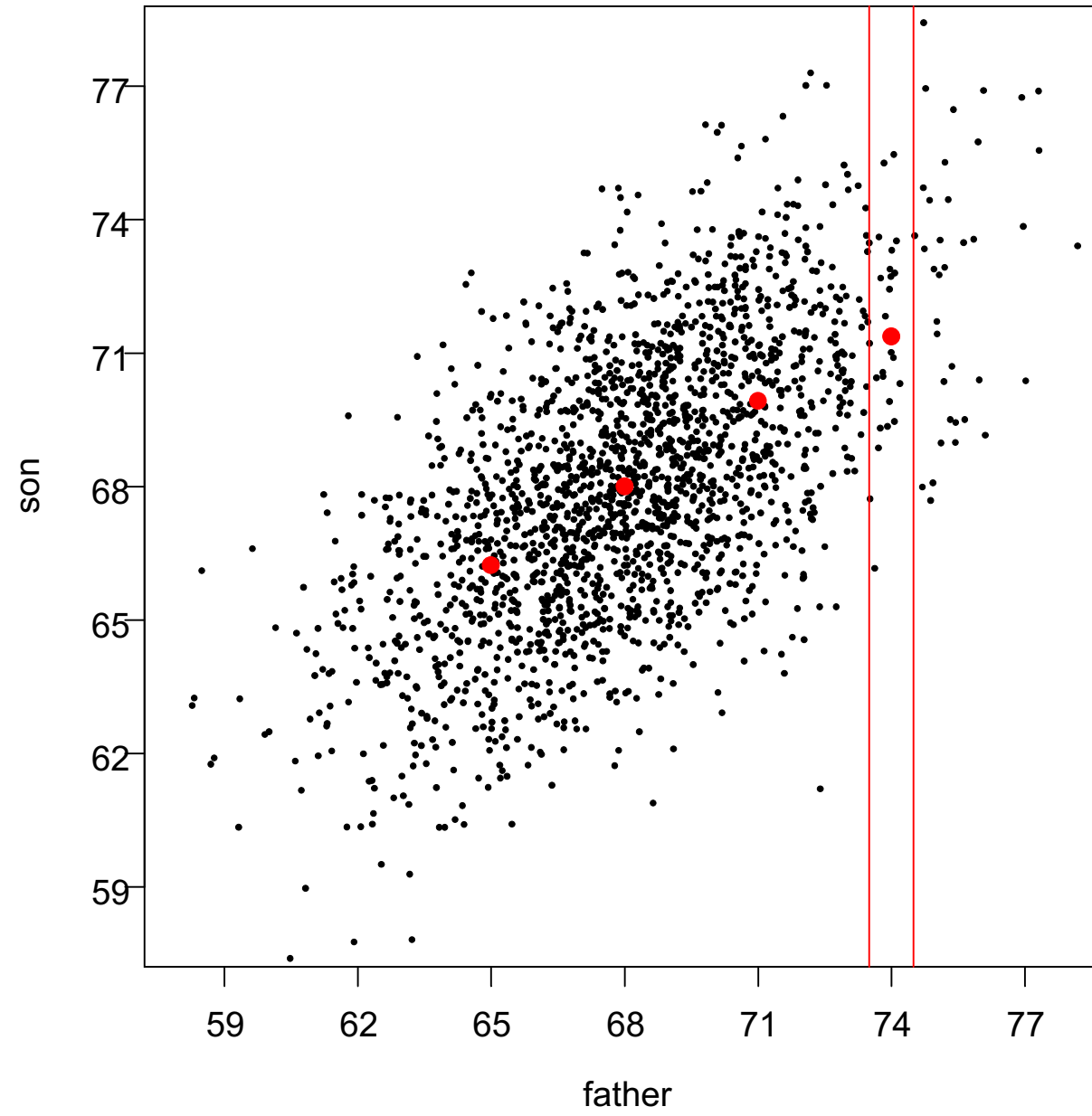


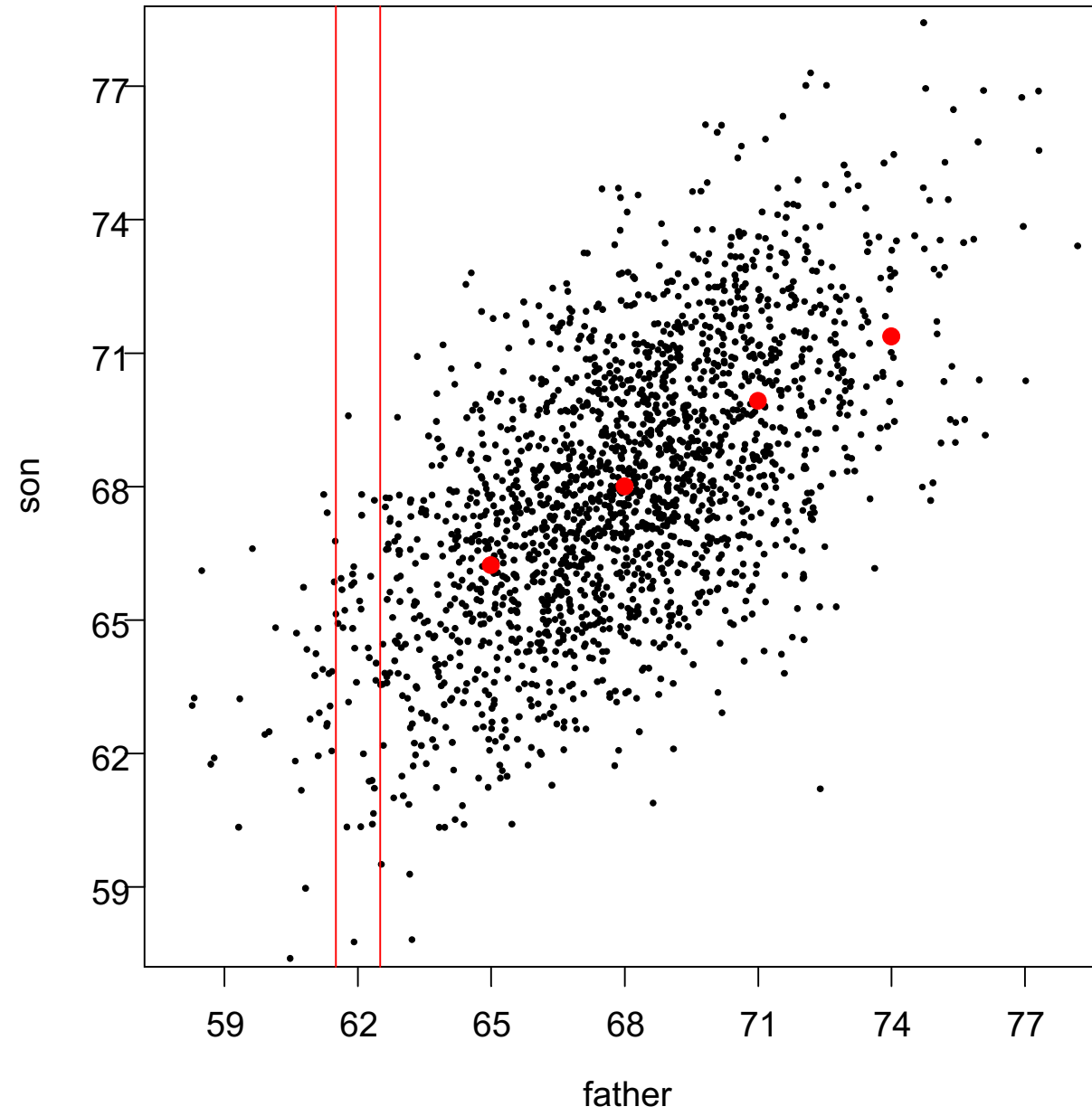


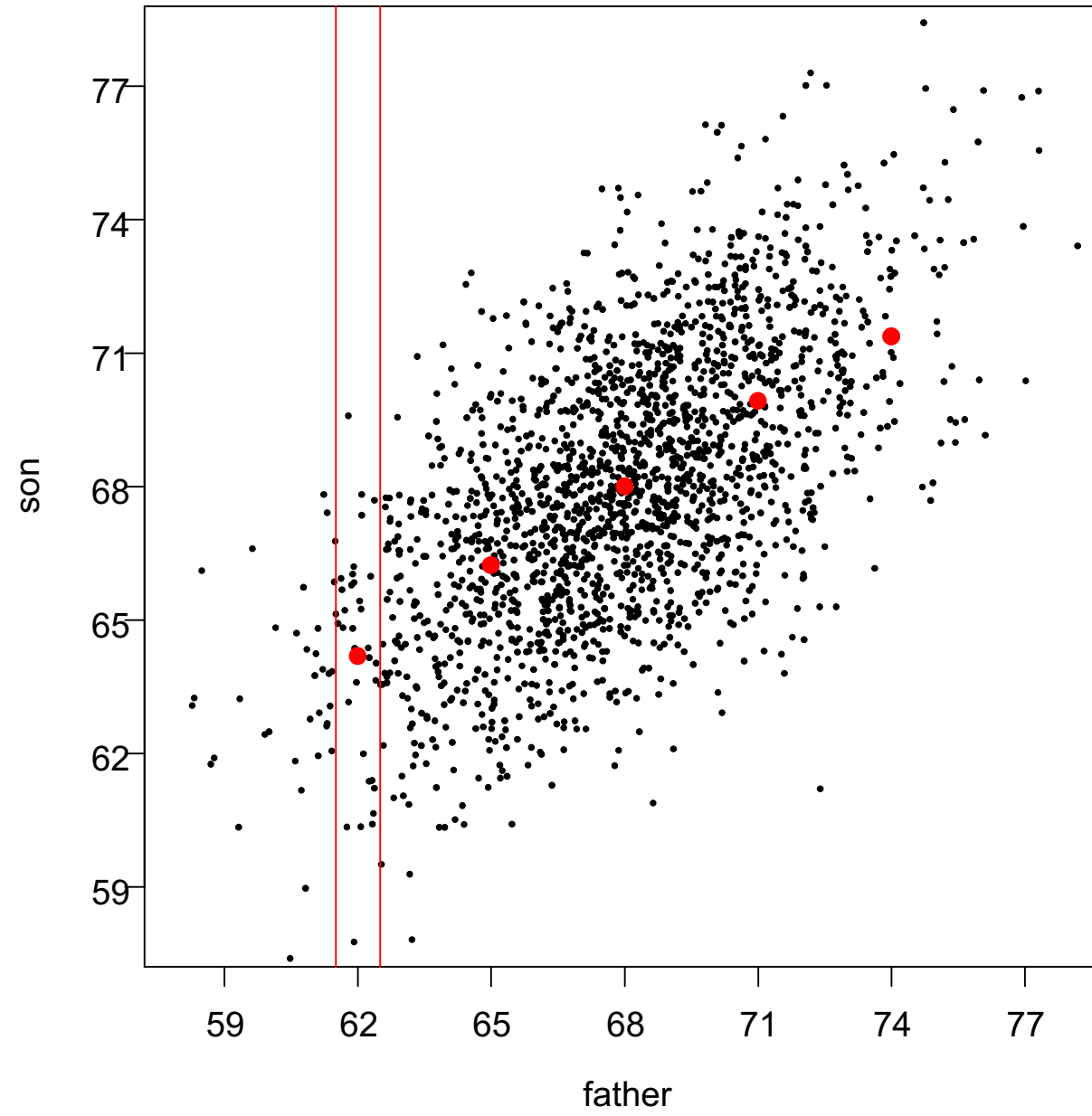


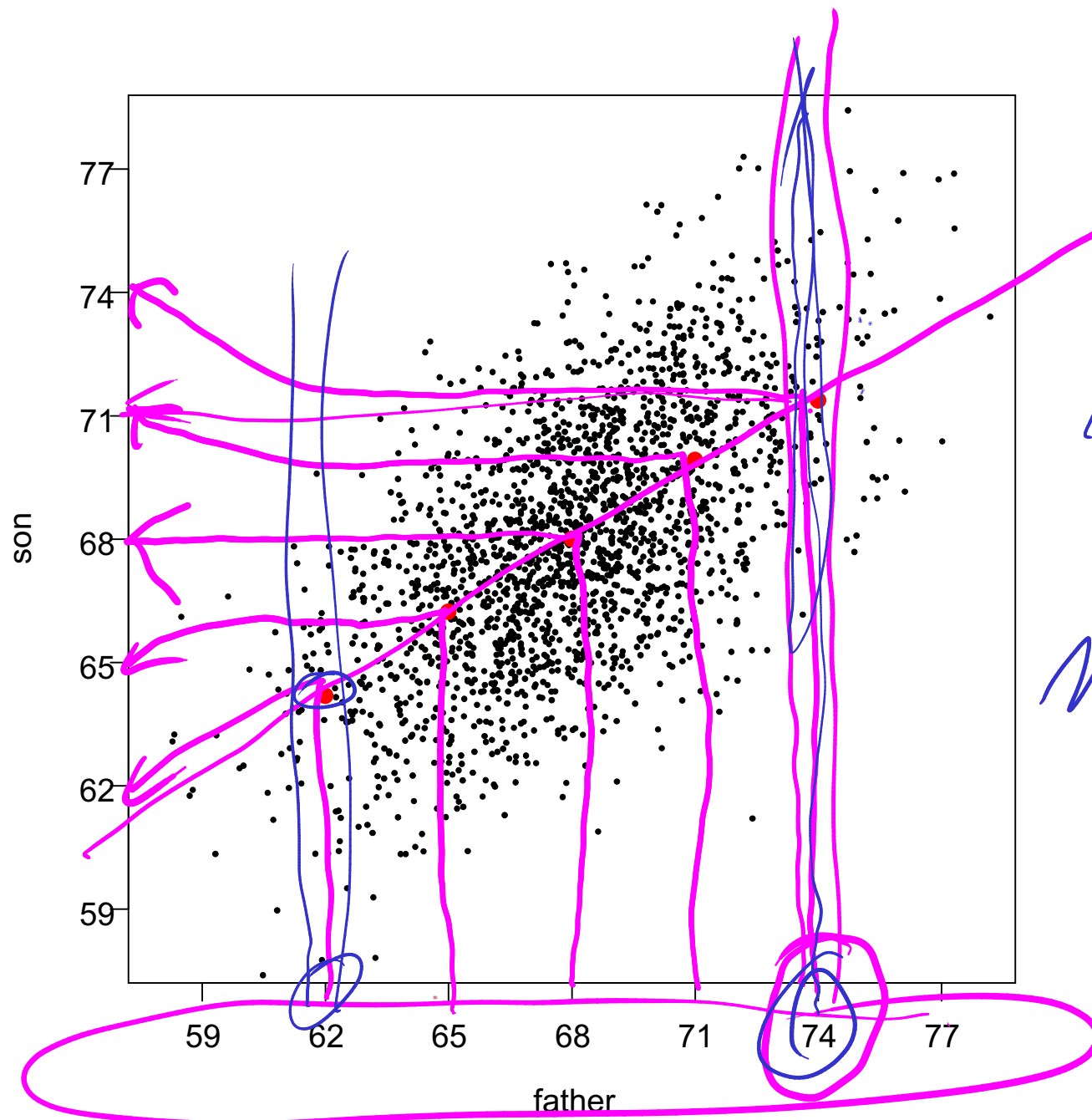






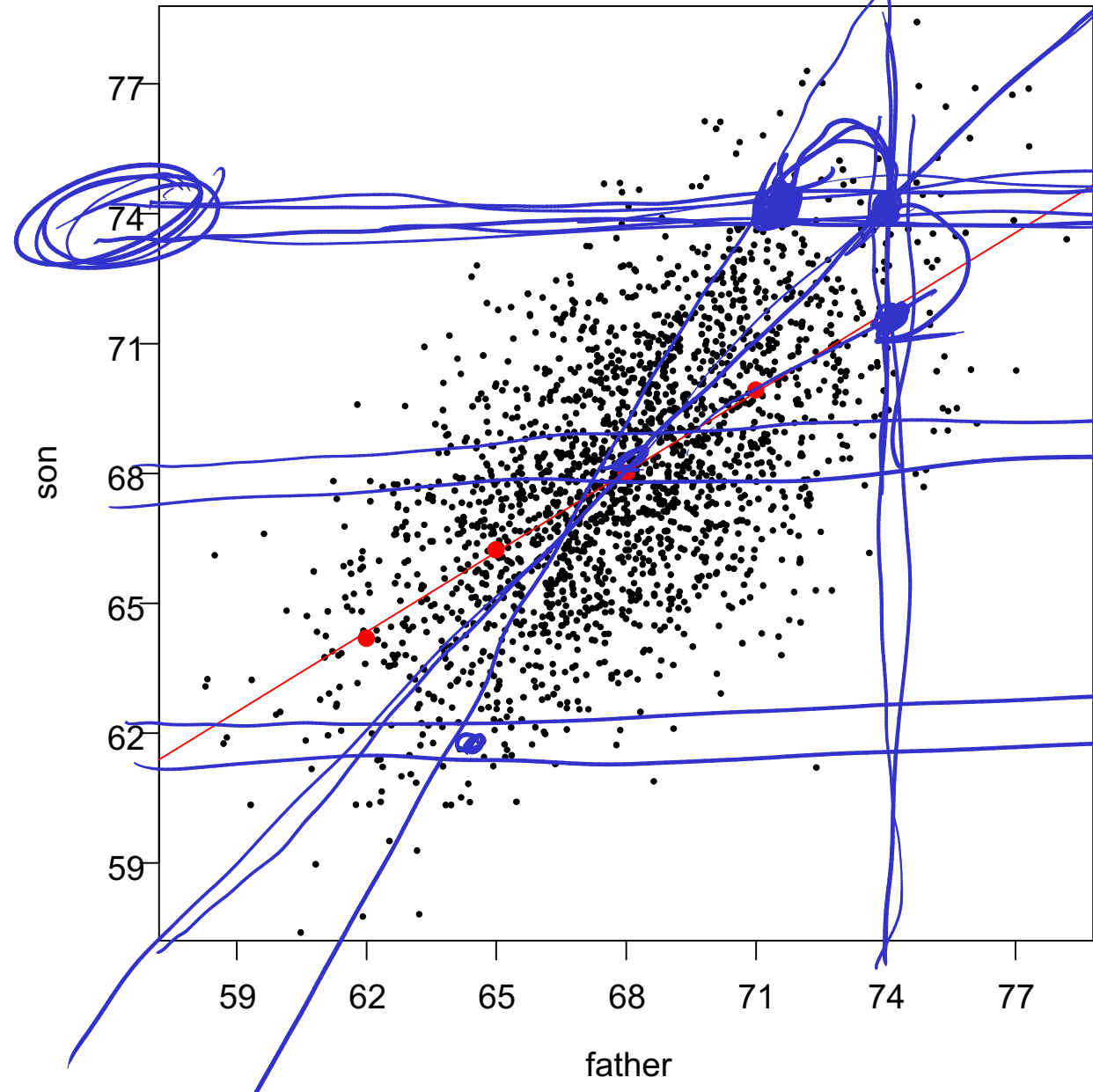




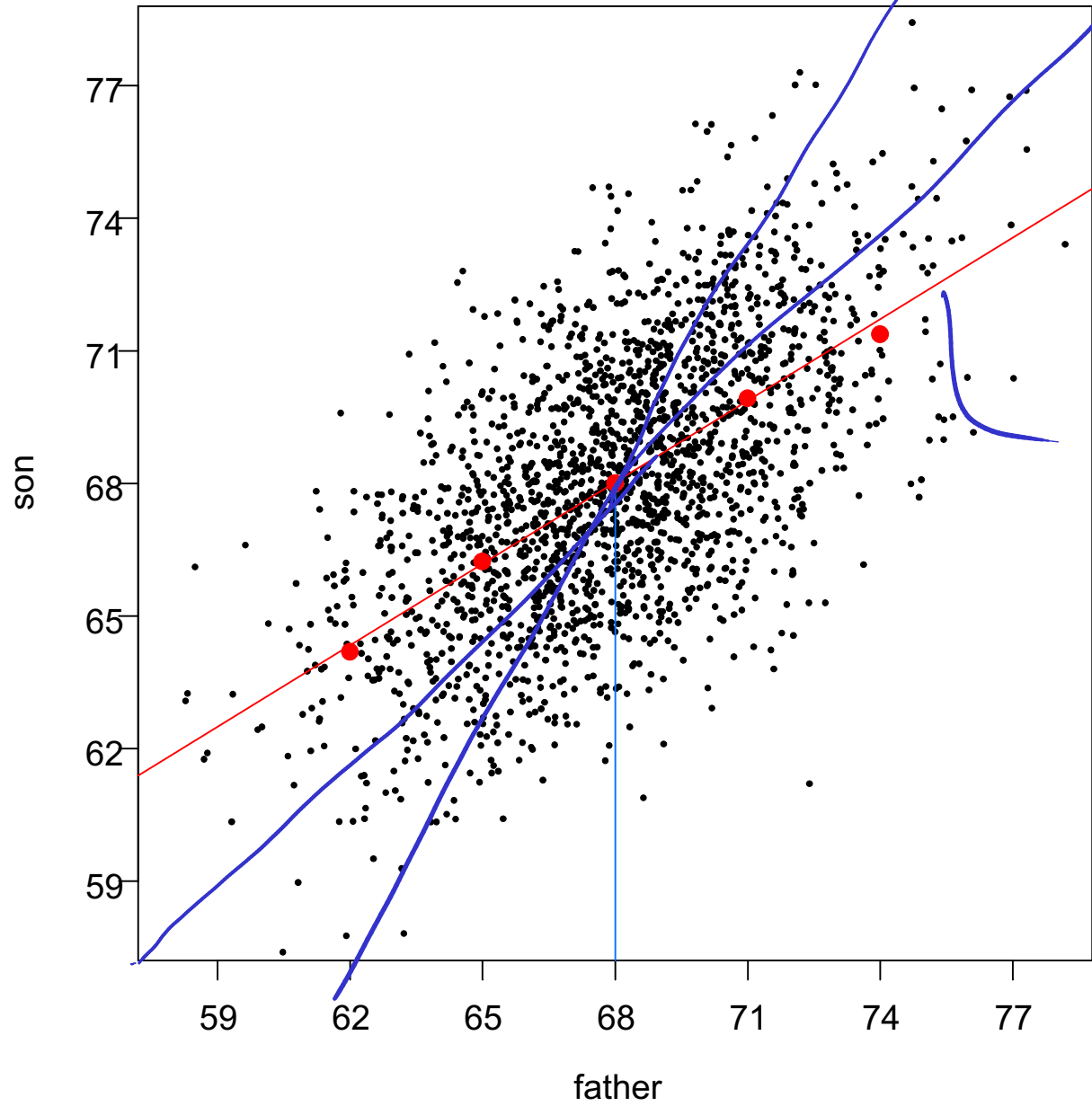


Regression to mediocrity

X or Y — pred X from Y
 $y = x$



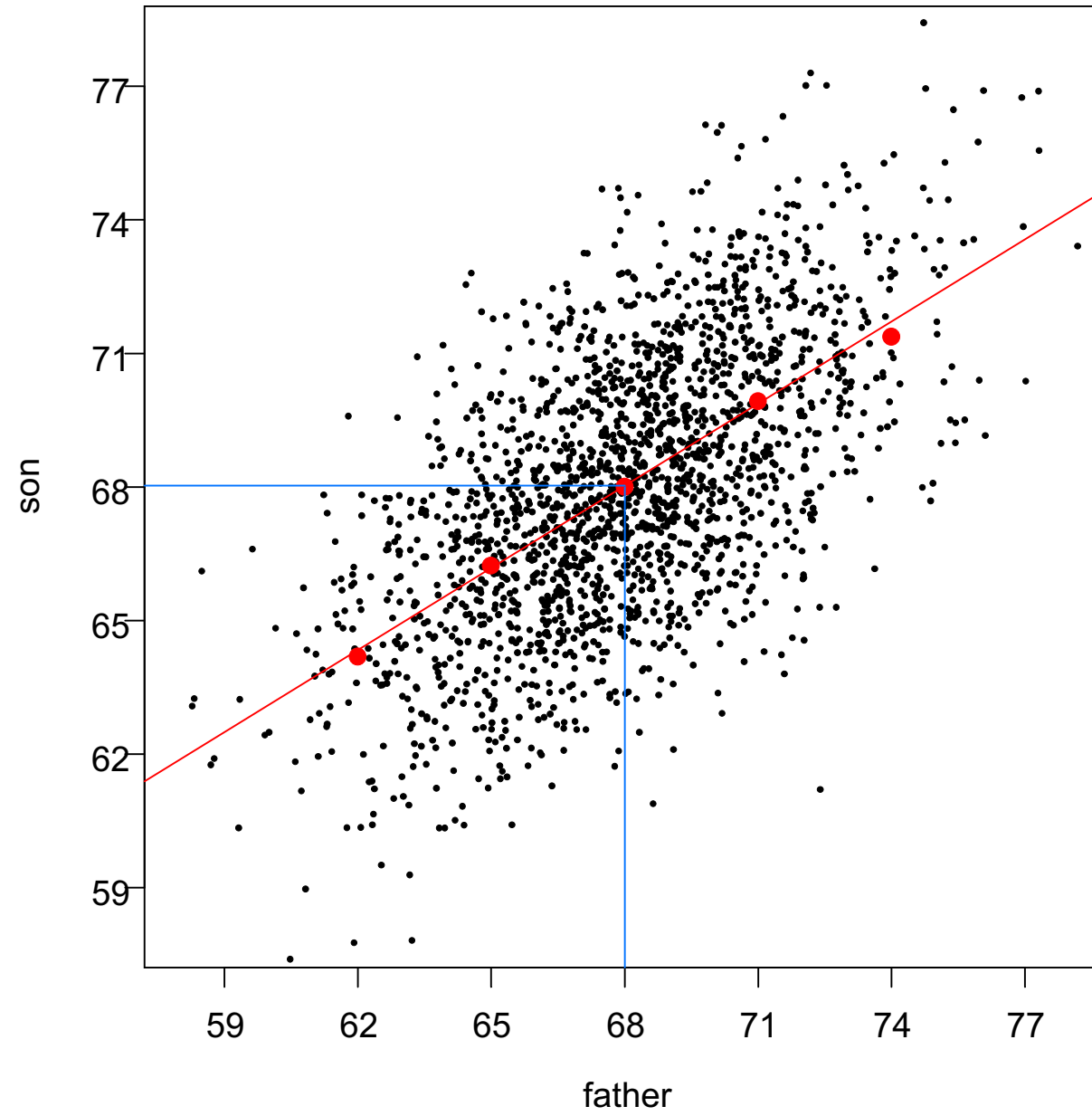
regression paradox

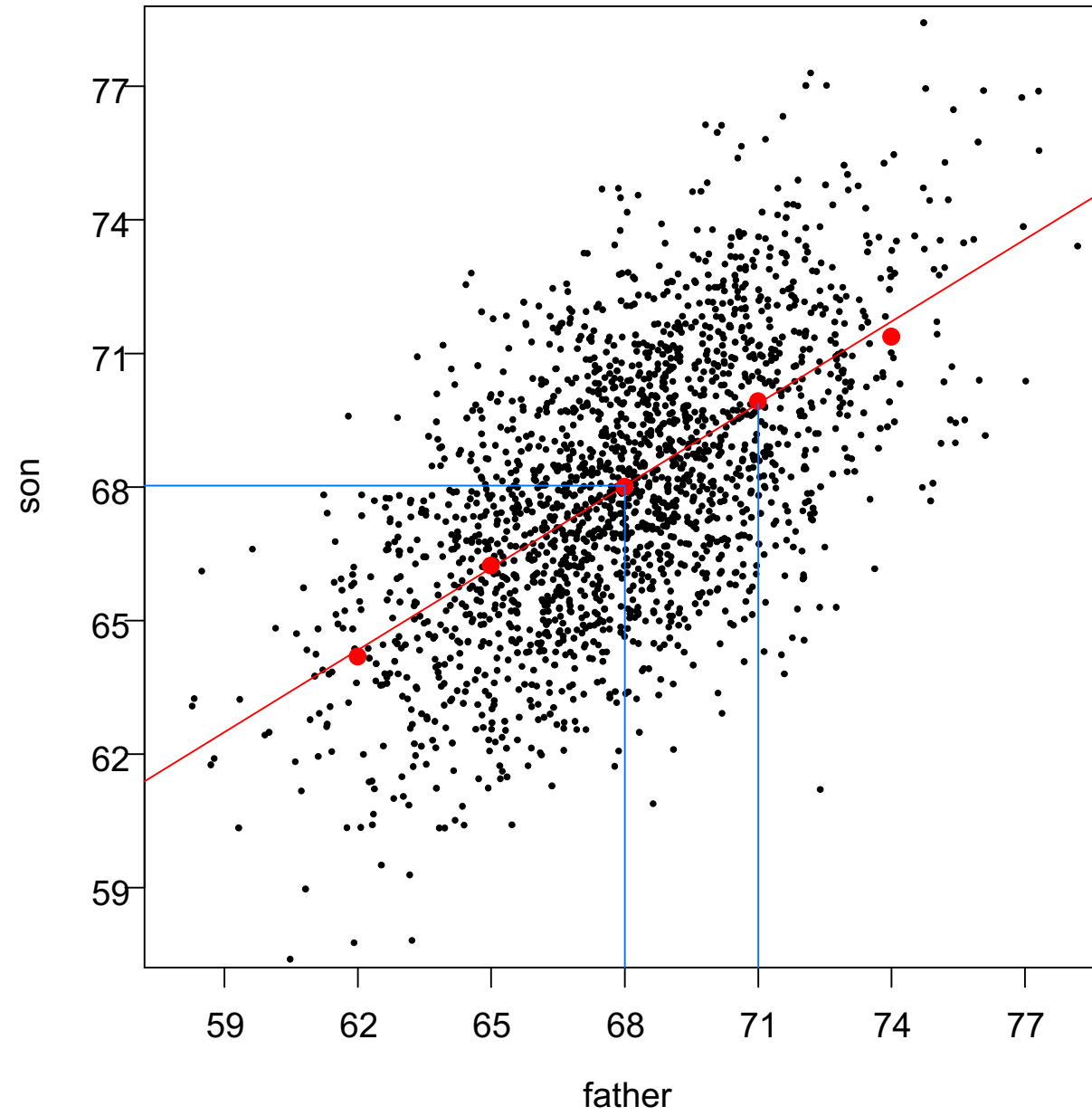


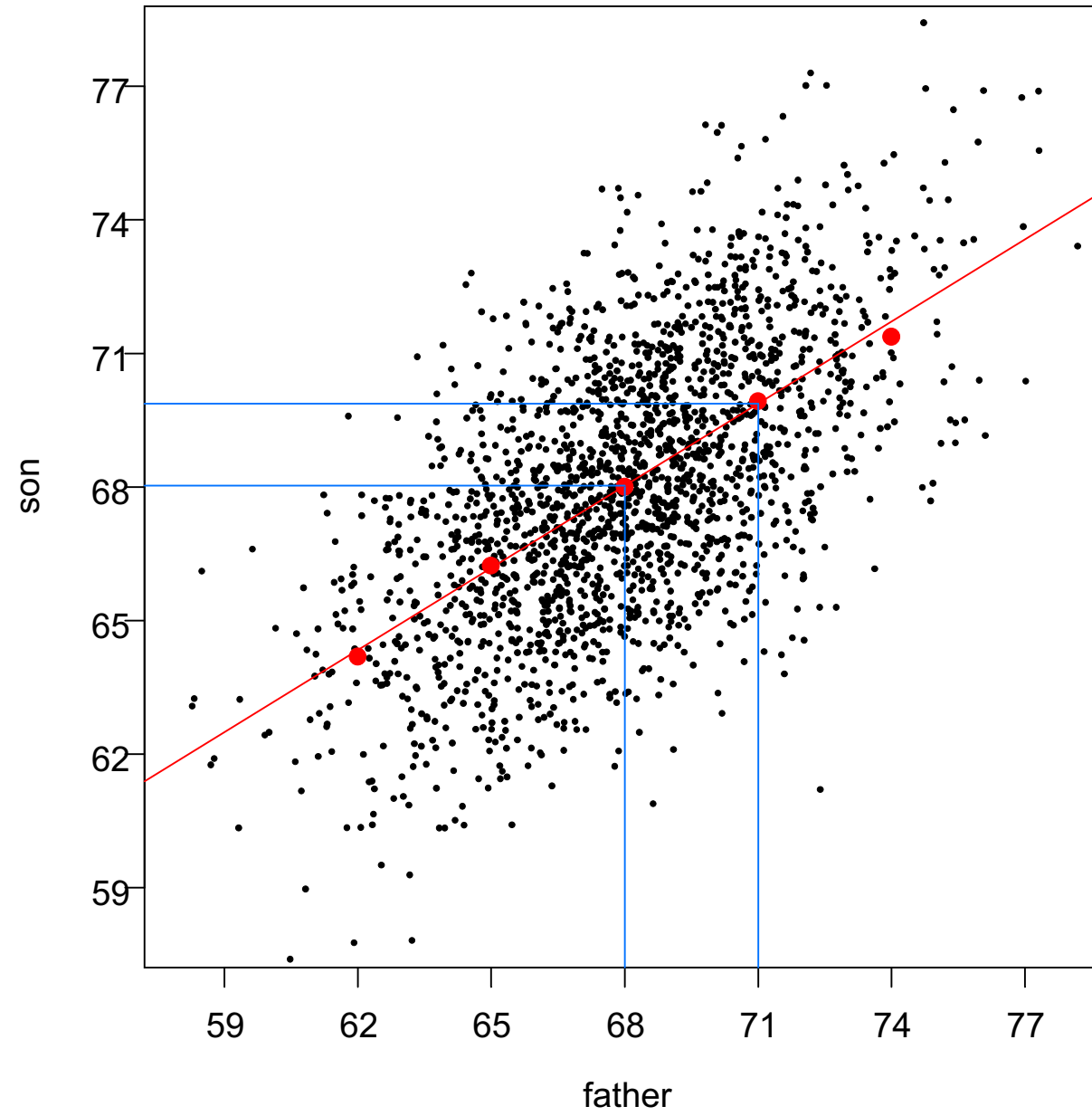
x only

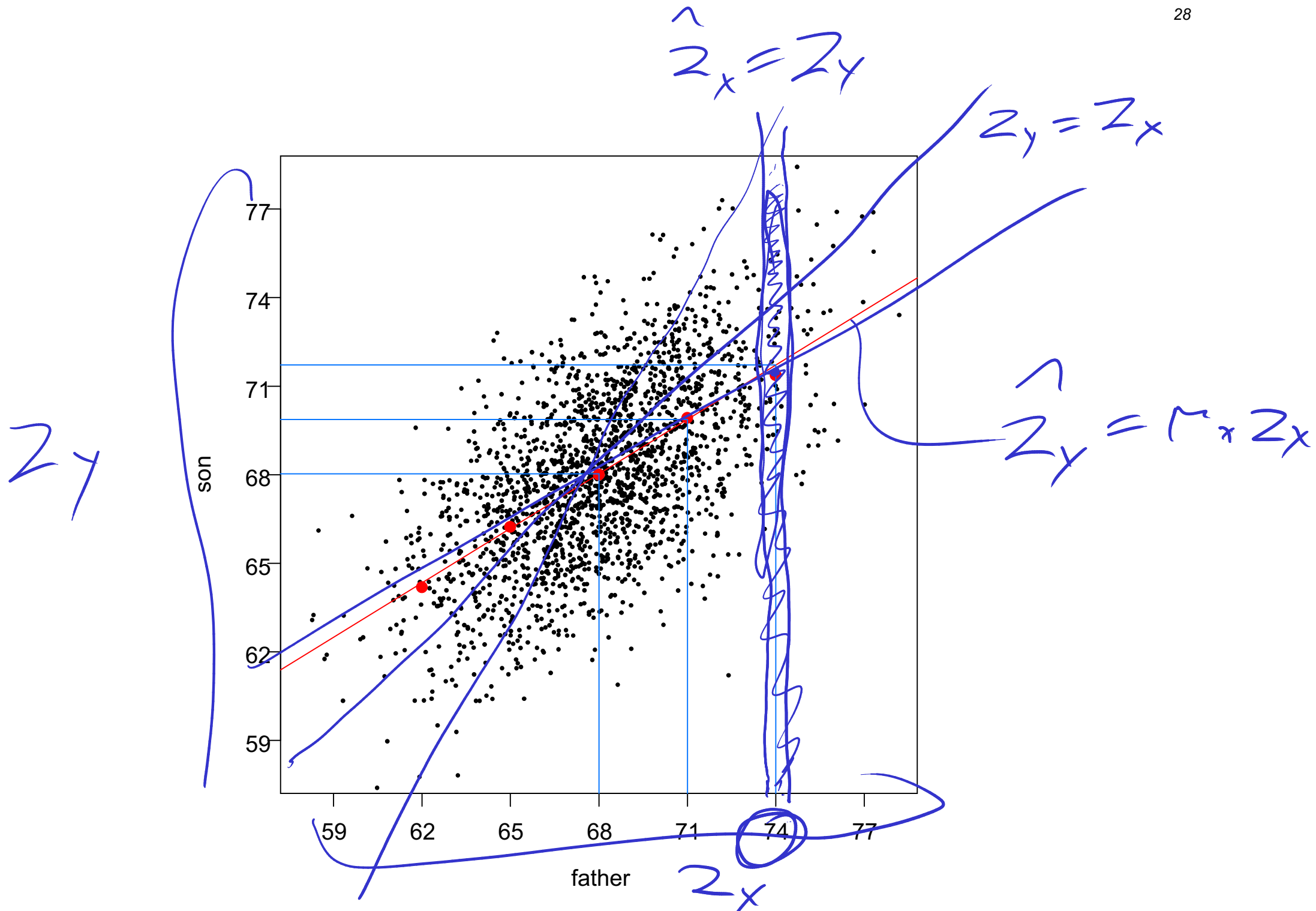
y only

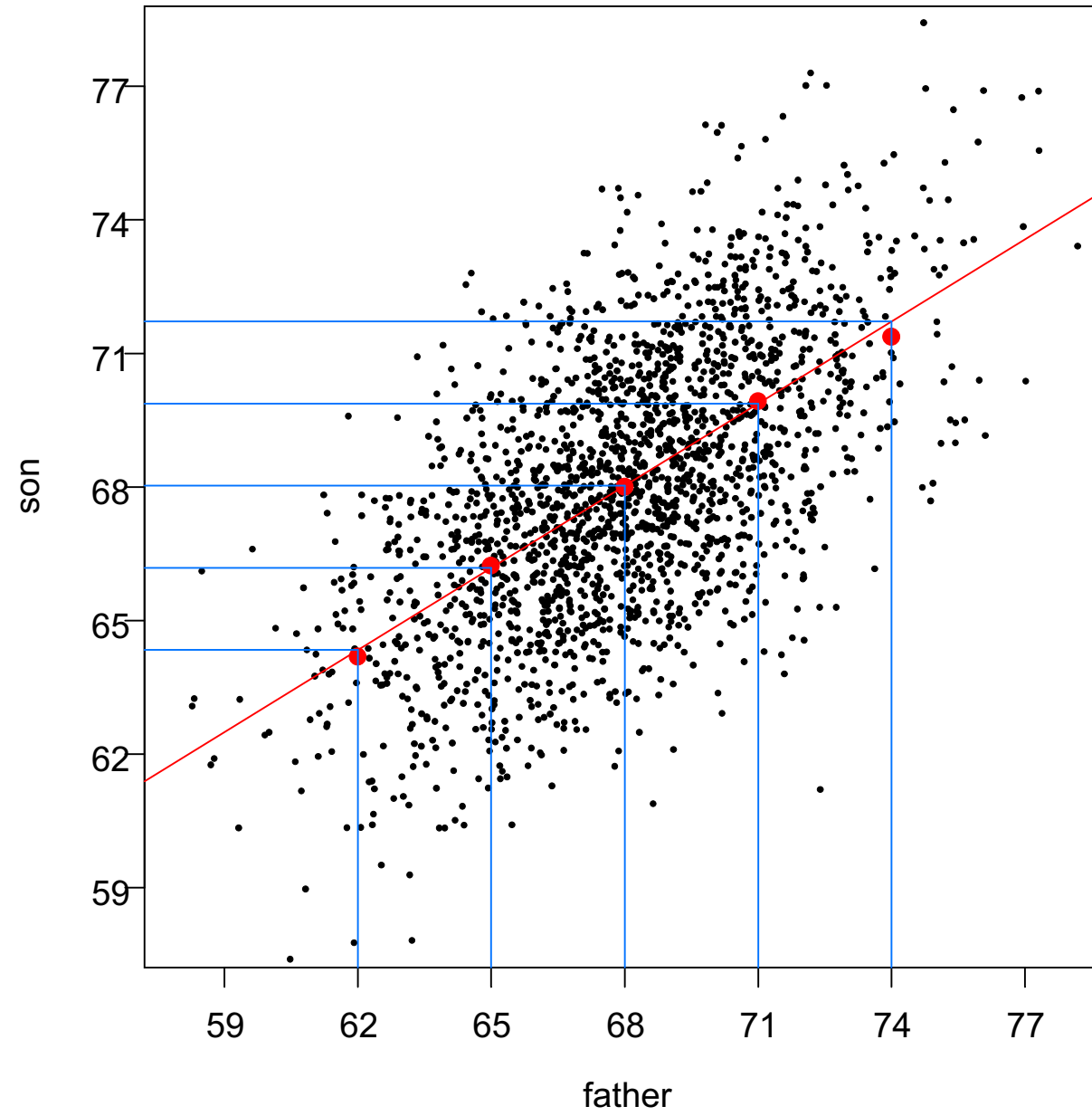
~~y only~~











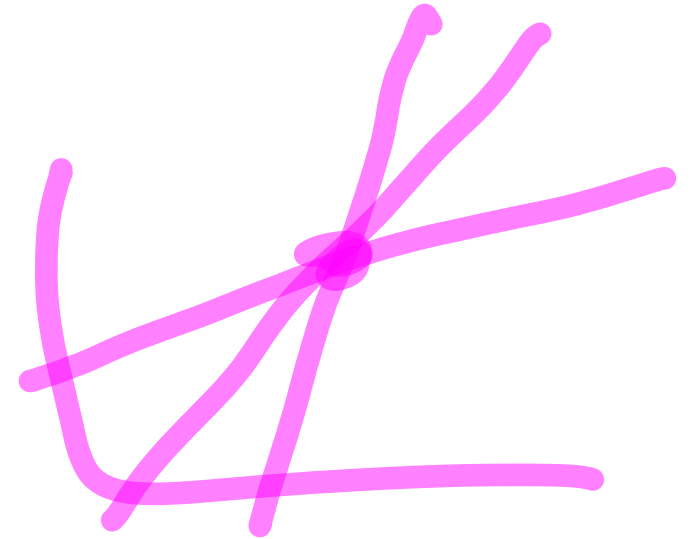
Father's height Son's expected height

68

Father's height Son's expected height

68

68



Father's height Son's expected height

68

68

$$\underline{71 = 68 + 3}$$

Father's height Son's expected height

$$\begin{array}{c} 68 \\ 71 = 68 + 3 \end{array}$$

$$\begin{array}{c} 68 \\ 69.8 = 68 + r \times 3 \end{array}$$

$$\begin{array}{l} r = 0.6 \\ r \times 3 = 1.8 \end{array}$$

Father's height Son's expected height

68

68

$$71 = 68 + 3$$

$$69.8 = 68 + r \times 3$$

$$74 = 68 + 6$$

$$71.6 = 68 + r \times 6$$

$$65 = 68 - 3$$

$$66.2 = 68 - r \times 3$$

$$62 = 68 - 6$$

$$64.4 = 68 - r \times 6$$

Correlation

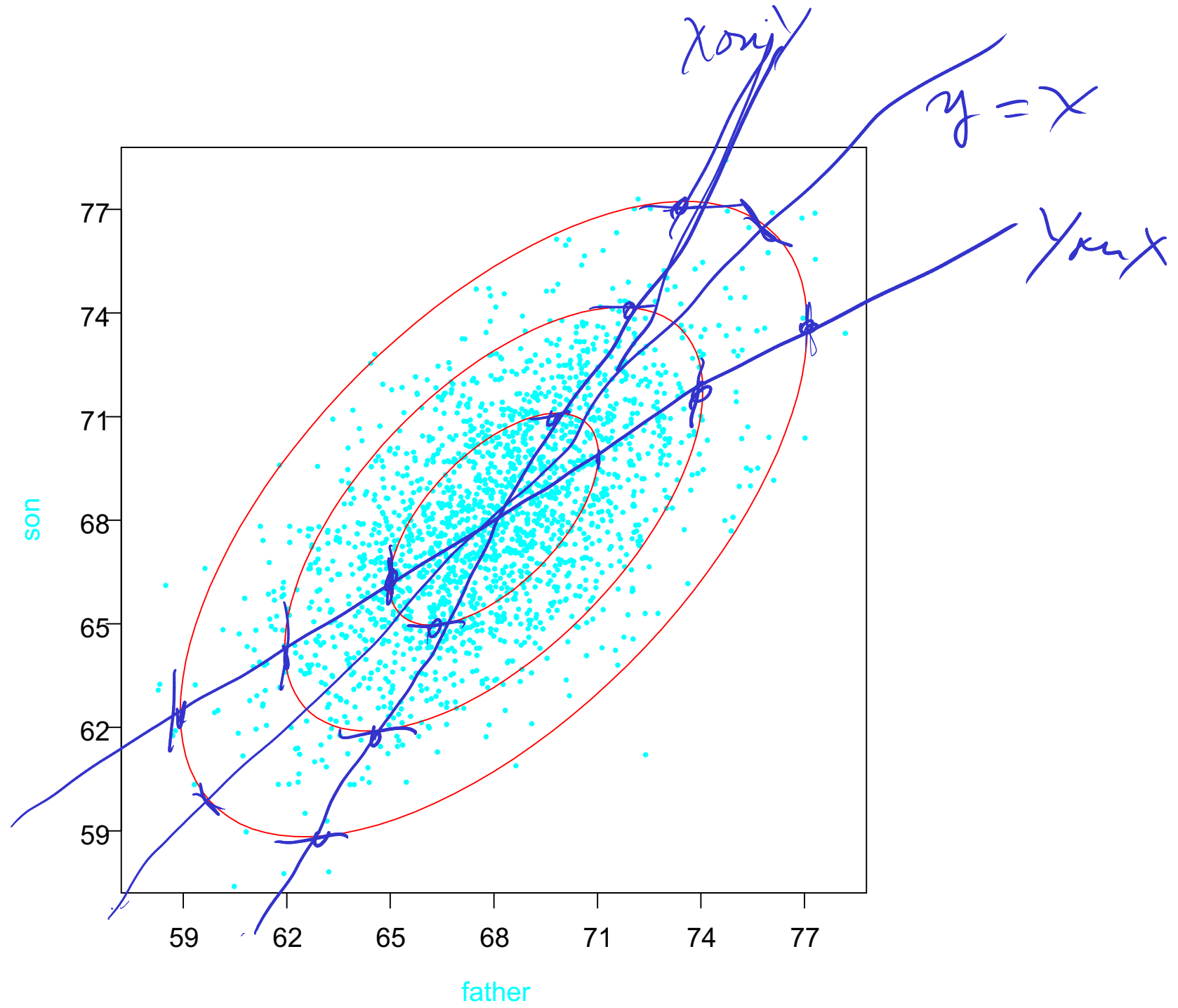
$$\hat{Y} = a + bX$$

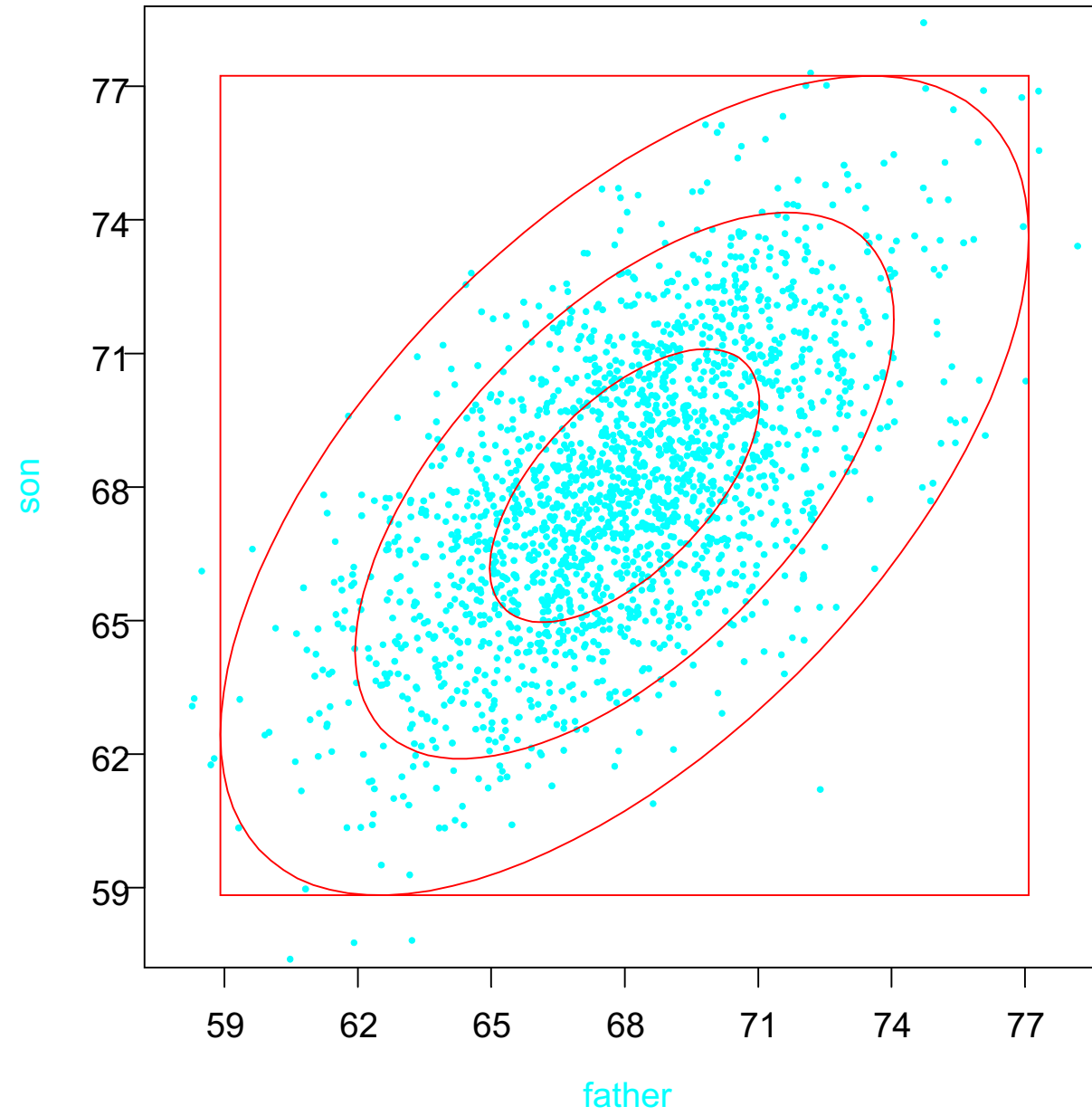
$$Z_Y = \frac{Y - \bar{Y}}{S_Y}$$

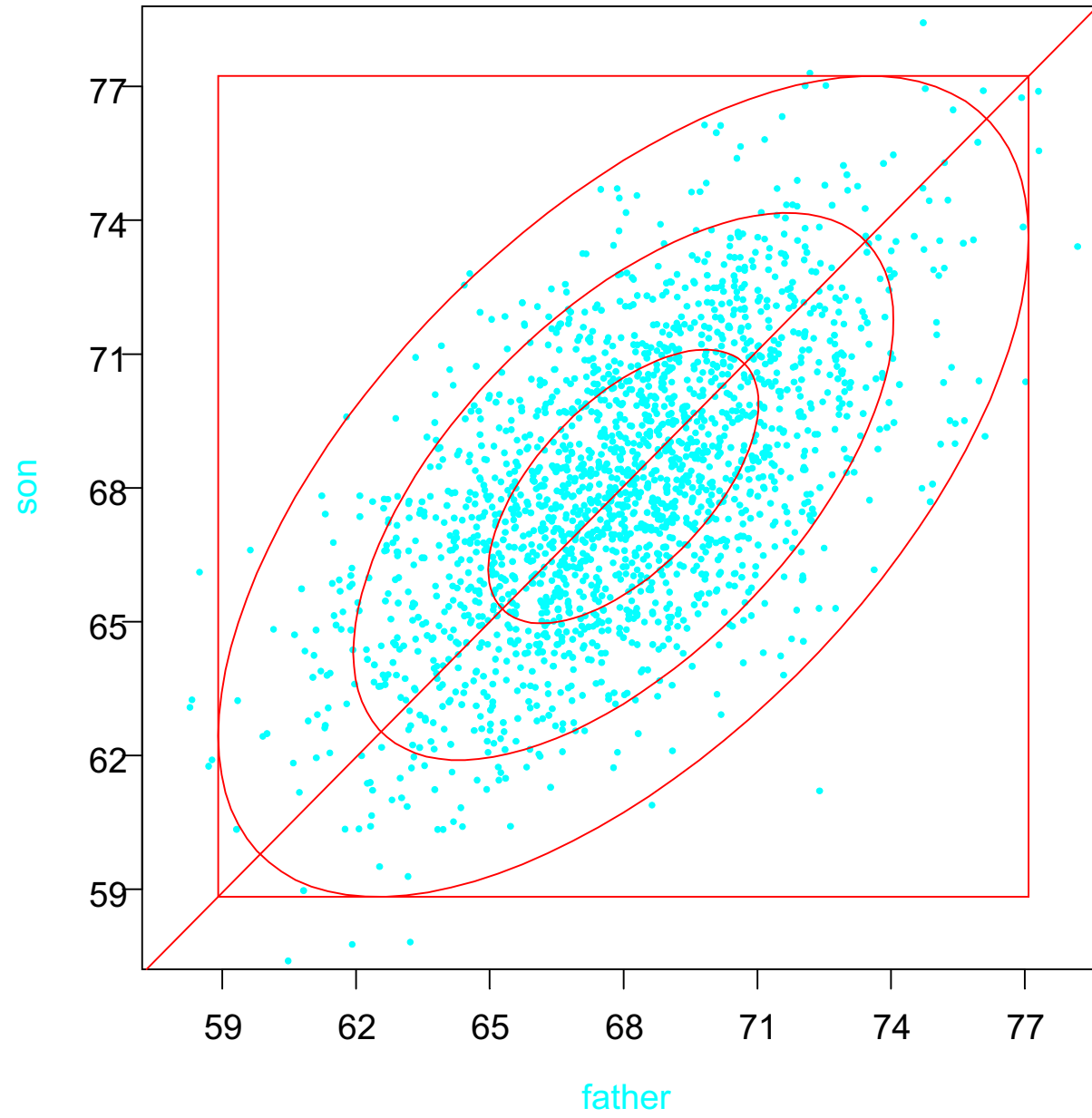
$$Z_X = \frac{X - \bar{X}}{S_X}$$

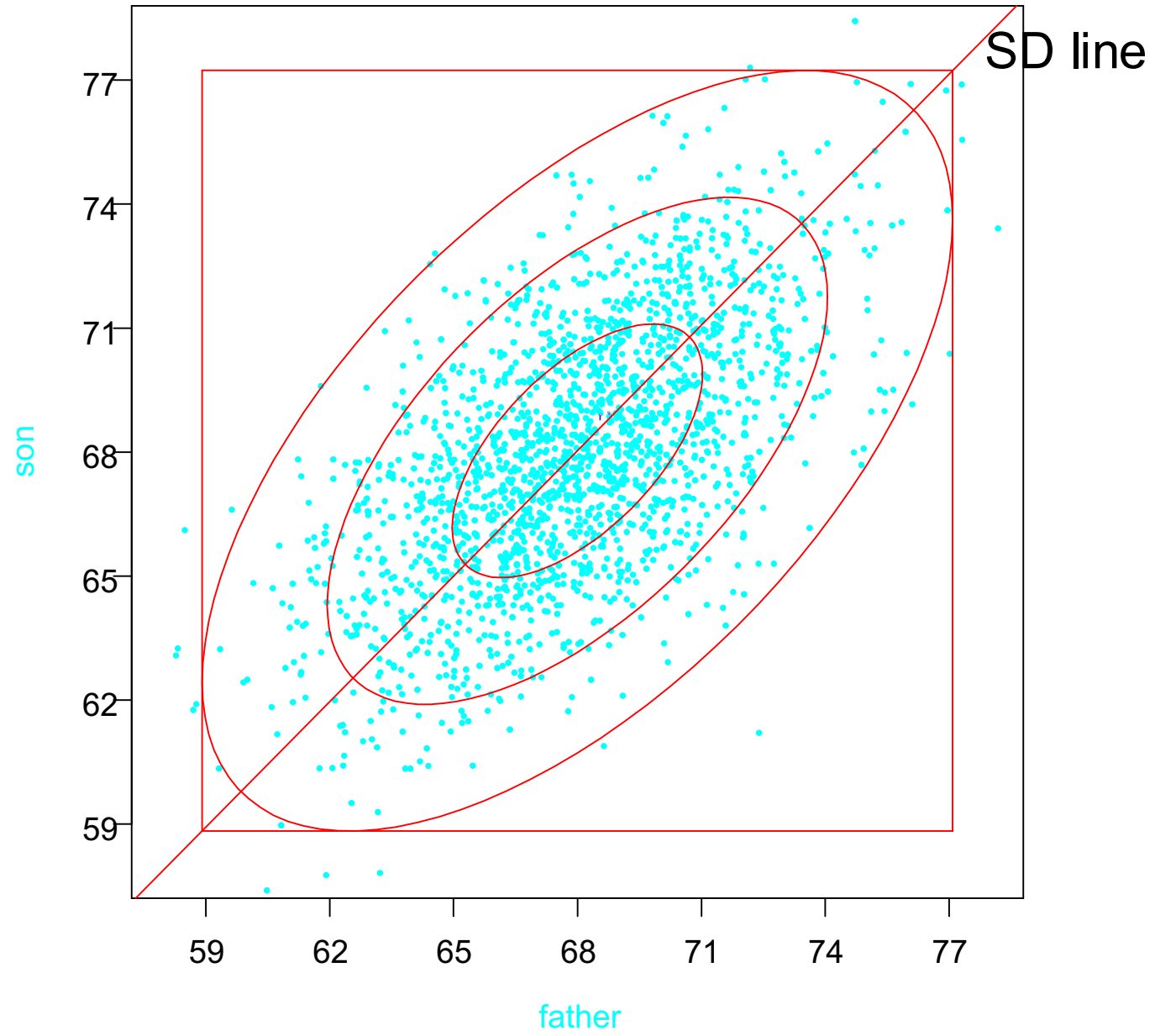
$$\hat{Z}_Y = 0 + r_{xy} Z_X$$

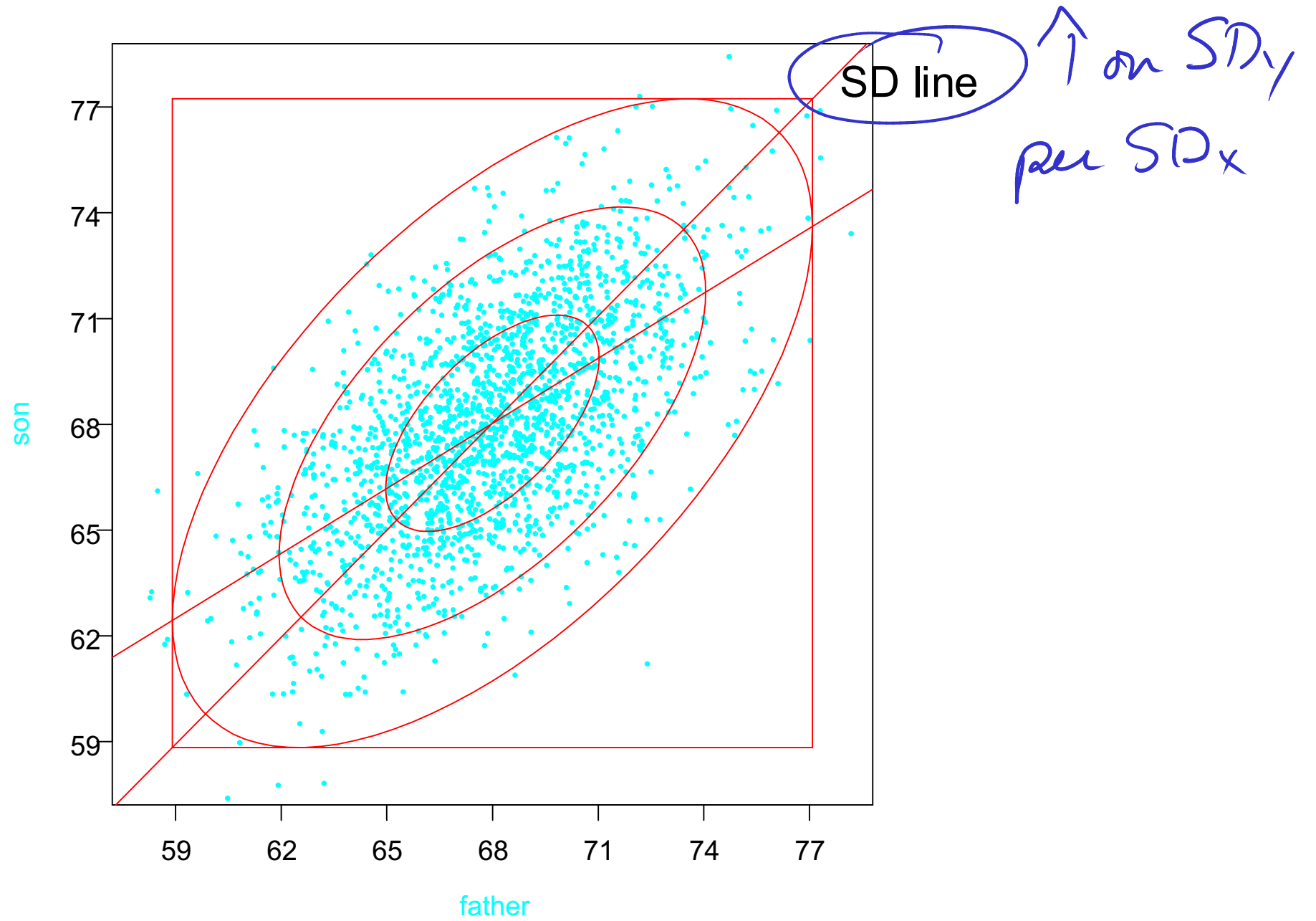
$$\hat{Z}_X = r_{xy} Z_Y$$

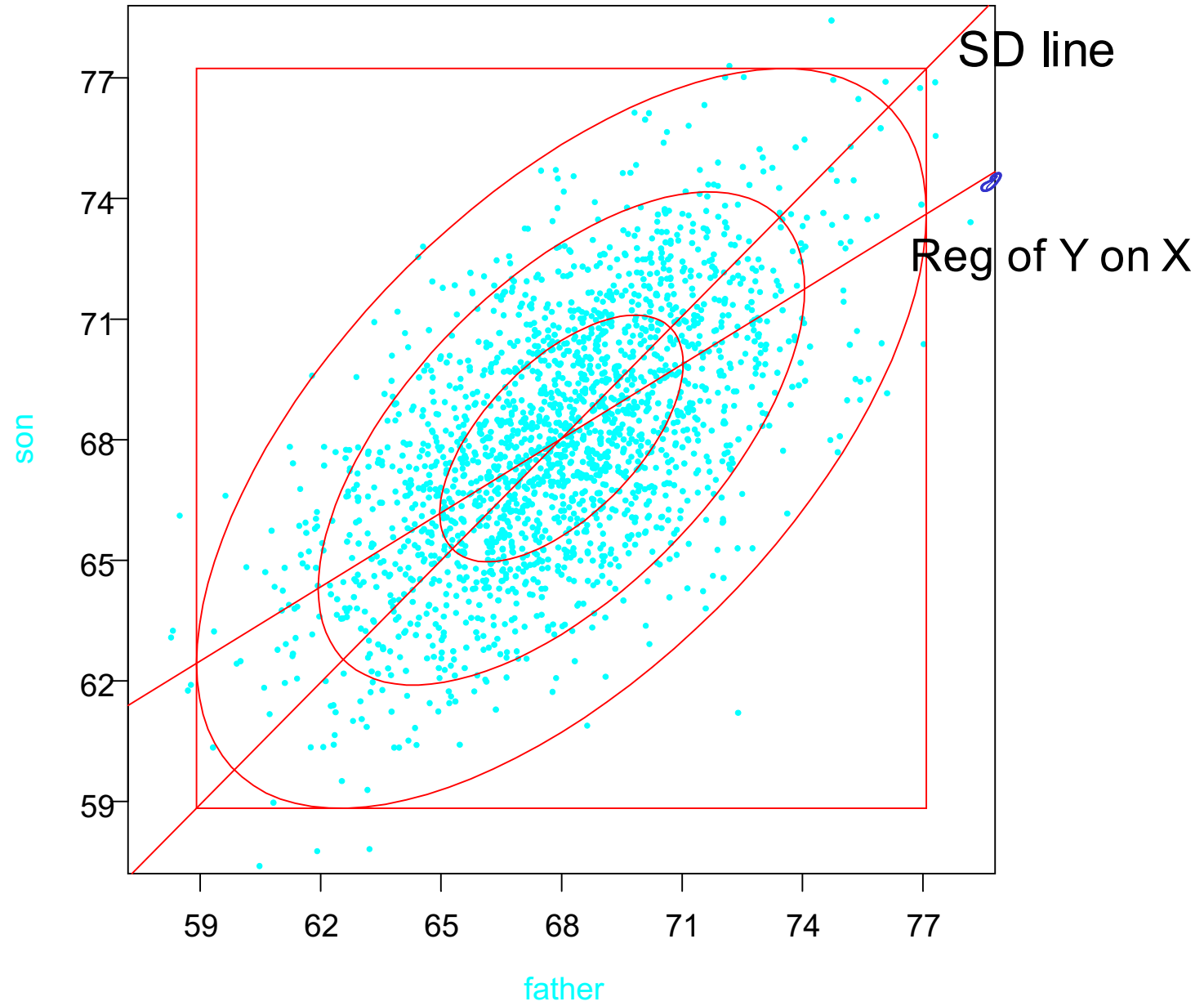


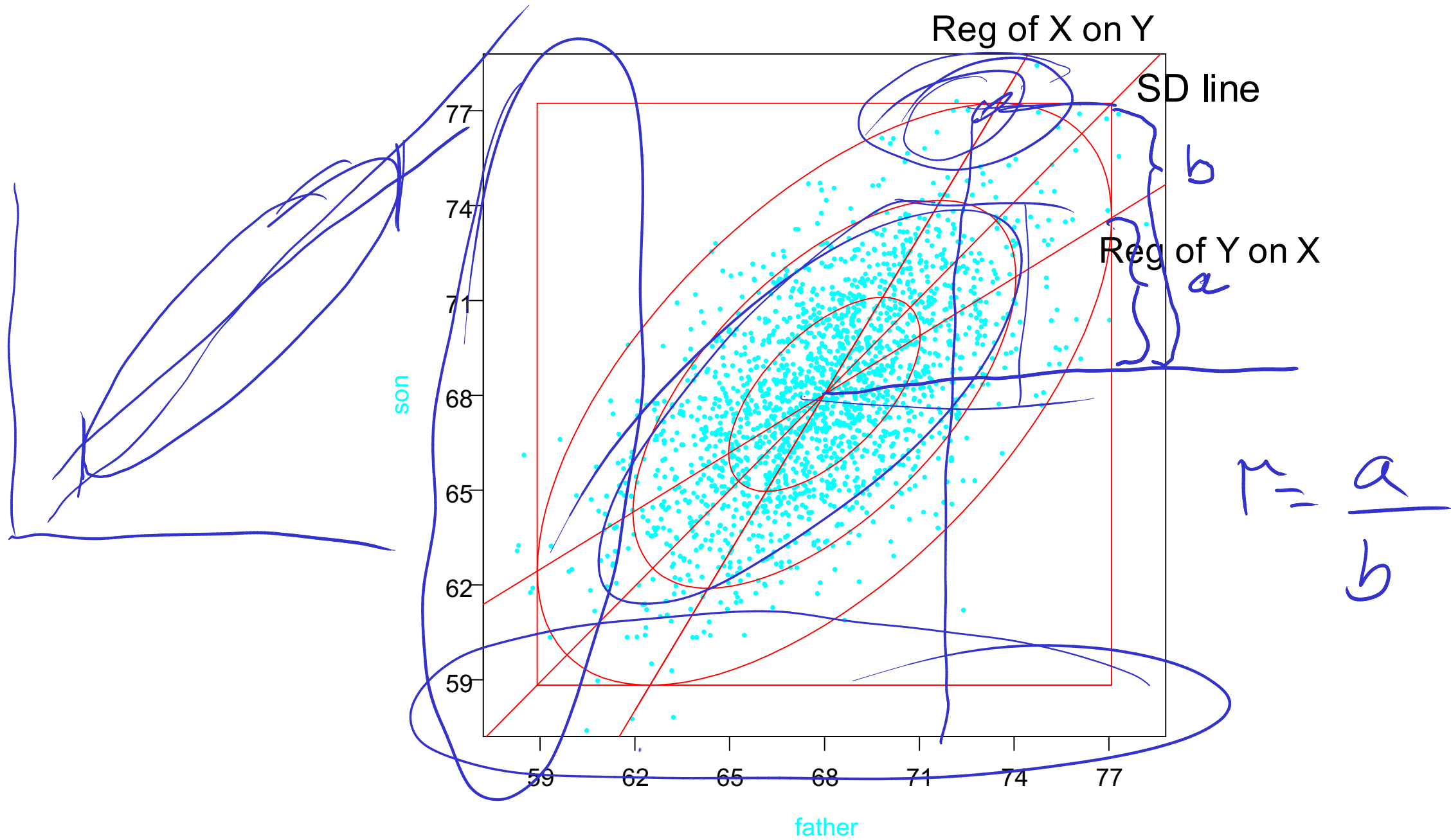












Regression Paradox:

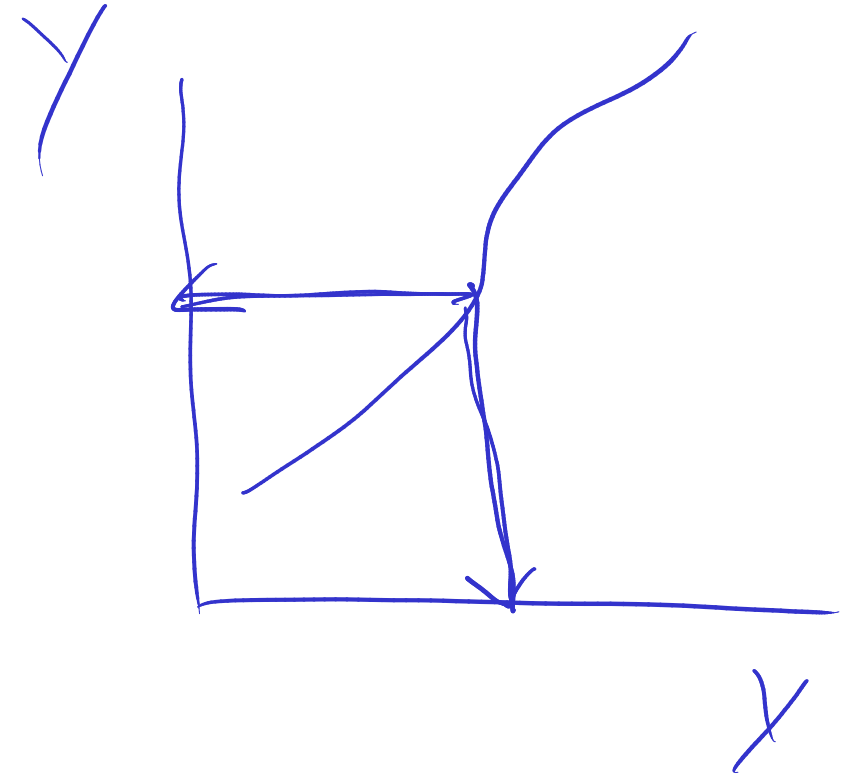
Mathematics:

$$f(X|Y) = f^{-1}(Y|X)$$

Statistics:

$$E(X|Y) \neq E^{-1}(Y|X)$$

unless X and Y are exact functions of each other.



In this example, the overall distribution of heights remains the same from generation to generation.

But, following individuals, it seems that heights are *regressing* to the overall mean.

Moreover, this works both forward and backward in time!

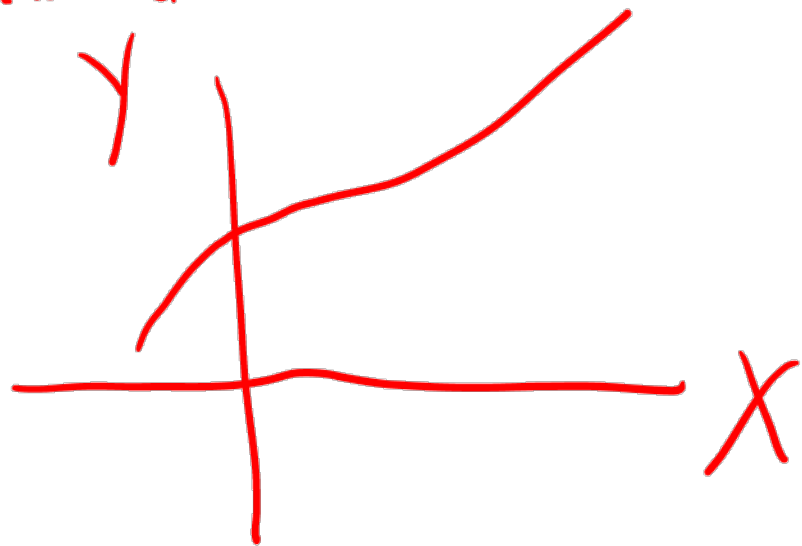
Regression Paradox

- The overall distribution of heights stays the same from generation to generation
- But when you follow individuals it looks like heights are being compressed towards the mean.
- Moreover, the same thing happens whether you go forward or backward in time.

Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

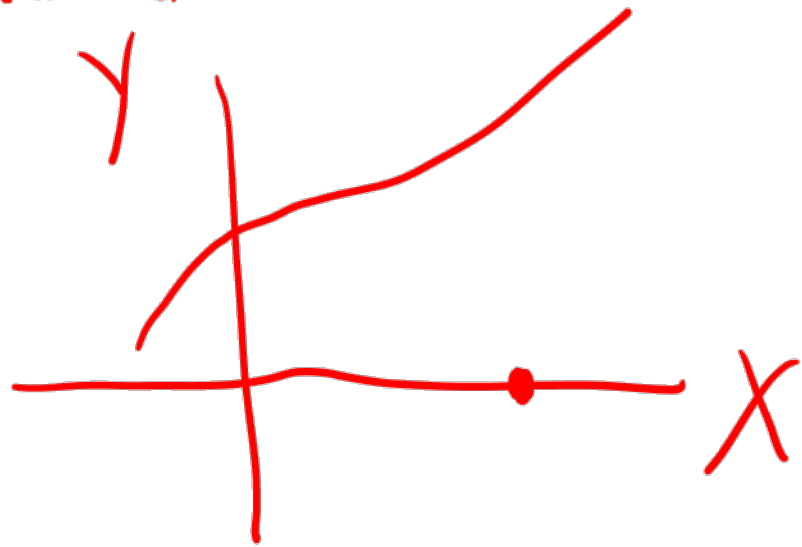
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

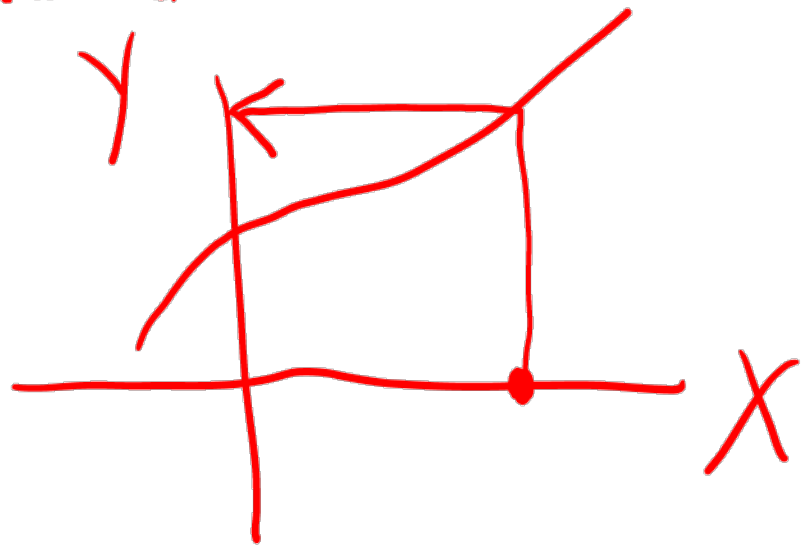
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

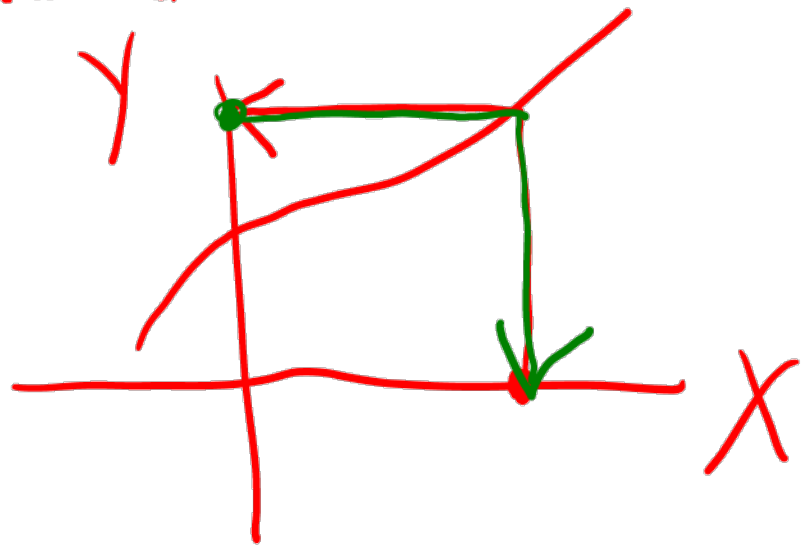
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

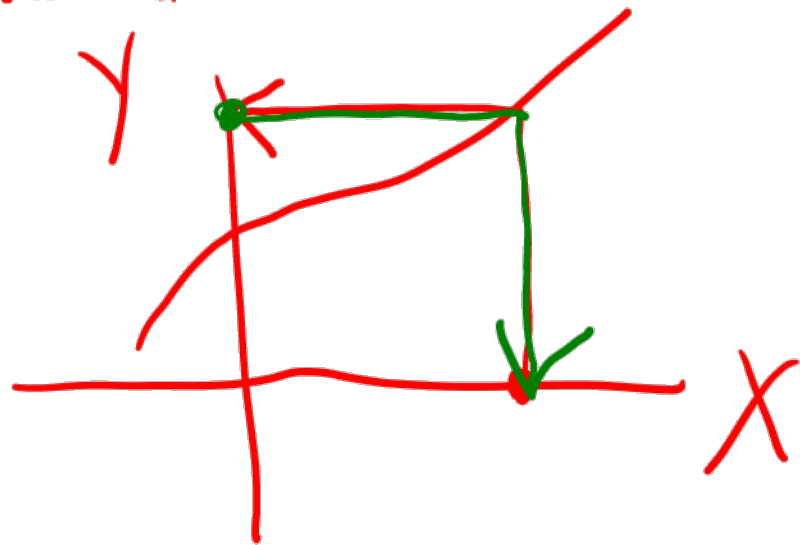
Mathematical inverse



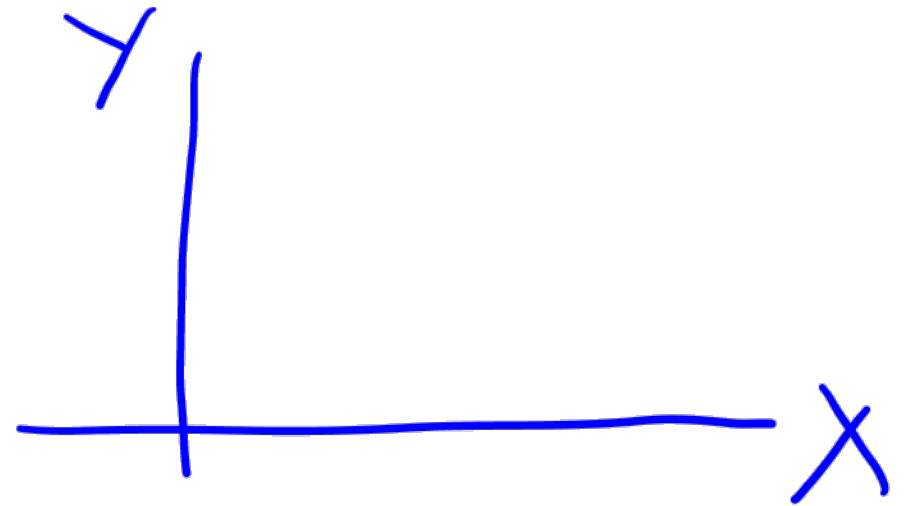
Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

Mathematical inverse



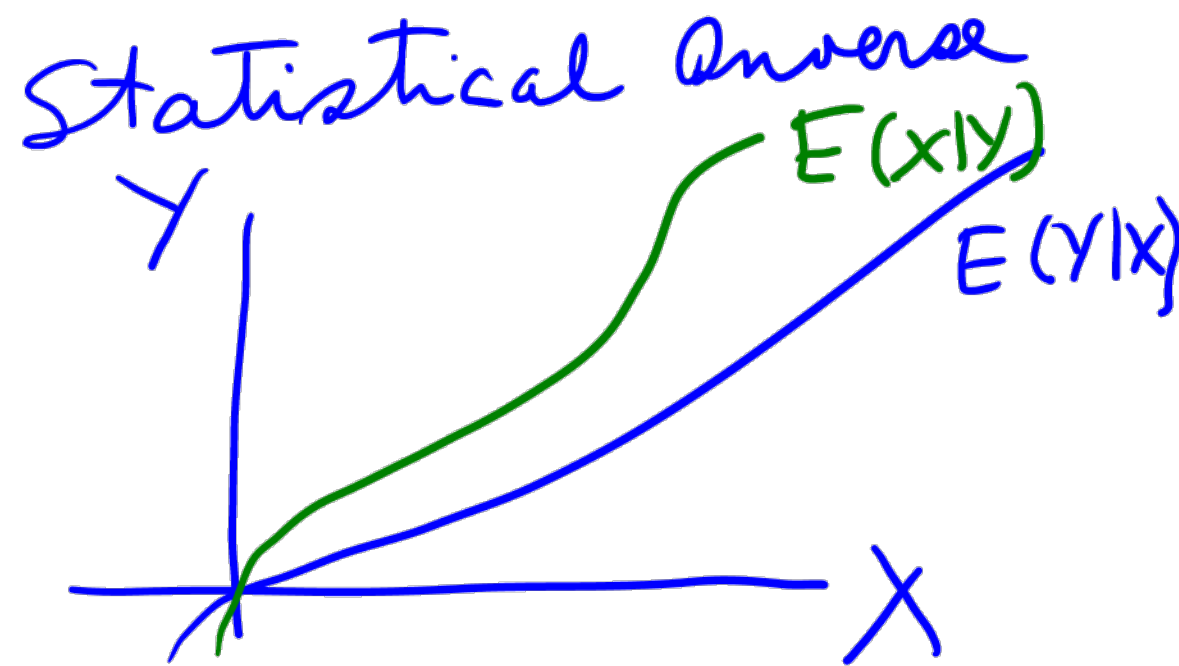
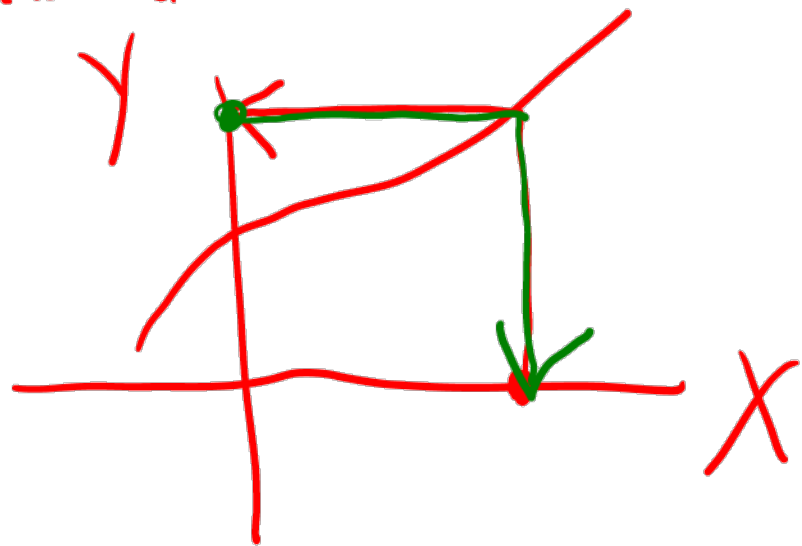
Statistical Inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

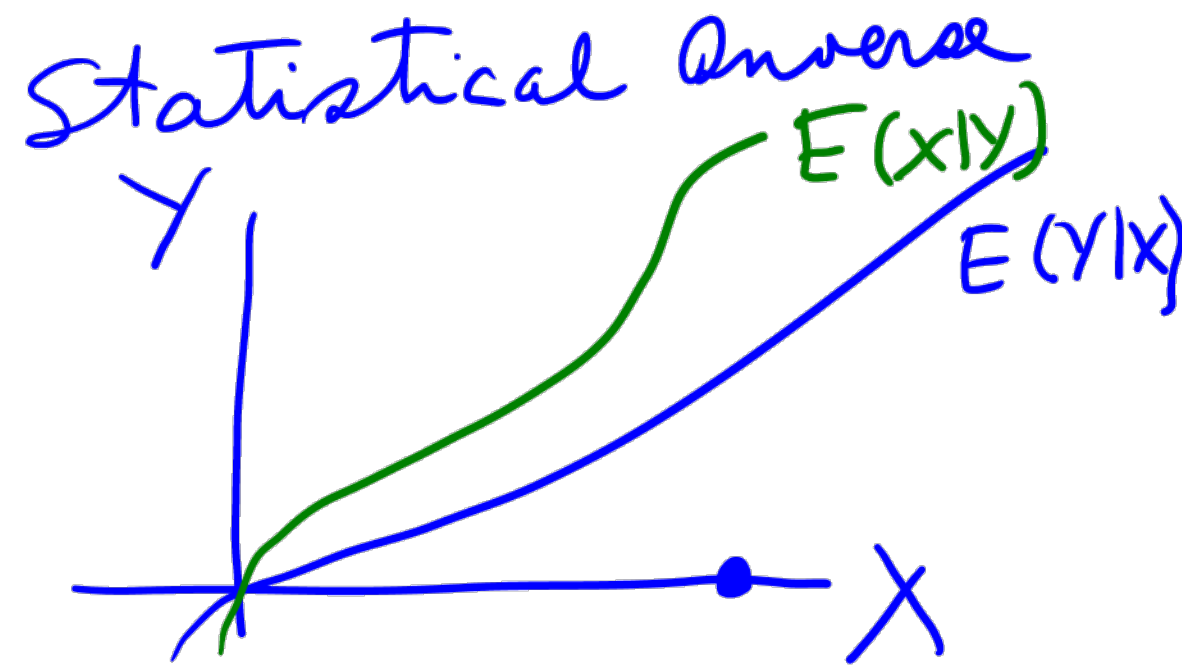
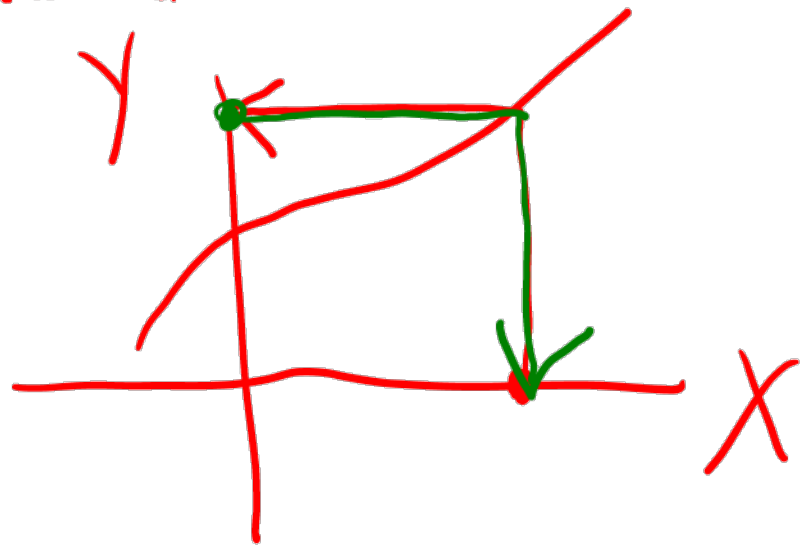
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

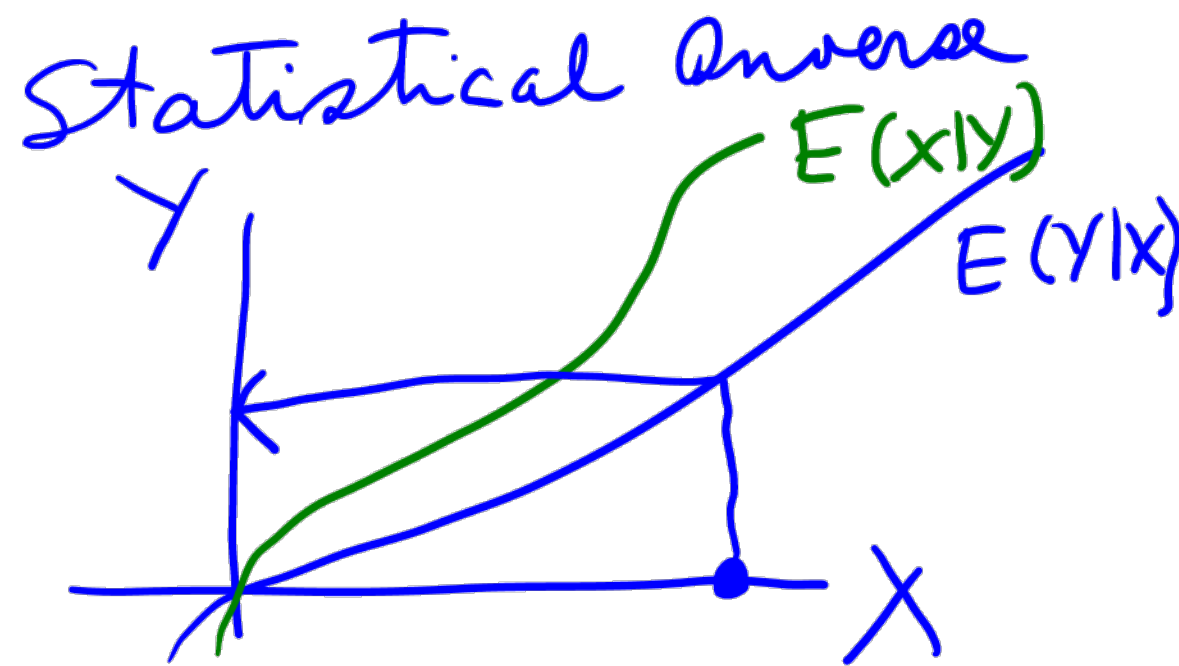
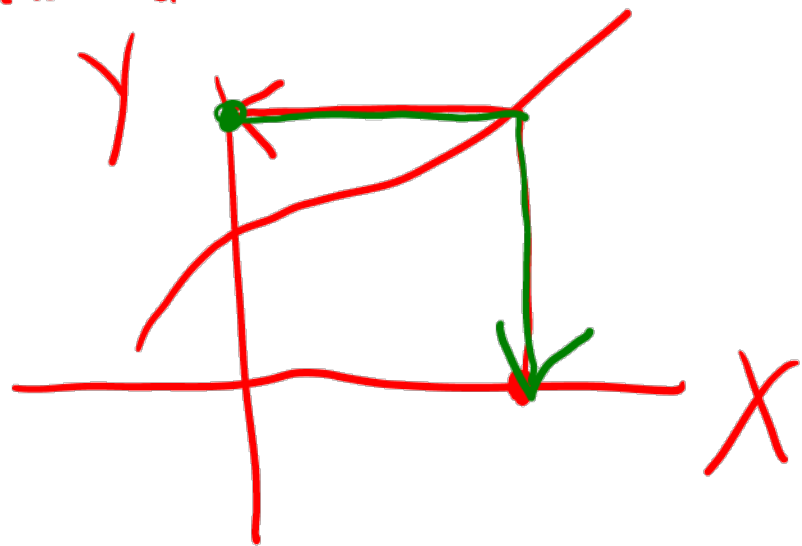
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

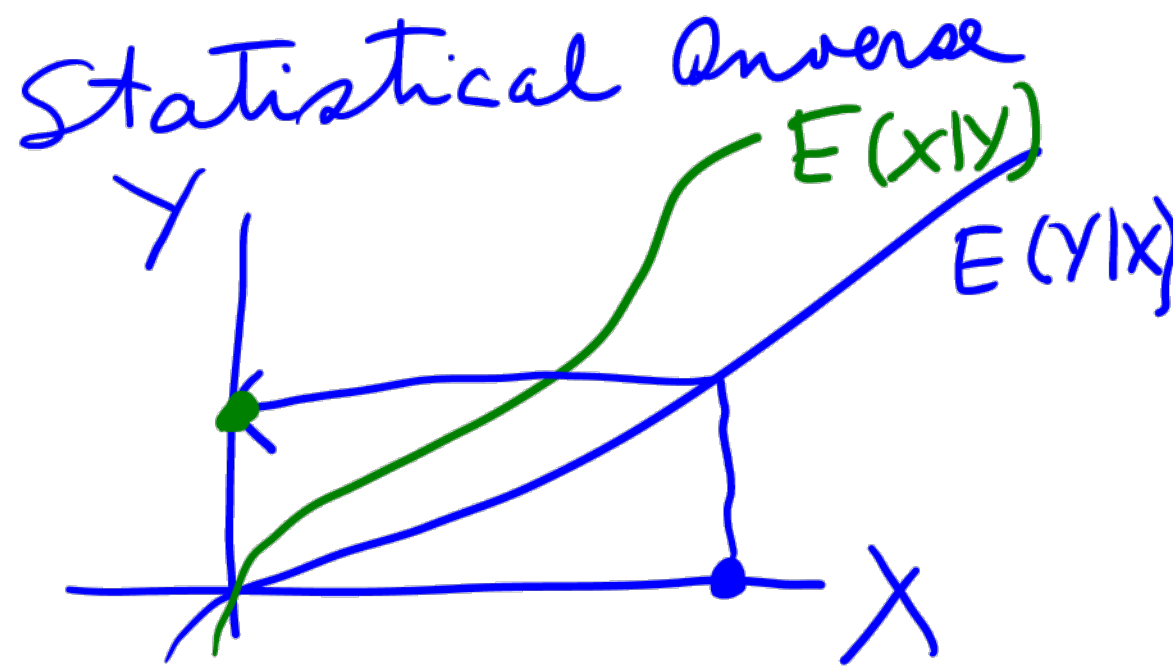
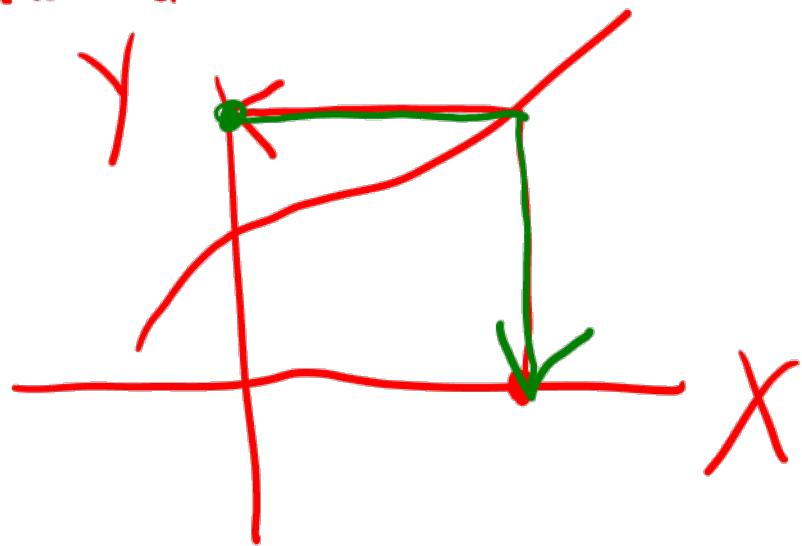
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

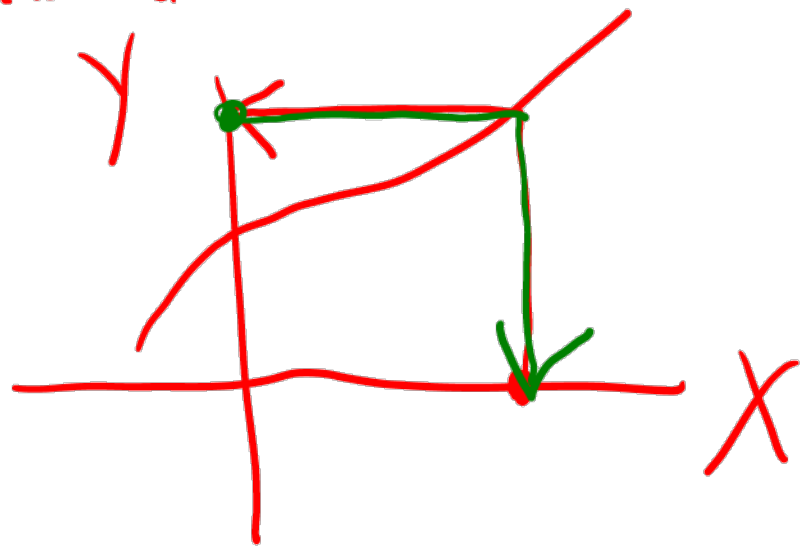
Mathematical inverse



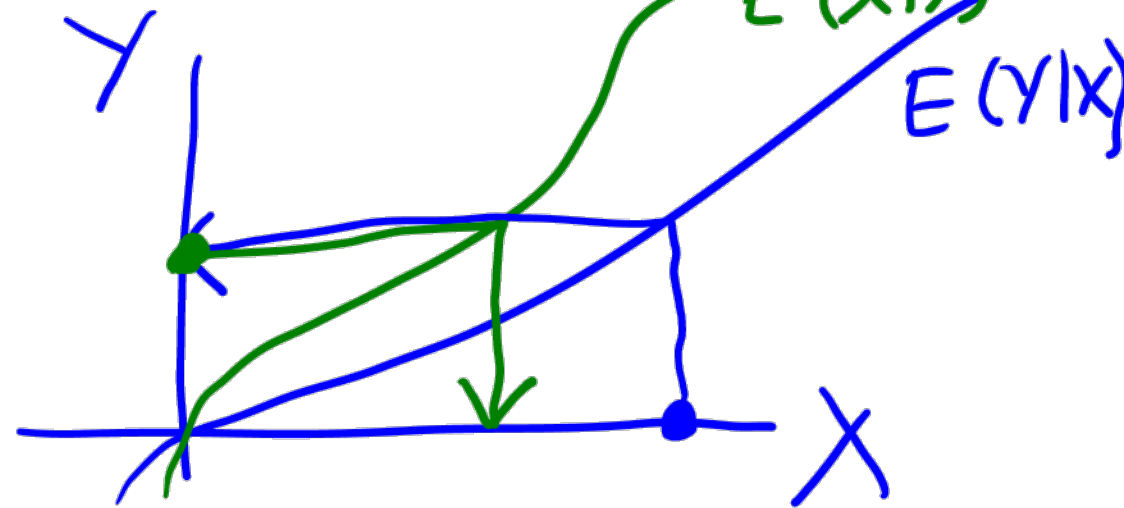
Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

Mathematical inverse



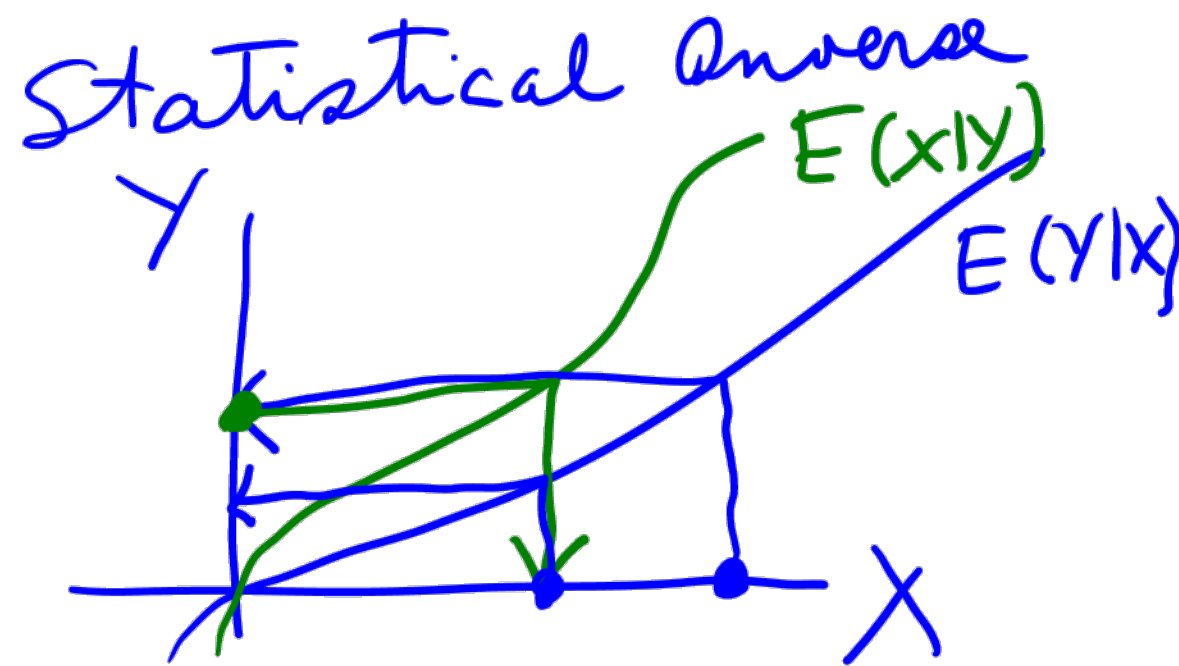
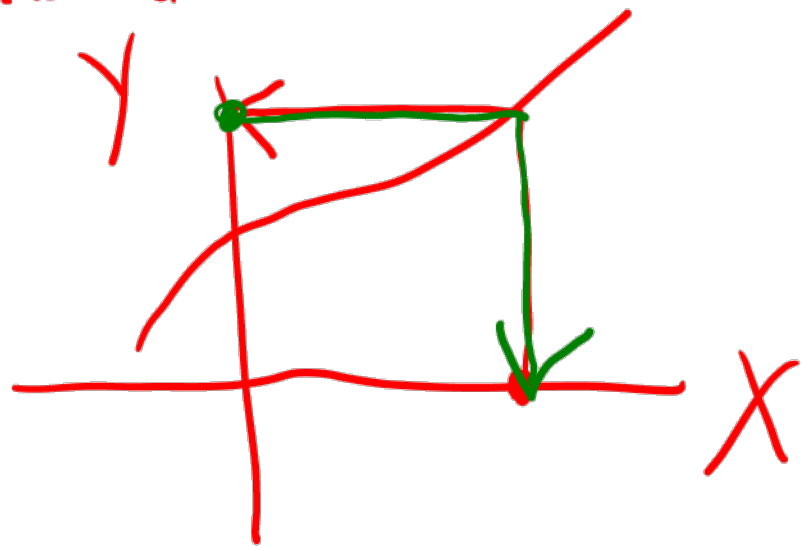
Statistical Inverse
 $E(X|Y)$
 $E(Y|X)$



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

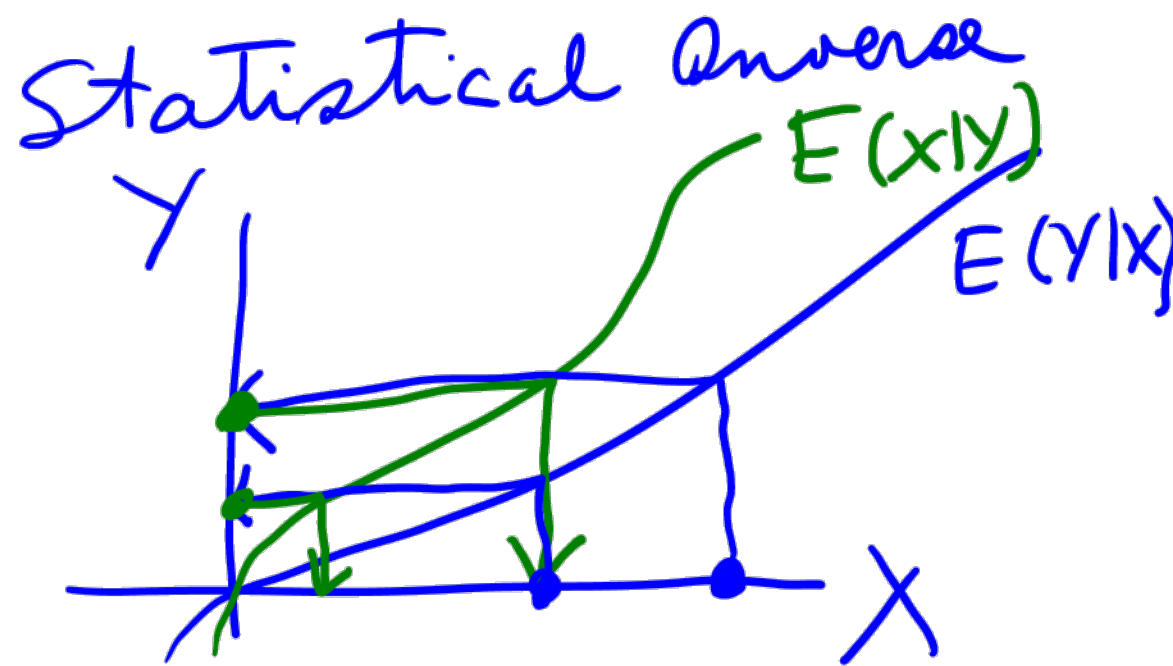
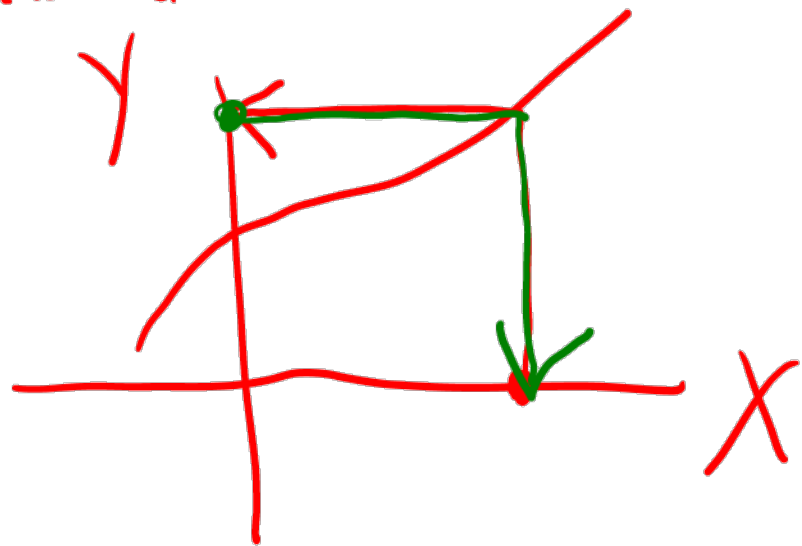
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

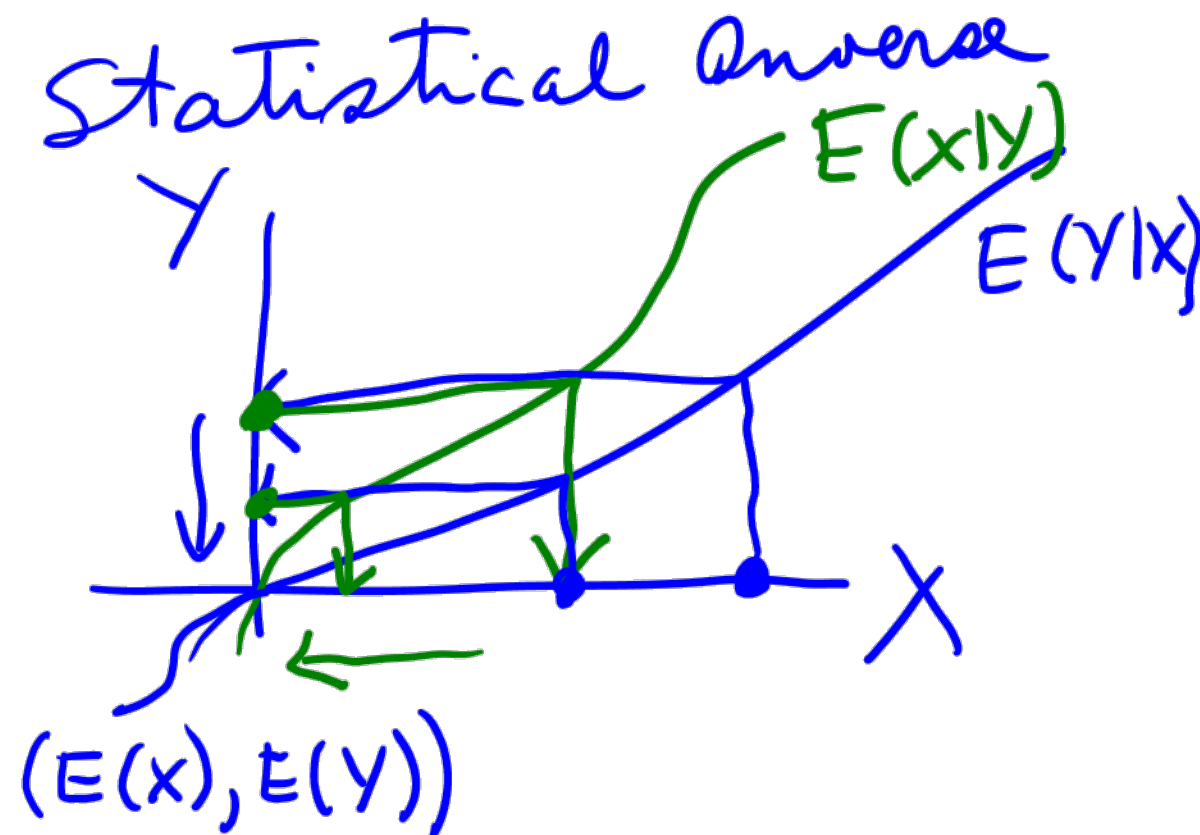
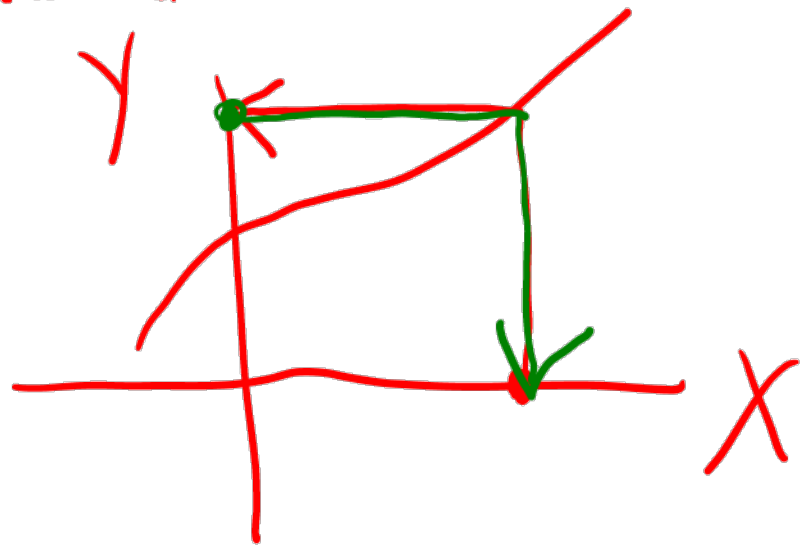
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

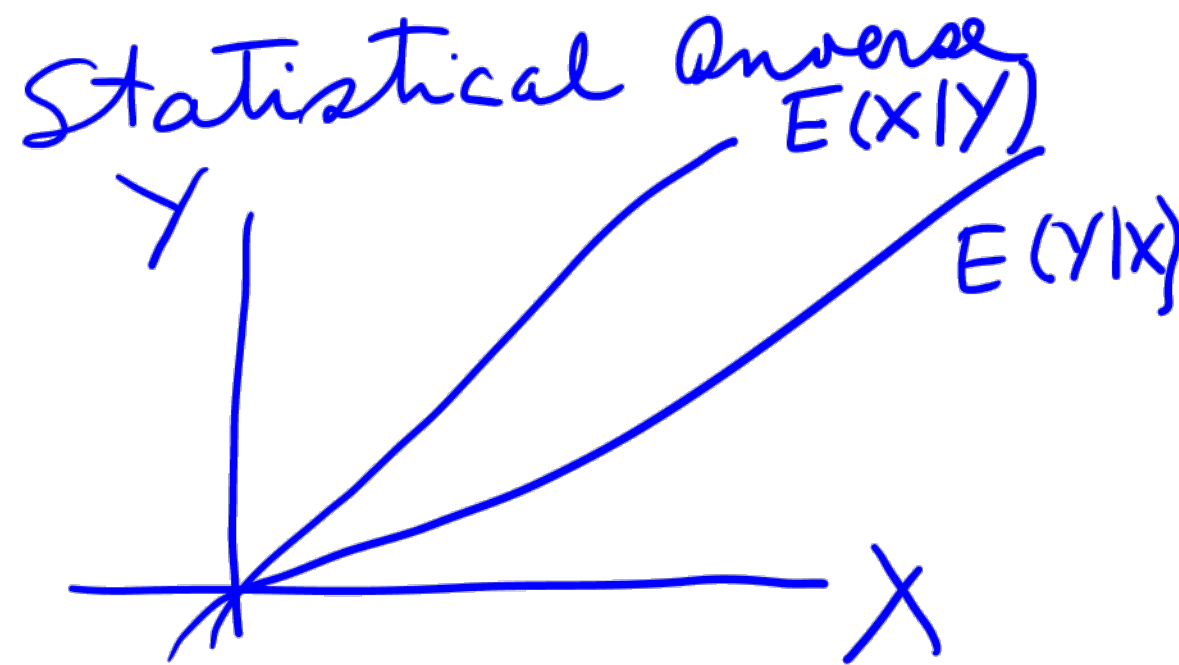
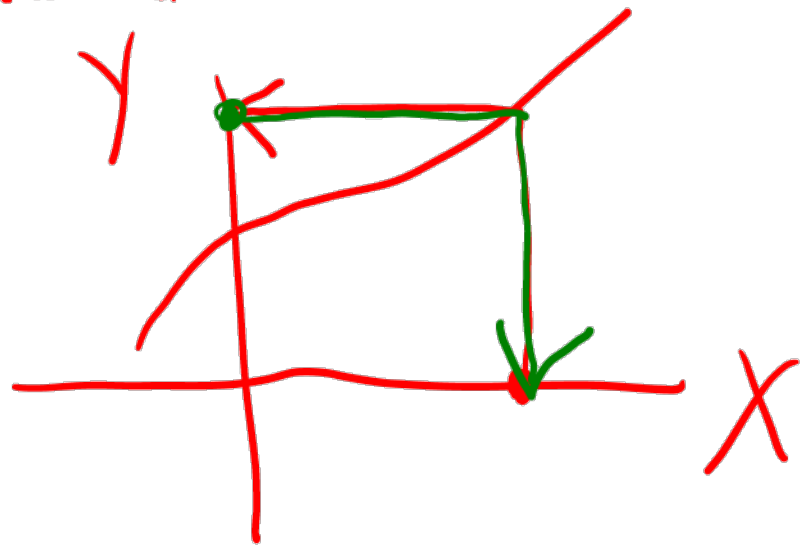
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

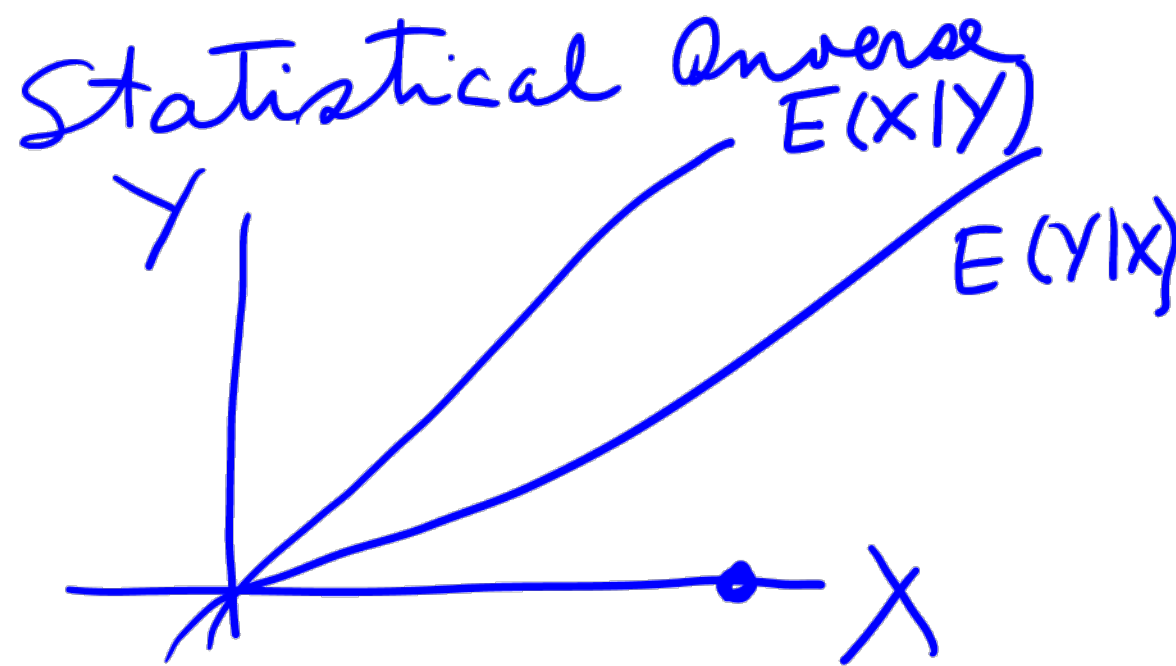
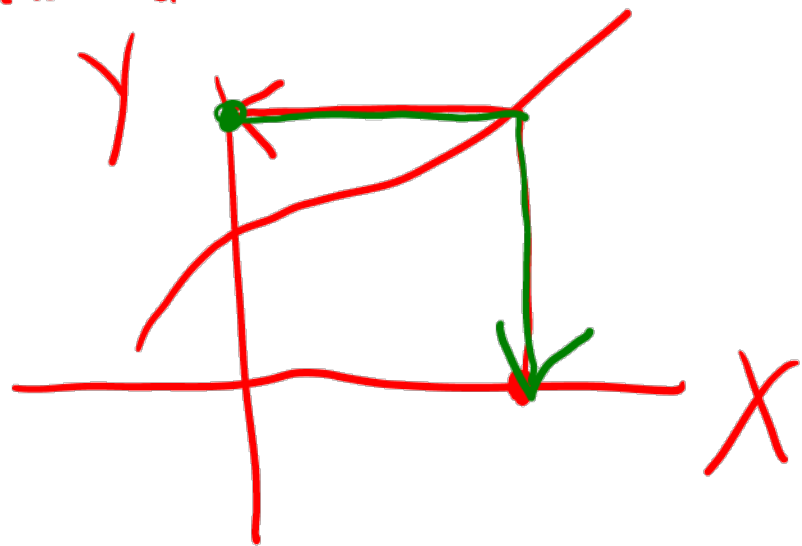
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

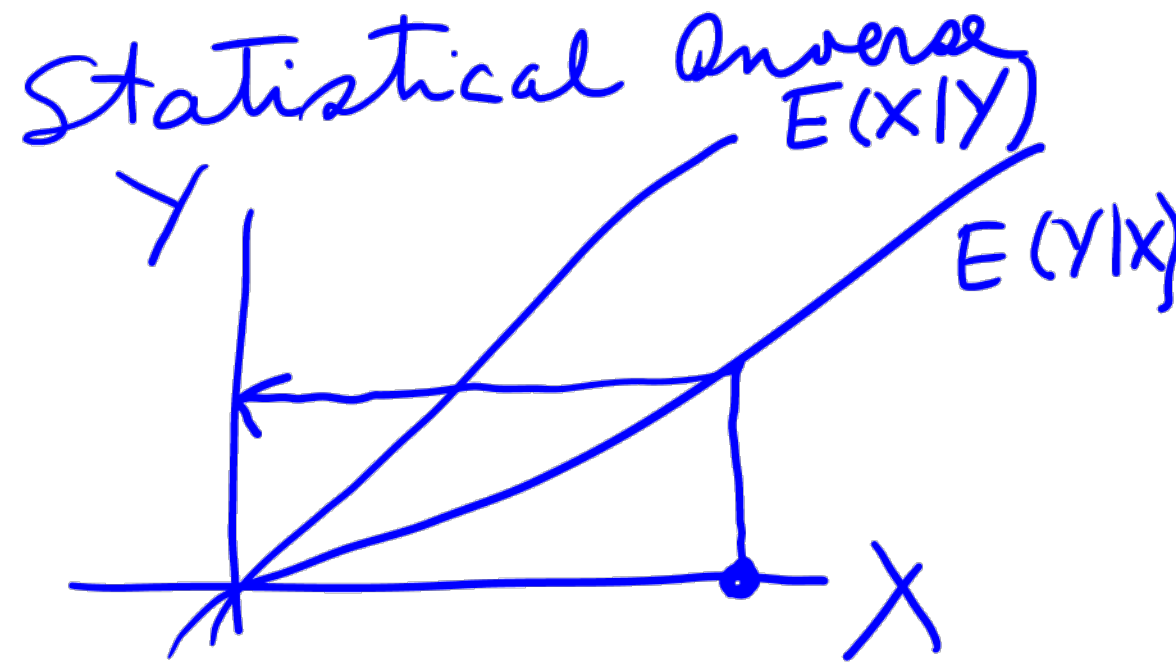
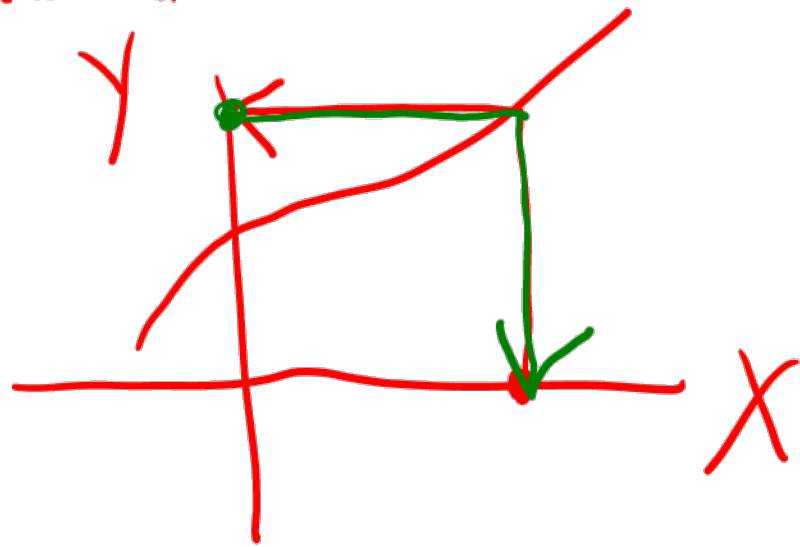
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

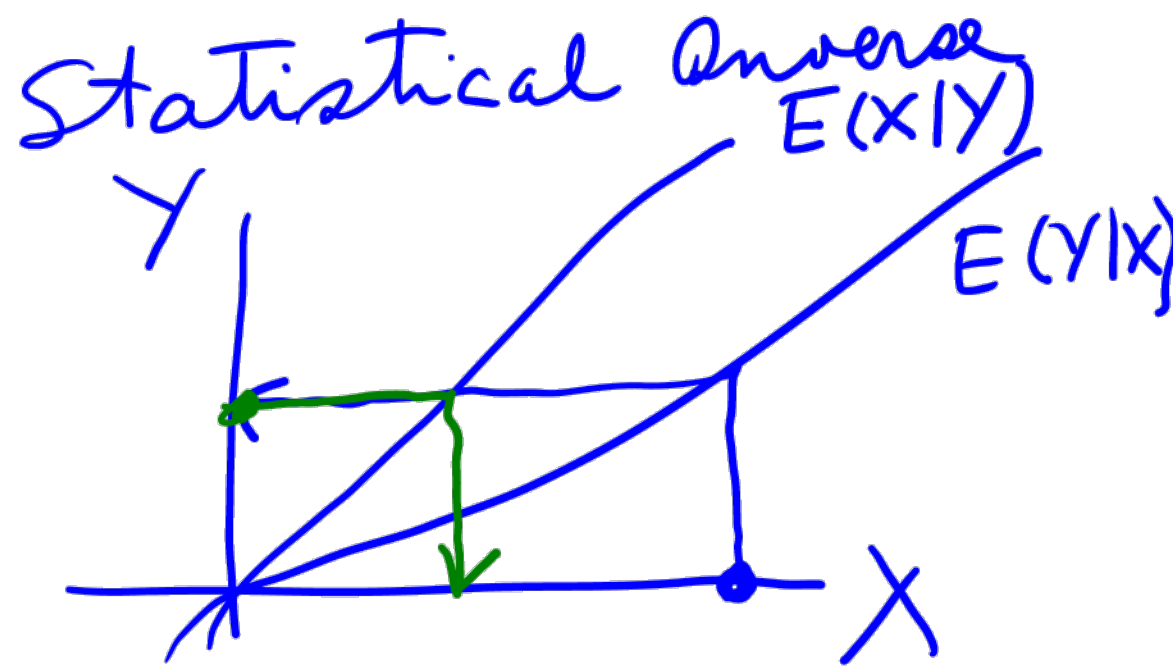
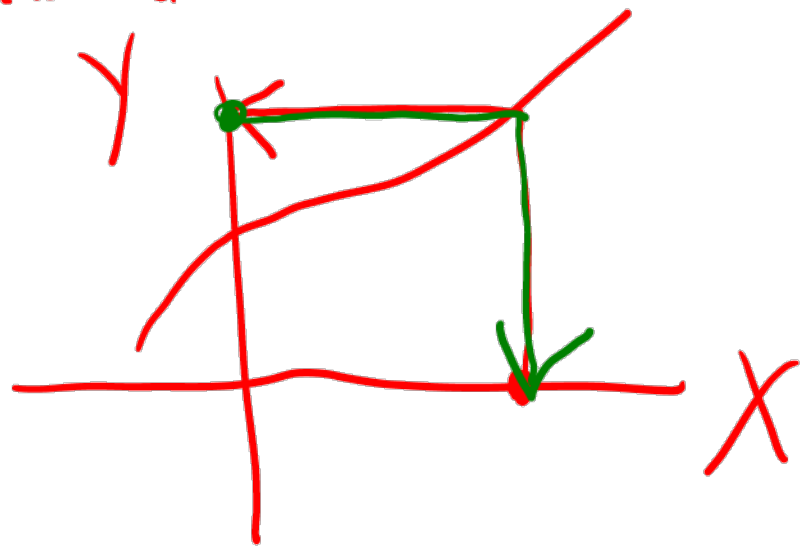
Mathematical inverse



Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

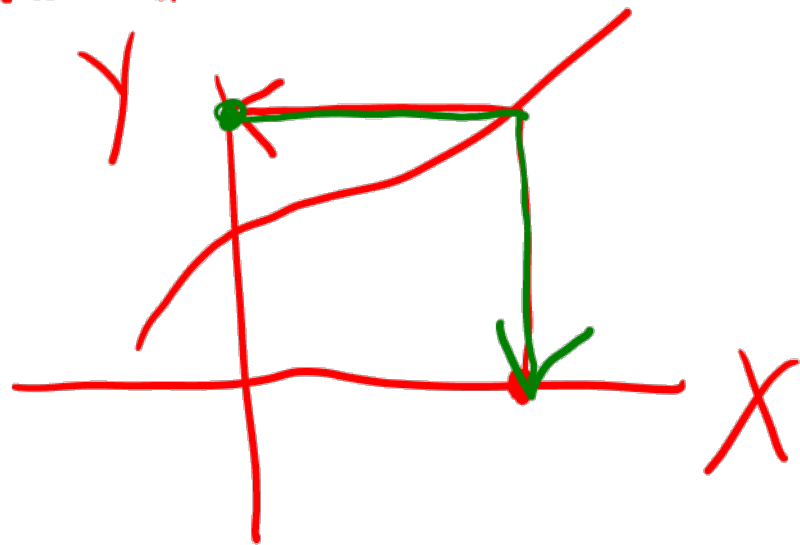
Mathematical inverse



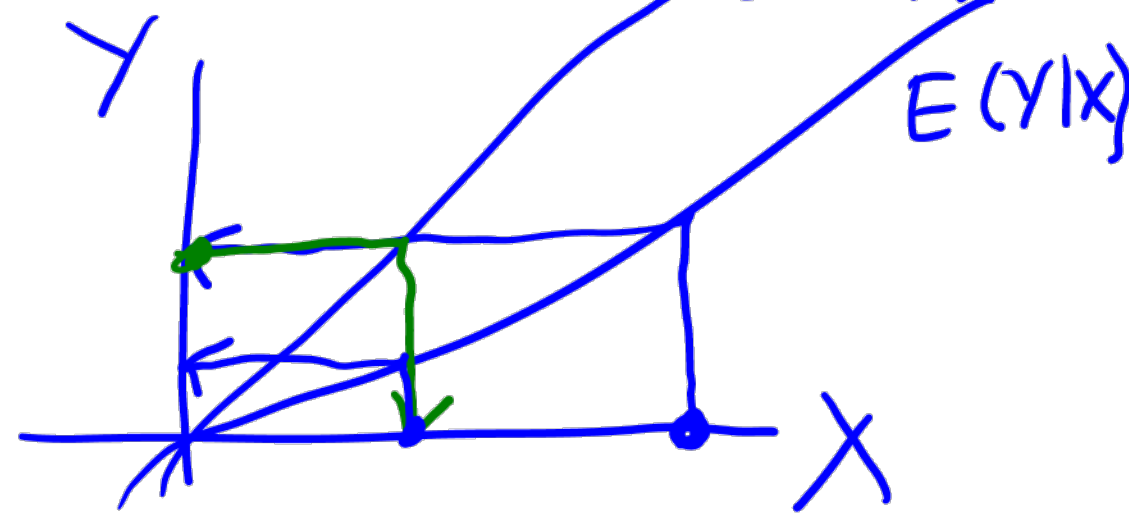
Simply

The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

Mathematical inverse



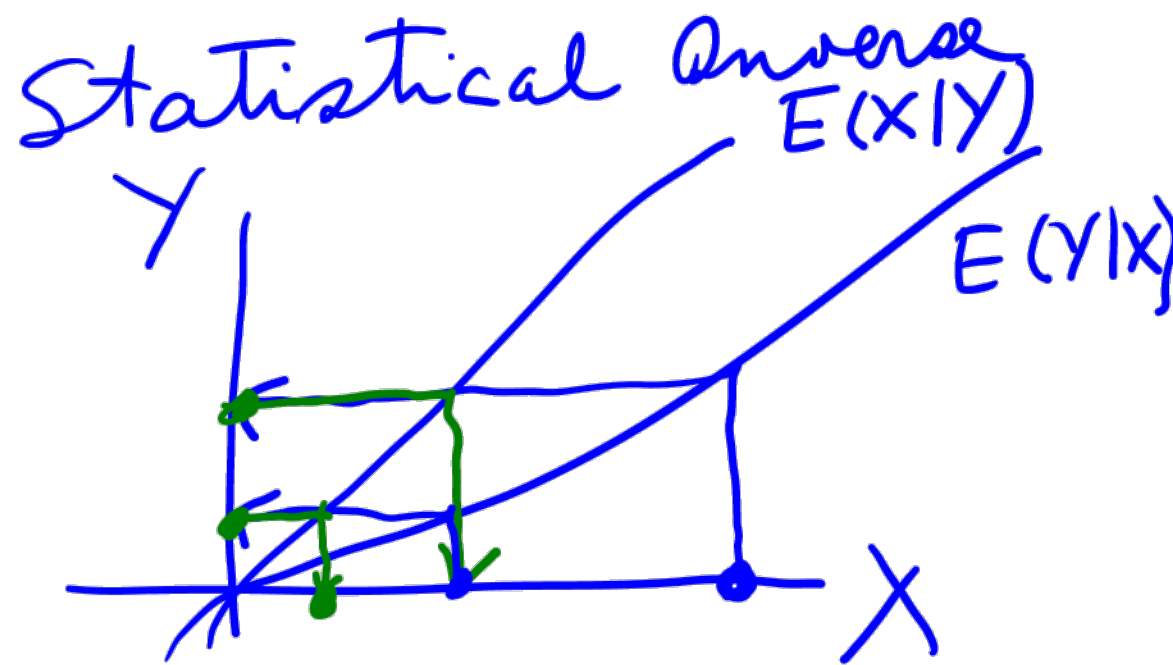
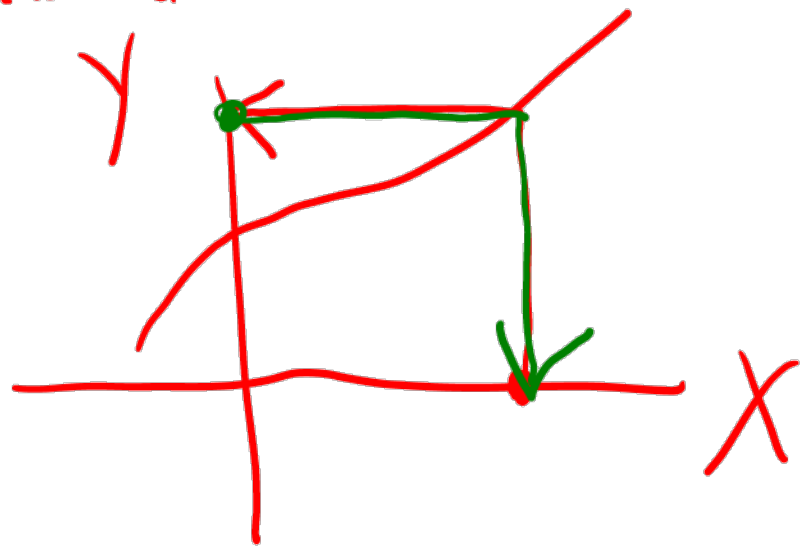
Statistical Inverse
 $E(X|Y)$



Simply

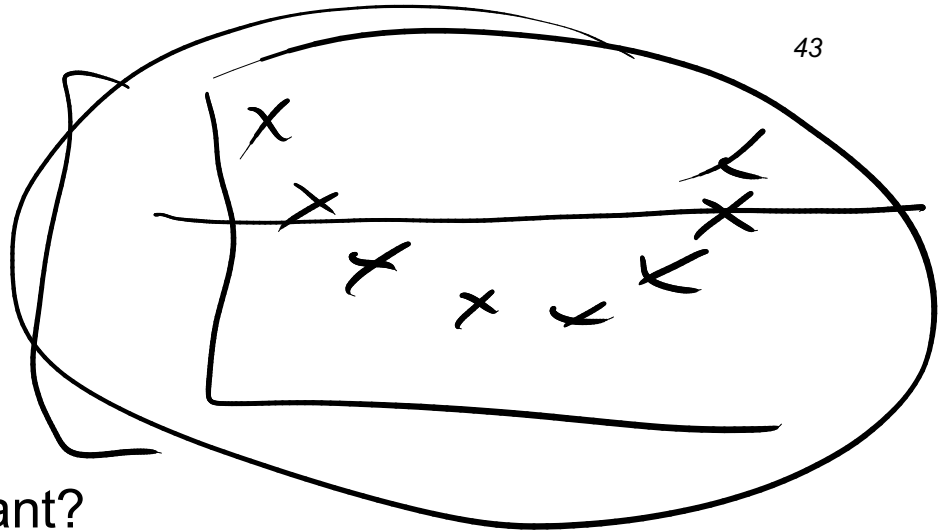
The regression of Y on X , $E(Y|X)$
is not the mathematical inverse of
the regression of X on Y , $E(X|Y)$

Mathematical inverse



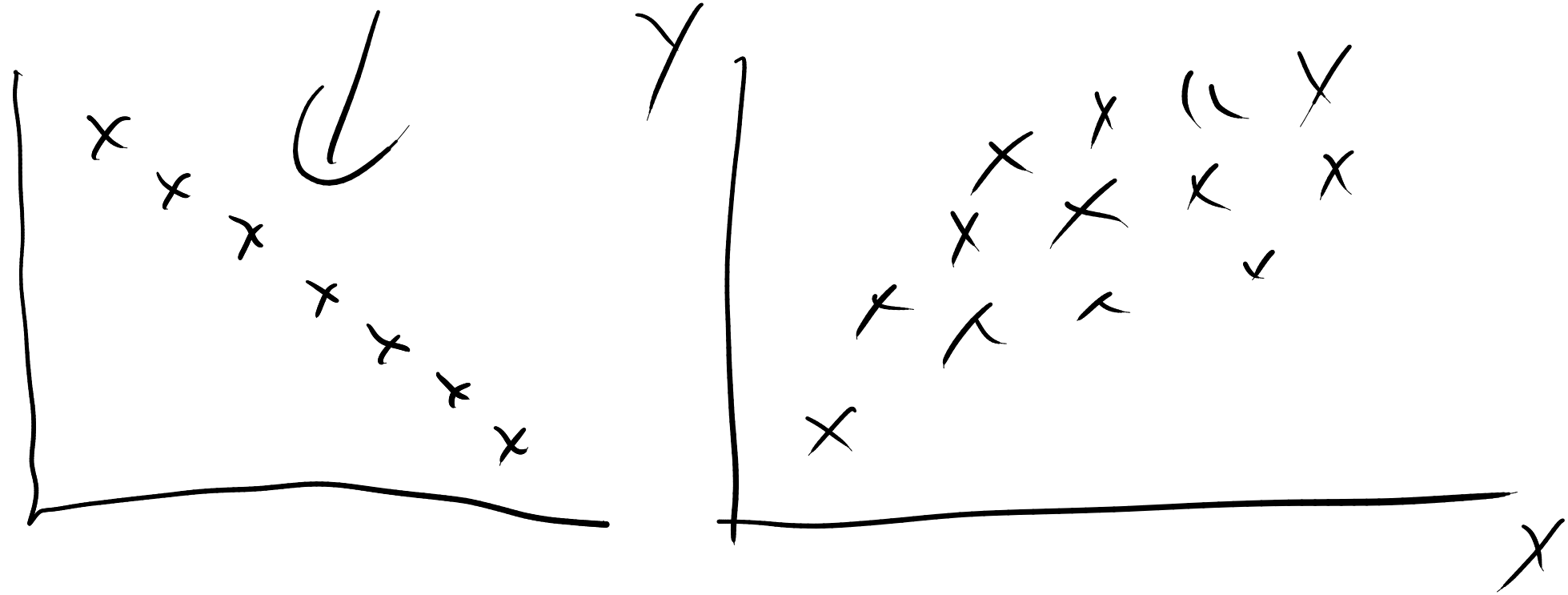
2 Eyeballing a scatterplot

correlation?



At a glance:

- What is the correlation?
- Is the relationship statistically significant?
- Approximate 95% confidence interval?



Some data:

1971 Canadian Occupational Prestige Data

- [1] Occupational title
- [2] Average education of incumbents, years
- [3] Average income of incumbents, dollars
- [4] Percent of incumbents who are women
- [5] Pineo-Porter prestige score for occupation
- [6] Canadian Census occupational code
- [7] Type of occupation
 - prof = professional and technical
 - wc = white collar
 - bc = blue collar
 - ? = missing (not classified)

Source: Census of Canada, 1971, in Fox (1997)

	Education	Income	PercFem	Prestige	Code	Type
GOV_ADMINISTRATORS	13.11	12351	11.16	68.8	1113	prof
GENERAL MANAGERS	12.26	25879	4.02	69.1	1130	prof
ACCOUNTANTS	12.77	9271	15.70	63.4	1171	prof
PURCHASING_OFFICERS	11.42	8865	9.11	56.8	1175	prof
CHEMISTS	14.62	8403	11.68	73.5	2111	prof
PHYSICISTS	15.64	11030	5.13	77.6	2113	prof
BIOLOGISTS	15.09	8258	25.65	72.6	2133	prof
ARCHITECTS	15.44	14163	2.69	78.1	2141	prof
CIVIL_ENGINEERS	14.52	11377	1.03	73.1	2143	prof
MINING_ENGINEERS	14.64	11023	0.94	68.8	2153	prof

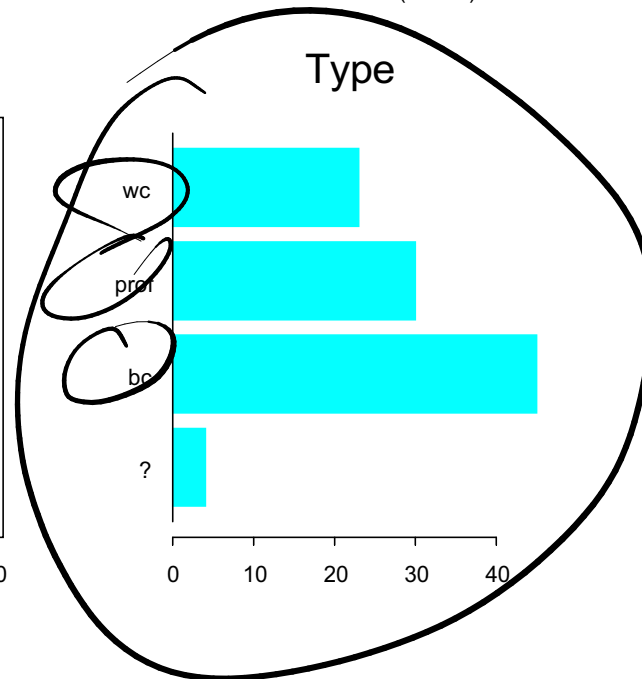
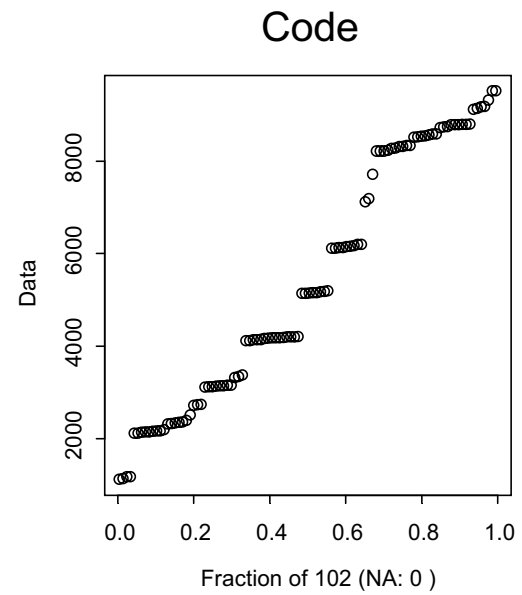
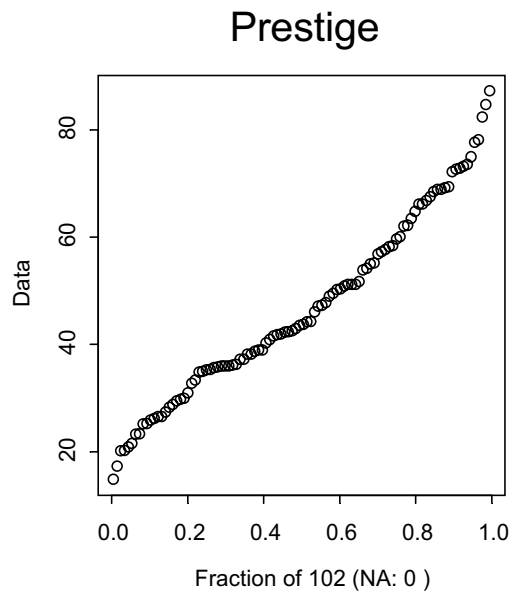
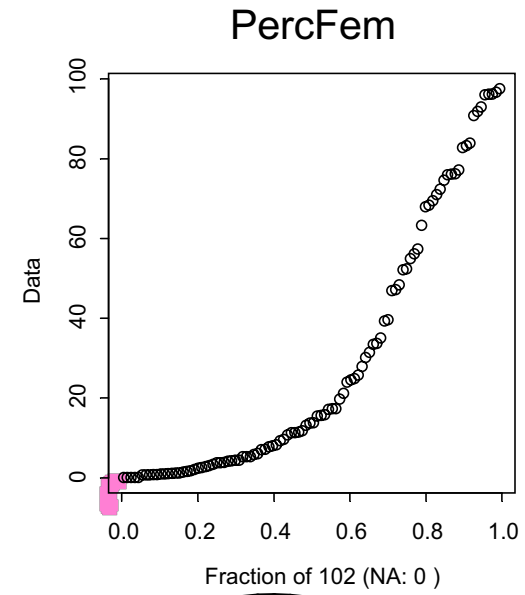
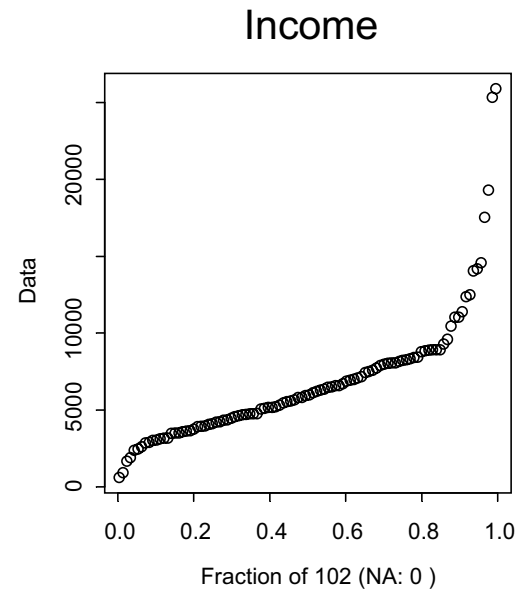
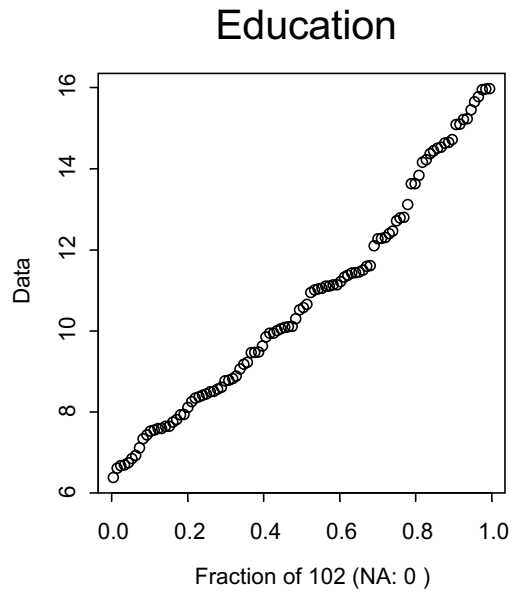
...

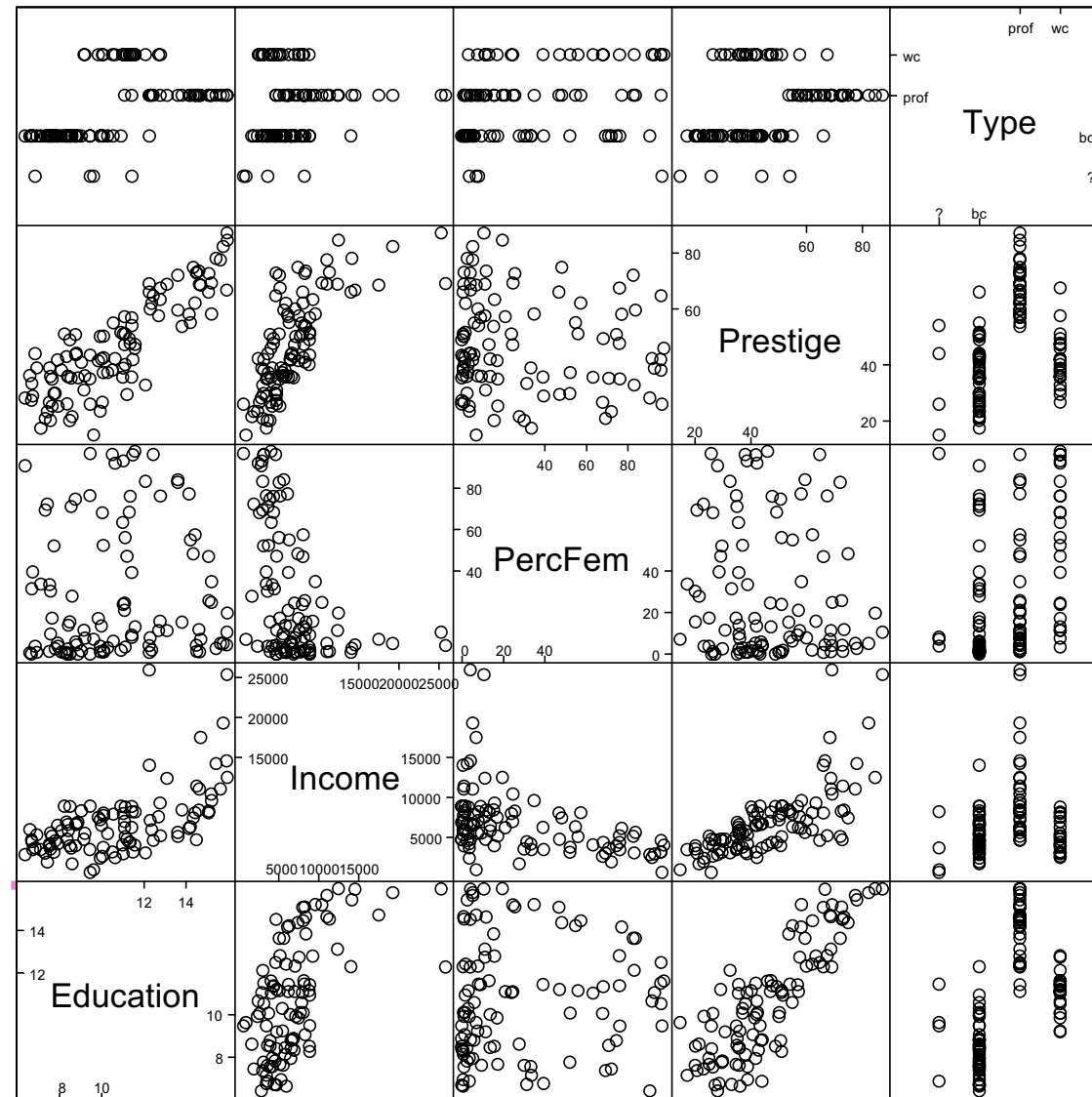
102 occupations

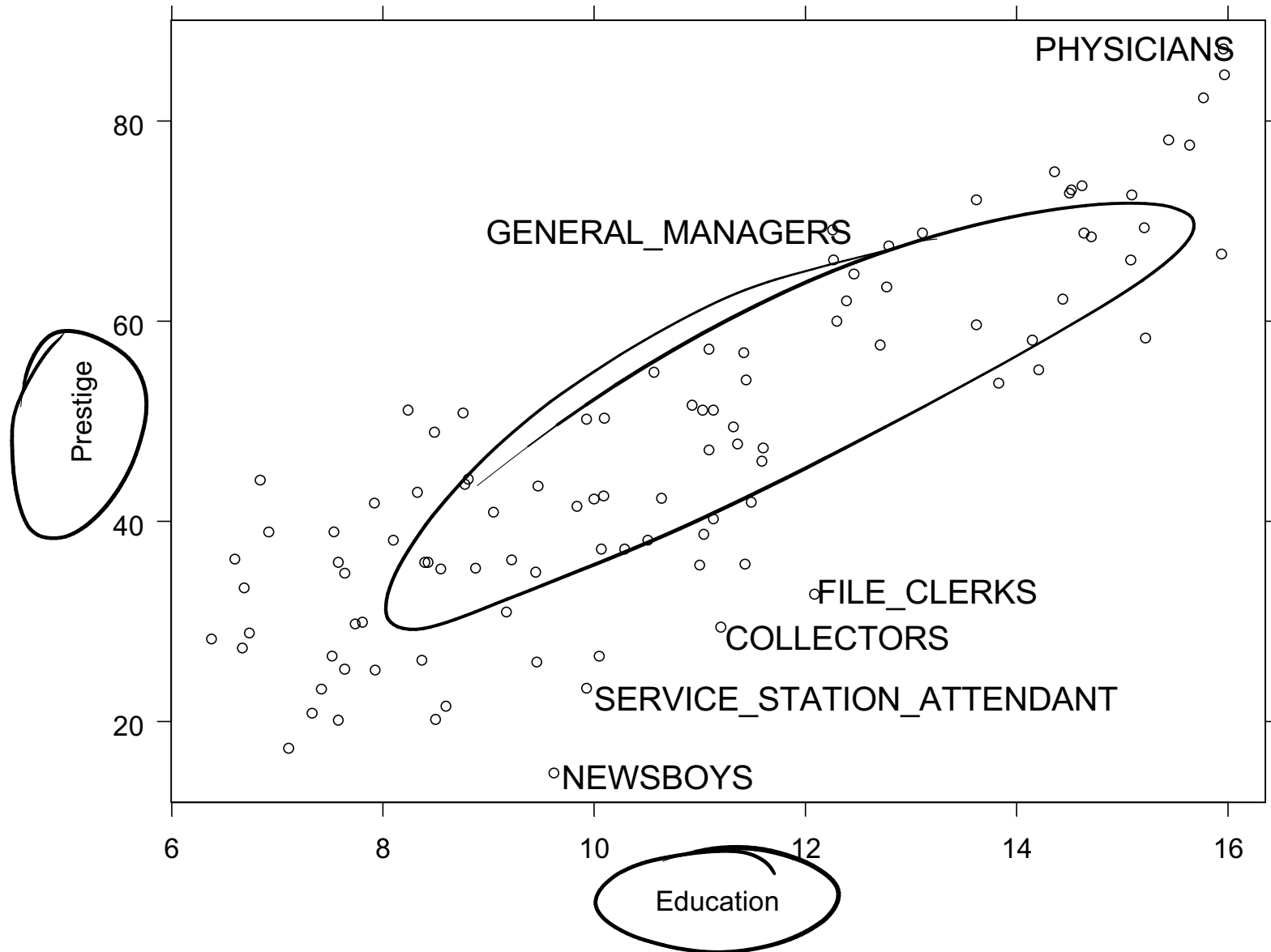
Summary:

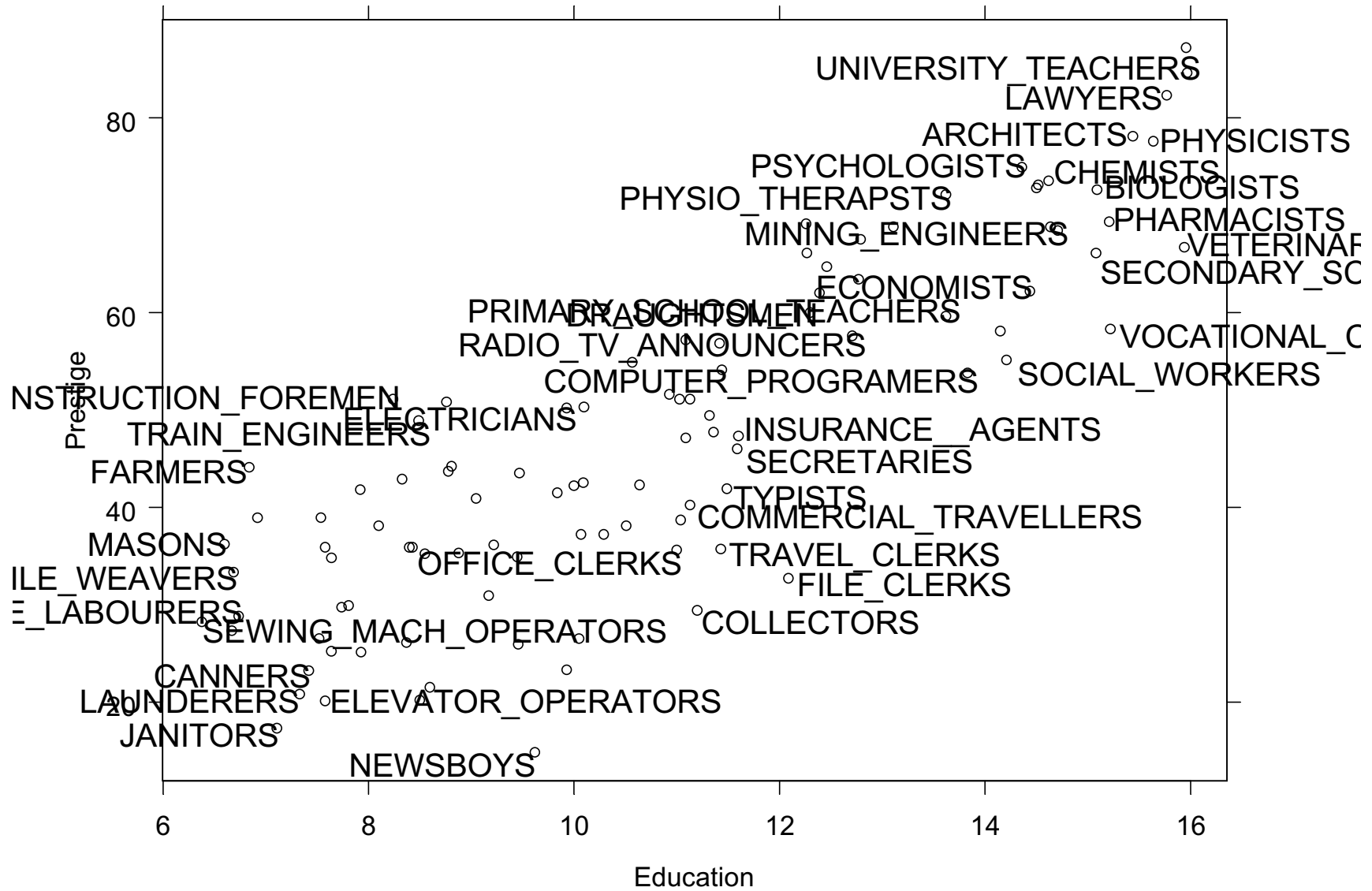
Education	Income	PercFem
Min.: 6.380	Min.: 611	Min.: 0.000
1st Qu.: 8.445	1st Qu.: 4106	1st Qu.: 3.592
Median:10.540	Median: 5930	Median:13.600
Mean:10.740	Mean: 6798	Mean:28.980
3rd Qu.:12.650	3rd Qu.: 8187	3rd Qu.:52.200
Max.:15.970	Max.:25880	Max.:97.510

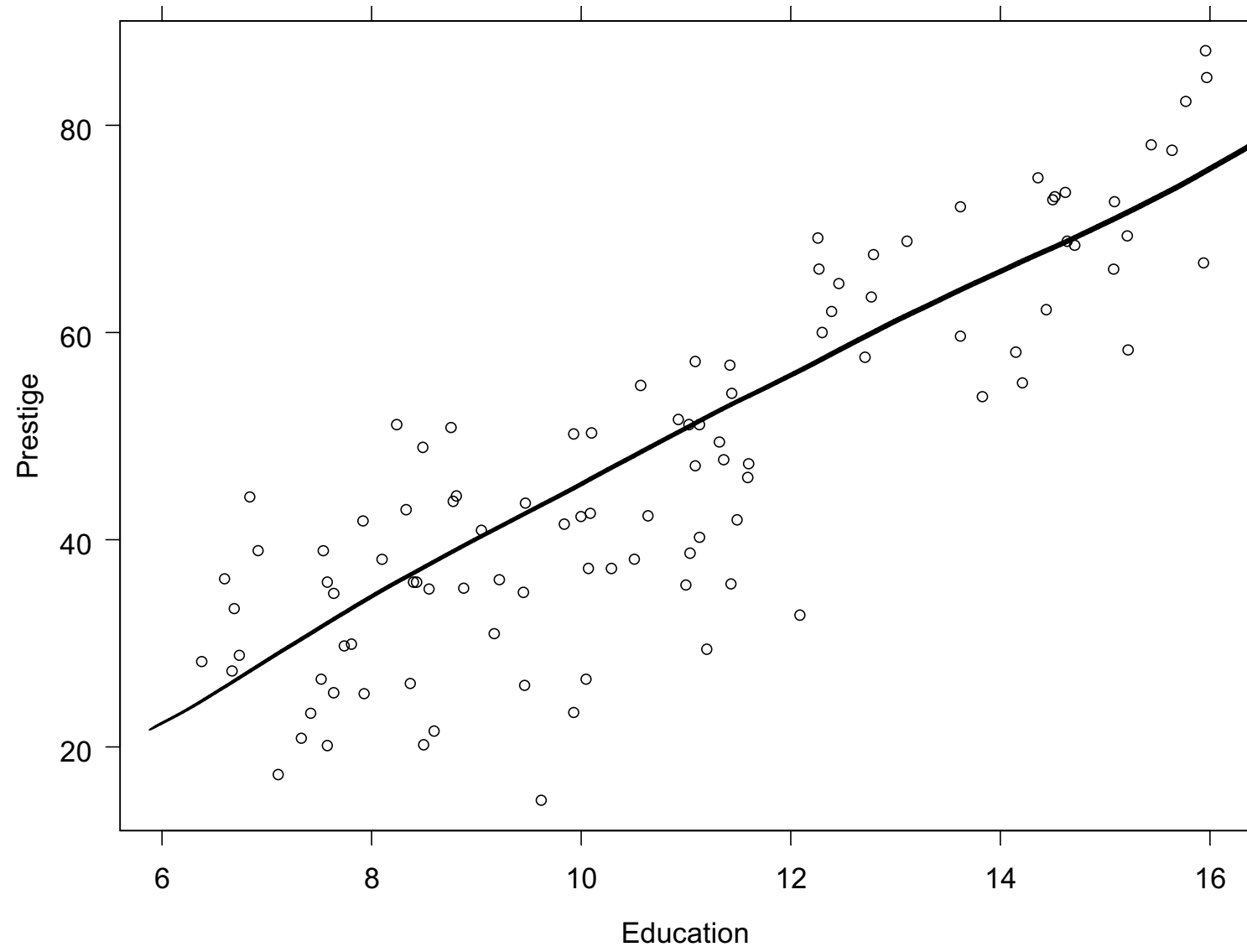
Prestige	Code	Type
Min.:14.80	Min.:1113	?: 4
1st Qu.:35.22	1st Qu.:3120	bc:45
Median:43.60	Median:5135	prof:30
Mean:46.83	Mean:5402	wc:23
3rd Qu.:59.28	3rd Qu.:8312	
Max.:87.20	Max.:9517	











$$\hat{y} = a + bX$$

Typical regression output:

```
> summary(fit <- lm(Prestige ~ Education, pdat))
```

```
Call: lm(formula = Prestige ~ Education, data = pdat)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-26.04  -6.523  0.6611  6.743  18.16
```

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-10.7320	3.6771	-2.9186	0.0043
Education	5.3609	0.3320	16.1478	0.0000

```
Residual standard error: 9.103 on 100 degrees of freedom
```

```
Multiple R-Squared: 0.7228
```

```
F-statistic: 260.8 on 1 and 100 degrees of freedom,  
the p-value is 0
```

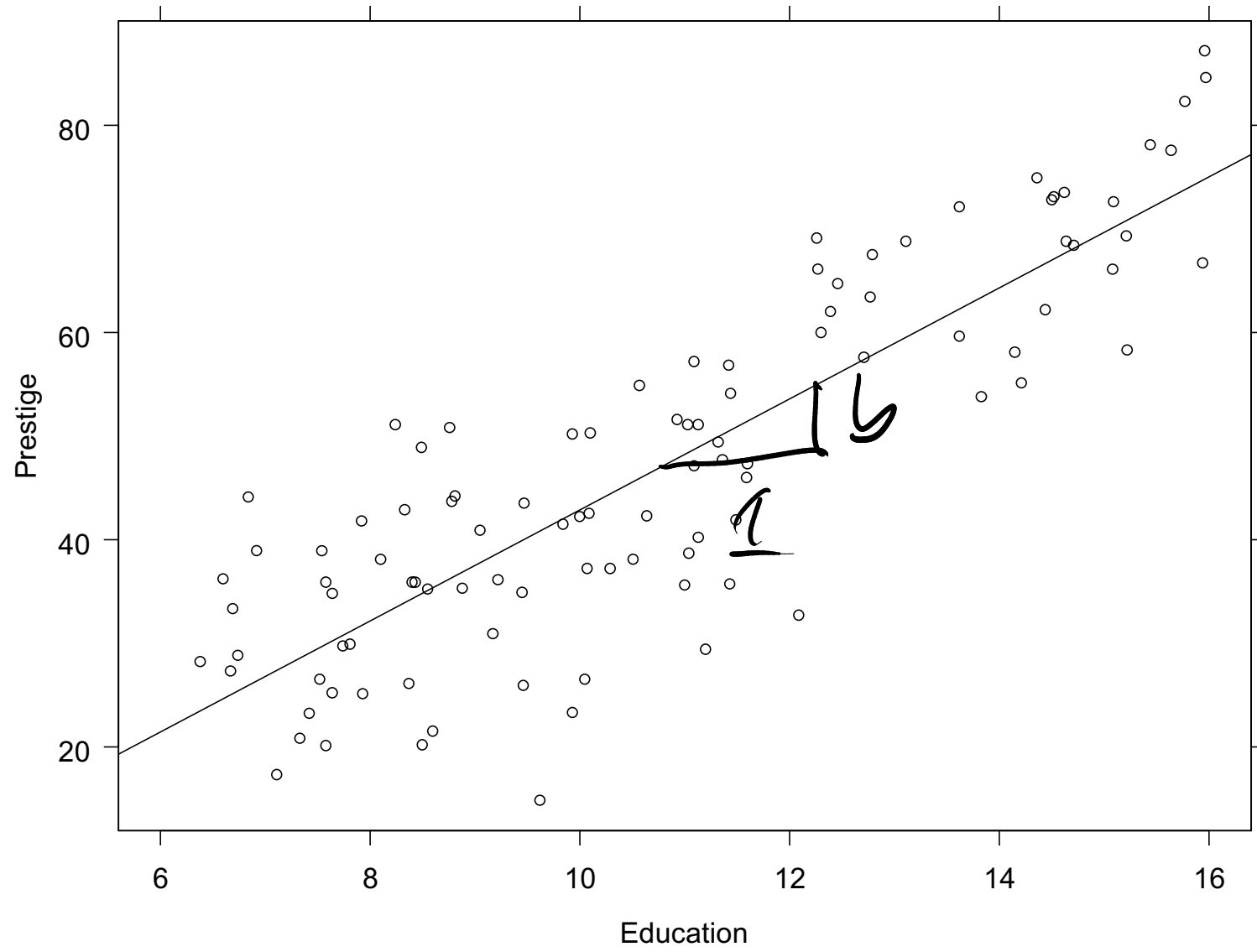
Se

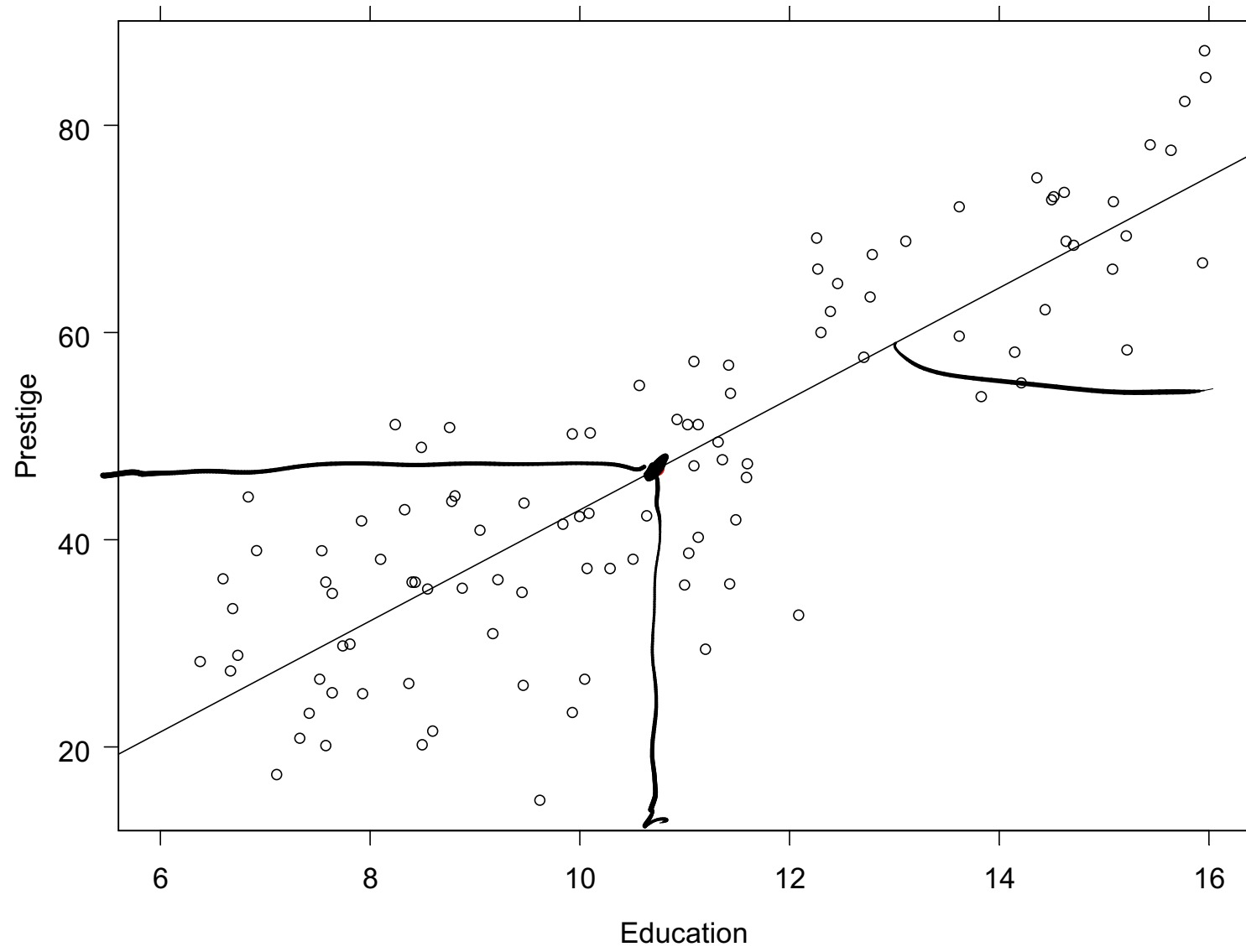
Slope of the least-squares line:

$$\begin{aligned}
 \hat{b} &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\
 &= \frac{[\sum (X_i - \bar{X})(Y_i - \bar{Y})] / [n - 1]}{[\sum (X_i - \bar{X})^2] / [n - 1]} \\
 &= \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)} \\
 &= \frac{s_{XY}}{s_X^2}
 \end{aligned}$$

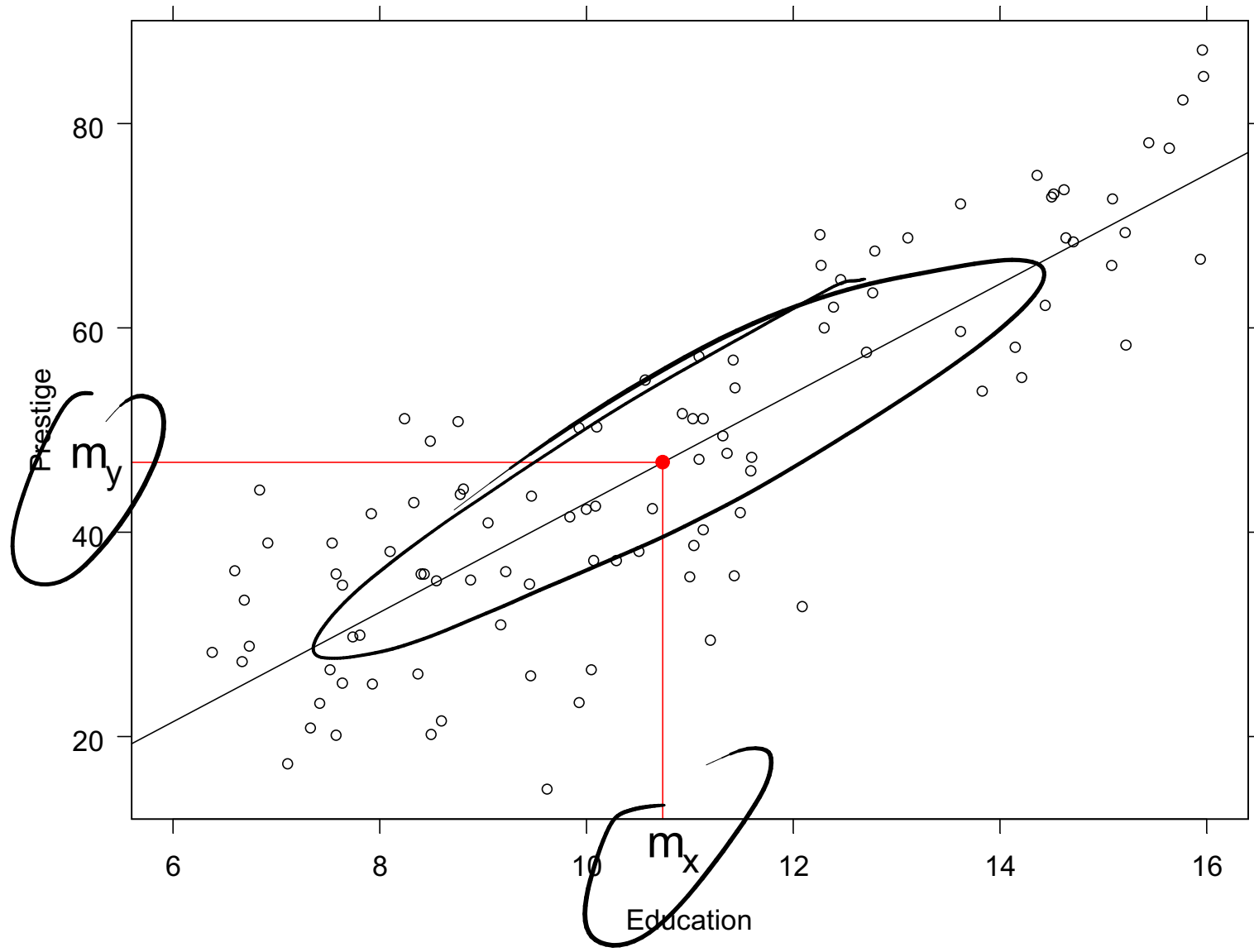
Intercept:

$$\bar{Y} = a + b\bar{X}$$





LS line



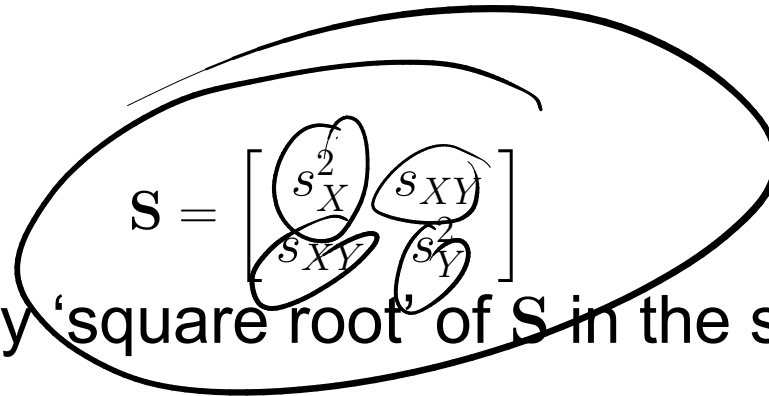
Data ellipse:

Data ellipse of radius r :

$$\mathcal{E}_r = \left\{ \begin{pmatrix} X \\ Y \end{pmatrix} : \left[\begin{pmatrix} X \\ Y \end{pmatrix} - \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \right]' \begin{bmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{bmatrix}^{-1} \left[\begin{pmatrix} X \\ Y \end{pmatrix} - \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \right] = r^2 \right\}$$

As a transformation of the unit circle:

Let



$$\mathbf{S} = \begin{bmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{bmatrix}$$

Let $\mathbf{S}^{1/2}$ be any 'square root' of \mathbf{S} in the sense that:

$$\mathbf{S} = \mathbf{S}^{1/2} \left(\mathbf{S}^{1/2} \right)'$$

A good one is the Choleski square root:

$$\mathbf{S}^{1/2} = \begin{bmatrix} s_X & 0 \\ s_{XY}/s_X & s_{Y \cdot X} \end{bmatrix} \mathbf{S}_e$$

where $s_{Y \cdot X}$ denotes the 'partial standard deviation' of Y adjusted for X :

$$s_{Y \cdot X} = \sqrt{s_Y^2 - s_{XY}^2/s_X^2}$$

Then

$$\mathcal{E}_r = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} + r \mathbf{S}^{1/2} \mathbf{U}$$

where \mathbf{U} is the unit circle.

If we let $SSE = \sum (Y_i - \hat{Y}_i)^2$, then:

$$s_{Y \cdot X} = \sqrt{\frac{SSE}{n-1}}$$

$$\mathbf{S} = \mathbf{A} \mathbf{A}^T$$

matrix
square
root

A lower Δ

$$\mathbf{S} = \mathbf{A} \mathbf{A}'$$

Spectral decomp:

$$\begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} = \text{Choleski?} \\ = \text{Sq. root?}$$

S is a var matrix

$$S = \Gamma \Lambda \Gamma'$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \\ = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$$

orthogonal diag with non-neg elements

$$\Lambda = \Lambda^{1/2} \Lambda^{1/2}$$

$$S = \underbrace{\Gamma \Lambda^{1/2}}_A \underbrace{\Lambda^{1/2} \Gamma'}_{A'}$$

$$S = A A'$$

The usual standard error of regression is:

So:

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

$$s_{Y \cdot X} \approx s_e$$

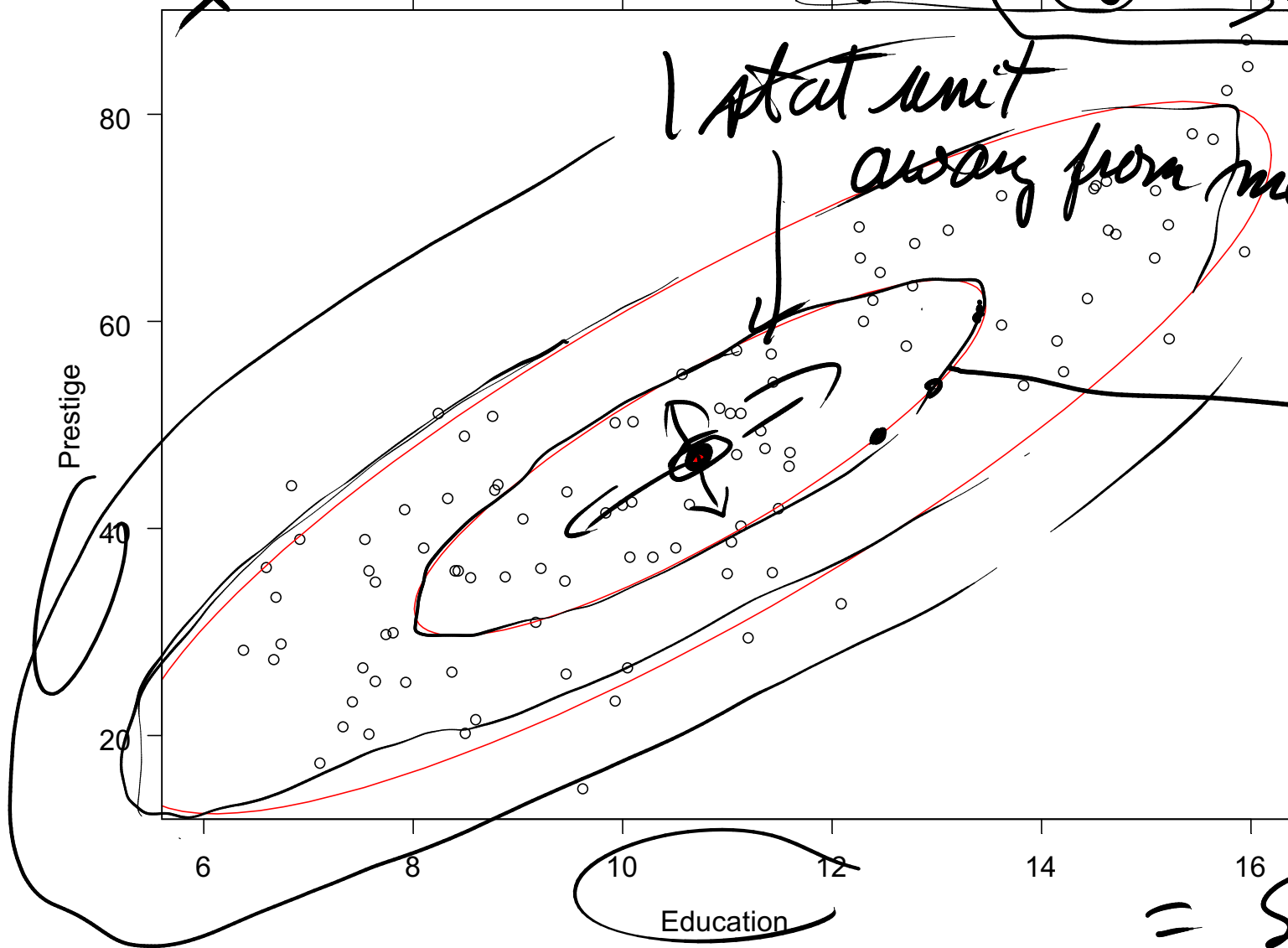
$$\hat{\sigma}_e = \sqrt{\frac{SSS}{n}}$$

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

$$\hat{\sigma}_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\frac{(x - \bar{x})^2}{S_x^2} = z_x^2$$

$$\epsilon_1 = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right]^T S^{-1} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right\}^T S^{-1} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$



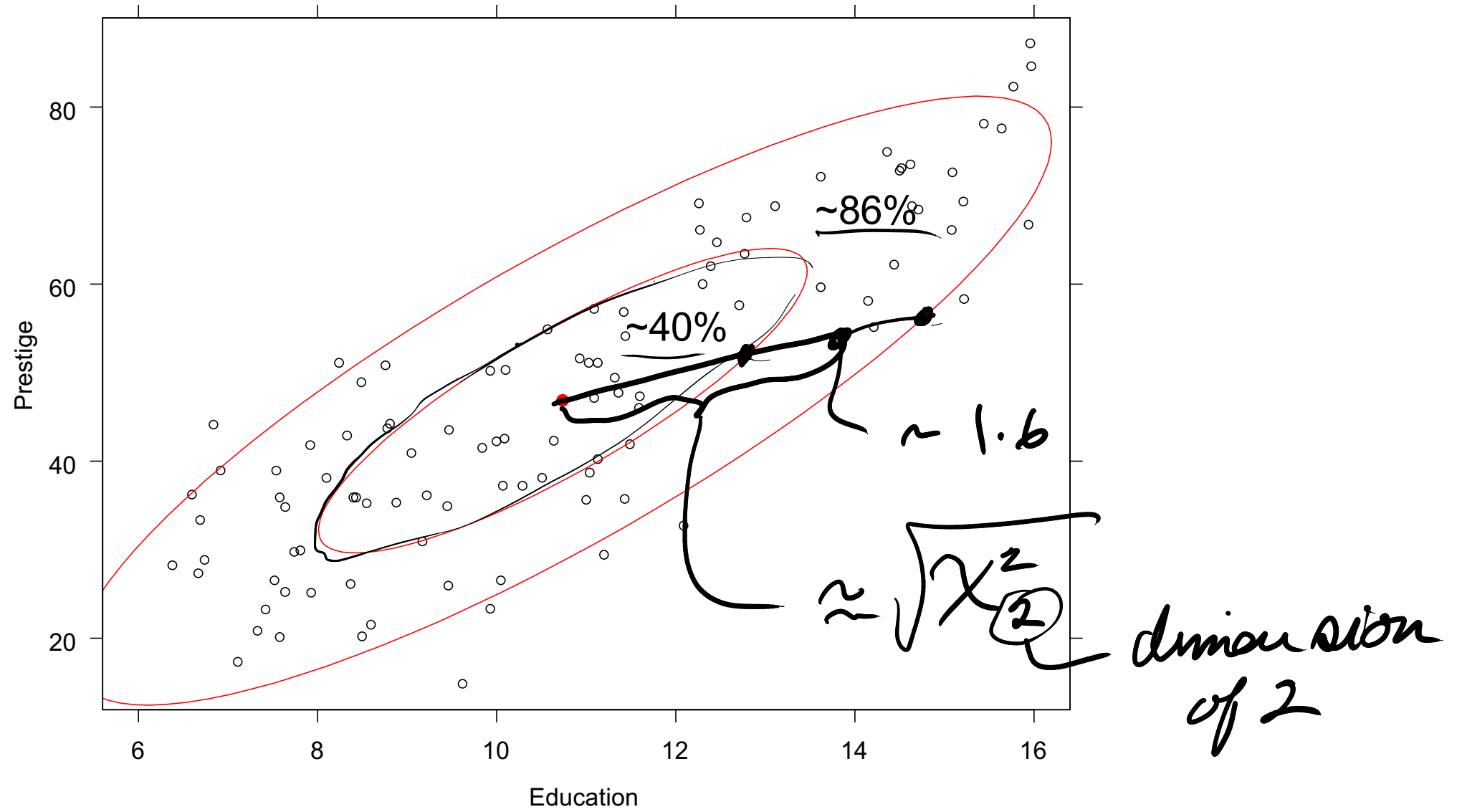
$$= 1$$

$$= z^2$$

$$= 3^2$$

ϵ_1
 $= \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : \right.$
 'Mahalanobis'
 distance

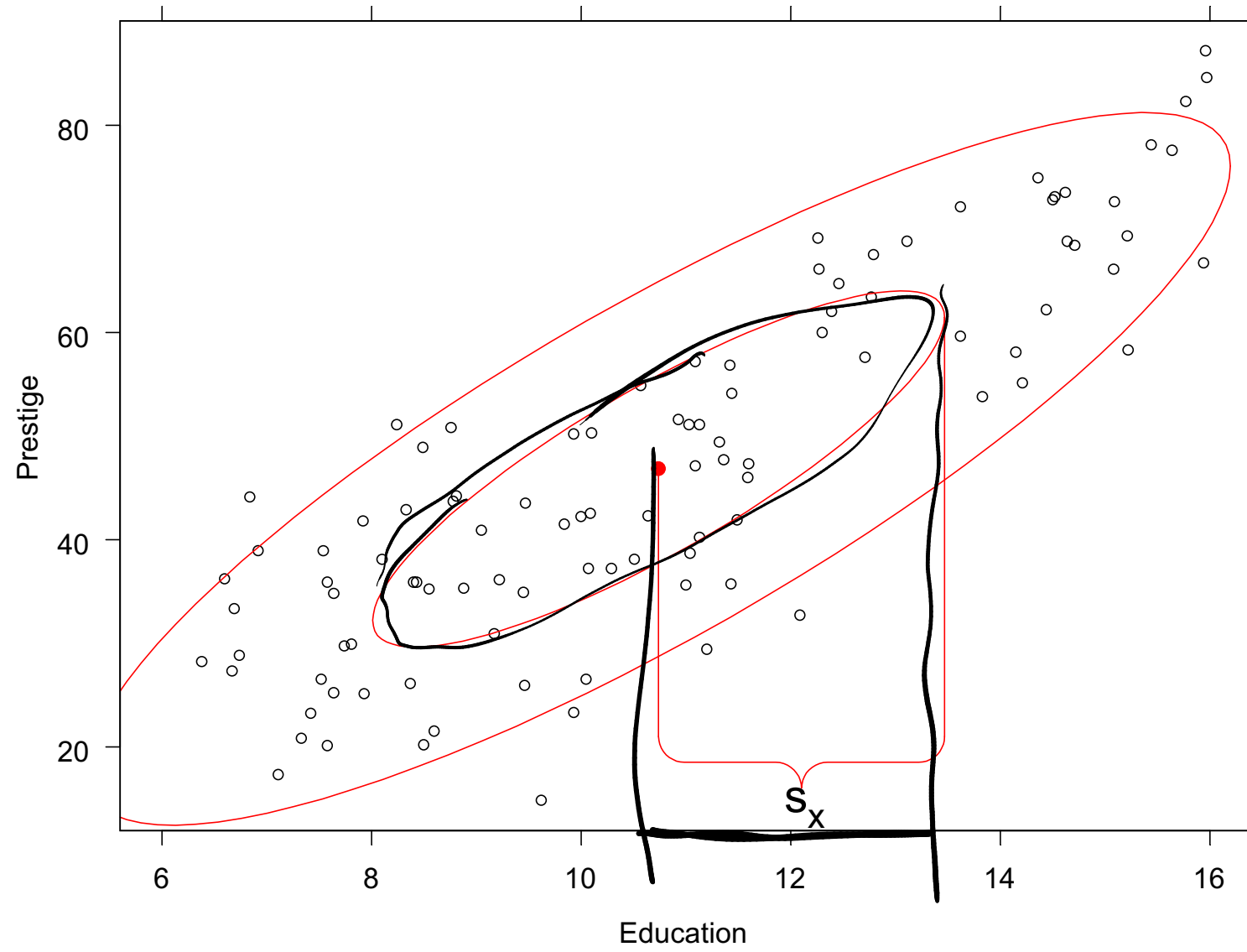
= Statistical distance

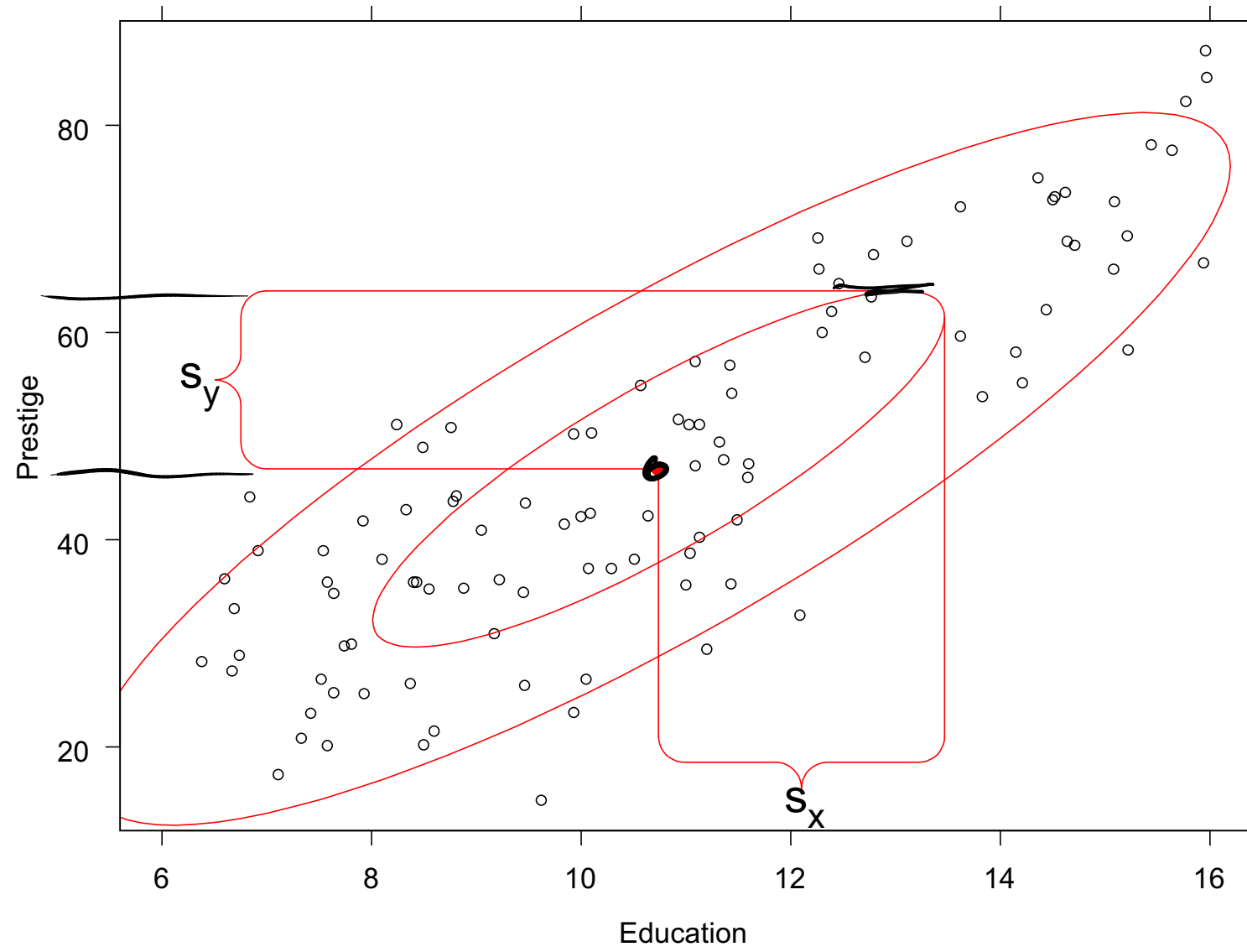


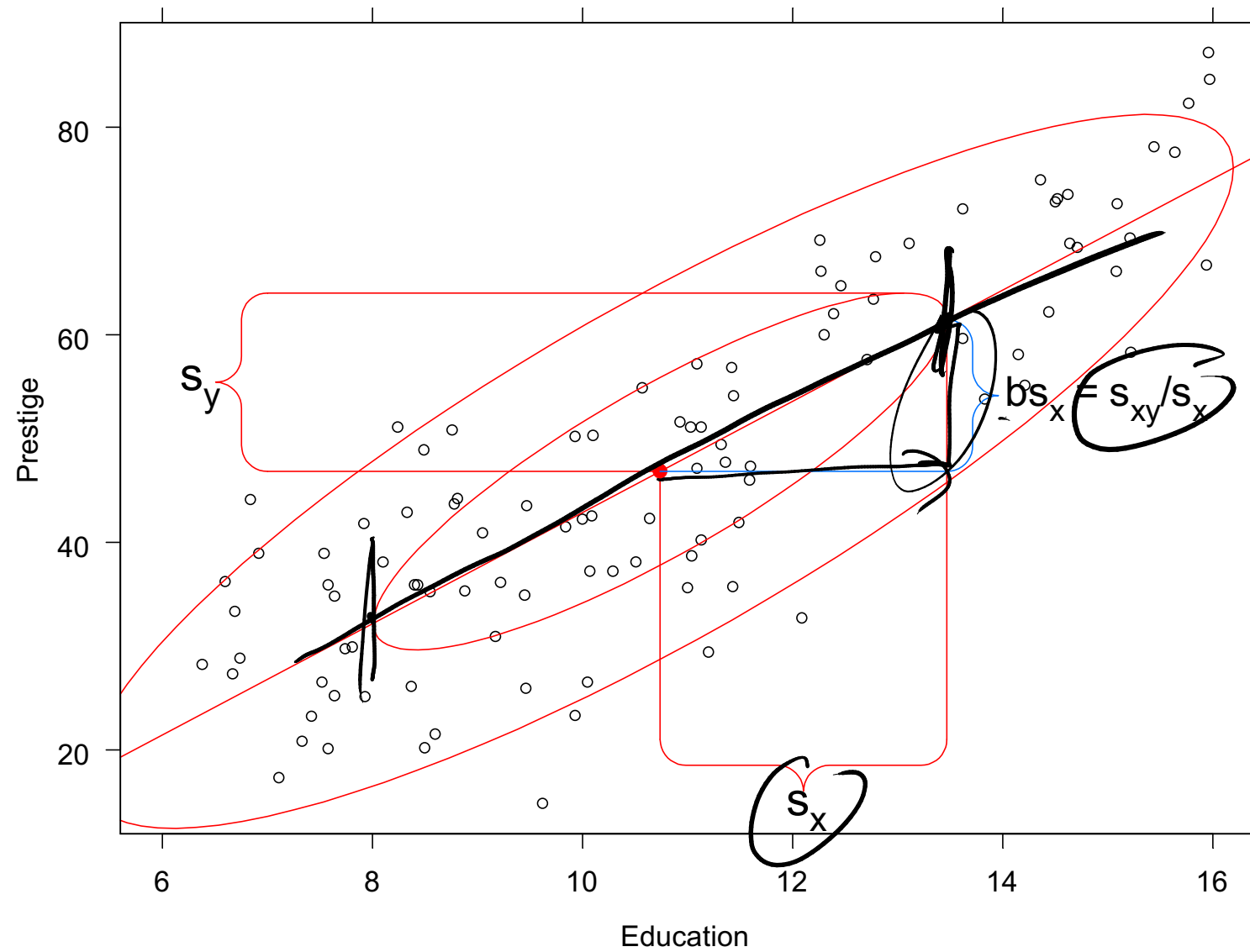
If the data cloud is approx. bivariate normal, then the proportion of points in \mathcal{E}_r is approx:

$$\Pr(\chi_2^2 \leq r^2)$$

where χ_2^2 has a χ^2 distribution with 2 degrees of freedom.

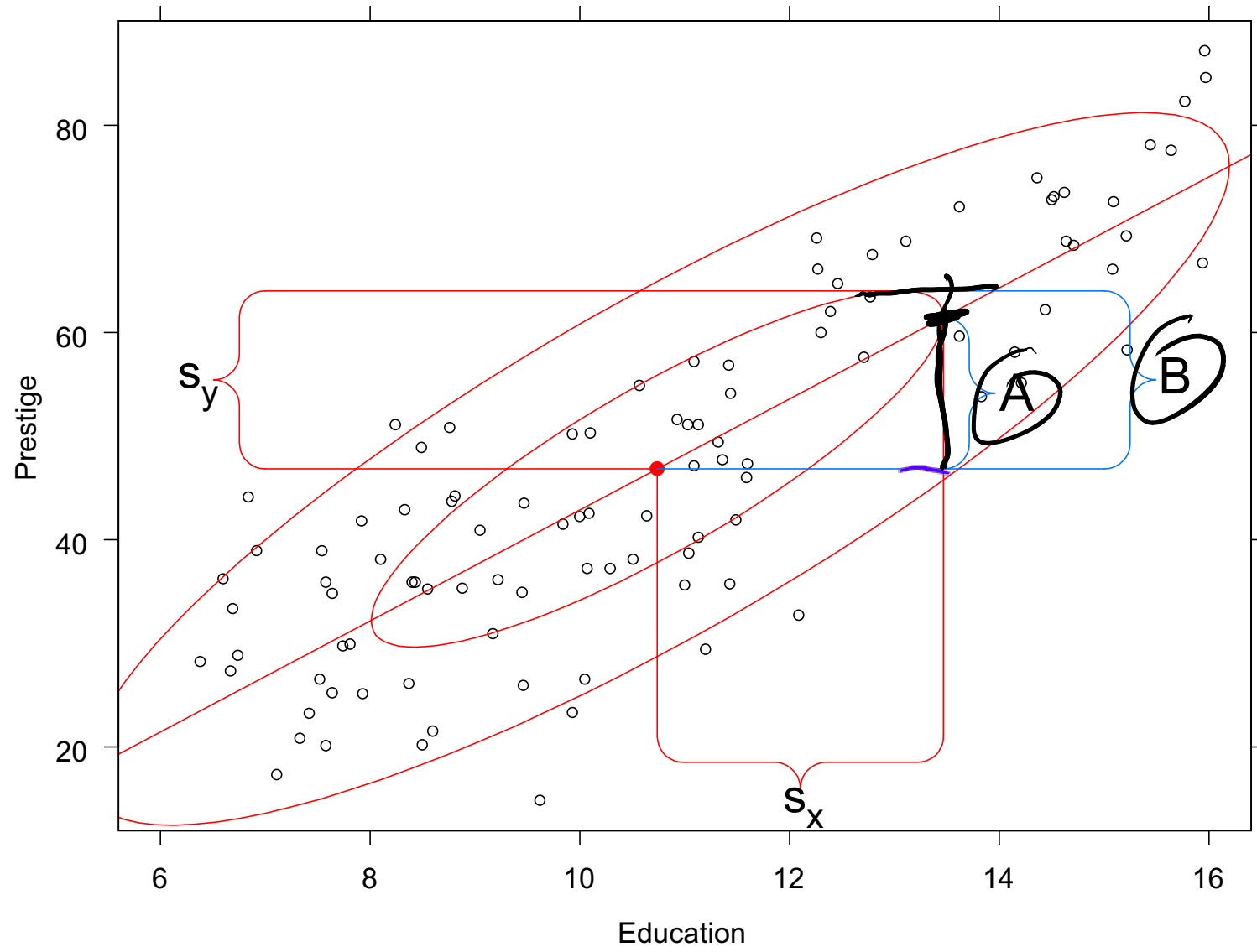


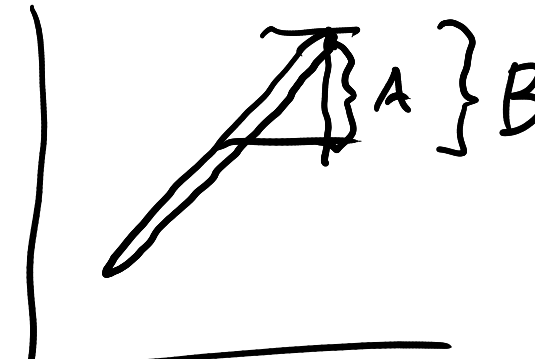
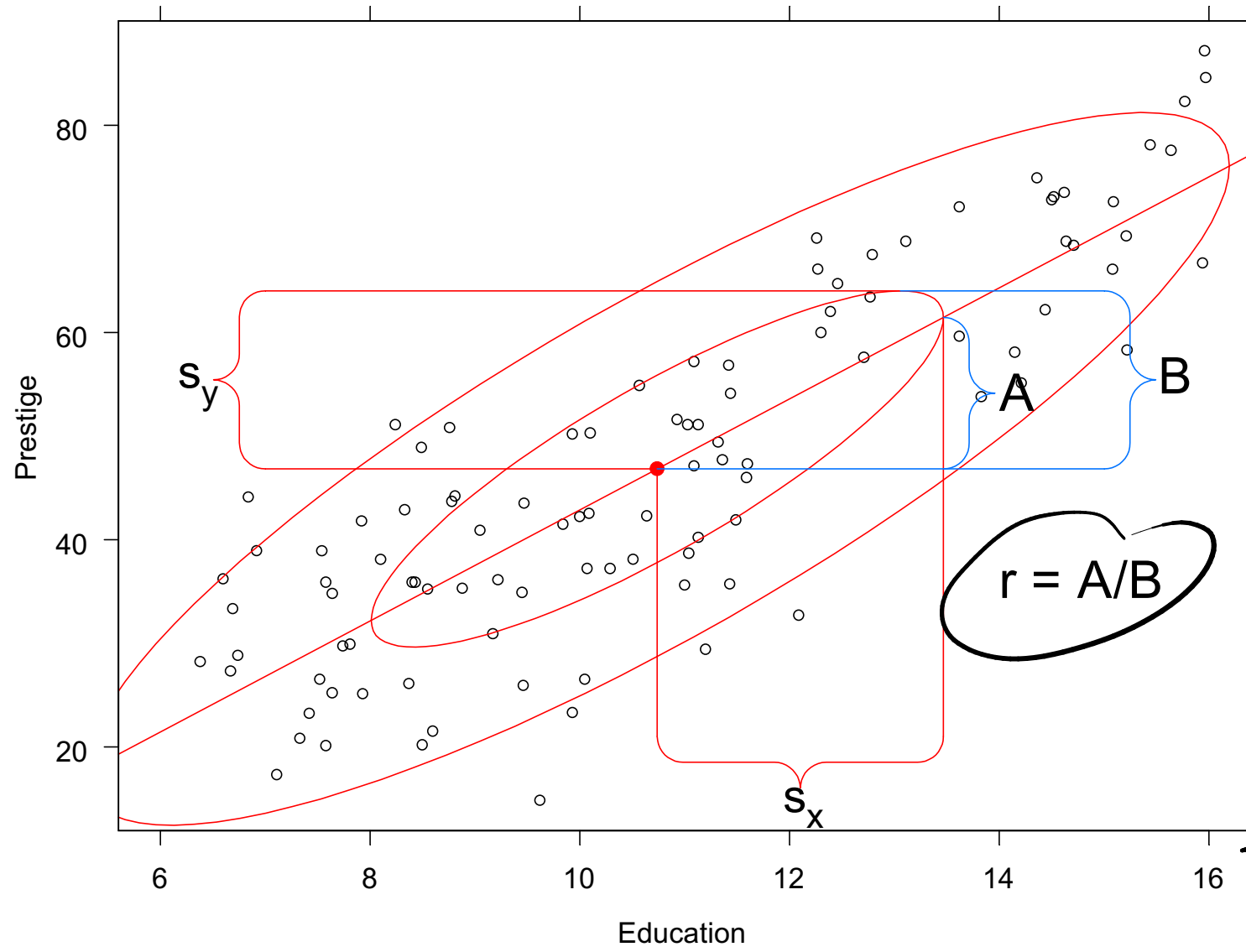




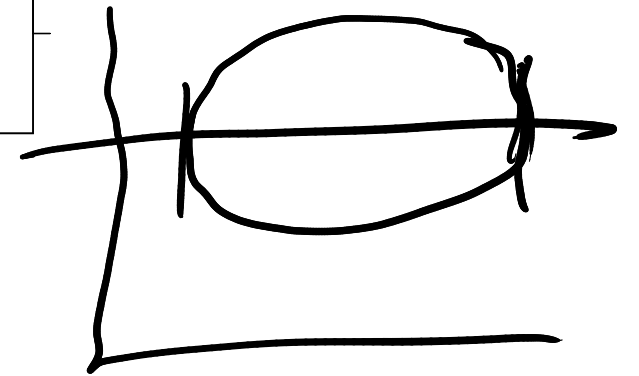
$$b = \frac{S_{xy}}{S_x^2}$$

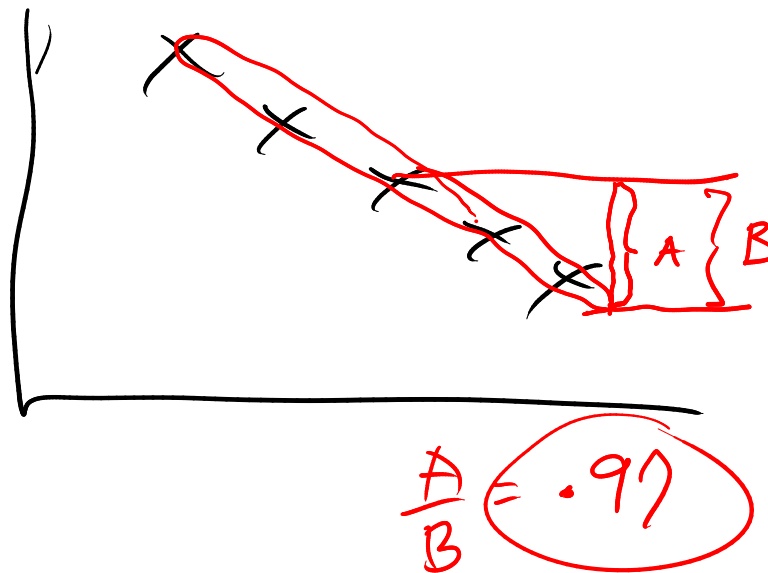
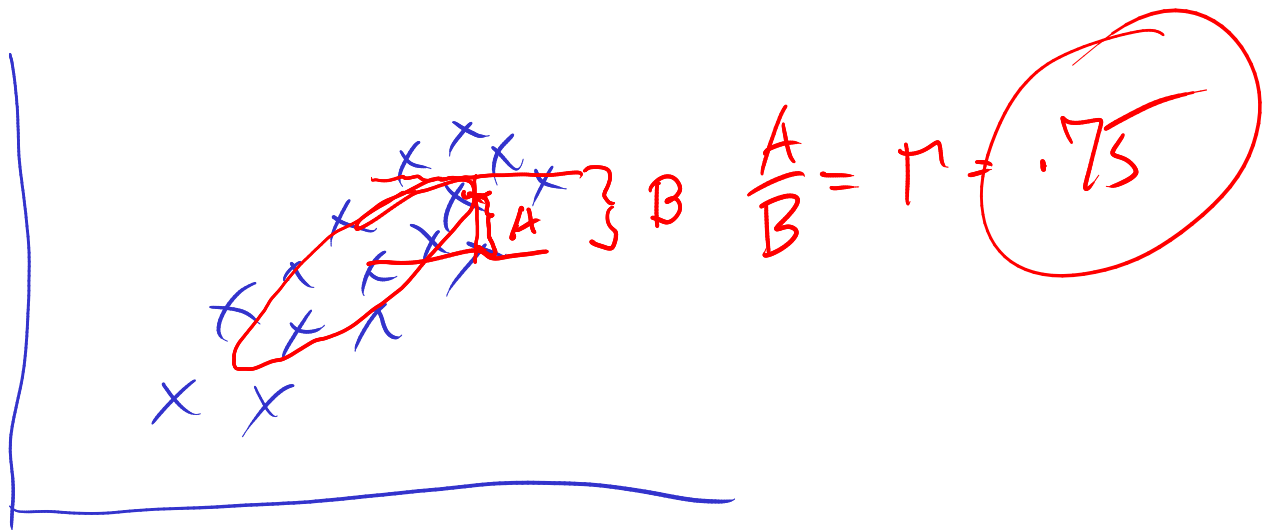
$$b S_x = \frac{S_{xy}}{S_x}$$





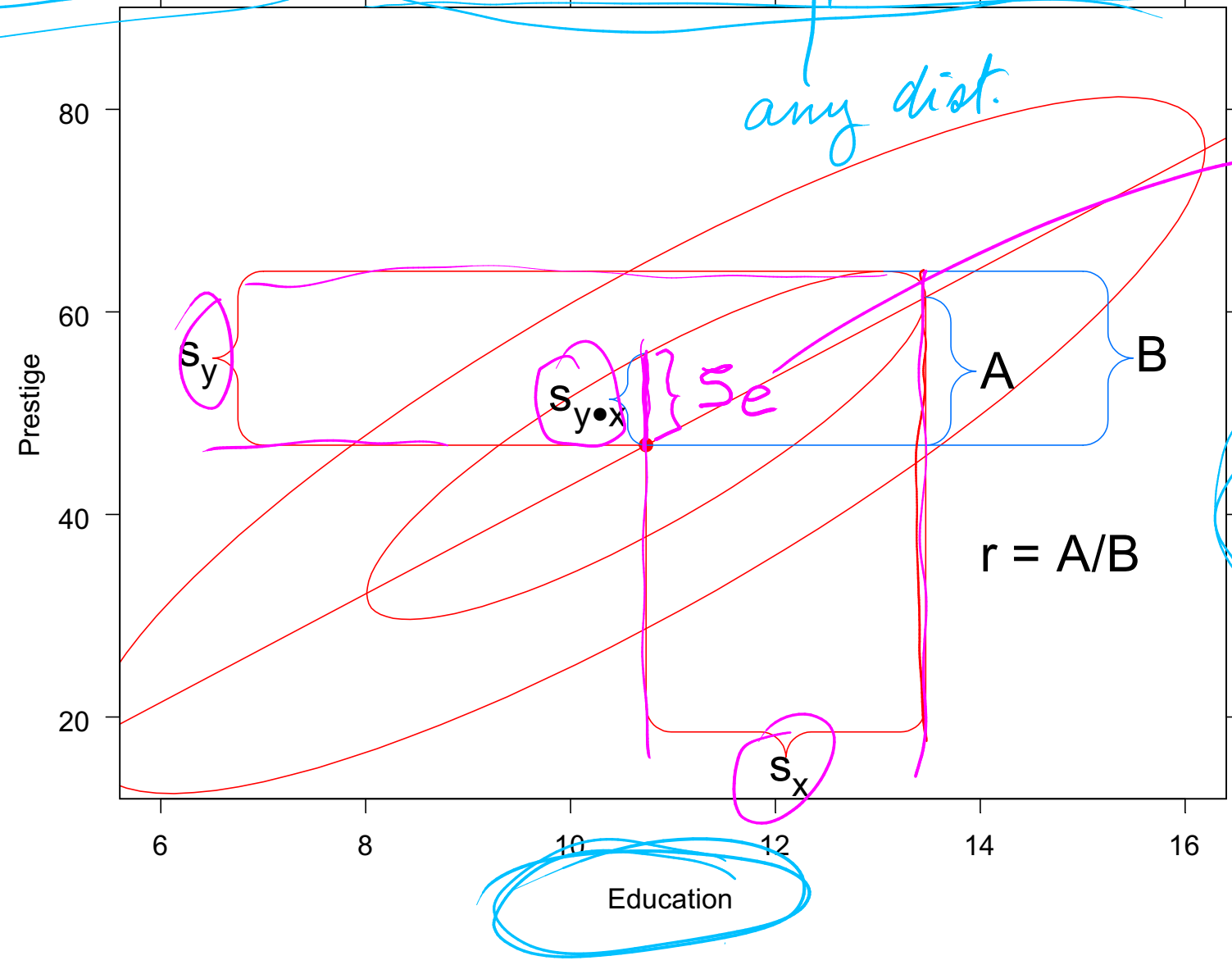
$\frac{A}{B}$ close to 1





$Y | X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

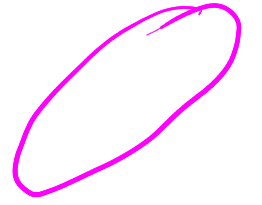
$Y = \beta_0 + \beta_1 X + \epsilon \sim N(0, \sigma^2)$

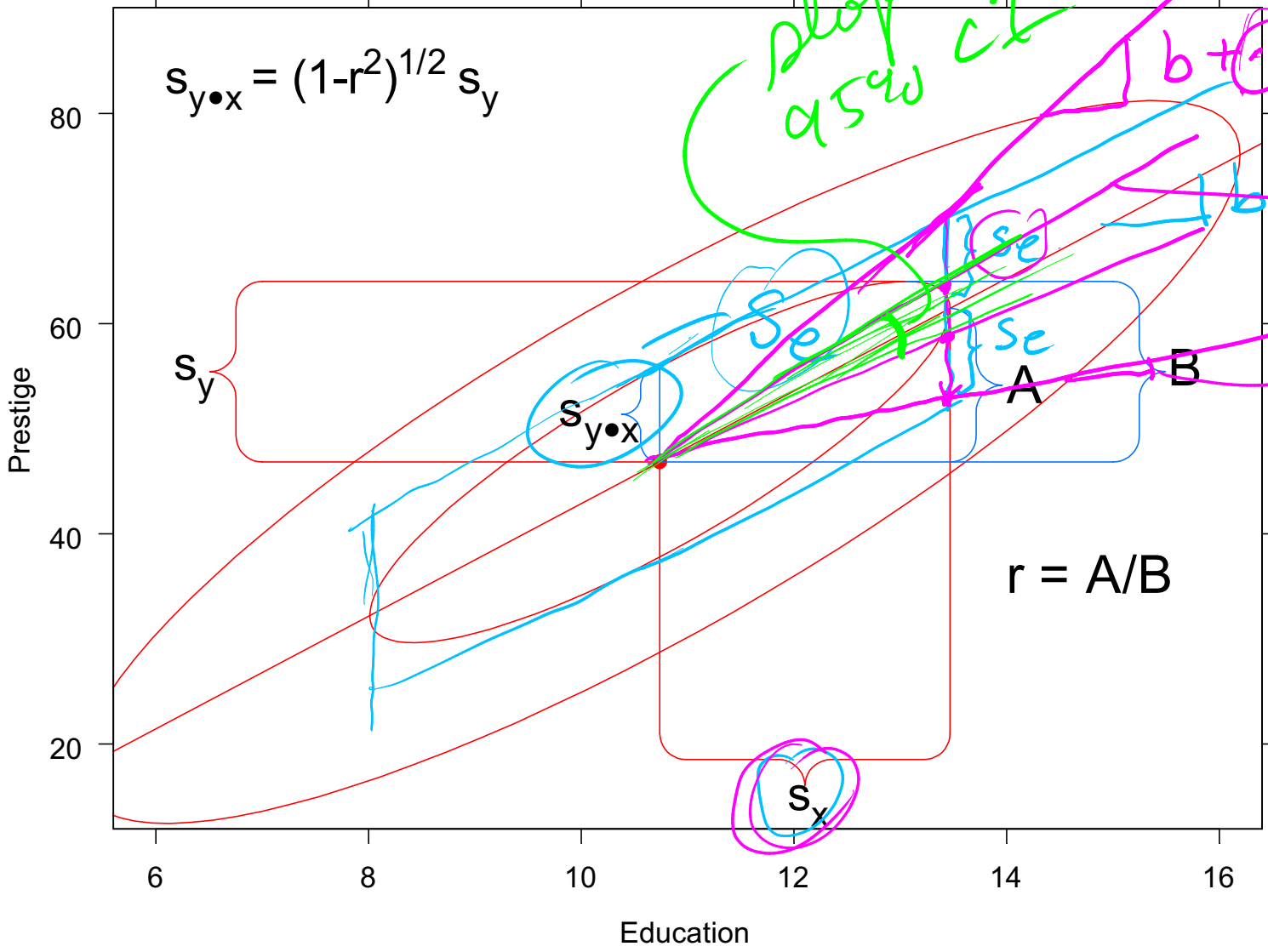


Res. SE

$$\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \begin{pmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{pmatrix}$$

equiv.





$$s_{y \cdot x} = (1-r^2)^{1/2} s_y$$

s_y

$s_{y \cdot x}$

s_x

$$r = A/B$$

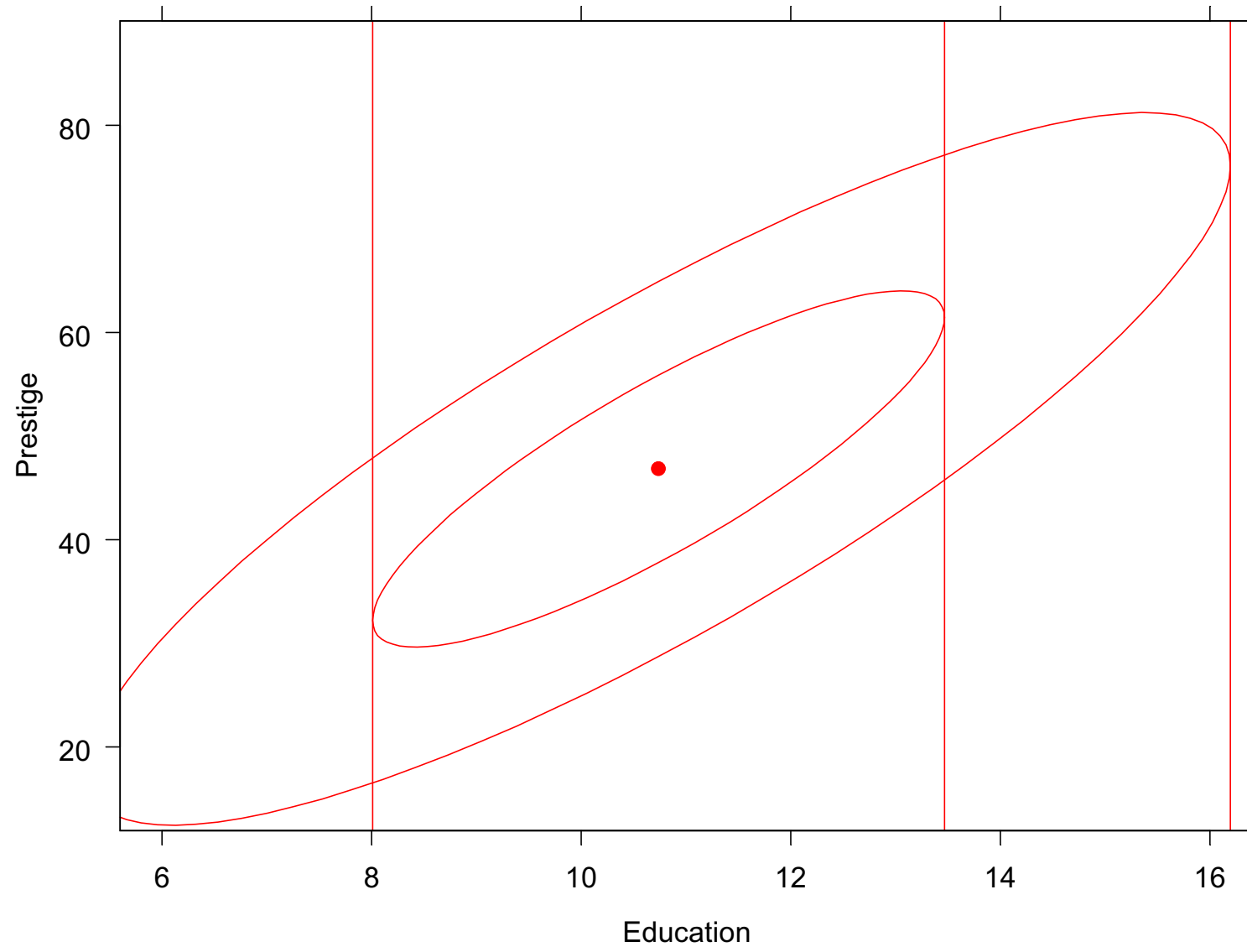
$$b \pm \frac{2}{\sqrt{n}} \frac{Se}{s_x}$$

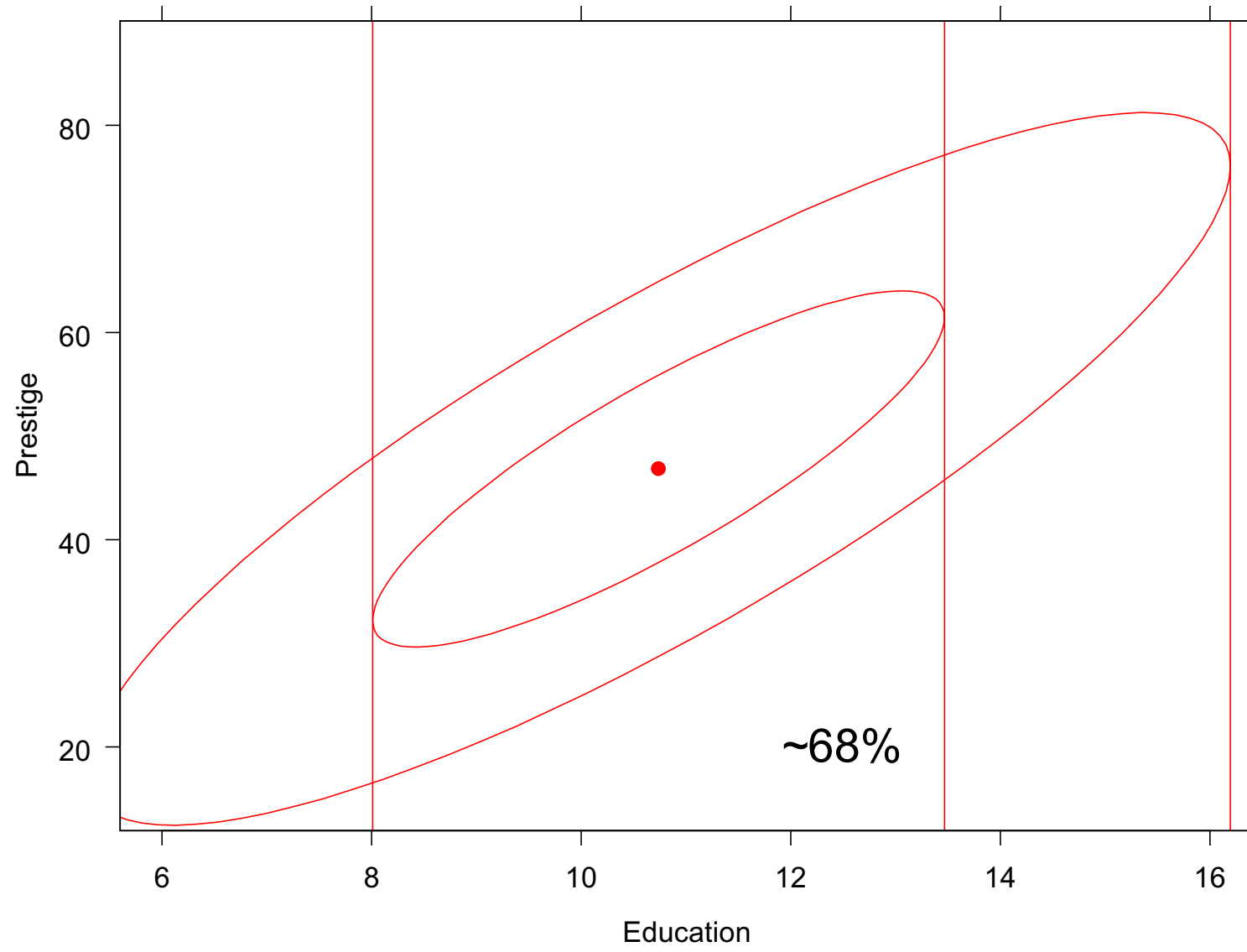
t

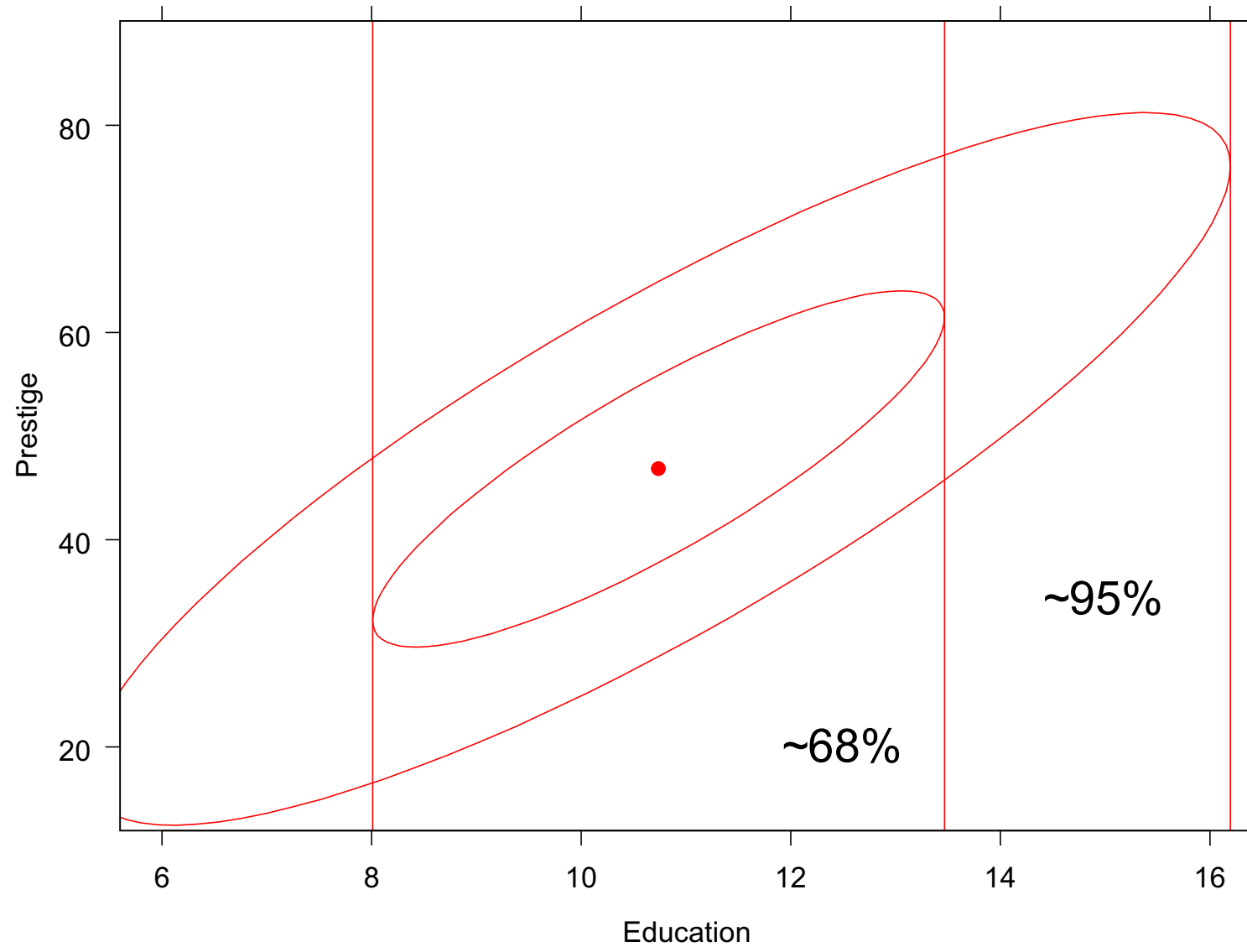
95% CI

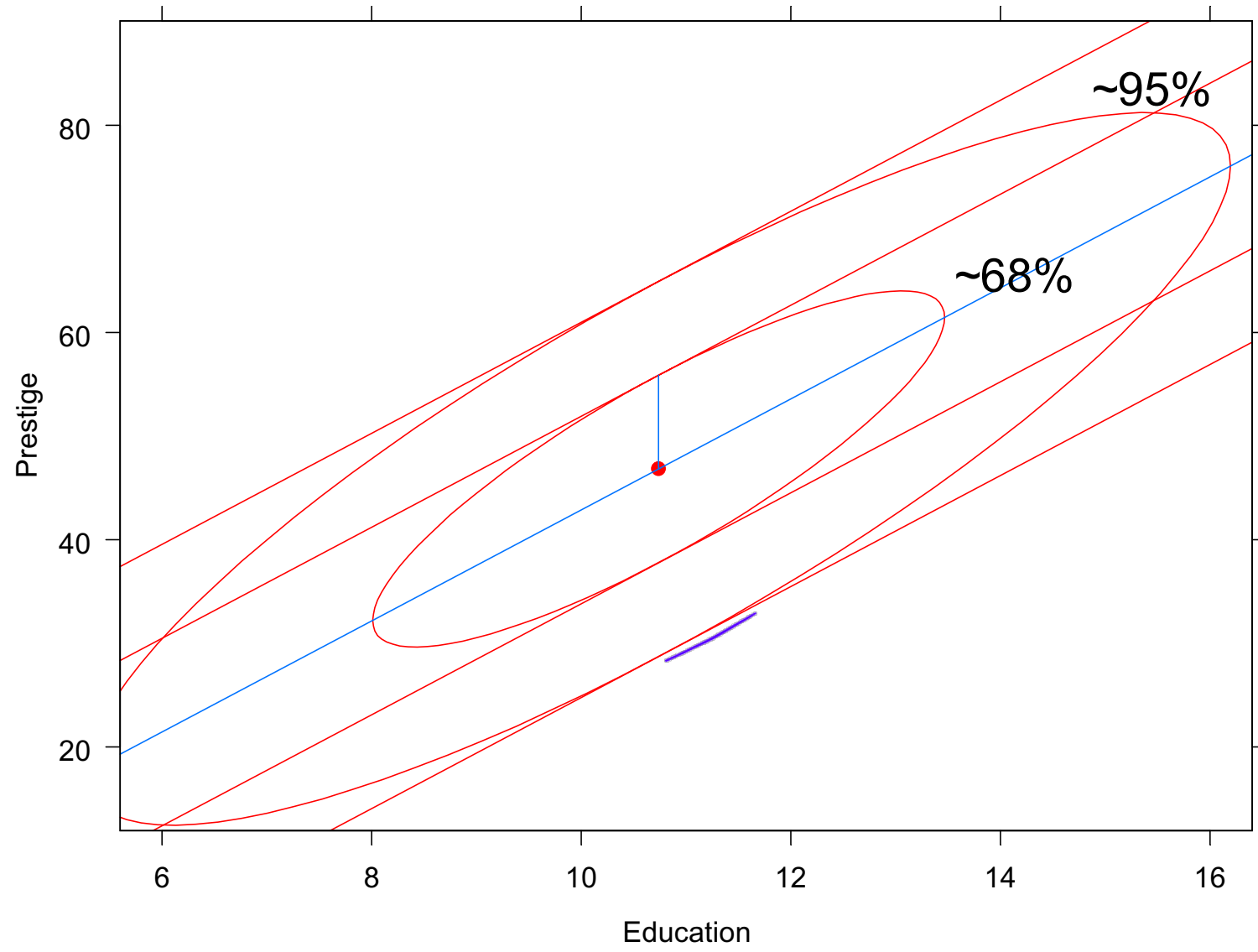
$$b - \frac{Se}{s_x}$$

slopes in 95% CI









If the data cloud is approx. bivariate normal, then the proportion of points in a *band* containing \mathcal{E}_r is approx:

$$\Pr(\chi_1^2 \leq r^2)$$

where χ_1^2 has a χ^2 distribution with 1 degrees of freedom.

3 Visual 95% Confidence Interval

95% Confidence interval for slope:

$$b \pm t_{0.975} \times SE(b)$$

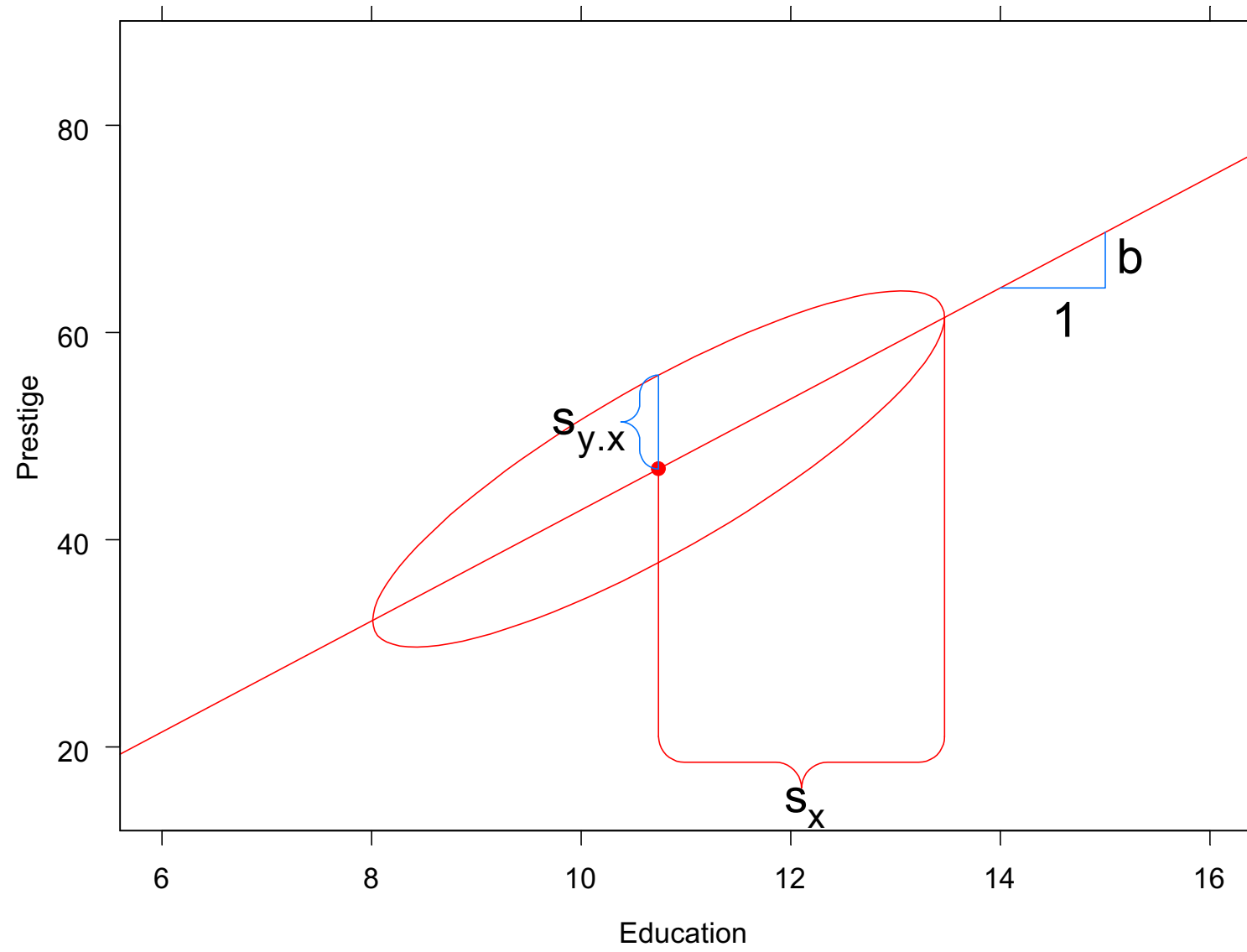
Now,

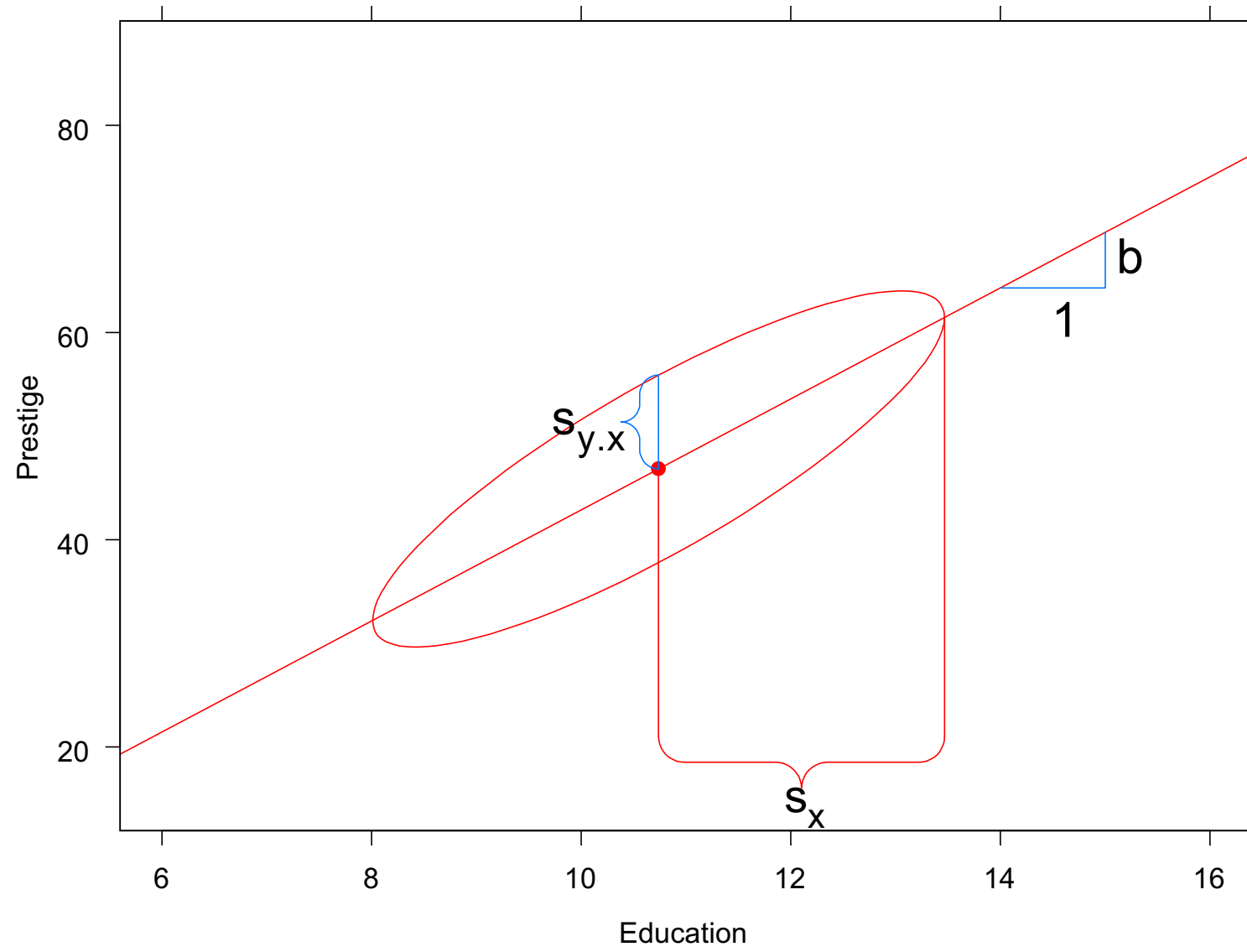
$$SE(b) = \frac{1}{\sqrt{n}} \frac{s_{y.x}}{s_x} \times \frac{\sqrt{n}}{\sqrt{n-2}}$$

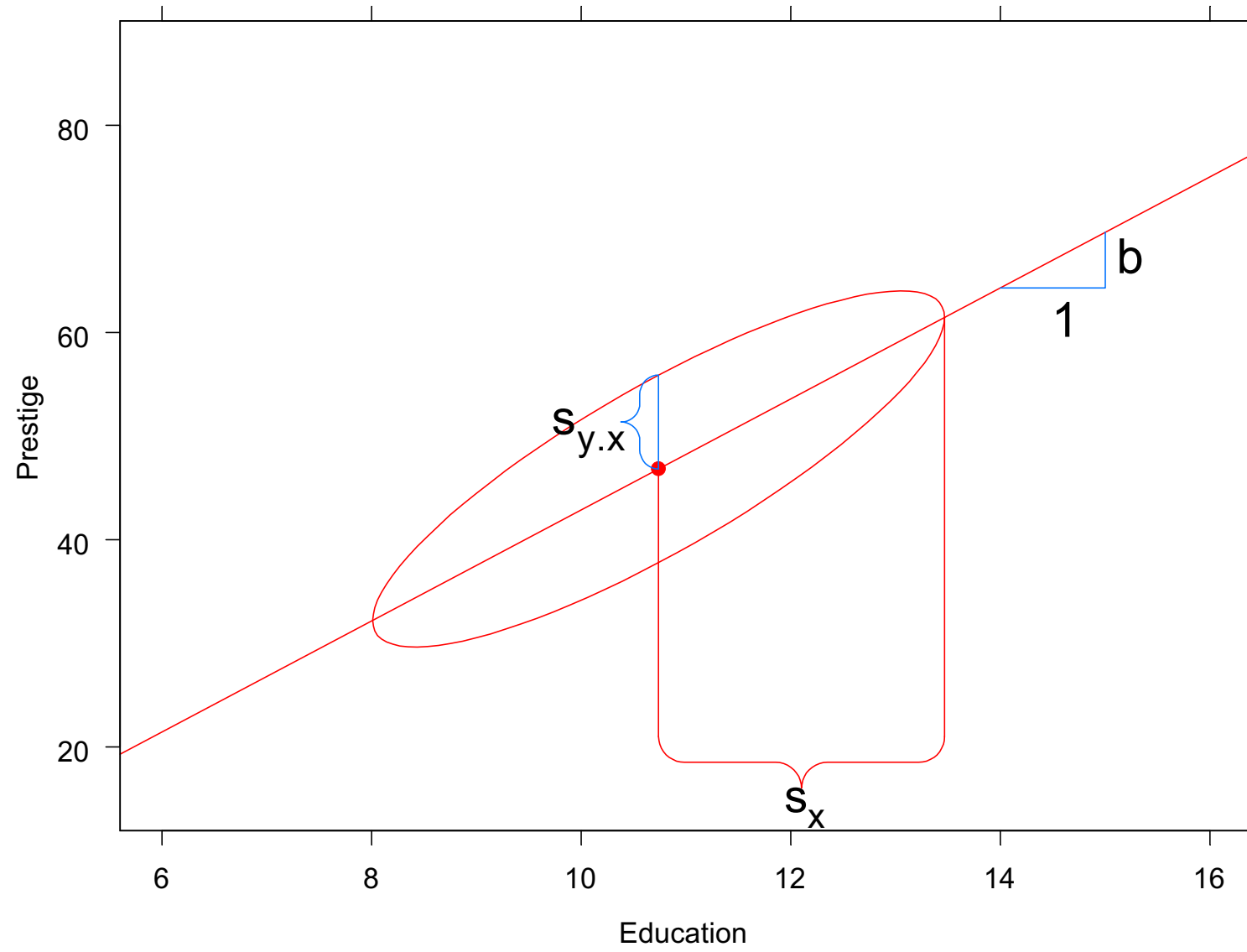
$$\simeq \frac{1}{\sqrt{n}} \frac{s_{y.x}}{s_x}$$

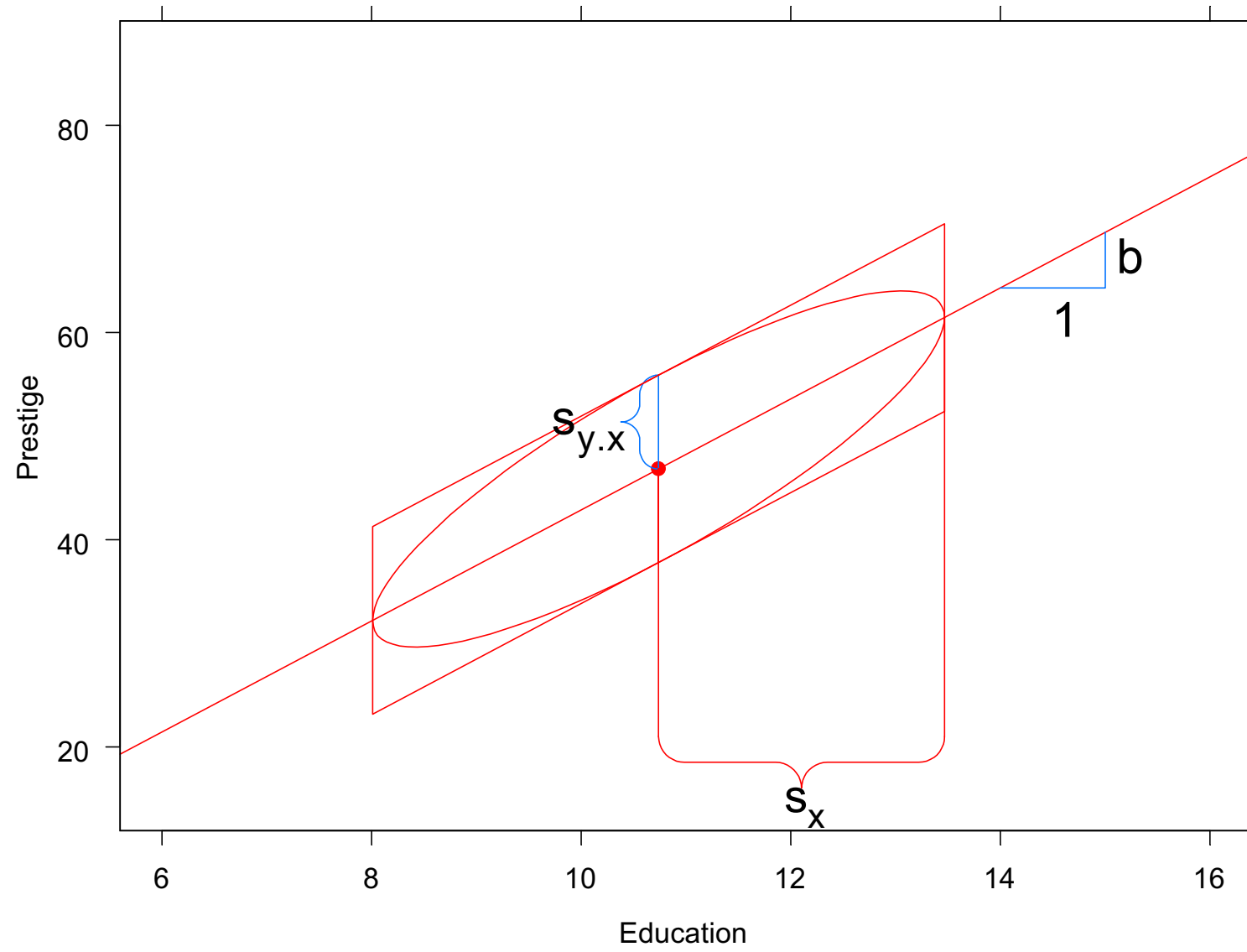
and taking $t_{0.975} \simeq 2$, we have an approximate 95% CI:

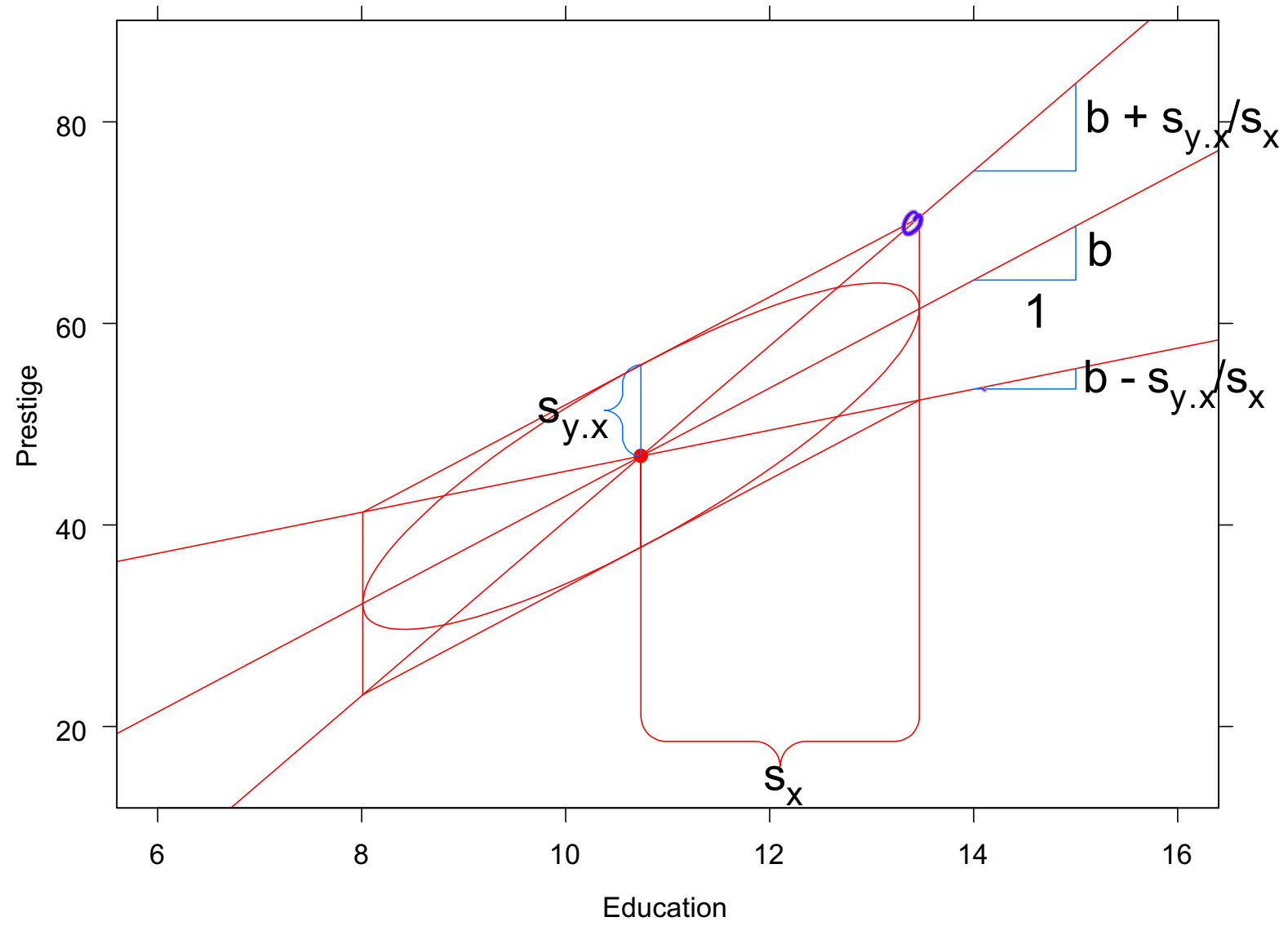
$$b \pm \frac{2}{\sqrt{n}} \times \frac{s_{y.x}}{s_x} \quad s_e$$

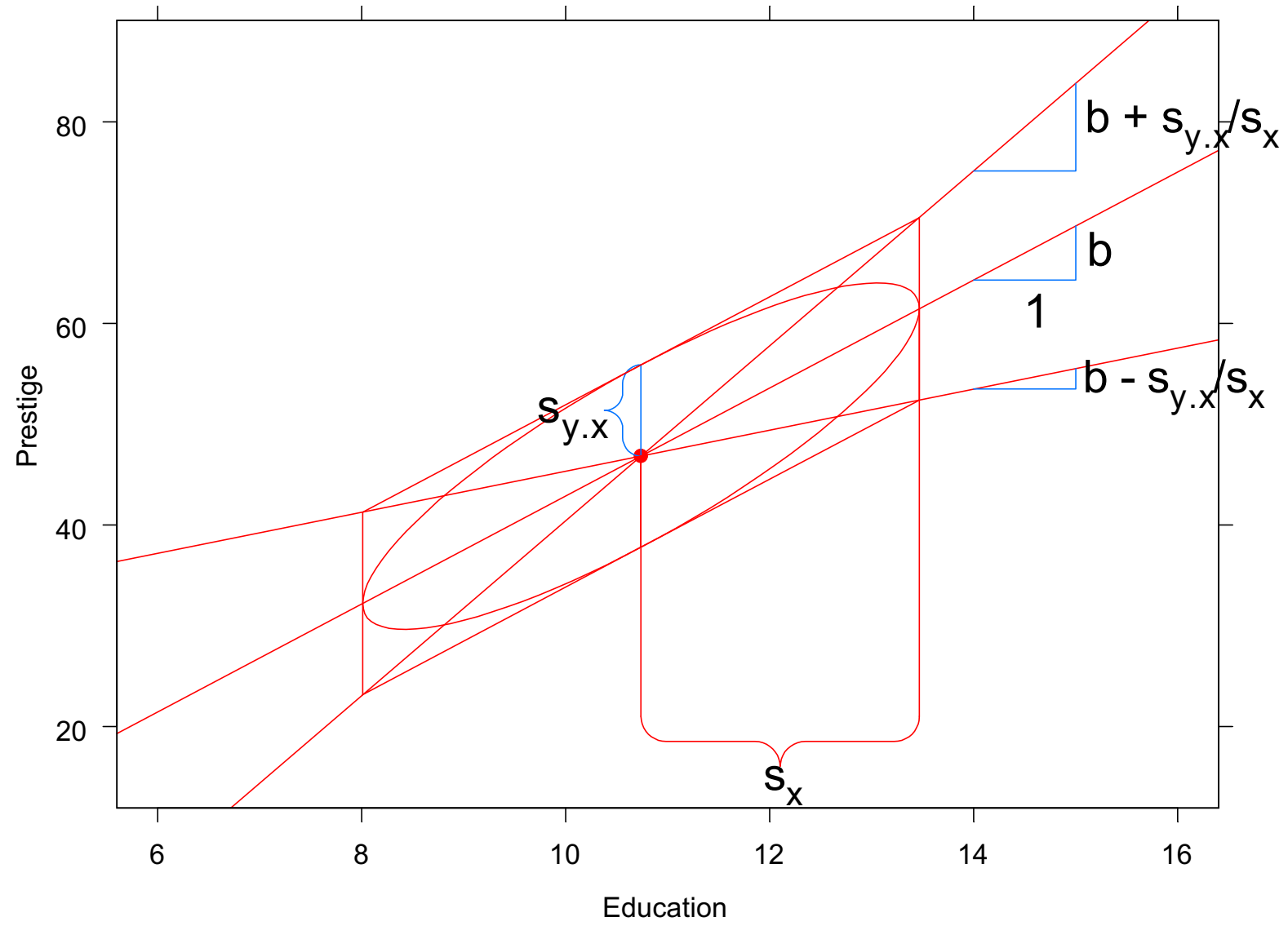


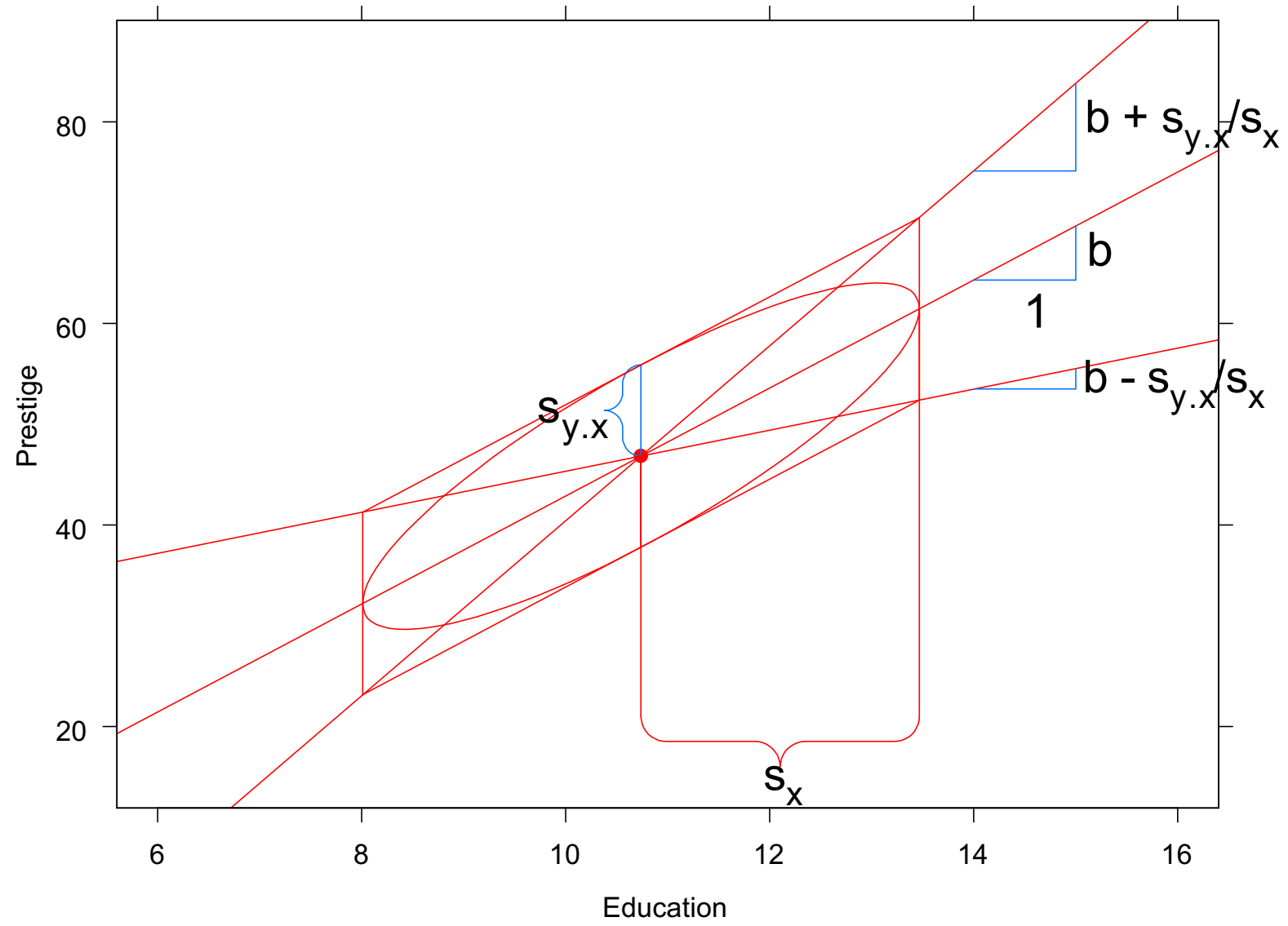


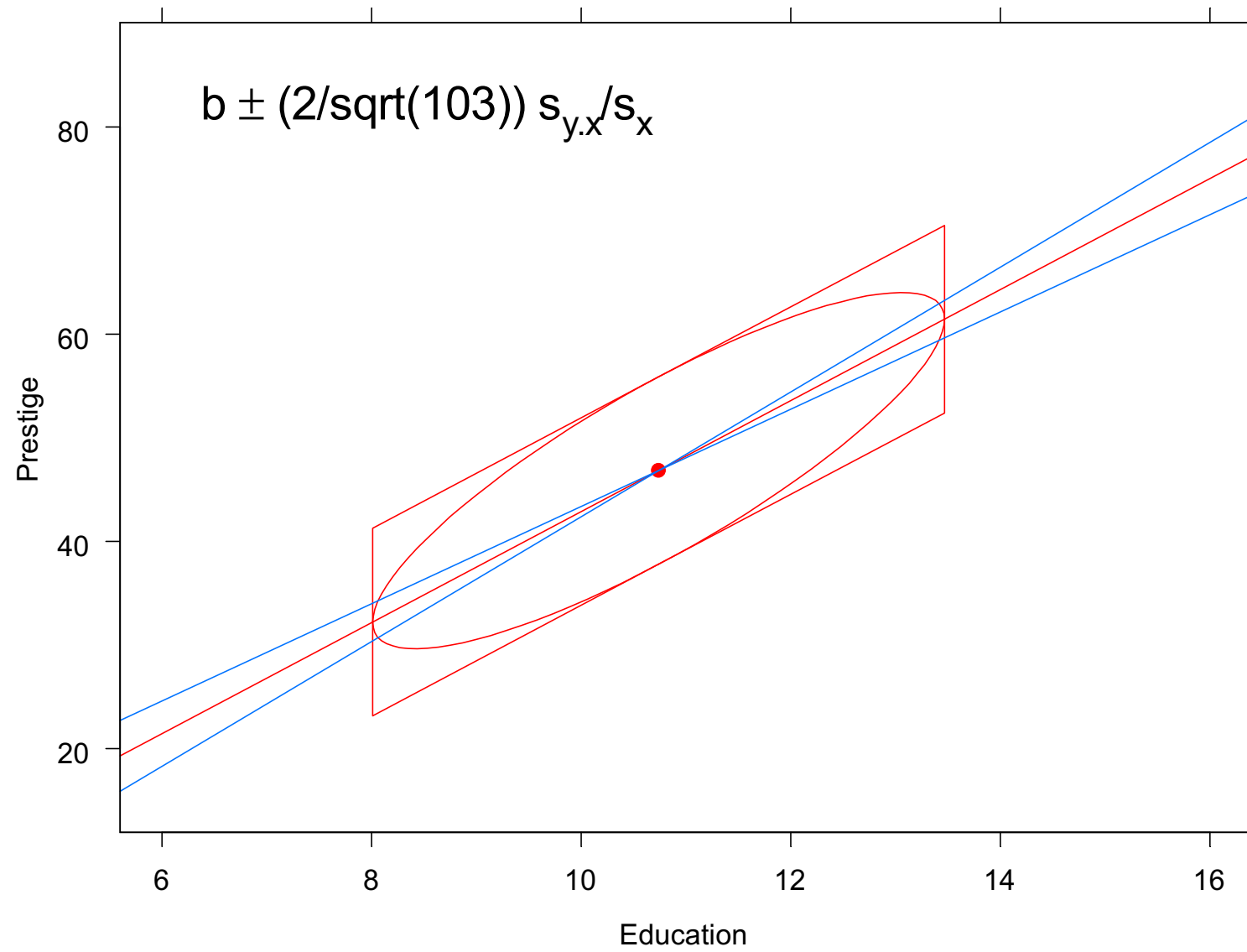


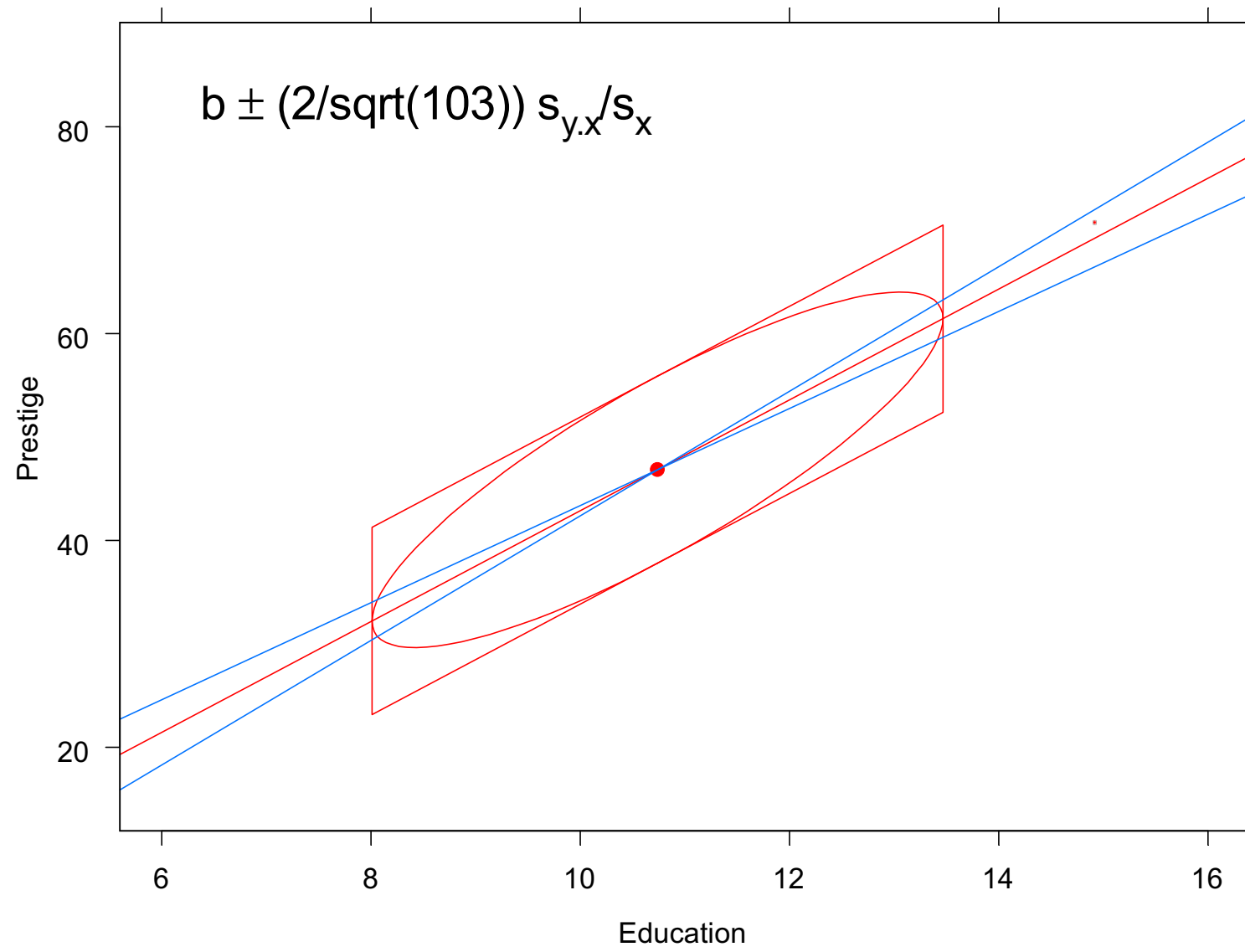








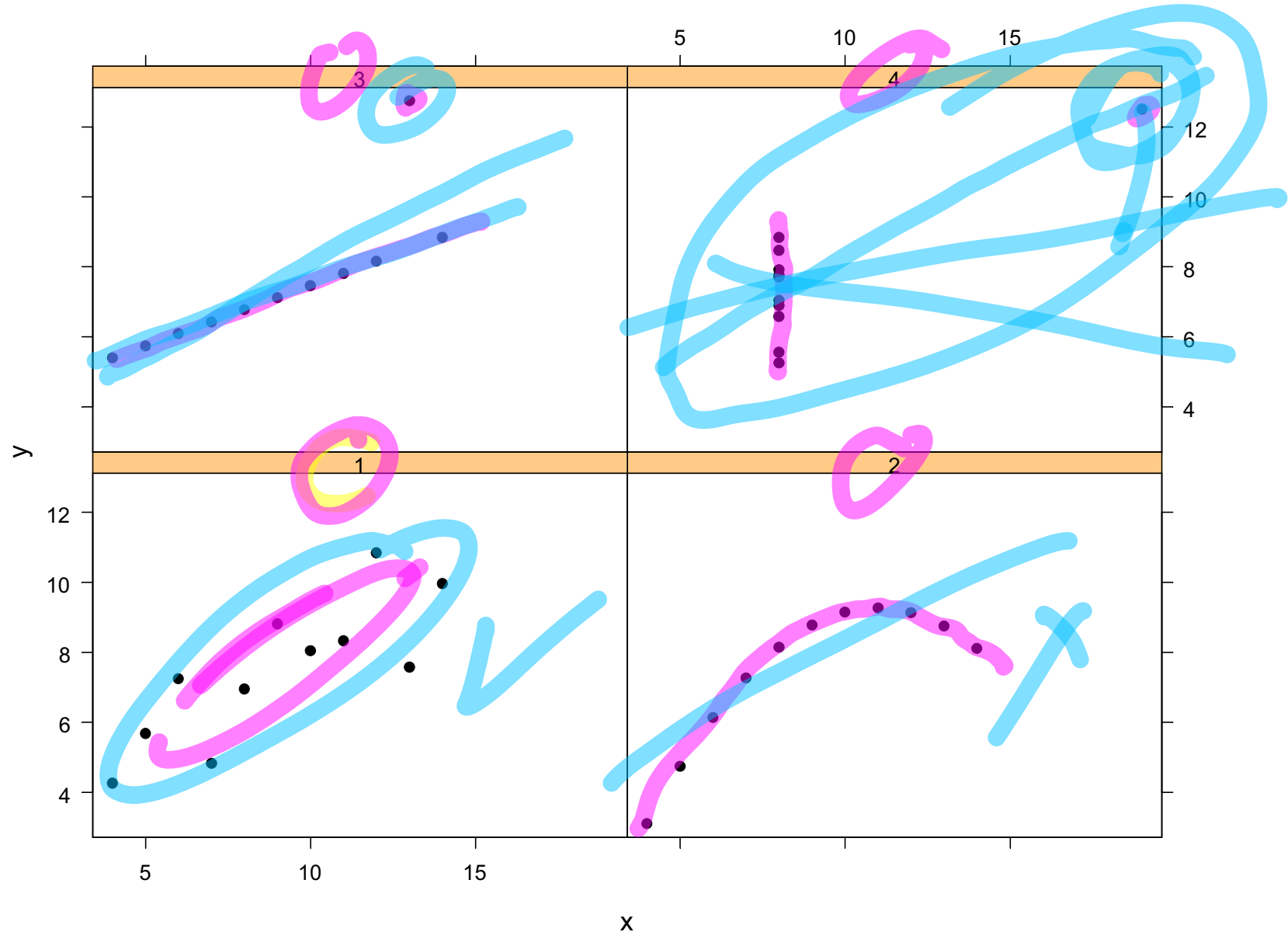




4 Anscombe Examples

Four datasets:

Same least-squares regression but very different stories



Same means, variances and covariances...
so same least-squares regression results:

```
Call: lm(formula = y ~ x, data = Anscombe, subset = type == 1)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-1.921  -0.4558  -0.04136  0.7094  1.839
```

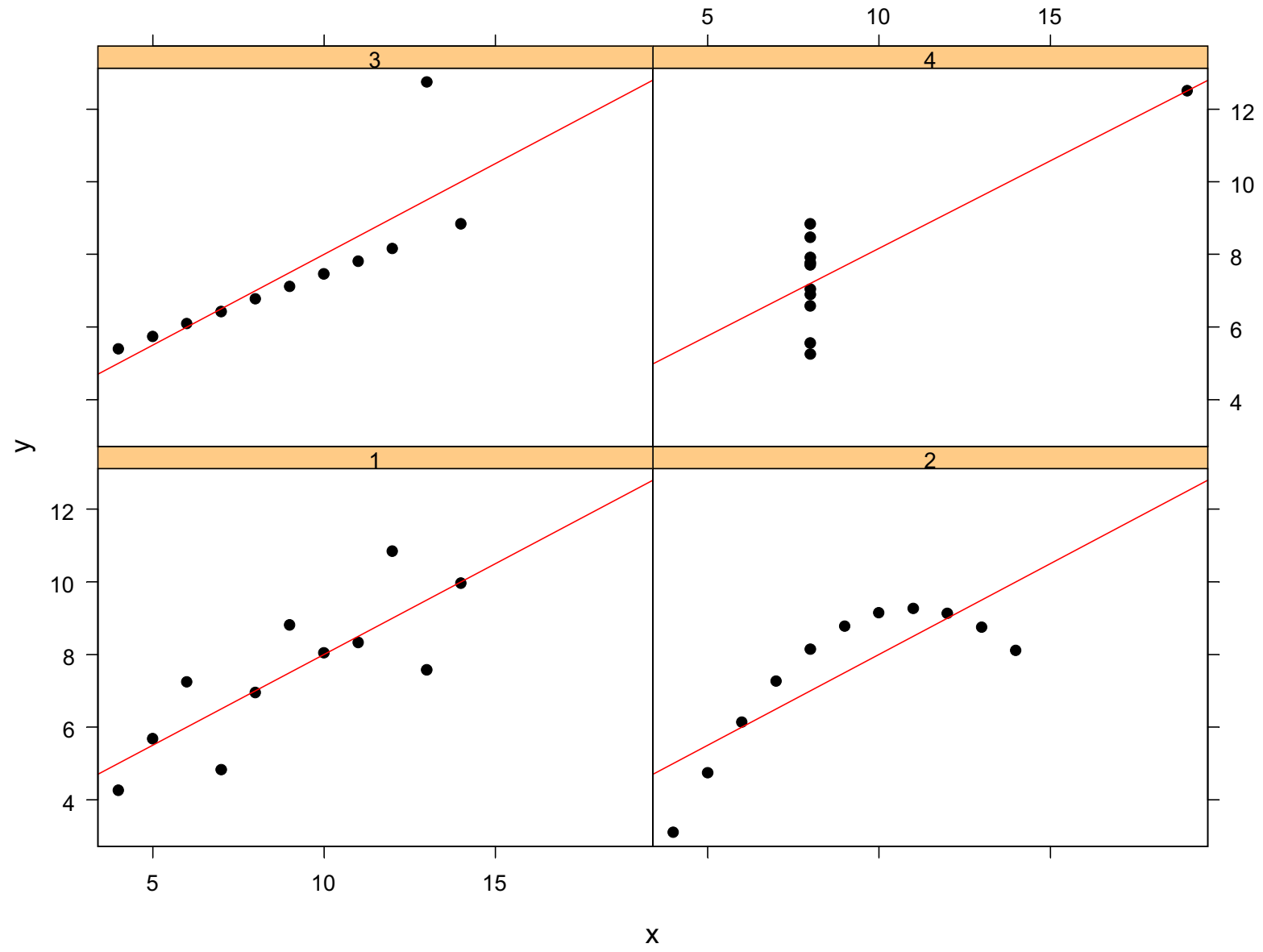
```
Coefficients:
```

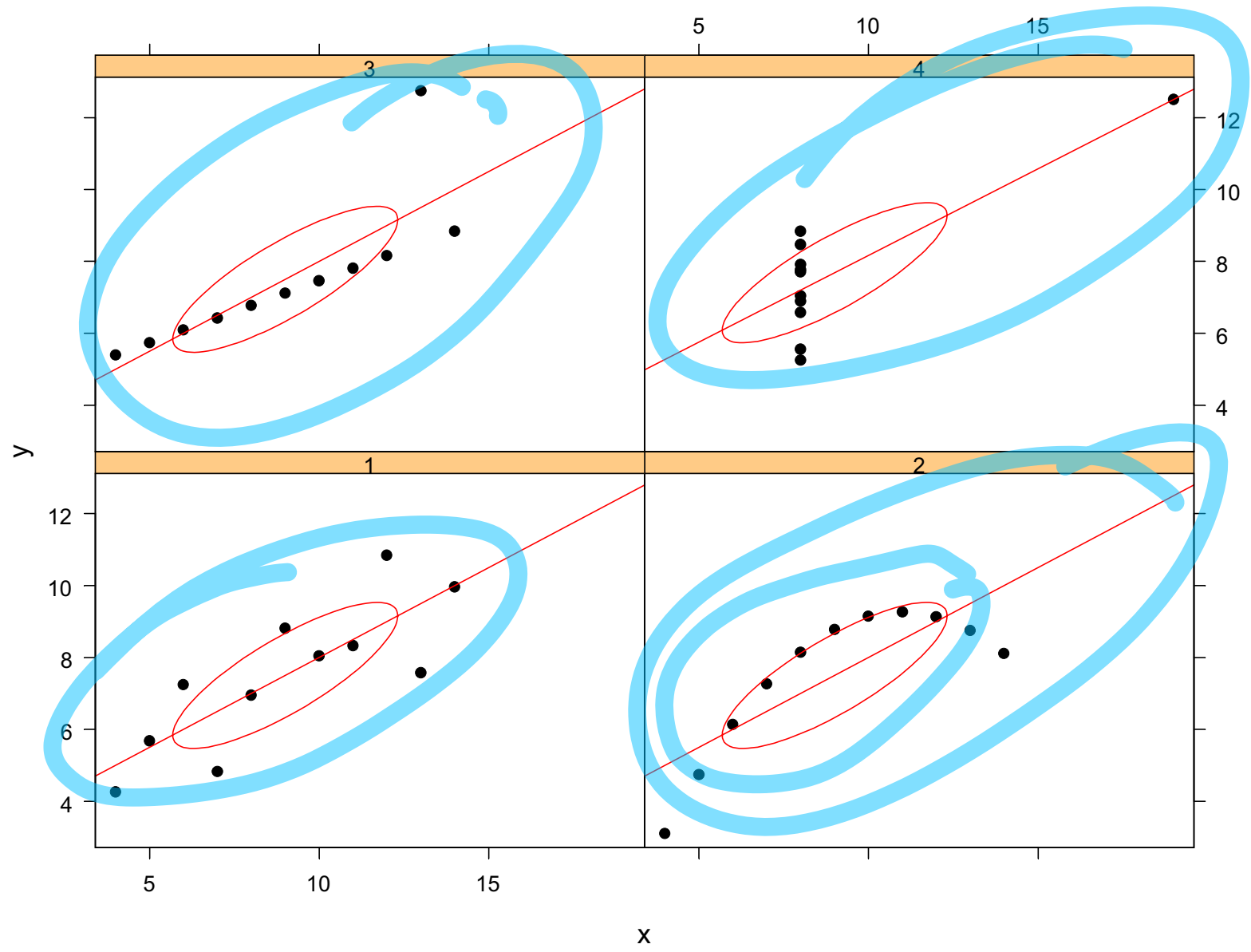
	Value	Std. Error	t value	Pr(> t)
(Intercept)	3.0001	1.1247	2.6673	0.0257
x	0.5001	0.1179	4.2415	0.0022

```
Residual standard error: 1.237 on 9 degrees of freedom
```

```
Multiple R-Squared: 0.6665
```

```
F-statistic: 17.99 on 1 and 9 degrees of freedom,  
the p-value is 0.00217
```





5 Influence and Andrews' Configuration

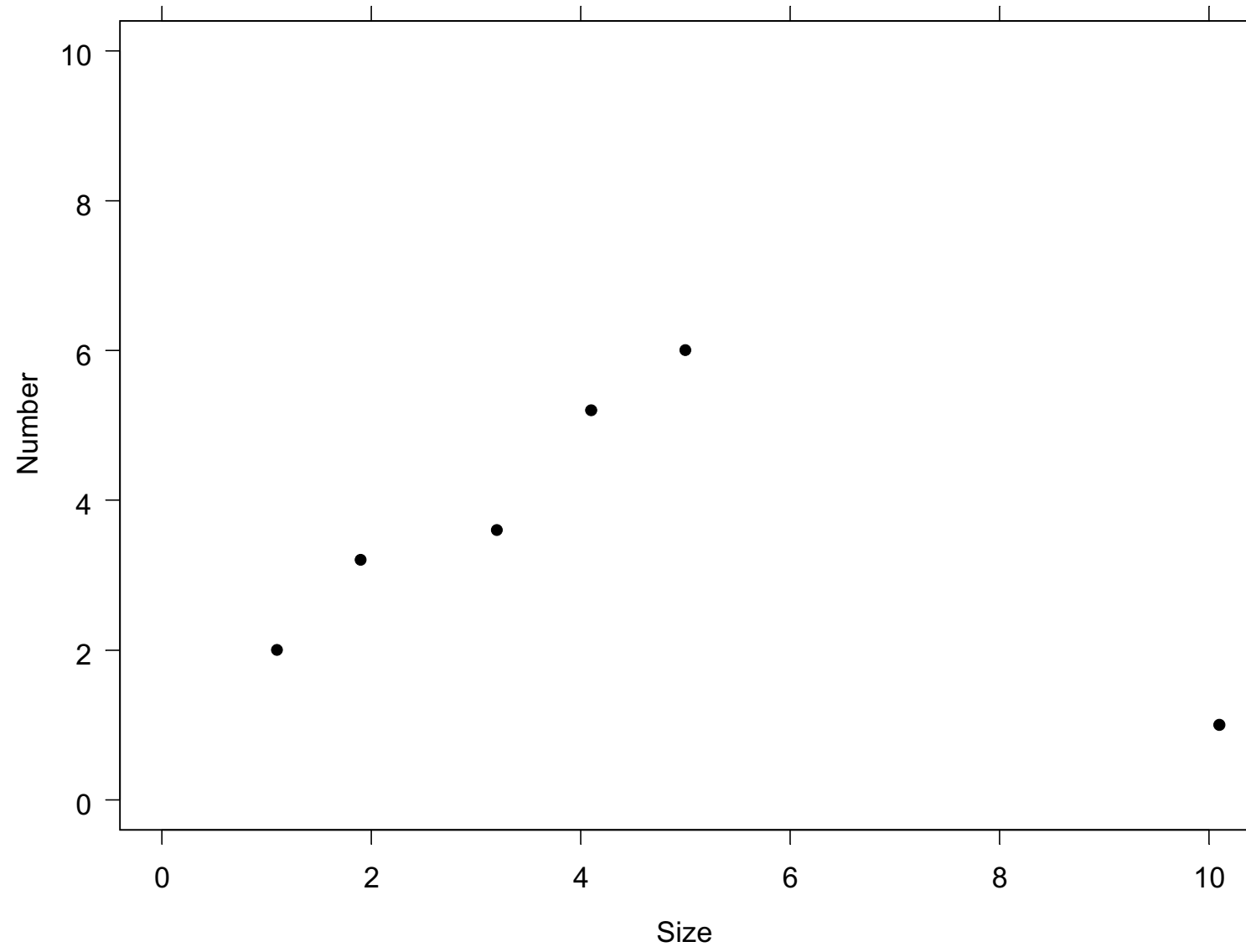
Darwin's data on Galapagos Islands:

X: size of island

Y: number of finch species

Hypothesis: X and Y should have a positive relationship if a larger area provides more evolutionary niches

Idealized version of Darwin's data = Andrews Configuration



Least-squares regression:

```
> fit <- lm(Number ~Size, Andrews)
> summary(fit)
```

```
Call: lm(formula = Number ~Size, data = Andrews)
```

```
Residuals:
```

```
      1      2      3      4      5      6
-1.969 -0.649 -0.05454 1.68 2.615 -1.623
```

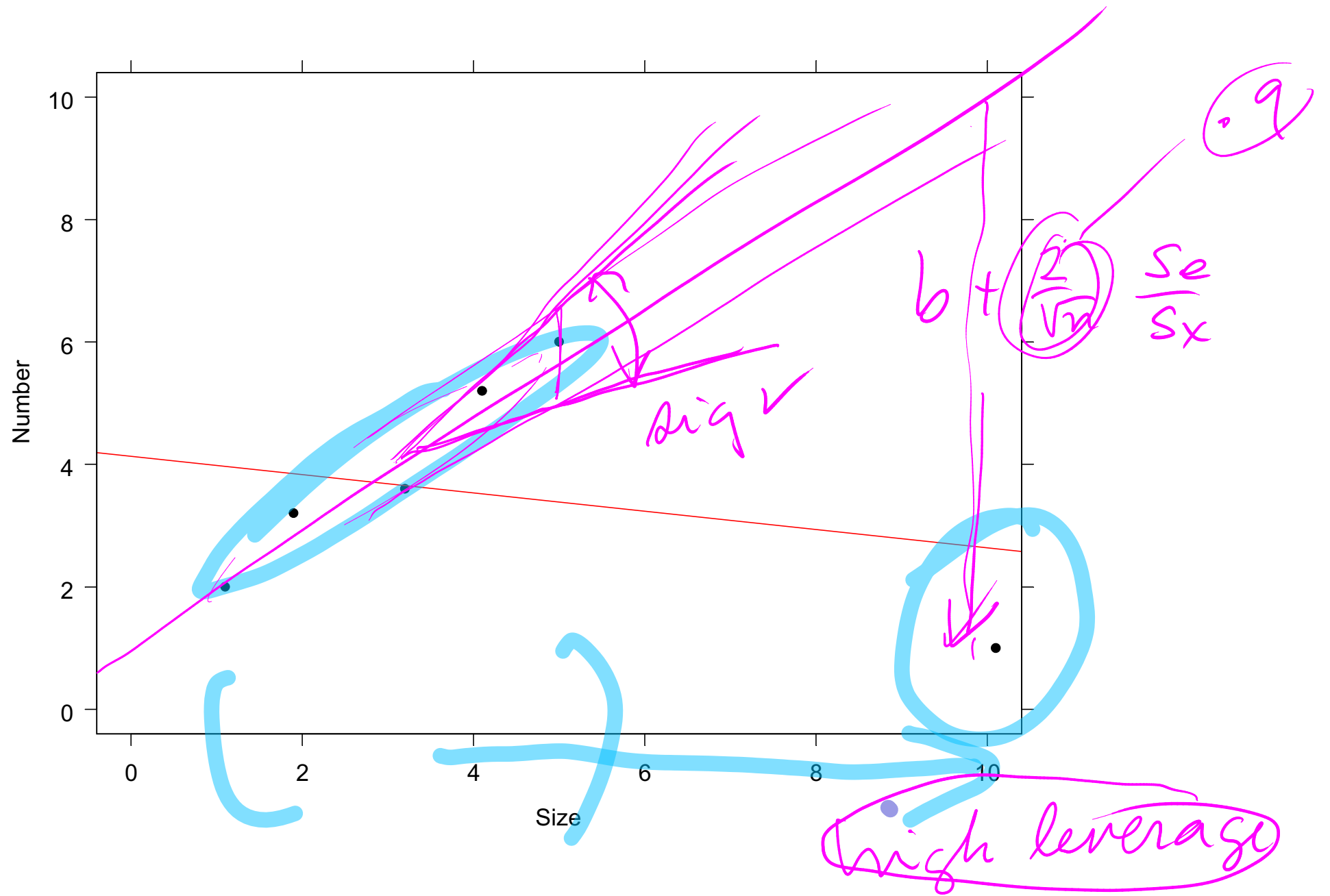
```
Coefficients:
```

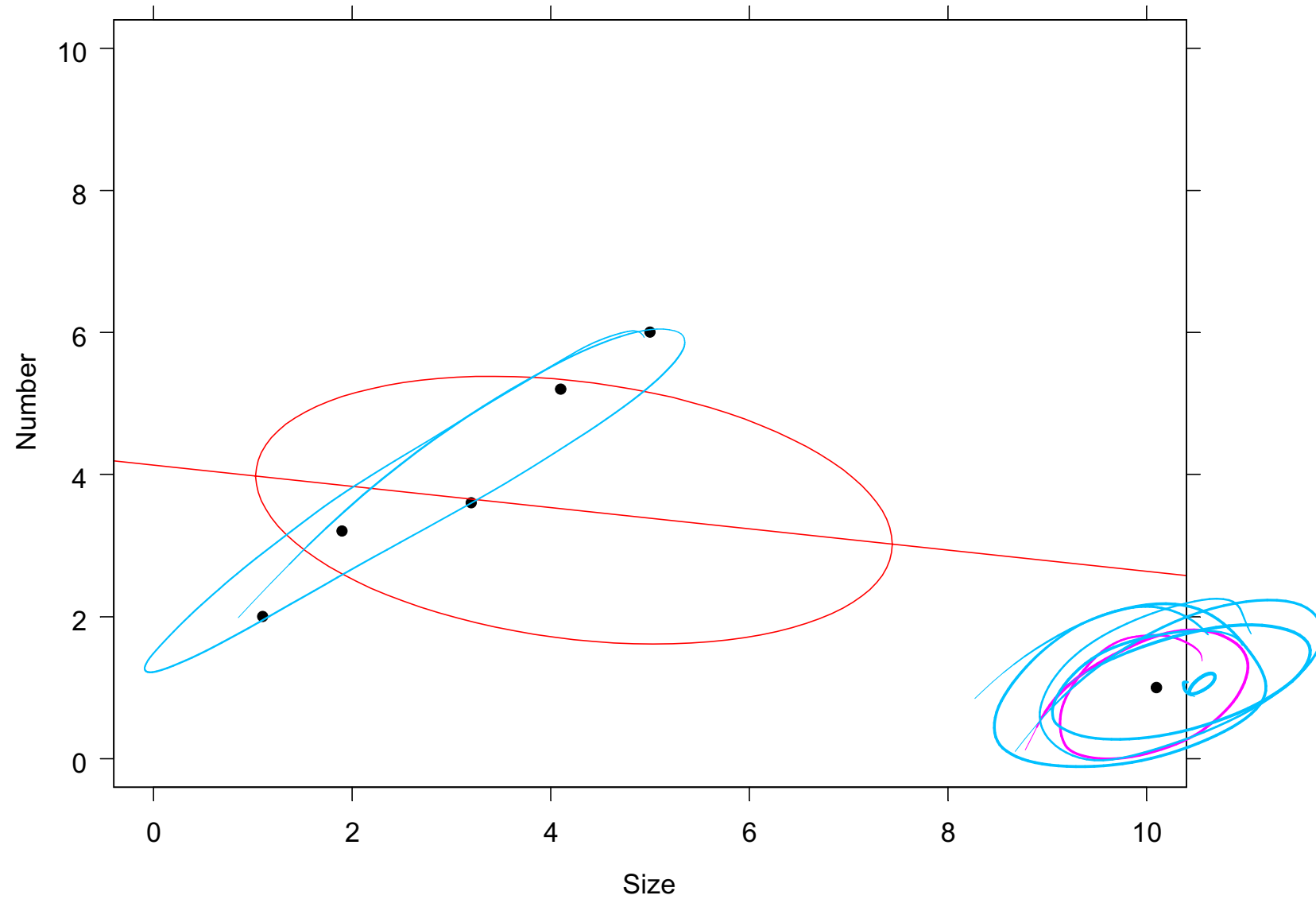
	Value	Std. Error	t value	Pr(> t)
(Intercept)	4.1331	1.4625	2.8261	0.0475
Size	-0.1496	0.2842	-0.5262	0.6266

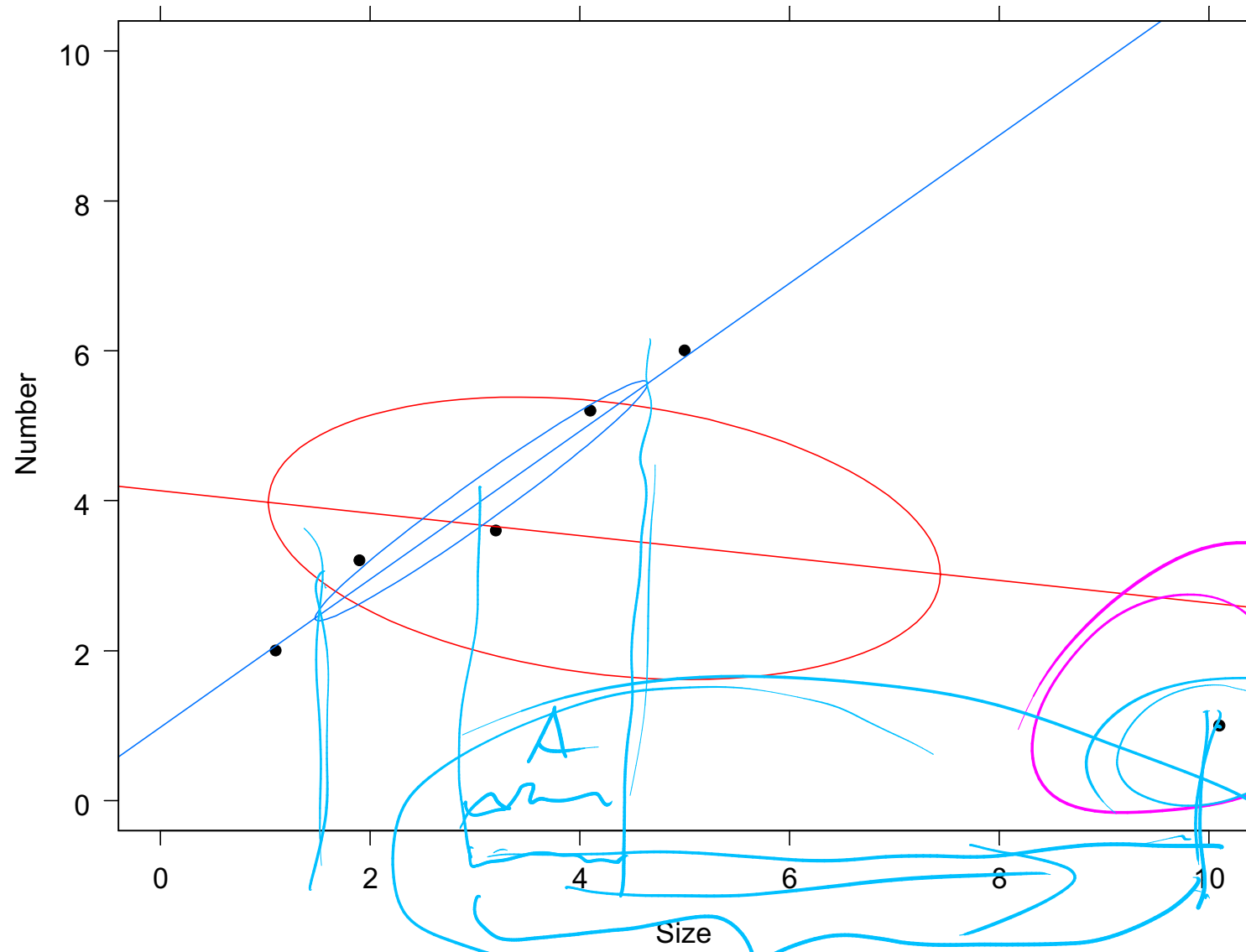
```
Residual standard error: 2.037 on 4 degrees of freedom
```

```
Multiple R-Squared: 0.06474
```

```
F-statistic: 0.2769 on 1 and 4 degrees of freedom,
the p-value is 0.6266
```

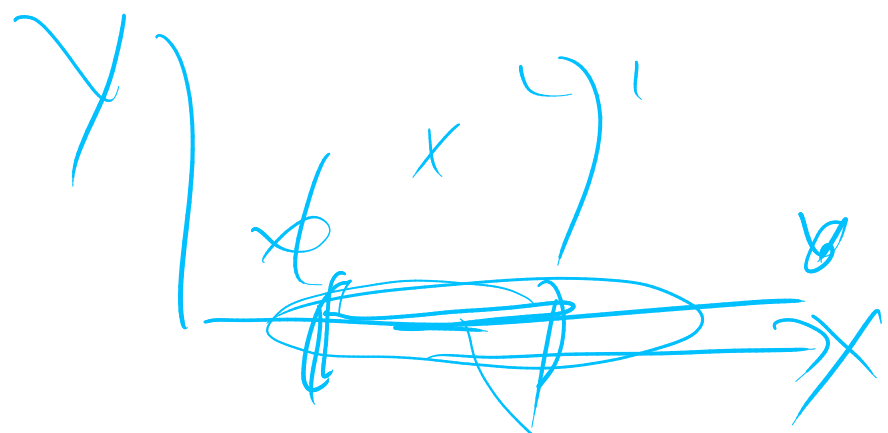







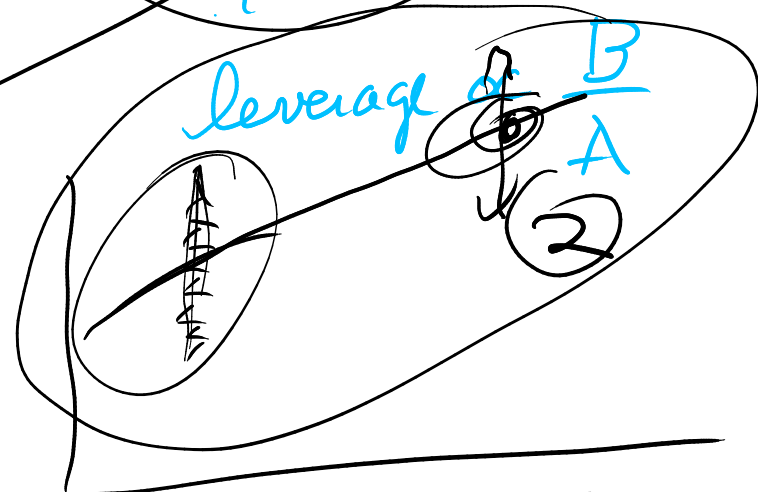
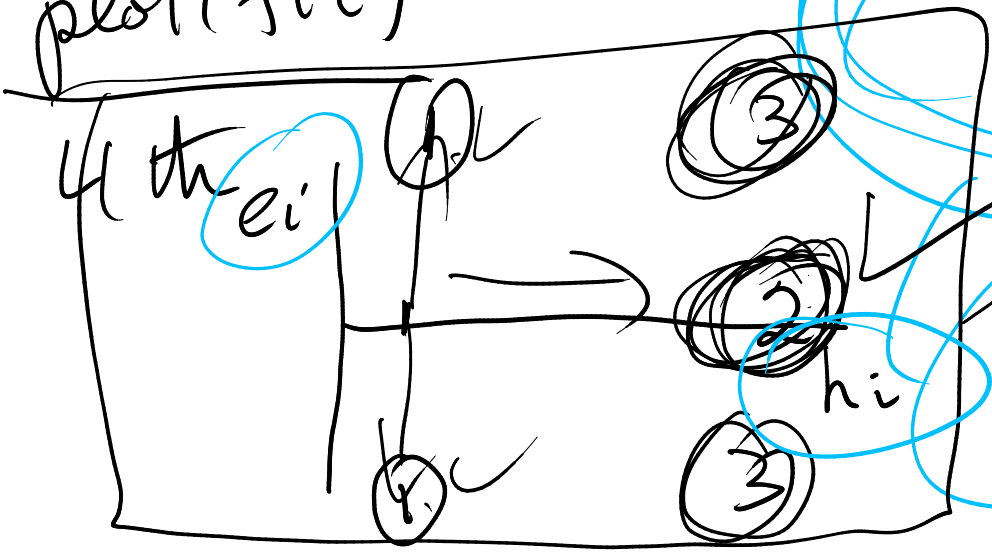
$$h_i = \frac{\frac{B}{A} - \frac{1}{2}}{1 + \frac{B}{A}}$$

$\frac{B}{A}$ or leverage



```
fit <- lm(y ~ x1 + x2, data)
```

```
plot(fit)
```



$$\hat{\beta} = (X'X)^{-1} X'Y$$

SE ~

$$SE(\hat{\beta}) = \frac{1}{\sqrt{n}} \frac{Se}{I_2} = \infty$$

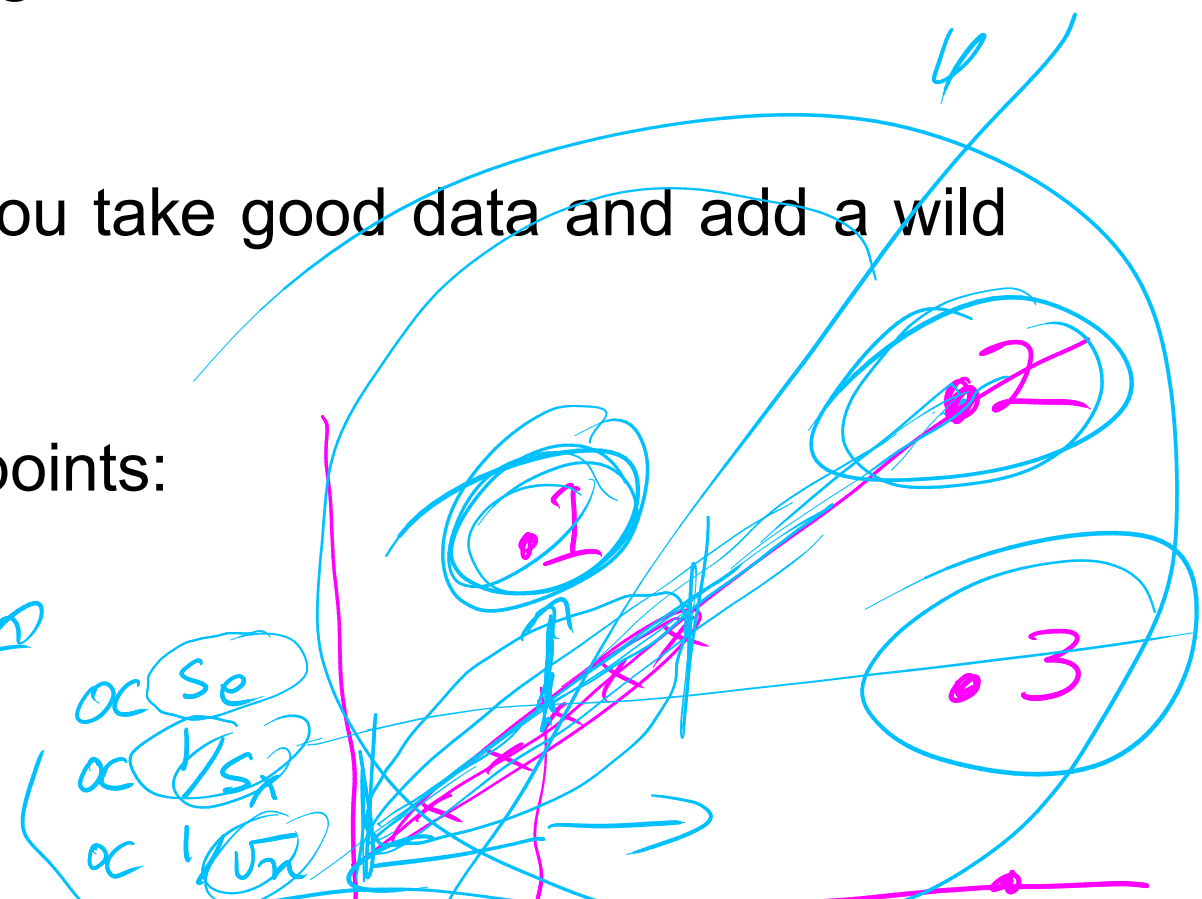
= 0

6 Influence and Leverage

What happens when you take good data and add a wild point?

Three archetypal wild points:

- Typical X, bad fit 1
- Unusual X, good fit 2
- Unusual X, bad fit 3

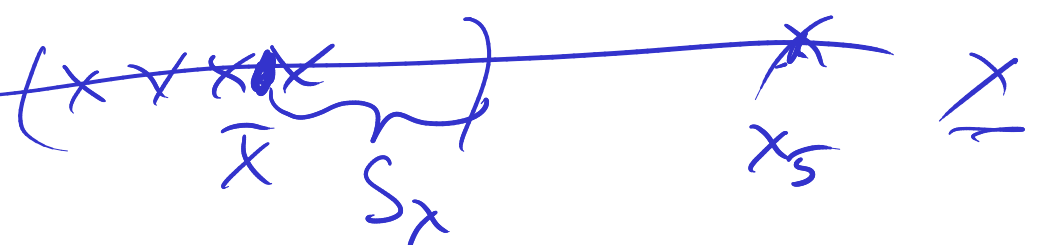


X	fit	$\hat{\beta}$	Se	S_x	$SE(\hat{\beta}) = \frac{1}{\sqrt{n}} \frac{Se}{S_x}$	$ t = \hat{\beta} / SE(\hat{\beta}) $	P
1	ok	bad	\uparrow	$\Delta \sim 0$	\uparrow	\downarrow	\uparrow
2	far	ok	$\Delta \sim 0$	\uparrow	\downarrow	\uparrow	\downarrow
3	far	bad	$\uparrow \downarrow$	\uparrow	\uparrow	\uparrow	\uparrow

fit <- lm(~)
plot(fit)

std
res.

t_i



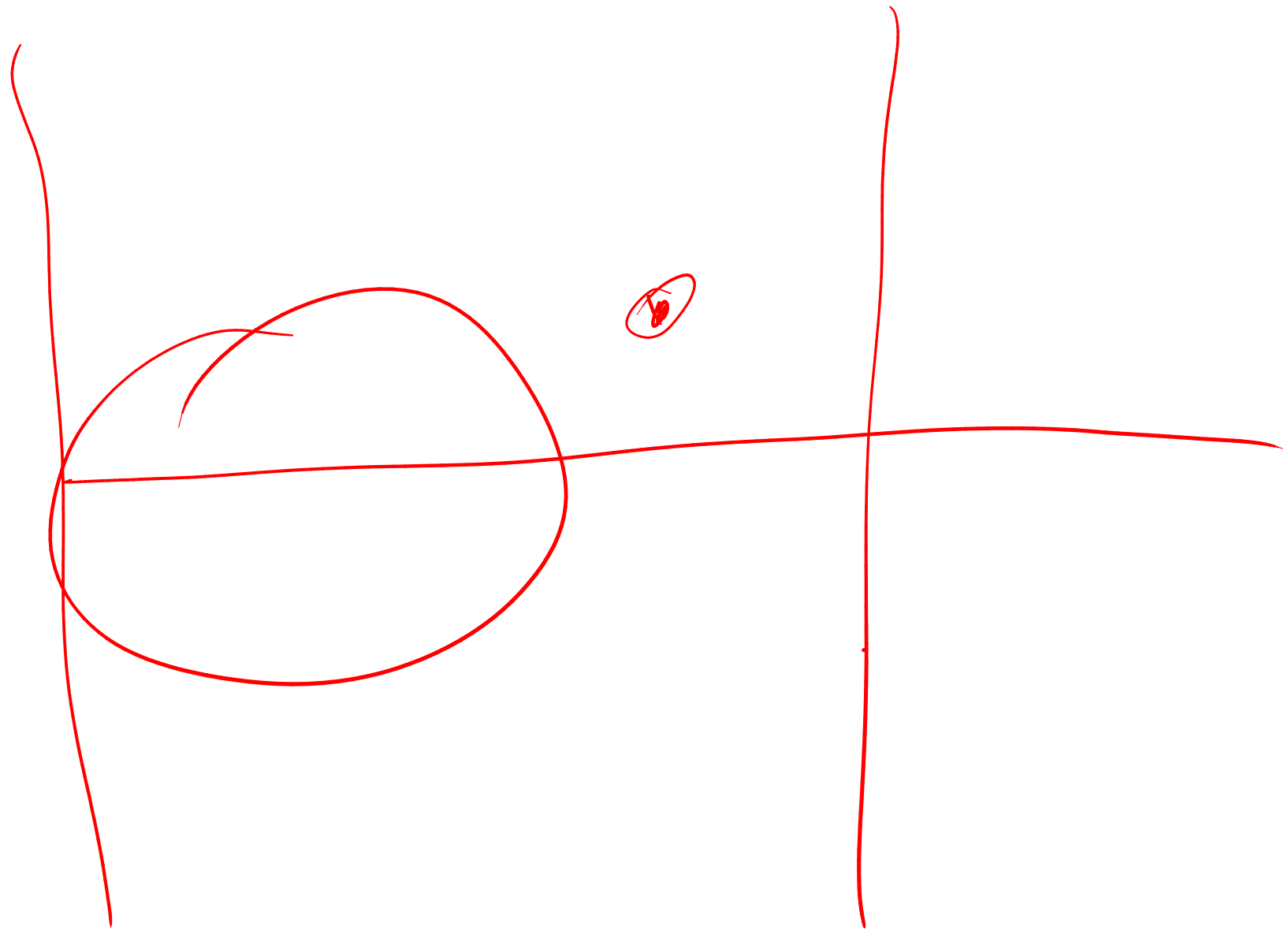
$$\Delta_5 = \frac{x_5 - \bar{x}}{s_x}$$

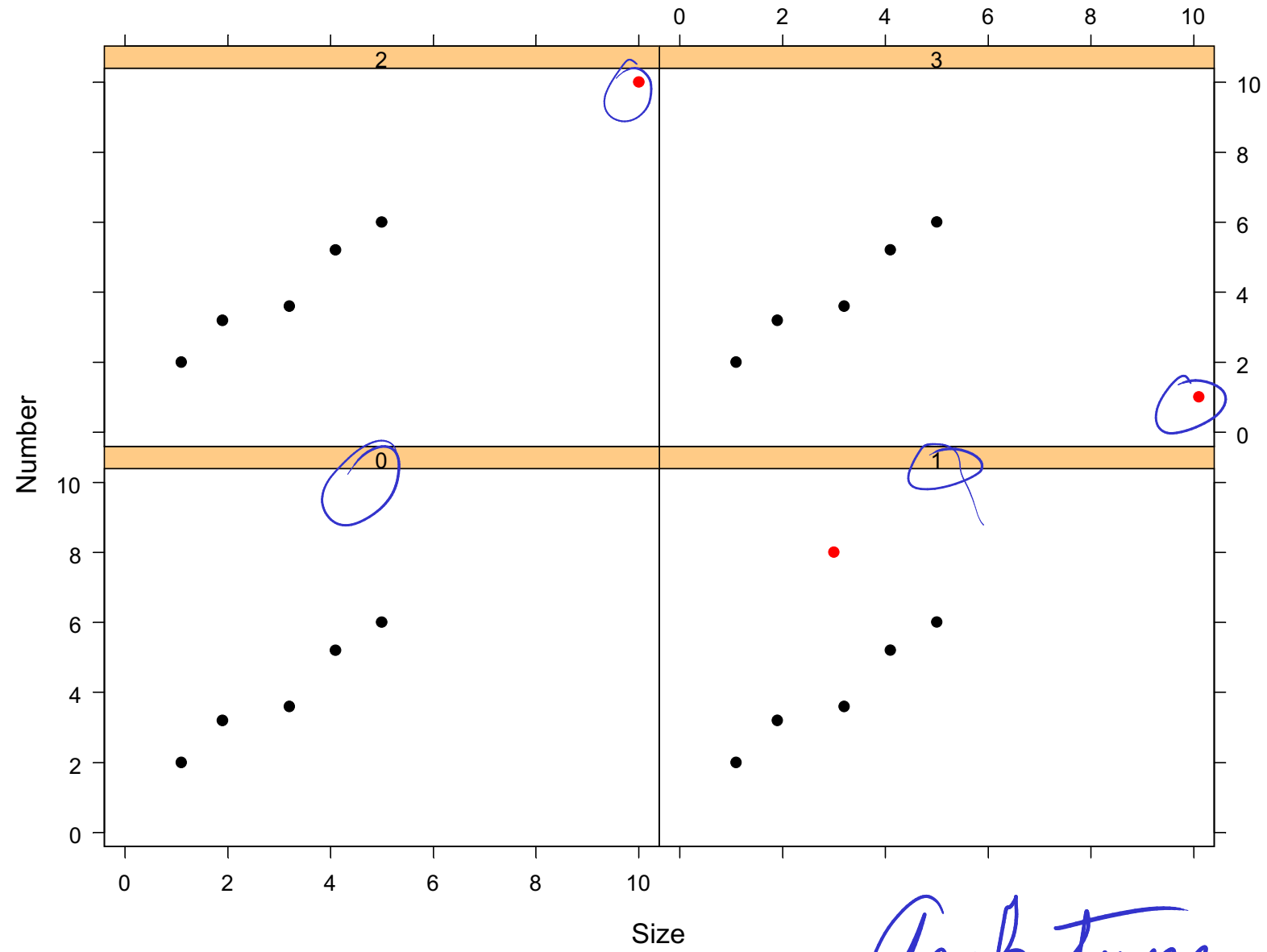
h_{ii} leverage

? Check: $\Delta_i = \frac{h_{ii} - \frac{1}{n}}{1 - h_{ii}}$

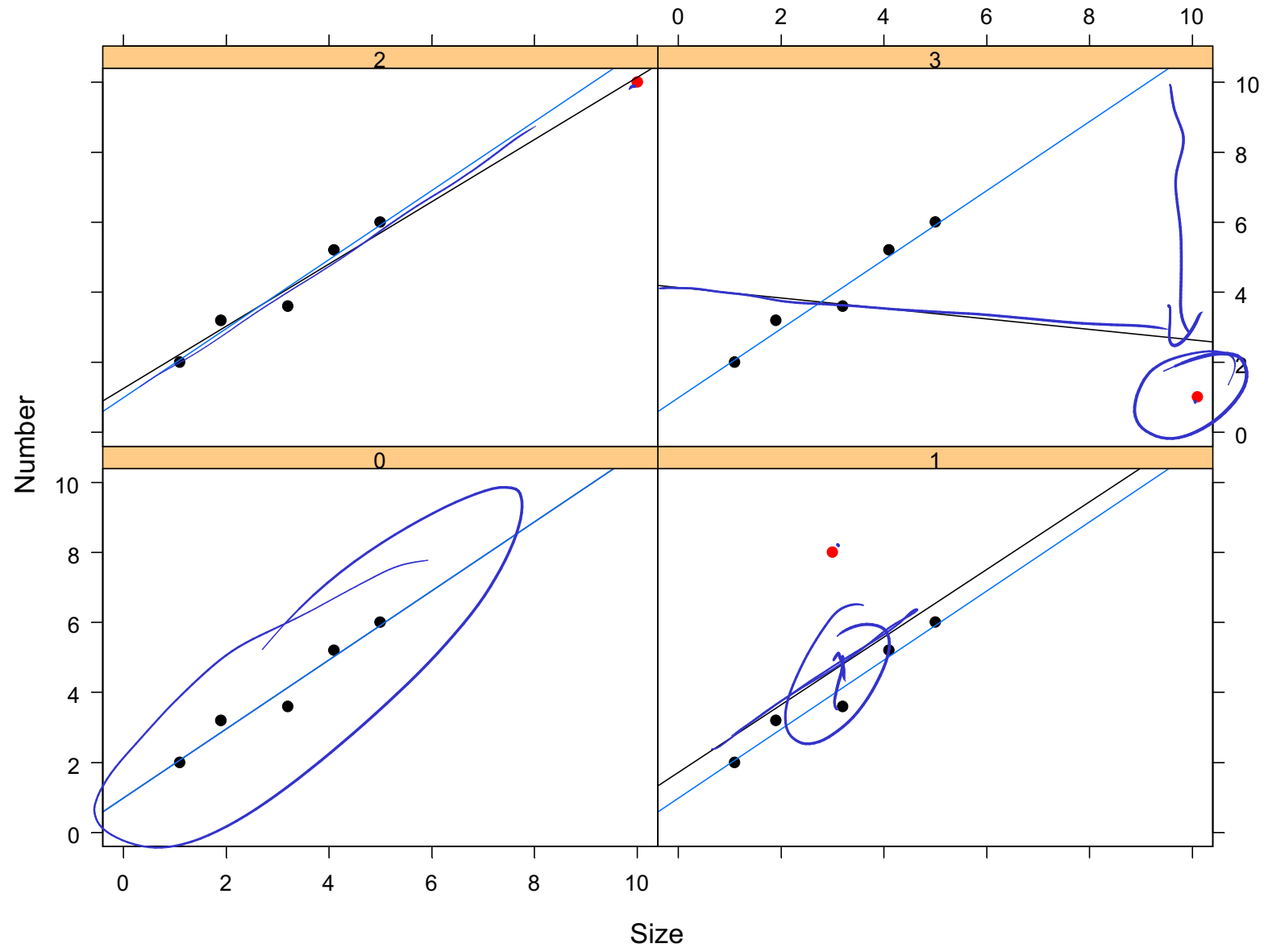
$$H = X(X'X)^{-1}X'$$

h_{ii} diag elts.

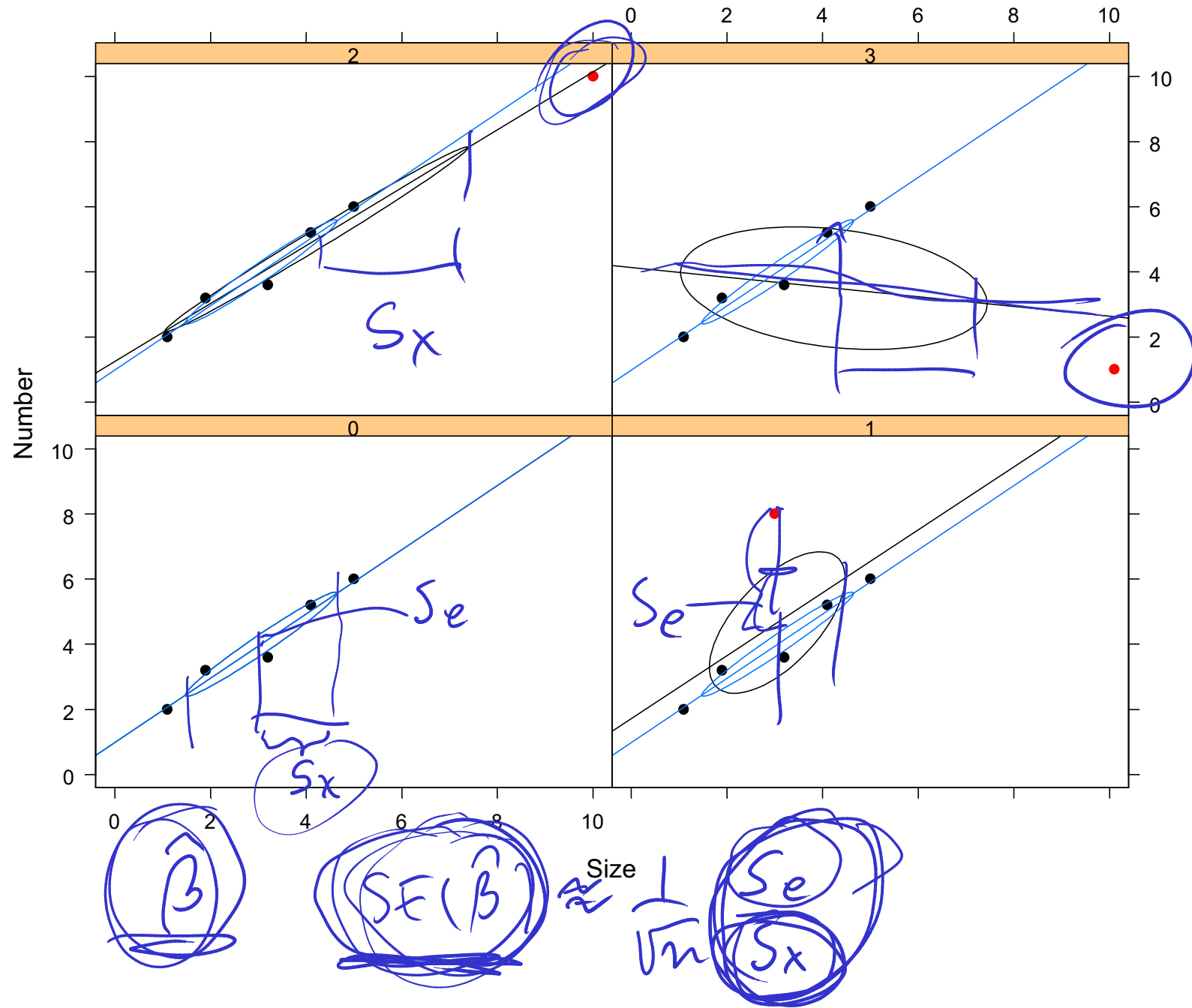


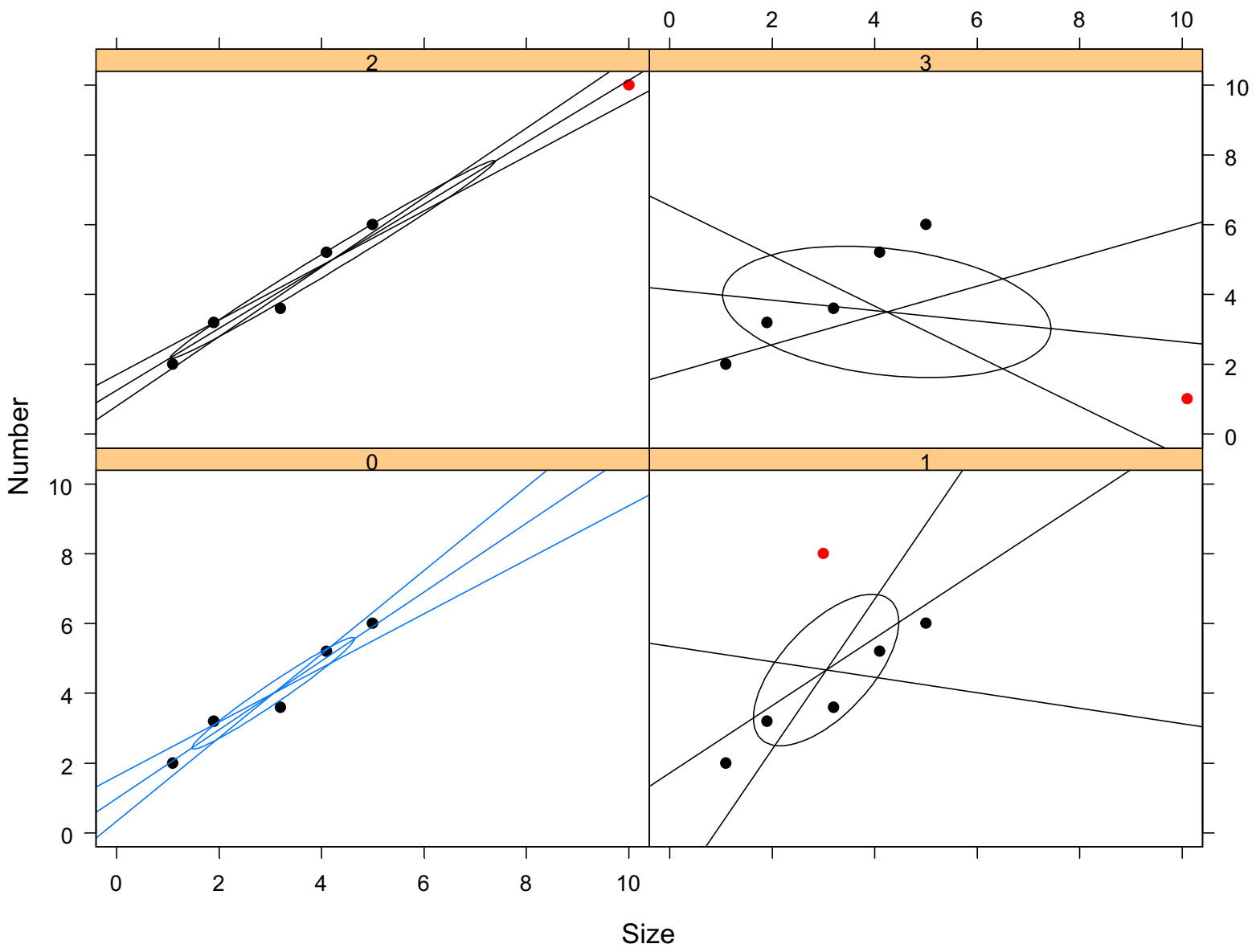


Archetypes



β_x





7 Measurement error

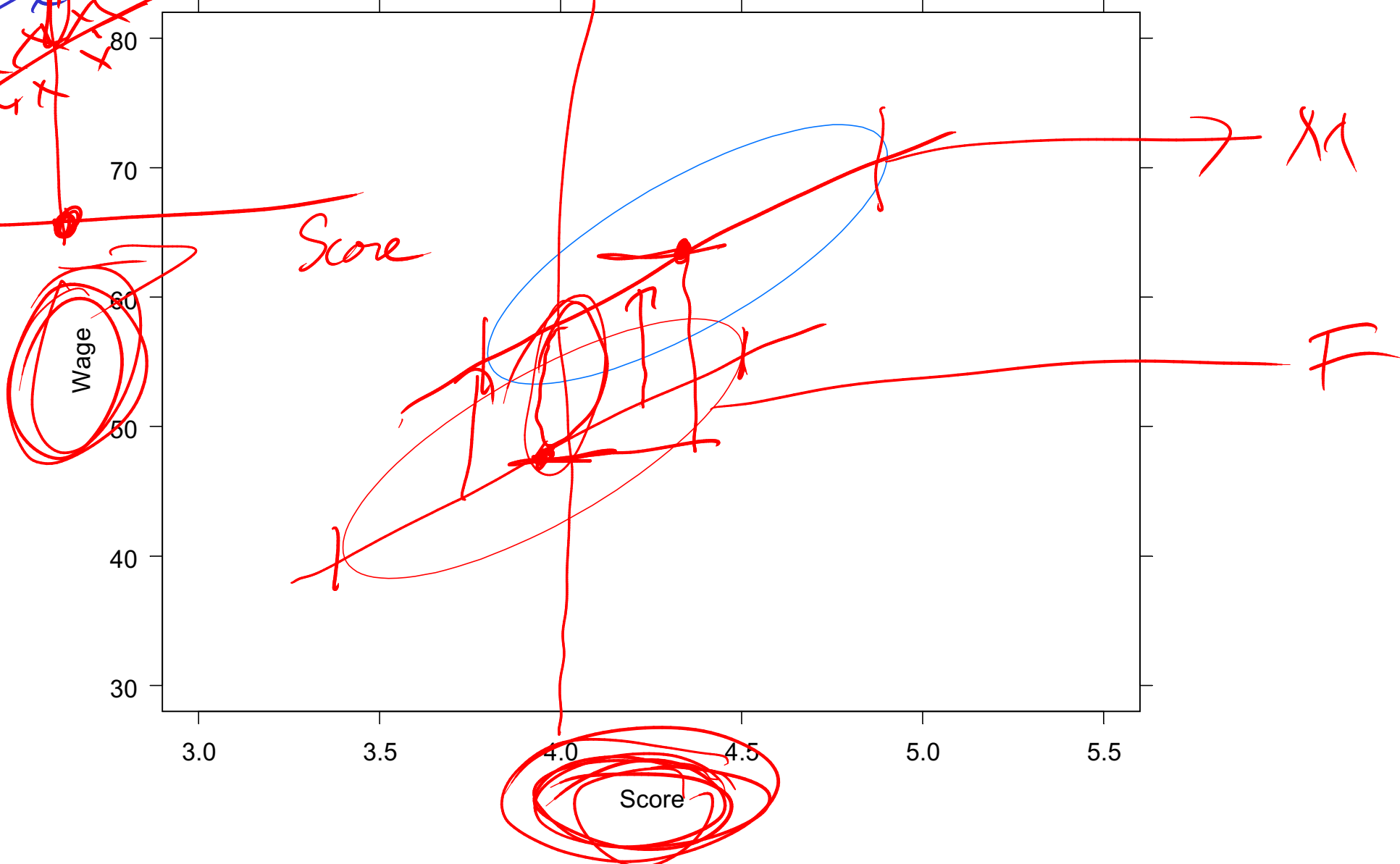
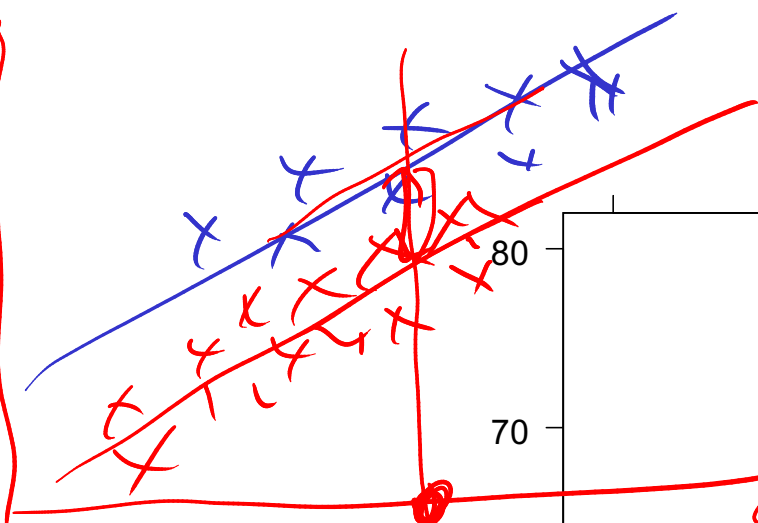
What happens when a covariate is measured with error?

Example: Pay Equity – Estimating the Gender Gap

Wages vs Job Value by Gender

$$Y \sim \text{Cont} + \text{Categ.} \\ \text{2 values}$$

\$



80

70

60

50

40

30

3.0

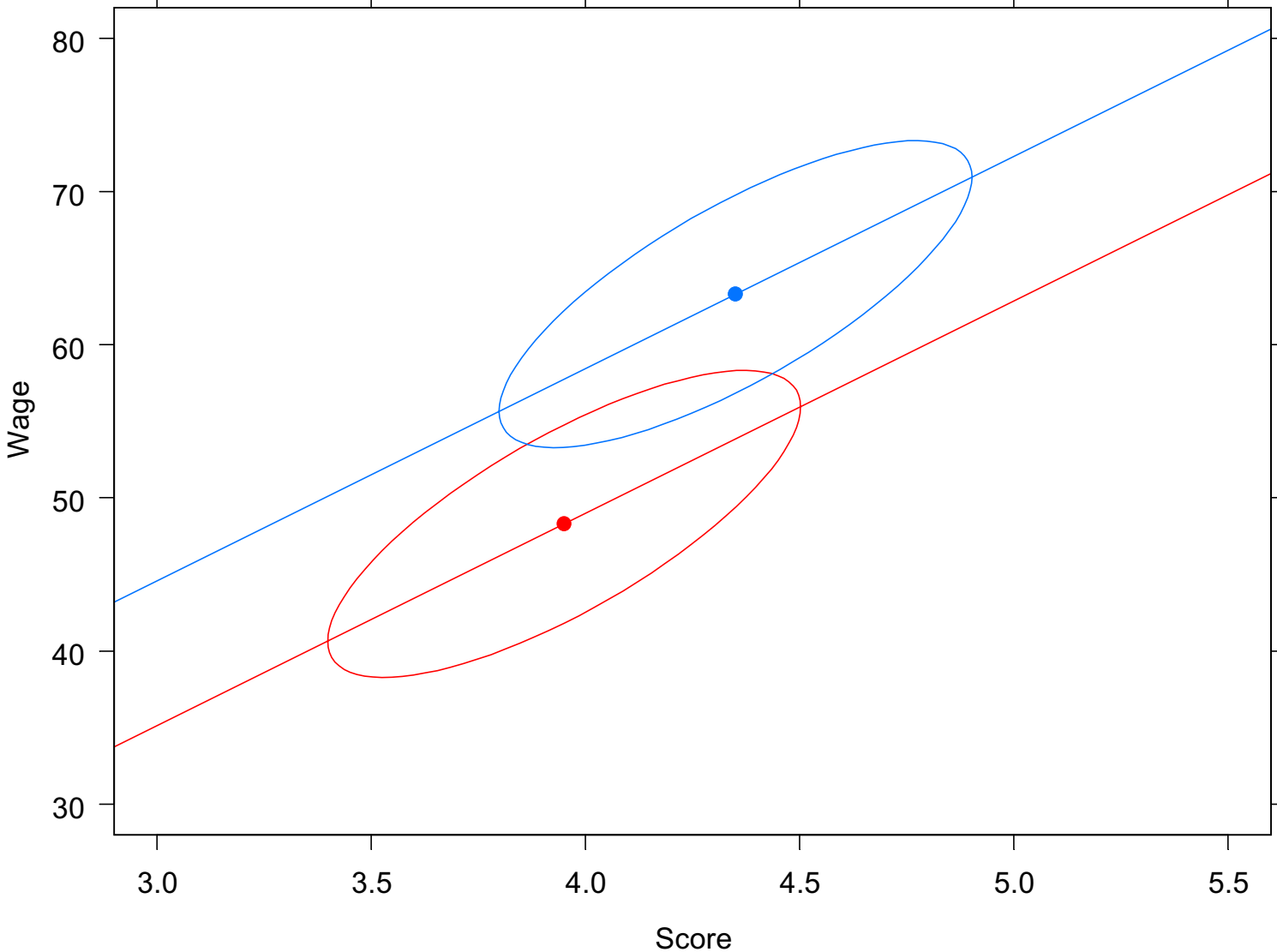
3.5

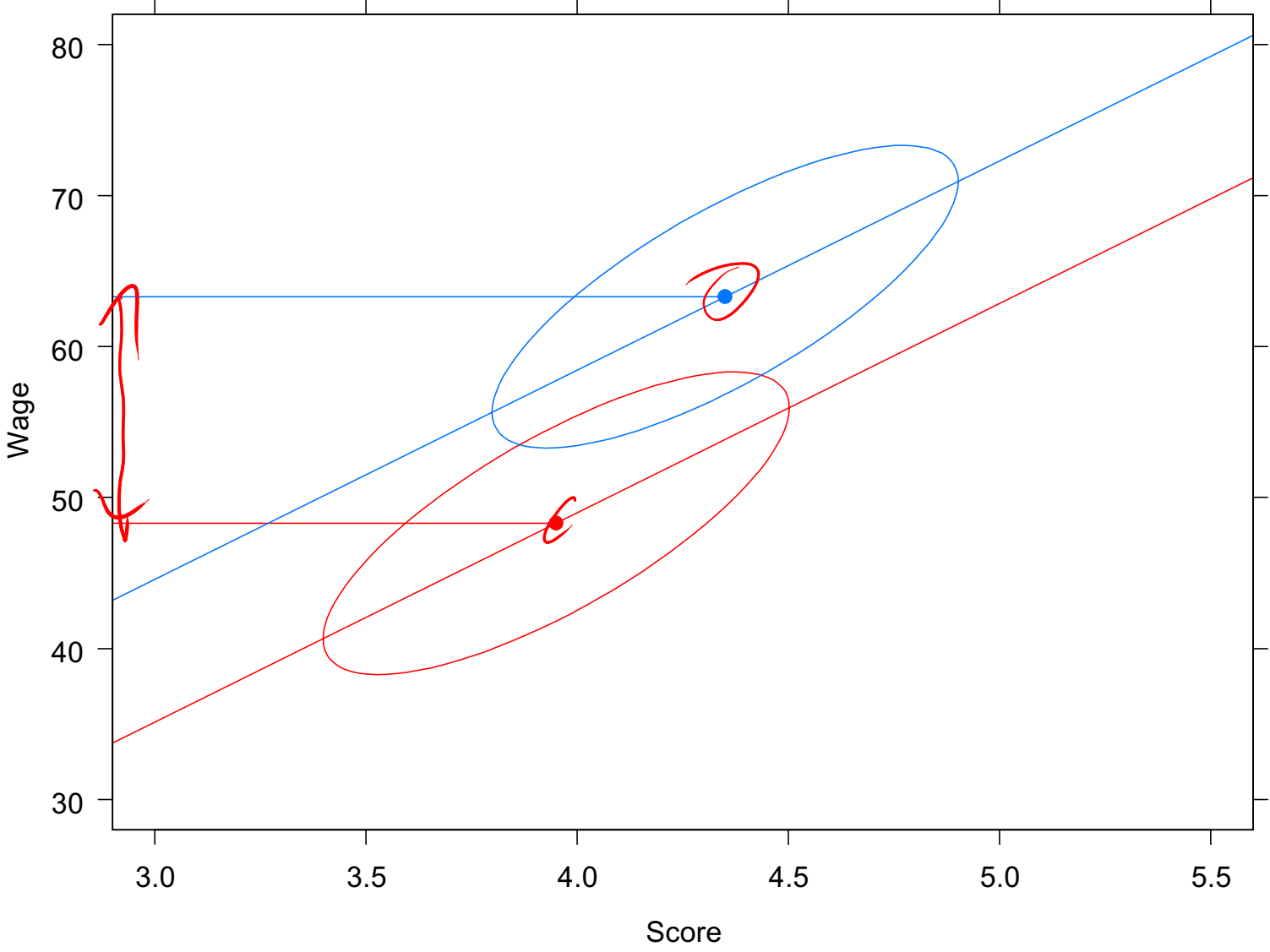
4.0

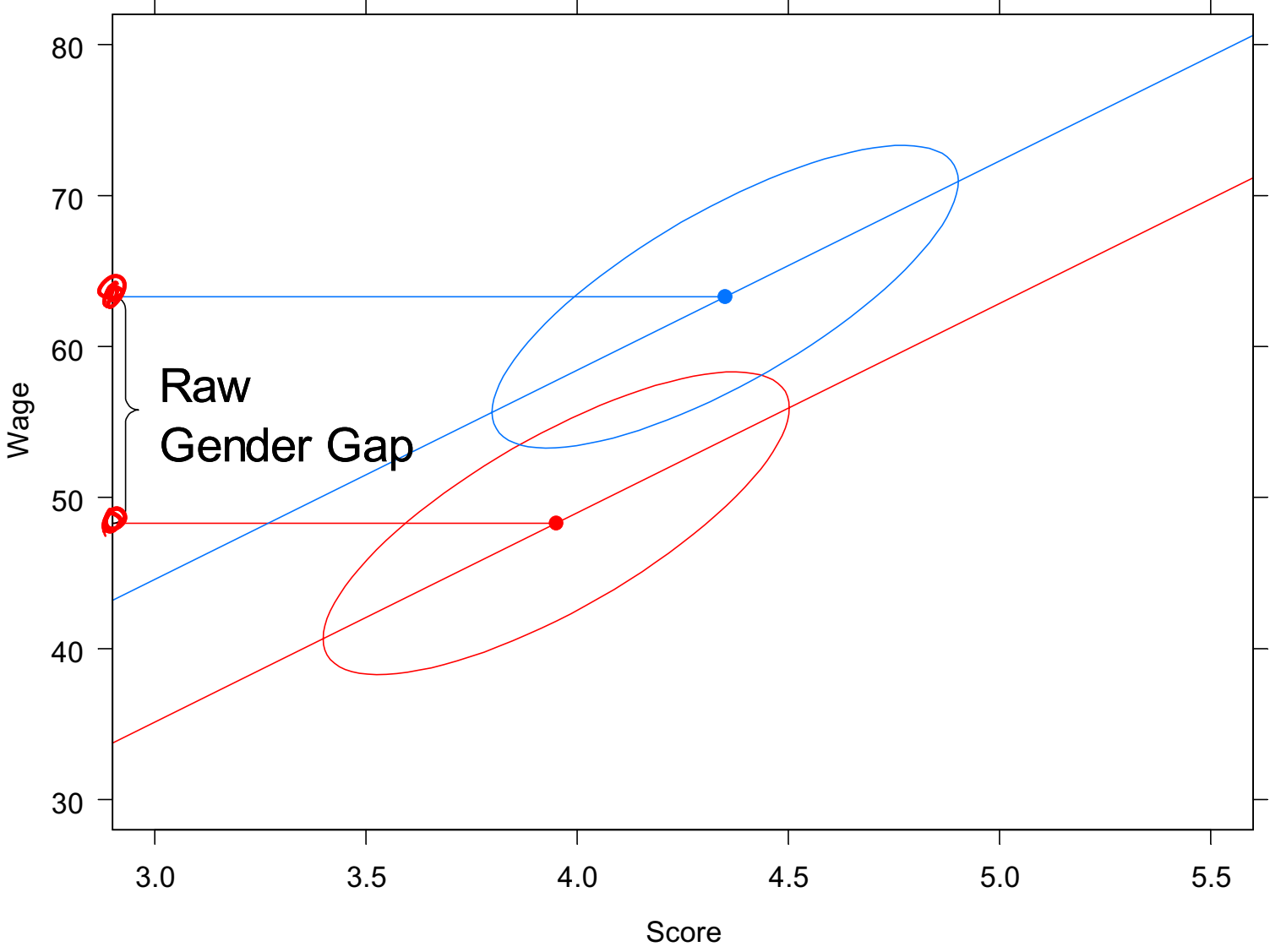
4.5

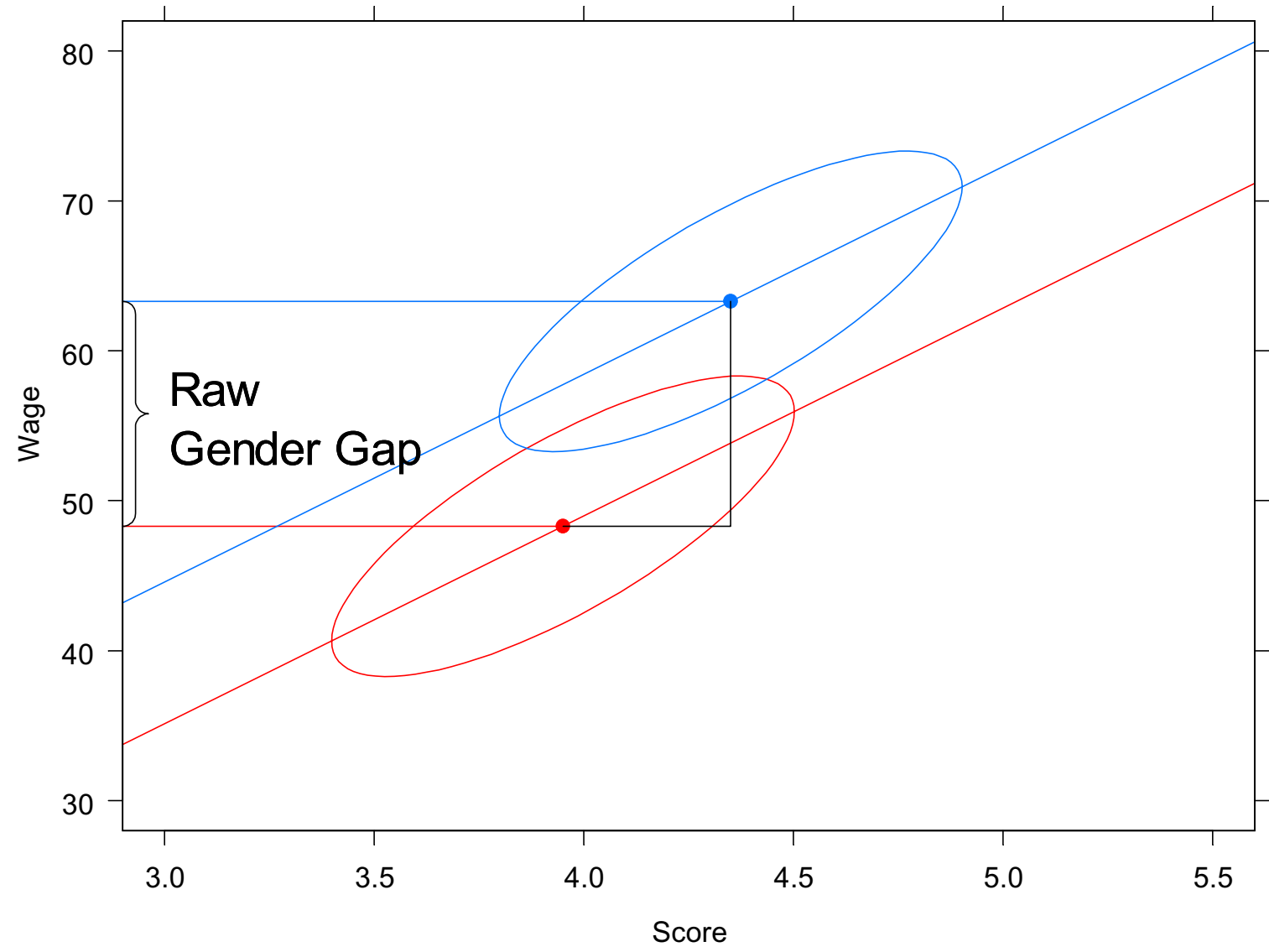
5.0

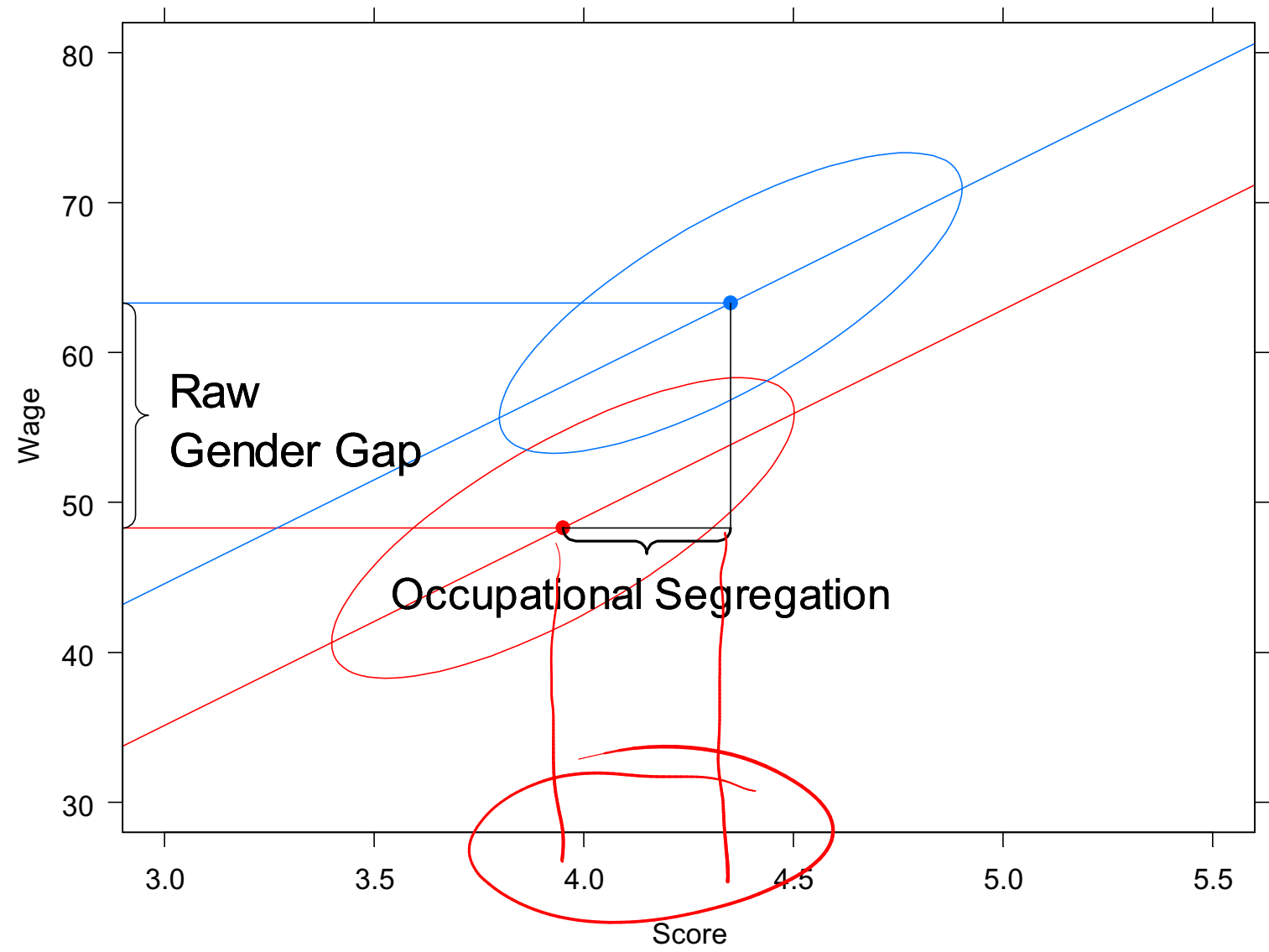
5.5

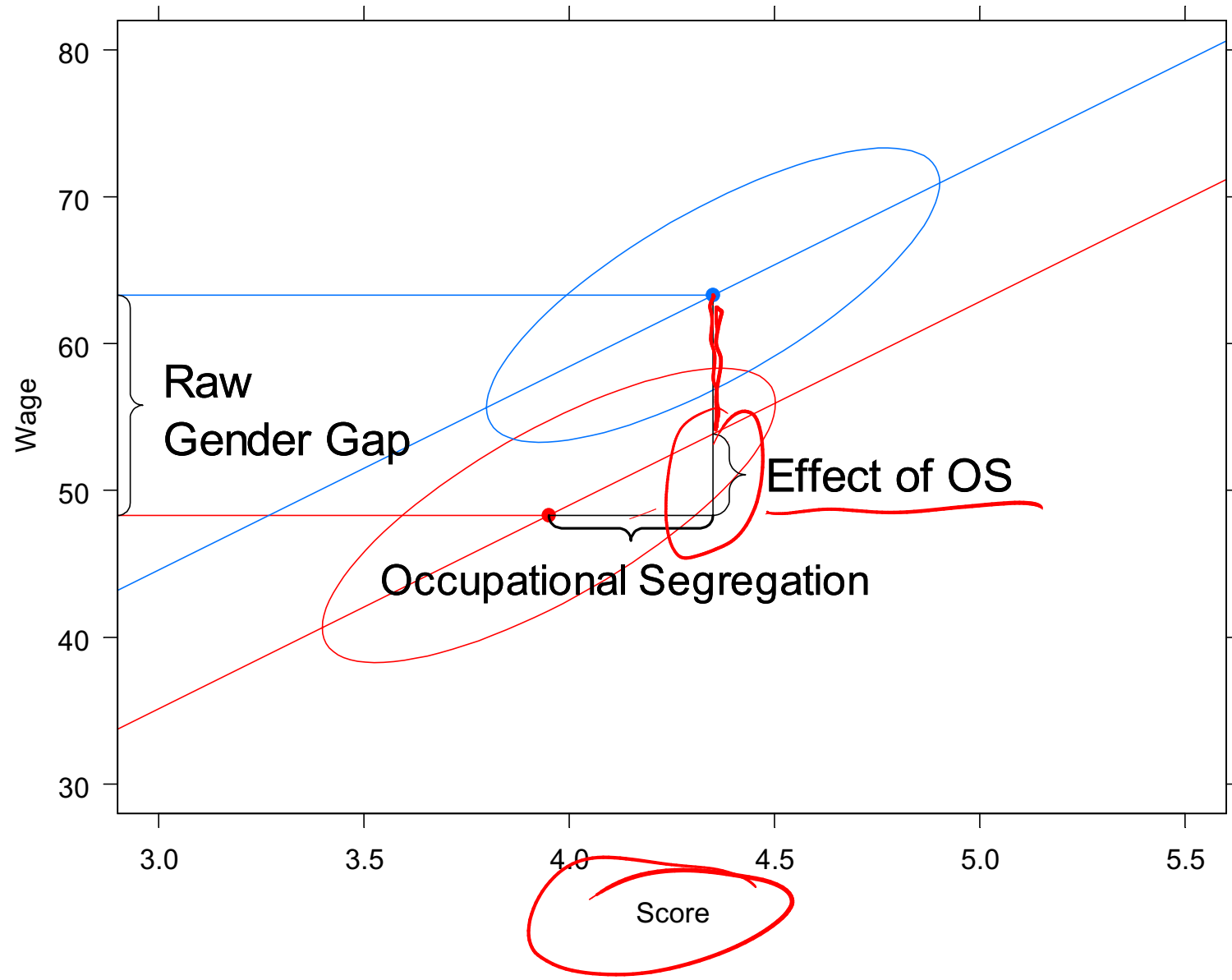


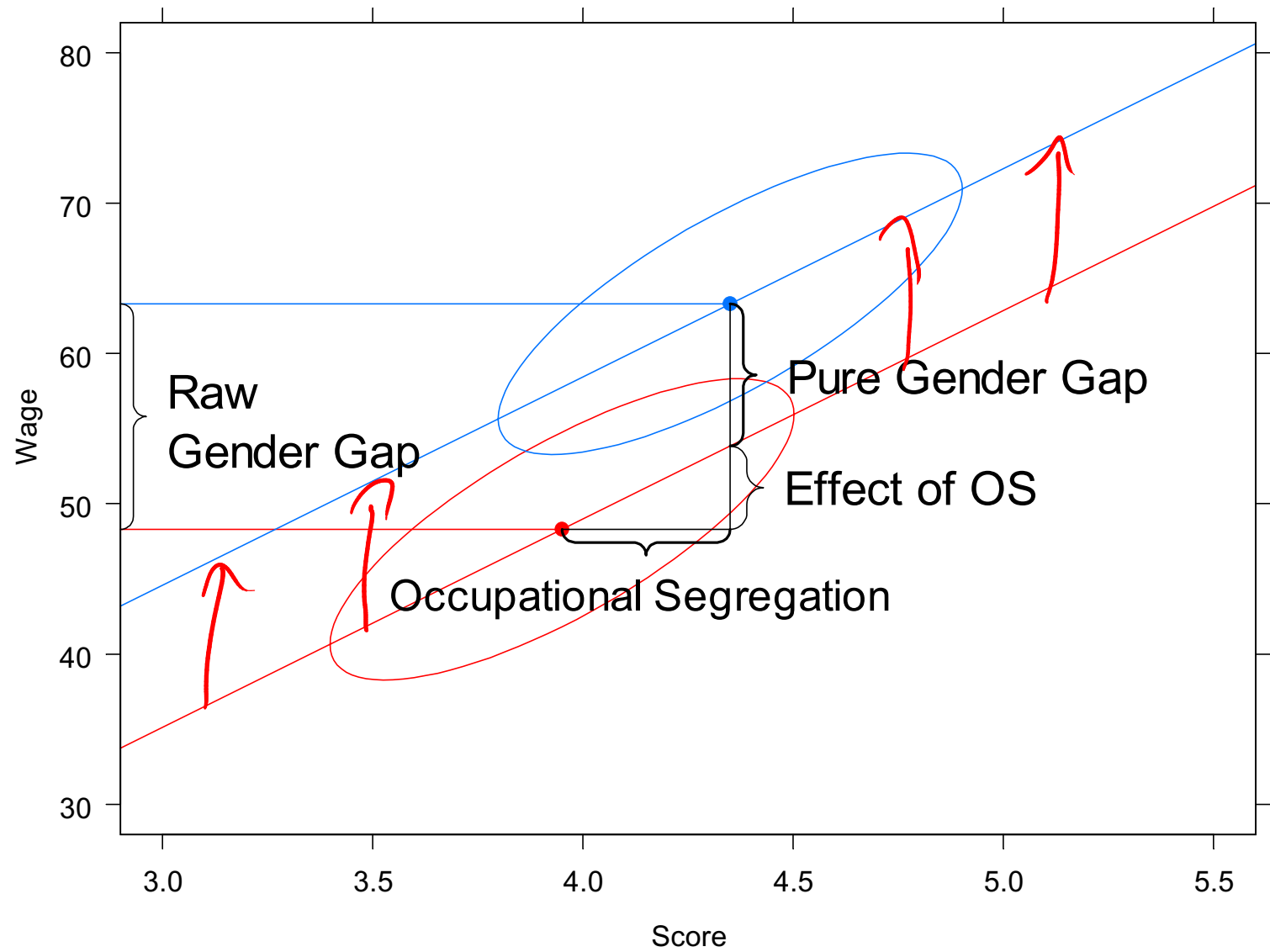


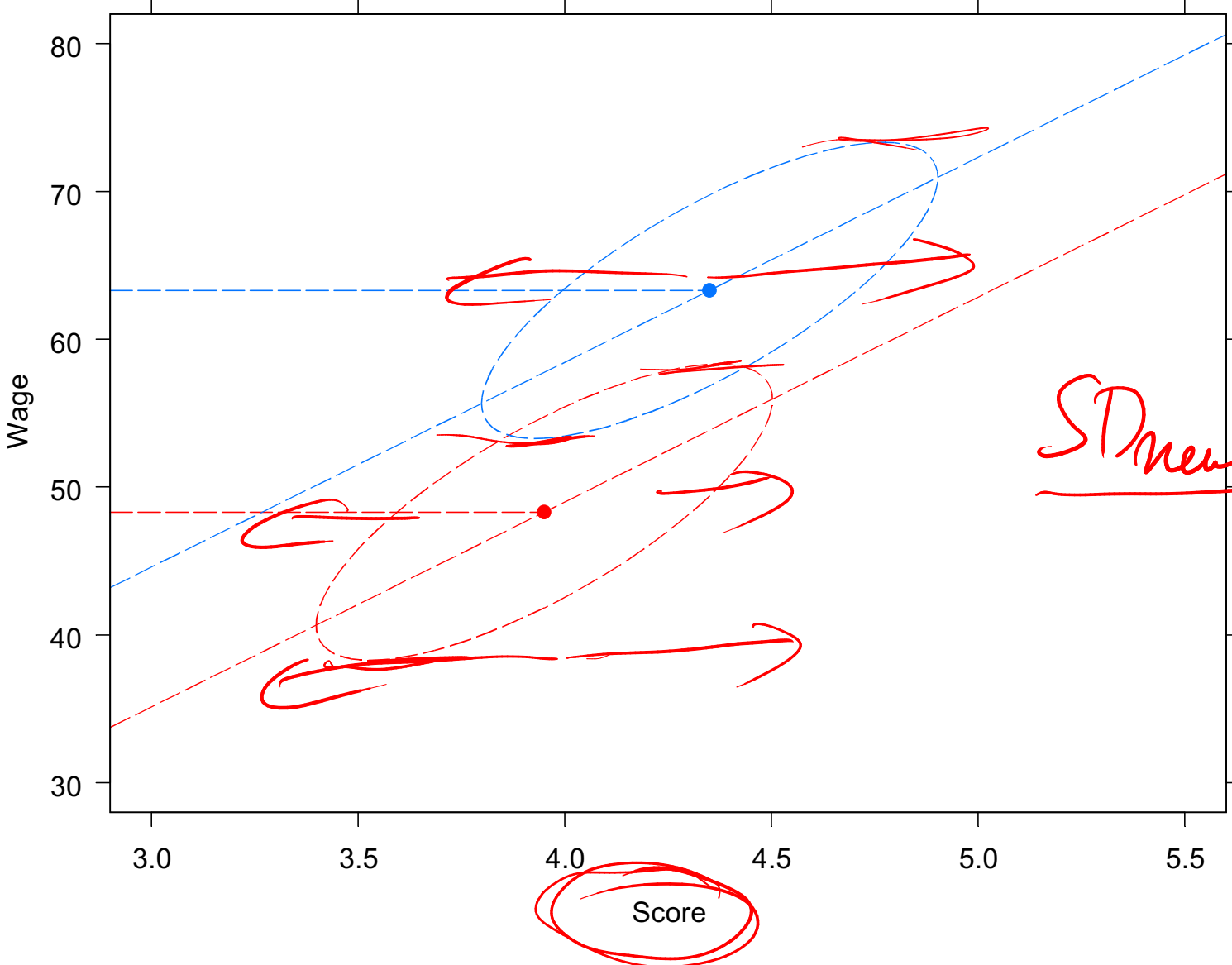




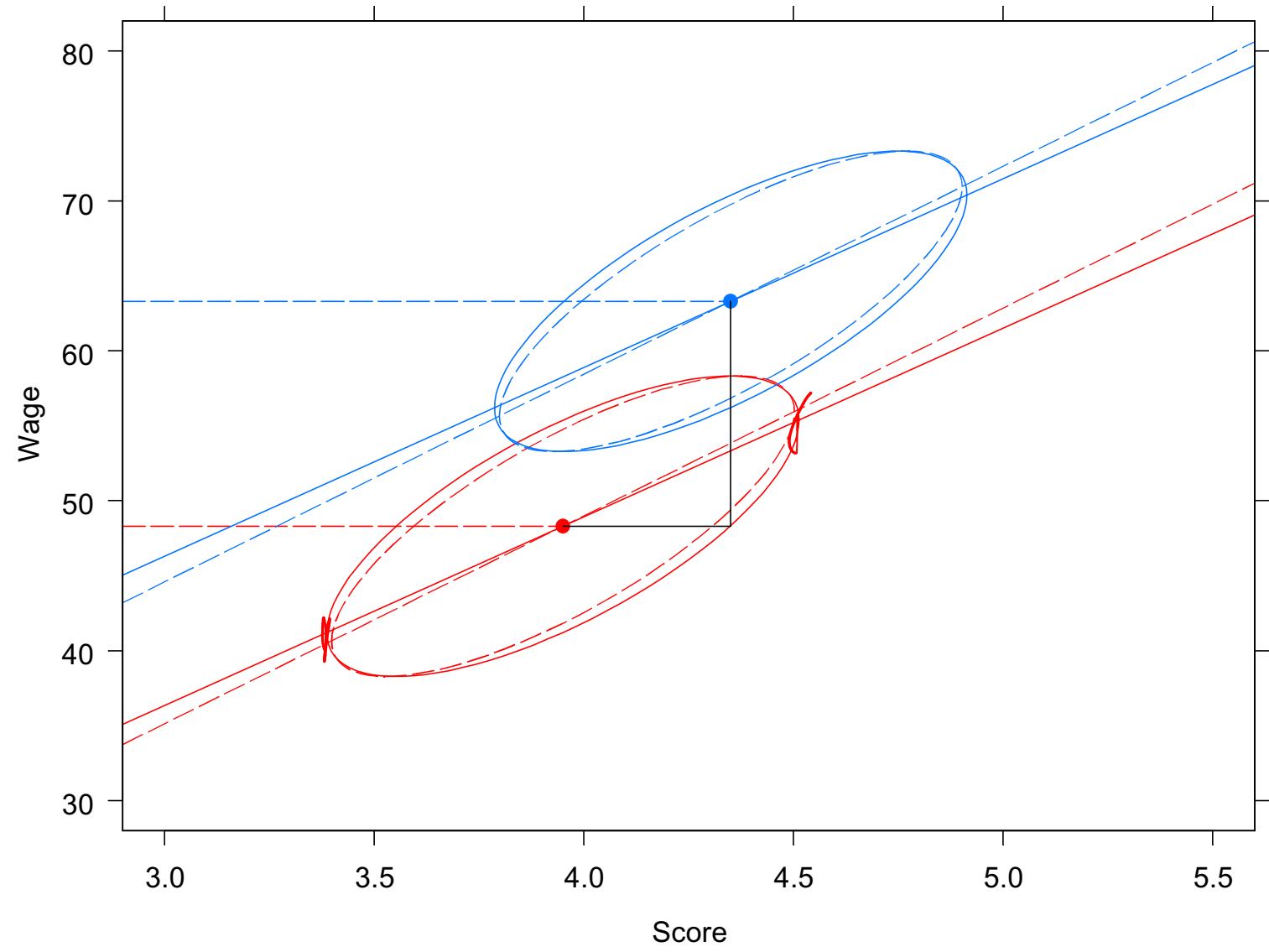


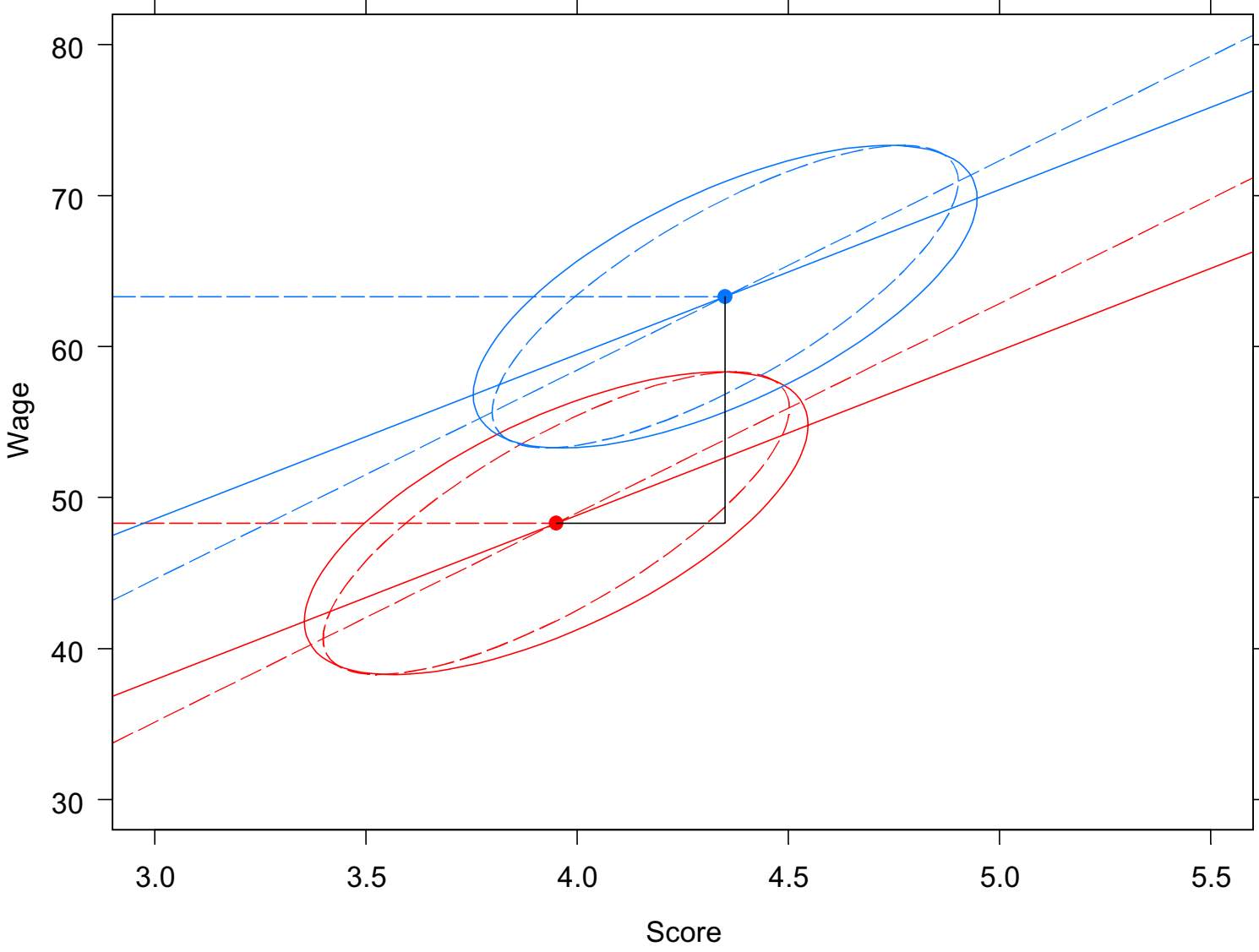


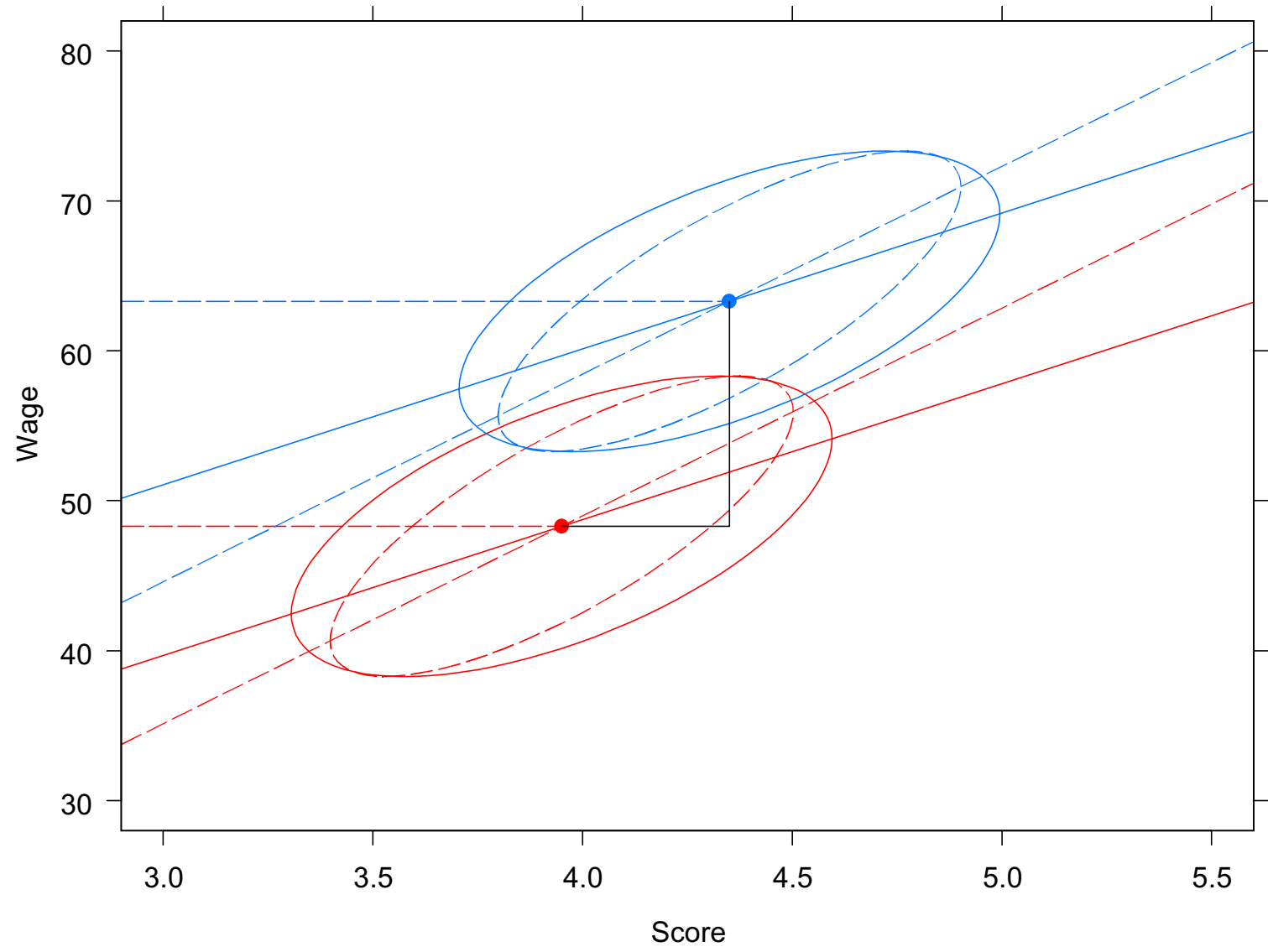




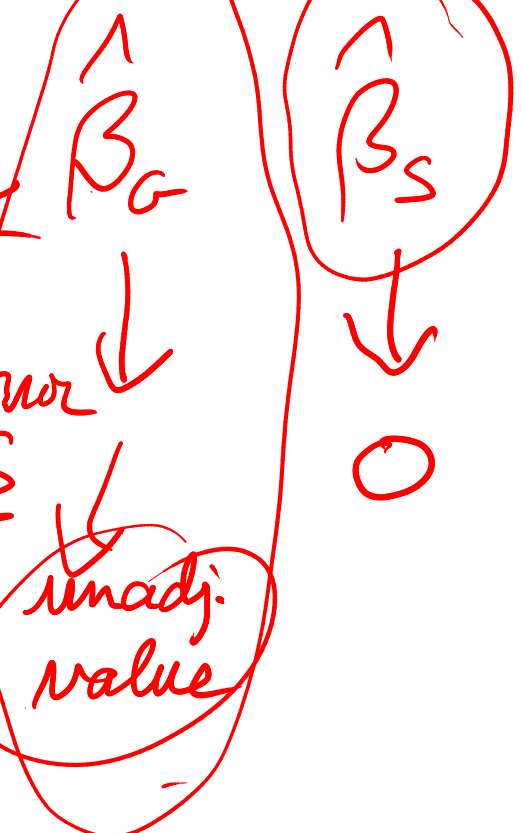
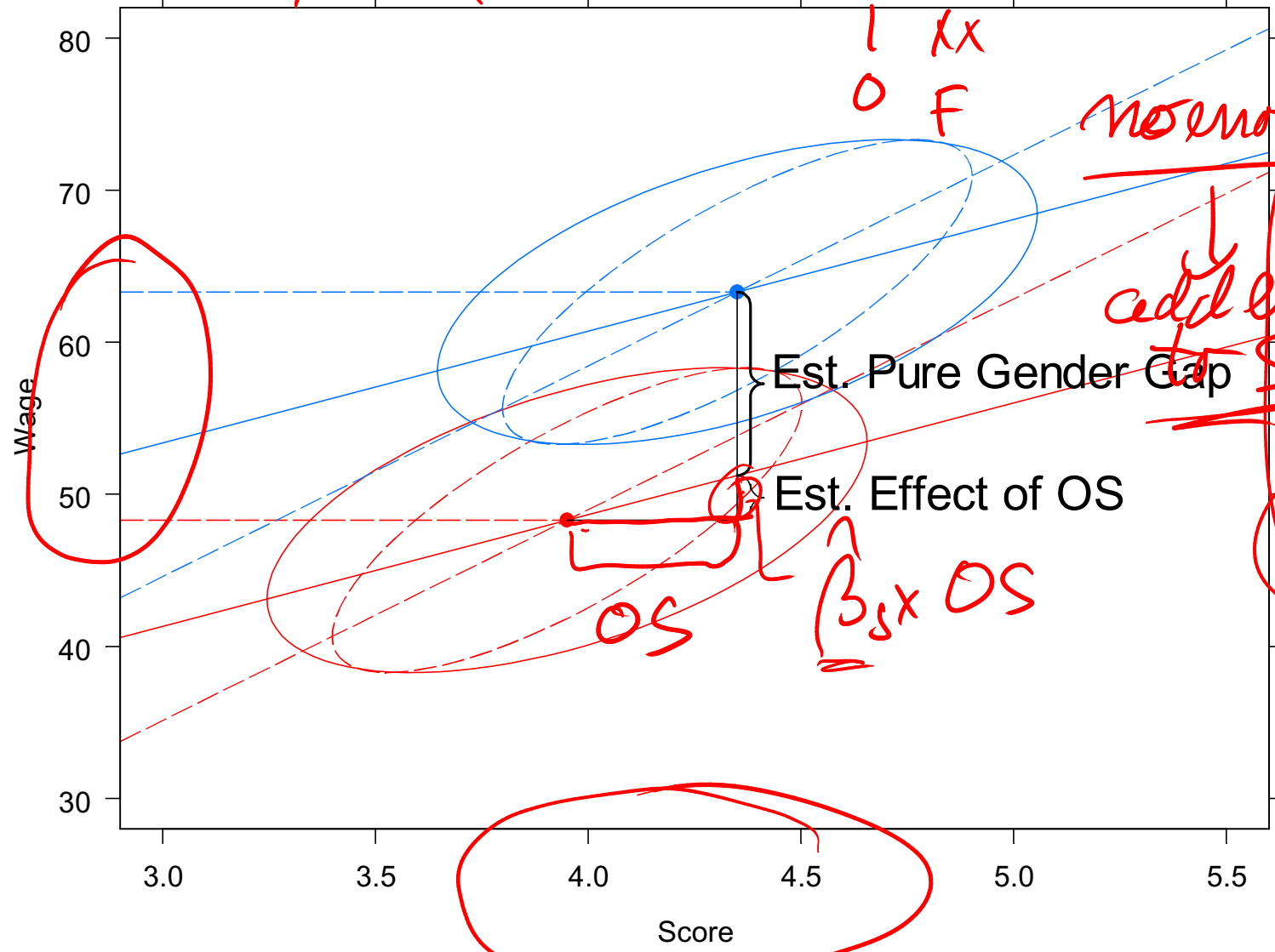
$$SD_{new}^2 = \sqrt{SD_{orig}^2 + SD_{err}^2}$$







$$E(Y) = \beta_0 + \beta_G G + \beta_S S$$



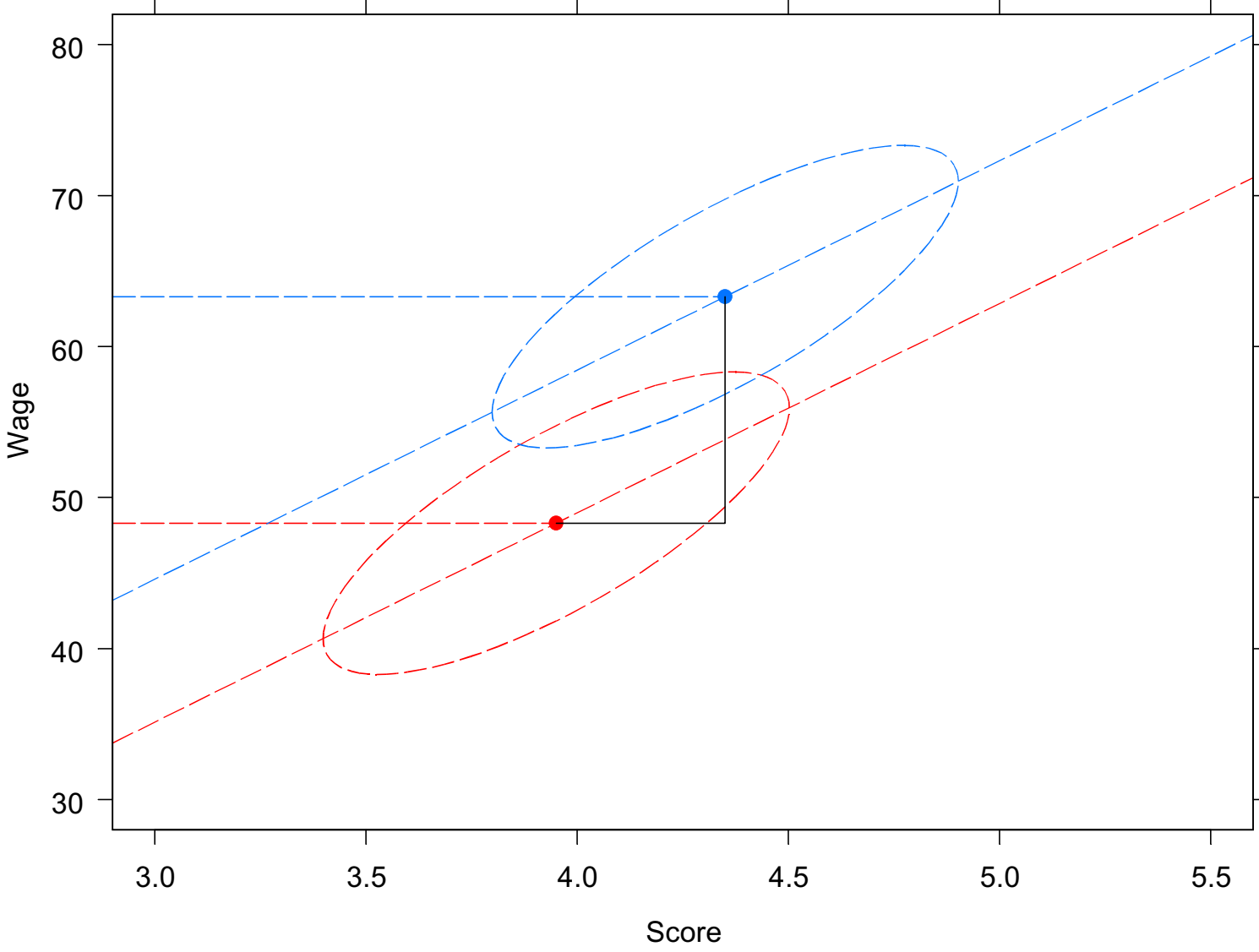
8 Bias in X: Job Evaluation Score

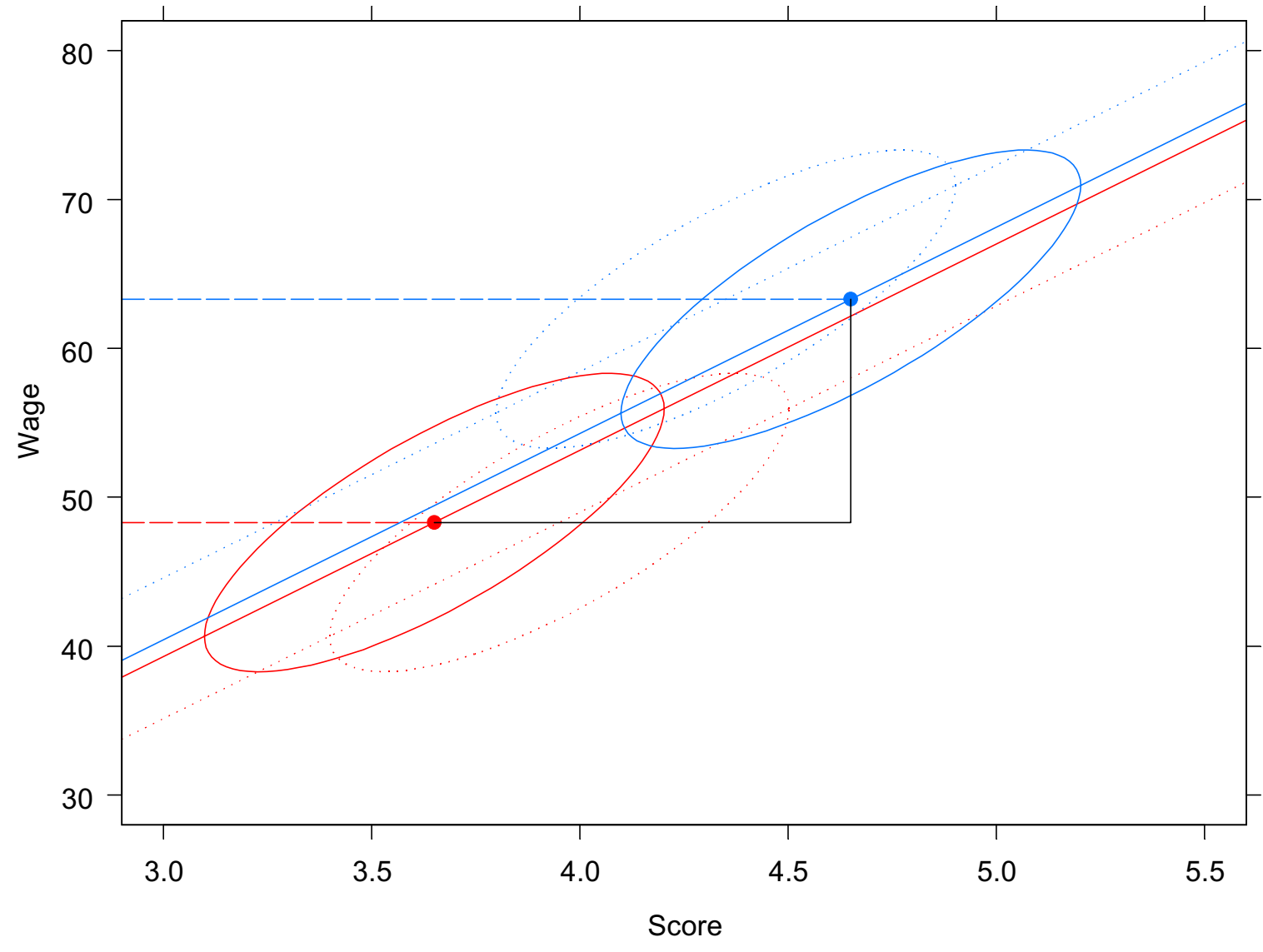
Measurement error leads to overestimation of gender gap



But bias might work in the opposite direction







9 Simpson's (Yule's) Paradox

Some types of association between variables:

- Marginal
- Conditional
- Partial
- Ecological

Paradoxes:

- Simpson's: Marginal and Conditional can have opposite signs
- Robinson's: Ecological and Conditional can have opposite signs

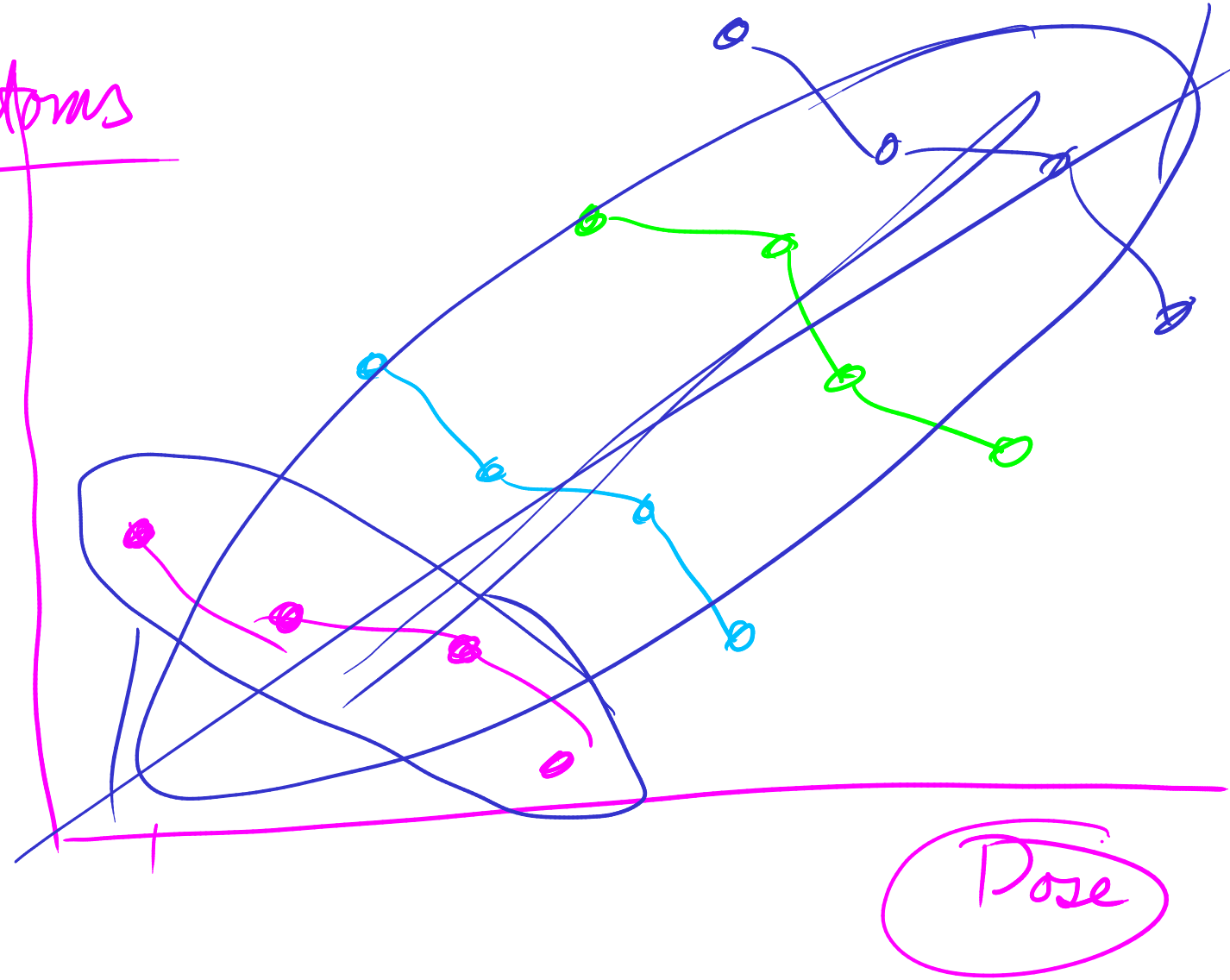
Example:

- Symptoms vs. Dosage of Drug: 6 patients at 4 dosages

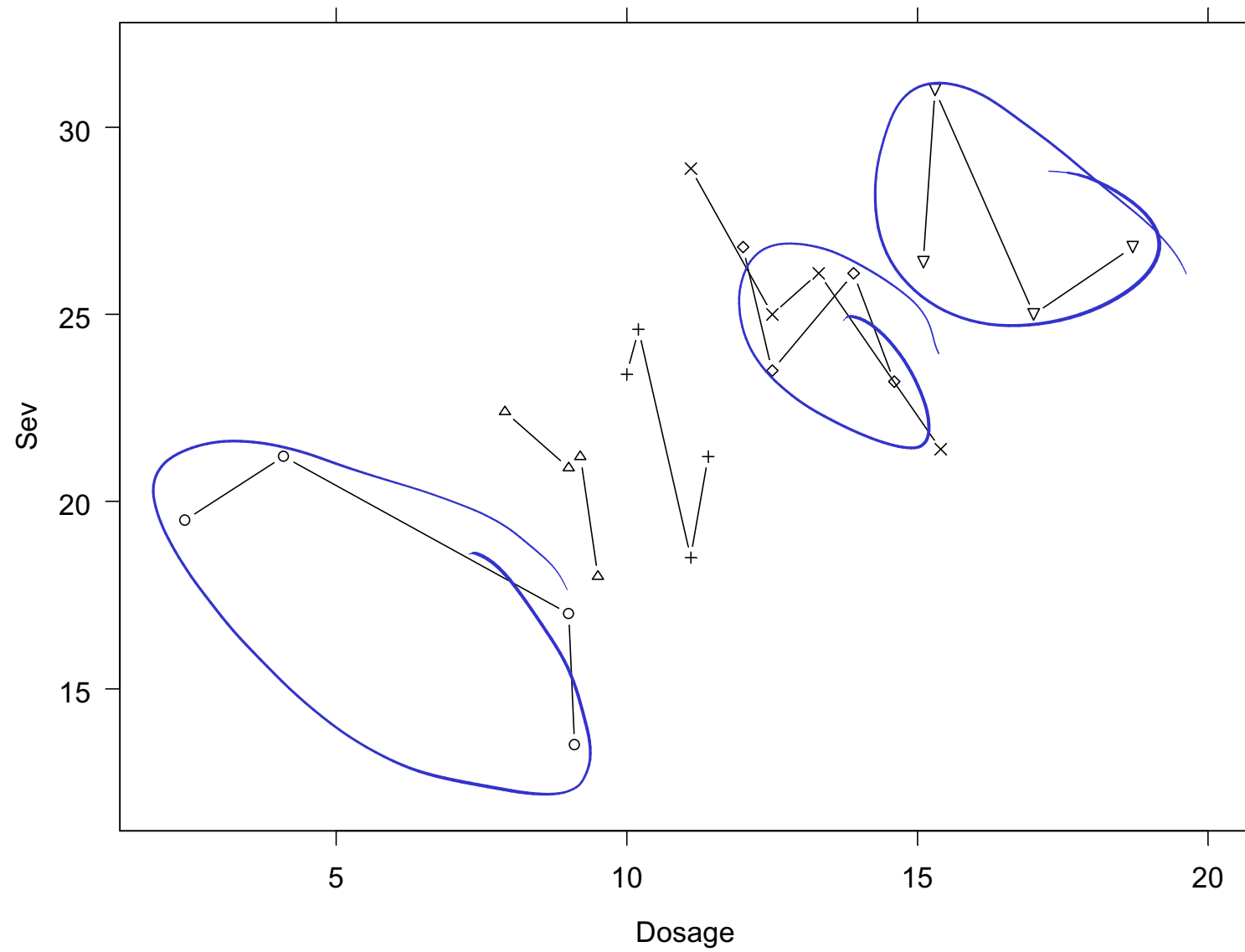
> Drug

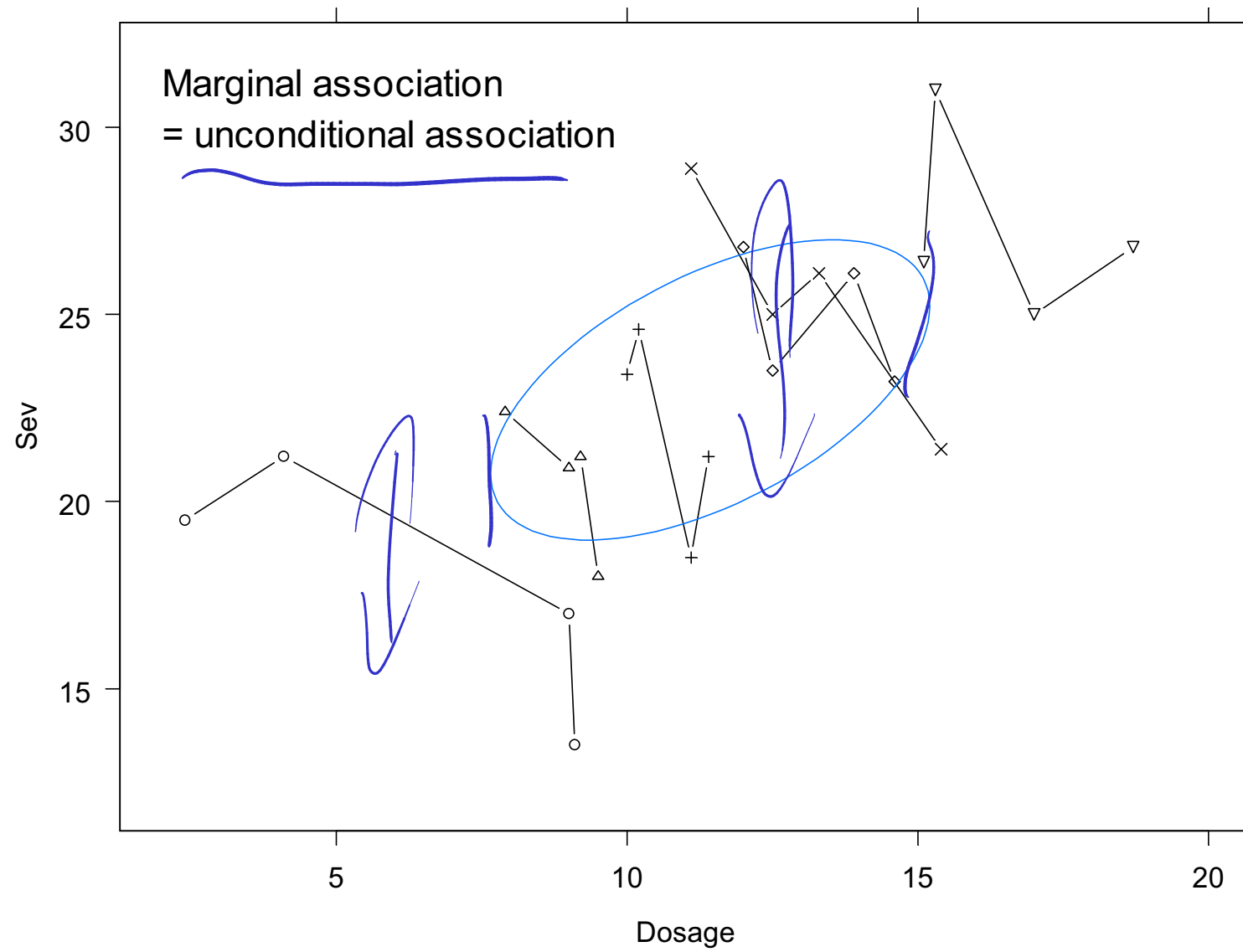
	ID	Dosage	Sev
1	A	10.3	22.4
2	A	10.6	22.8
3	A	11.3	21.7
4	A	11.3	20.9
5	B	11.1	23.1
6	B	11.3	22.7
7	B	11.3	22.8
8	B	11.4	22.0
9	C	11.4	23.3
10	C	11.5	23.6
11	C	11.6	22.1
12	C	11.6	22.8
13	D	11.6	24.7
14	D	11.8	23.8
15	D	11.9	24.0
16	D	12.2	22.8
17	E	11.7	24.2
18	E	11.8	23.4
19	E	12.0	24.0

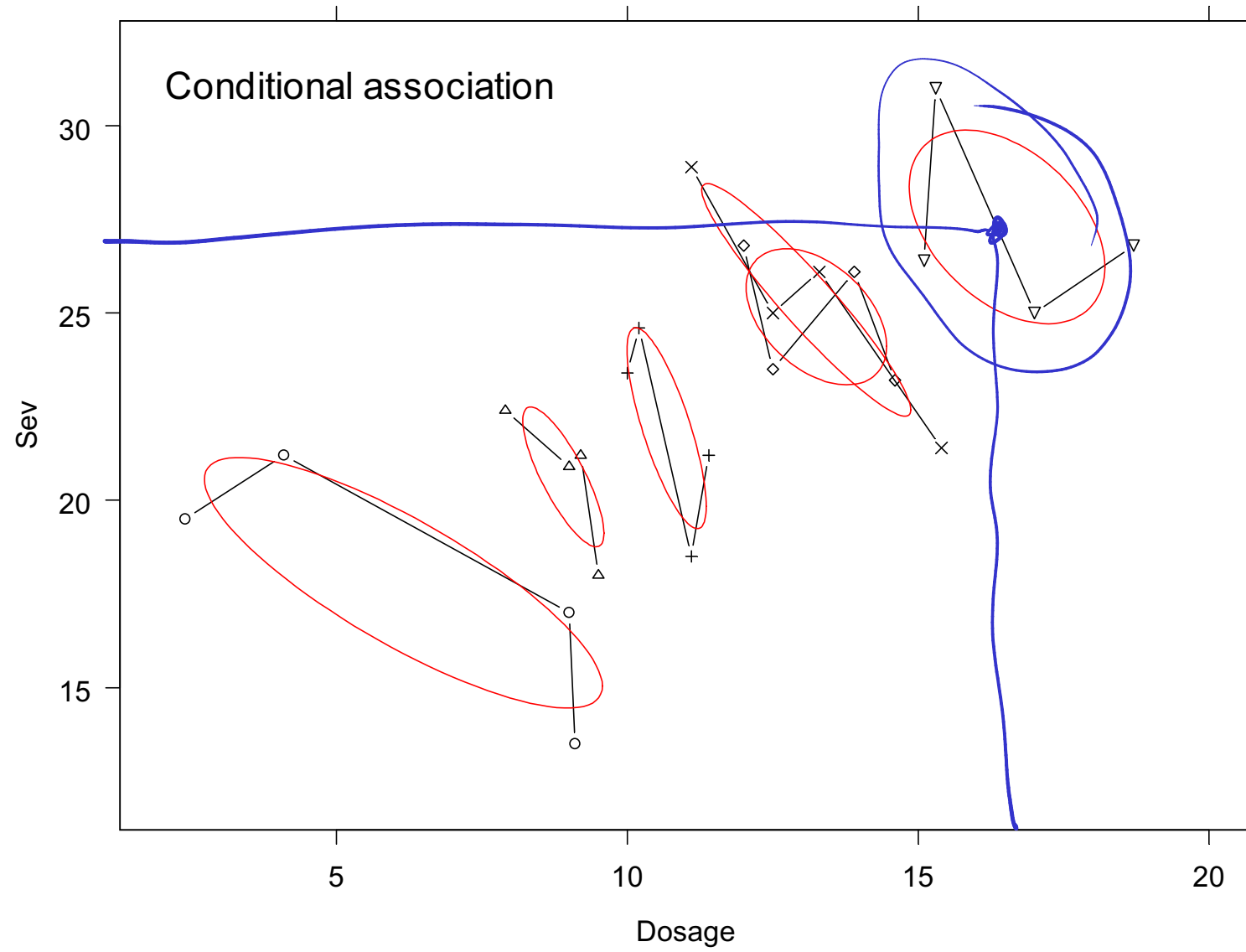
Symptoms



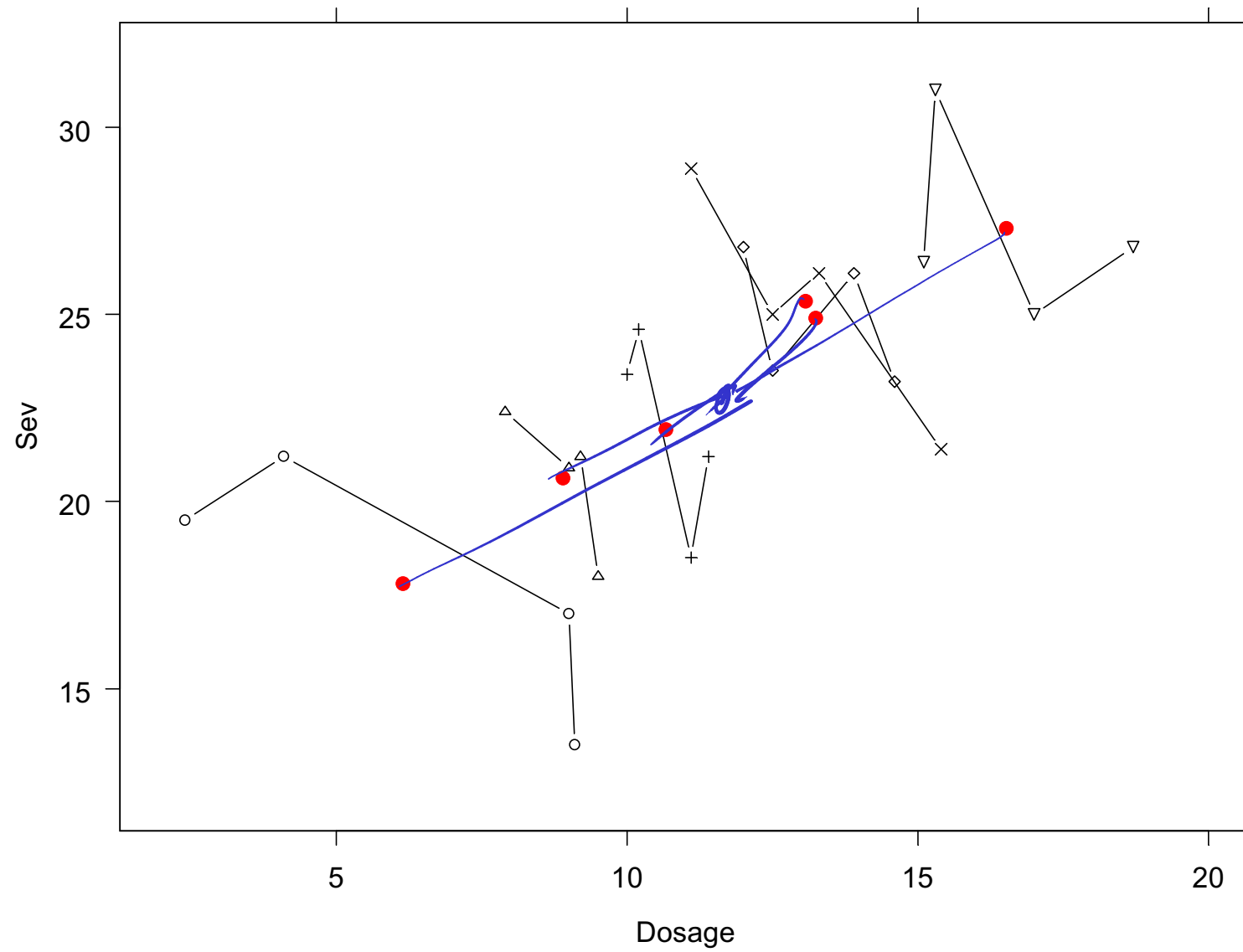
20	E	12.1	23.3
21	F	12.2	24.1
22	F	12.2	25.3
23	F	12.4	23.7
24	F	12.7	24.2

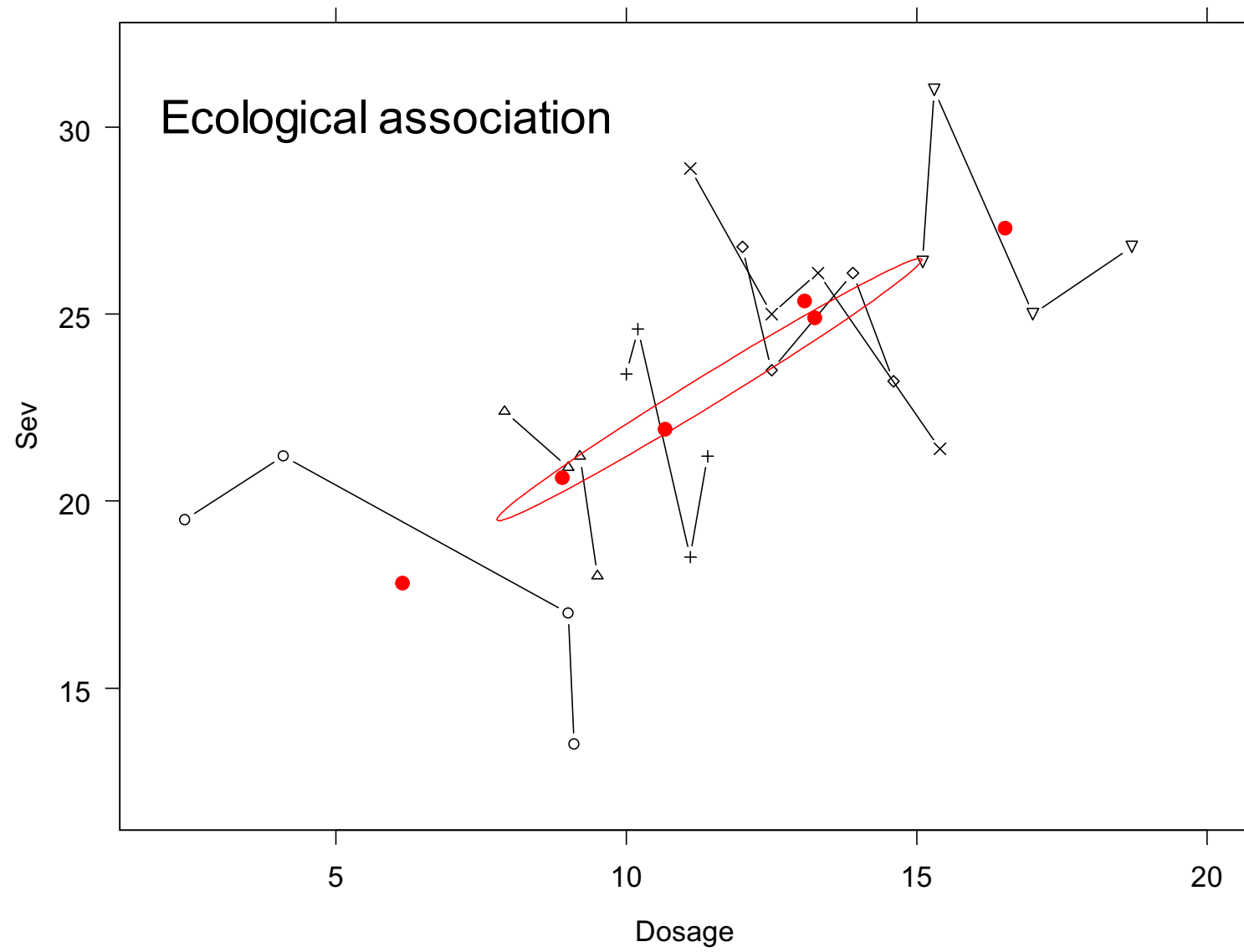






$Y \sim D + 10$





Simple regression = marginal association

```
> fit <- lm(Sev ~ Dosage, Drug)
> summary(fit)
```

```
Call: lm(formula = Sev ~ Dosage, data = Drug)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-8.103  -1.688   0.5618   1.902   6.112
```

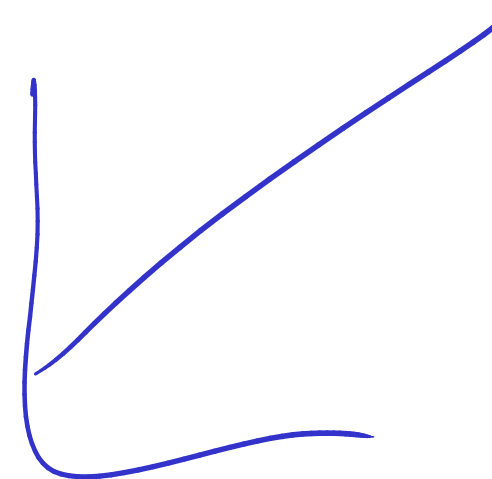
```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	16.2110	2.2596	7.1744	0.0000
Dosage	0.5925	0.1881	3.1500	0.0046

```
Residual standard error: 3.406 on 22 degrees of freedom
```

```
Multiple R-Squared: 0.3108
```

```
F-statistic: 9.922 on 1 and 22 degrees of freedom,
the p-value is 0.004648
```



Multiple regression = Analysis of covariance (with these data)

```
> fit <- lm(Sev ~ Dosage + ID, Drug)
> summary(fit)
```

```
Call: lm(formula = Sev ~ Dosage + ID, data = Drug)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.005	-1.696	0.519	1.429	2.489

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	34.2840	2.9082	11.7886	0.0000
Dosage	-0.9888	0.2520	-3.9230	0.0011
ID1	2.7720	0.7748	3.5776	0.0023
ID2	1.9424	0.4797	4.0490	0.0008
ID3	2.4207	0.4005	6.0434	0.0000

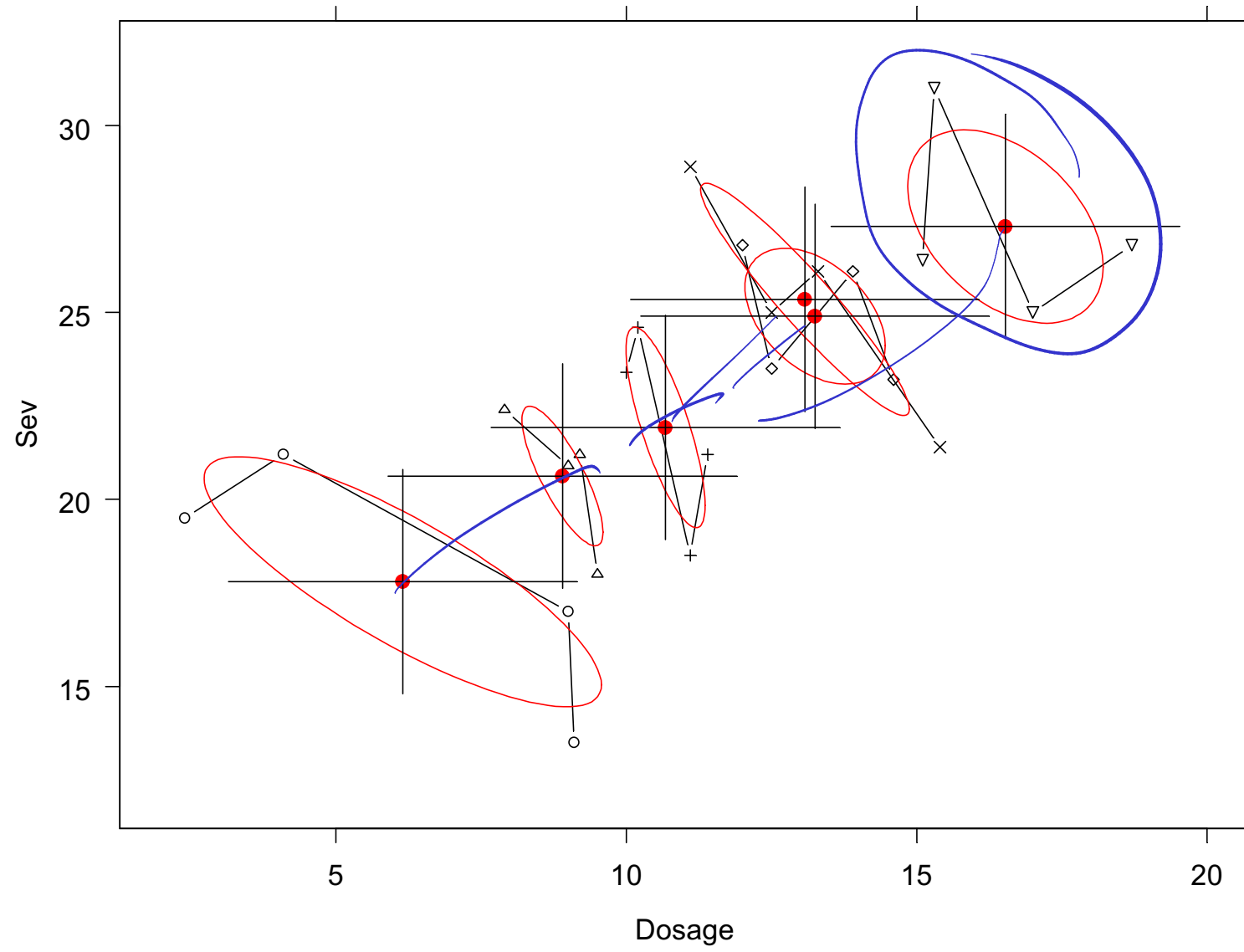
same as
coefficient
using a
simple reg.
on AVP

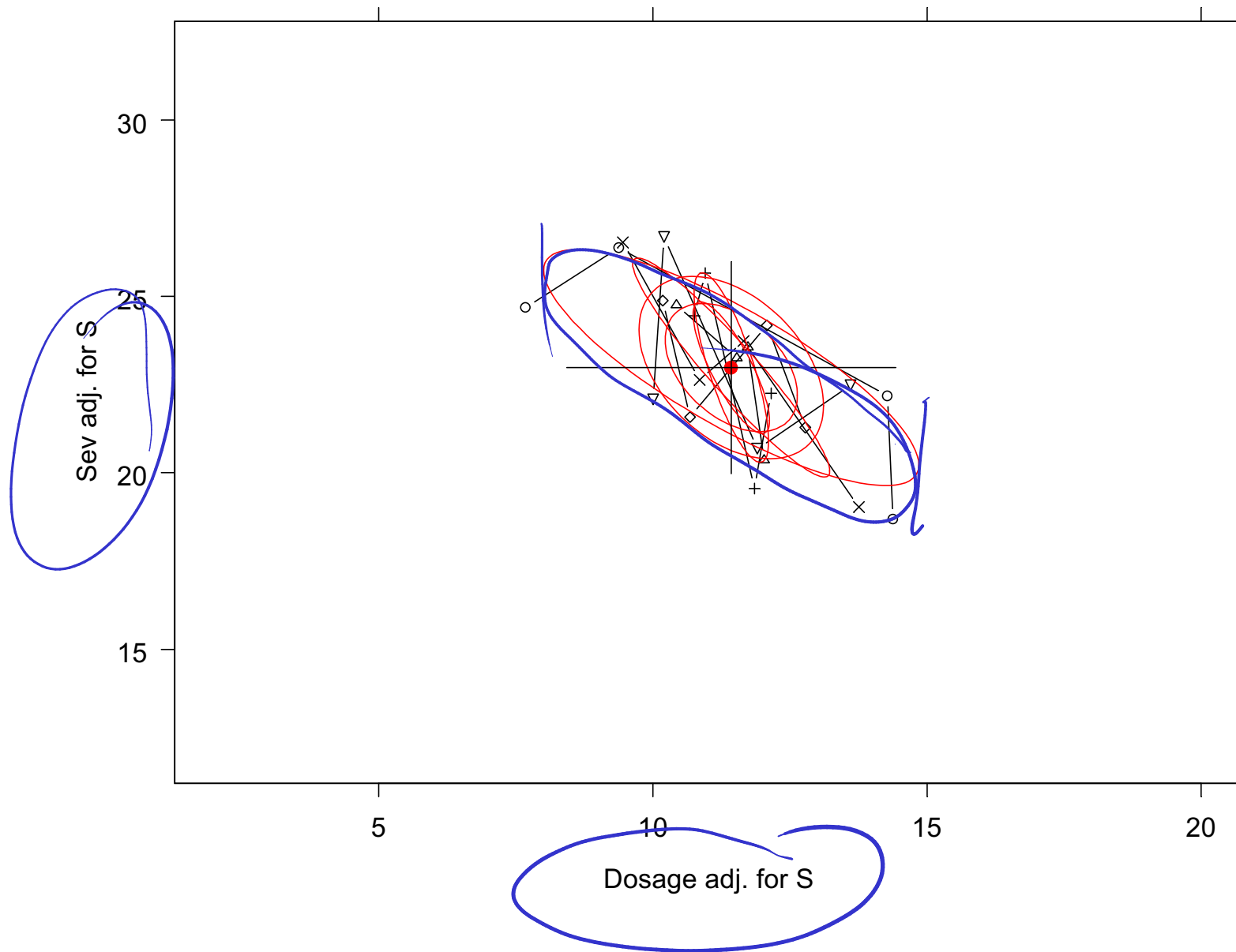
ID4	1.3970	0.2829	4.9377	0.0001
ID5	1.8710	0.3130	5.9769	0.0000

Residual standard error: 1.96 on 17 degrees of freedom

Multiple R-Squared: 0.8236

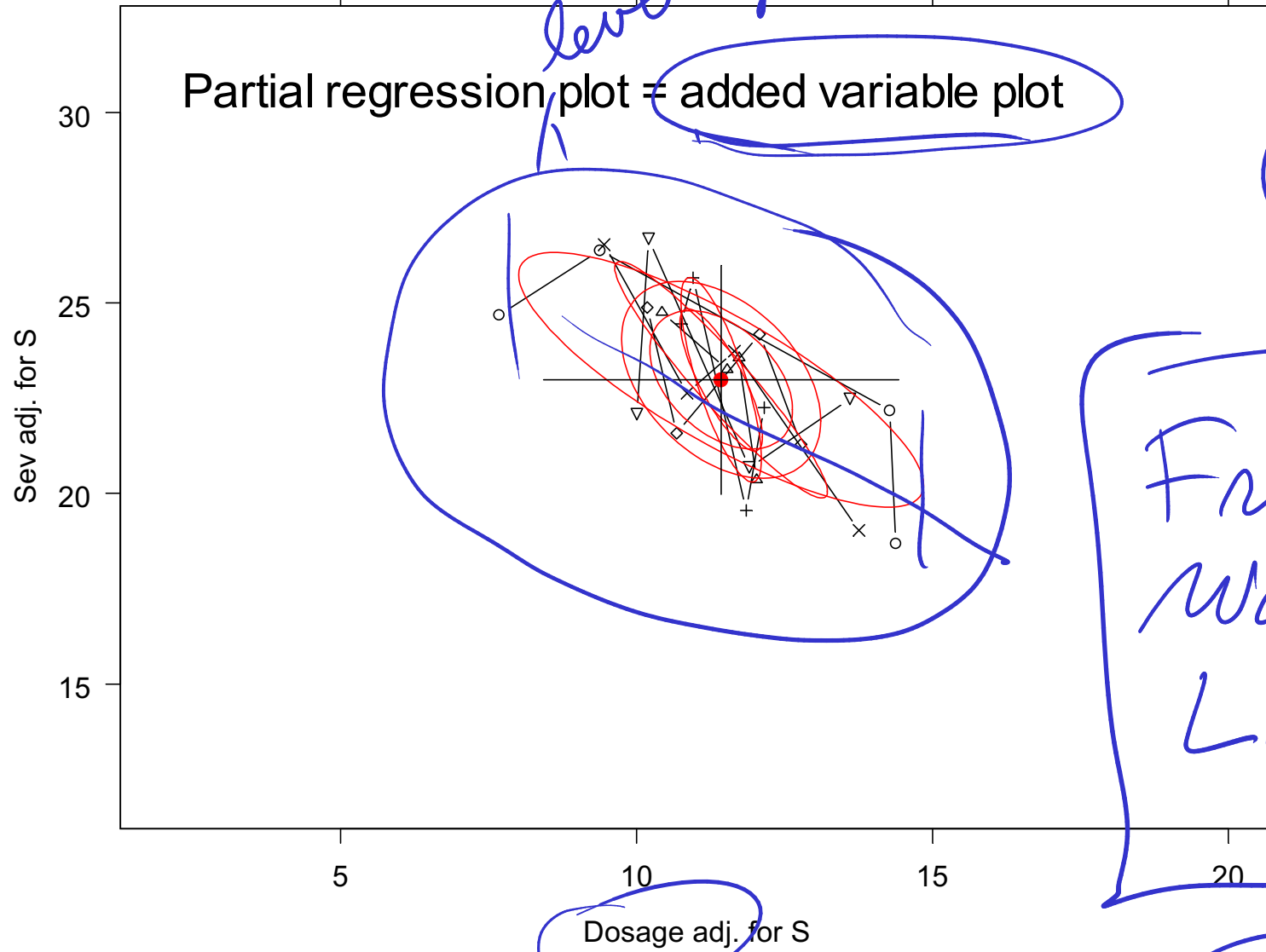
F-statistic: 13.23 on 6 and 17 degrees of freedom,
the p-value is 0.00001393





PRLP
leverage

Partial regression plot = added variable plot

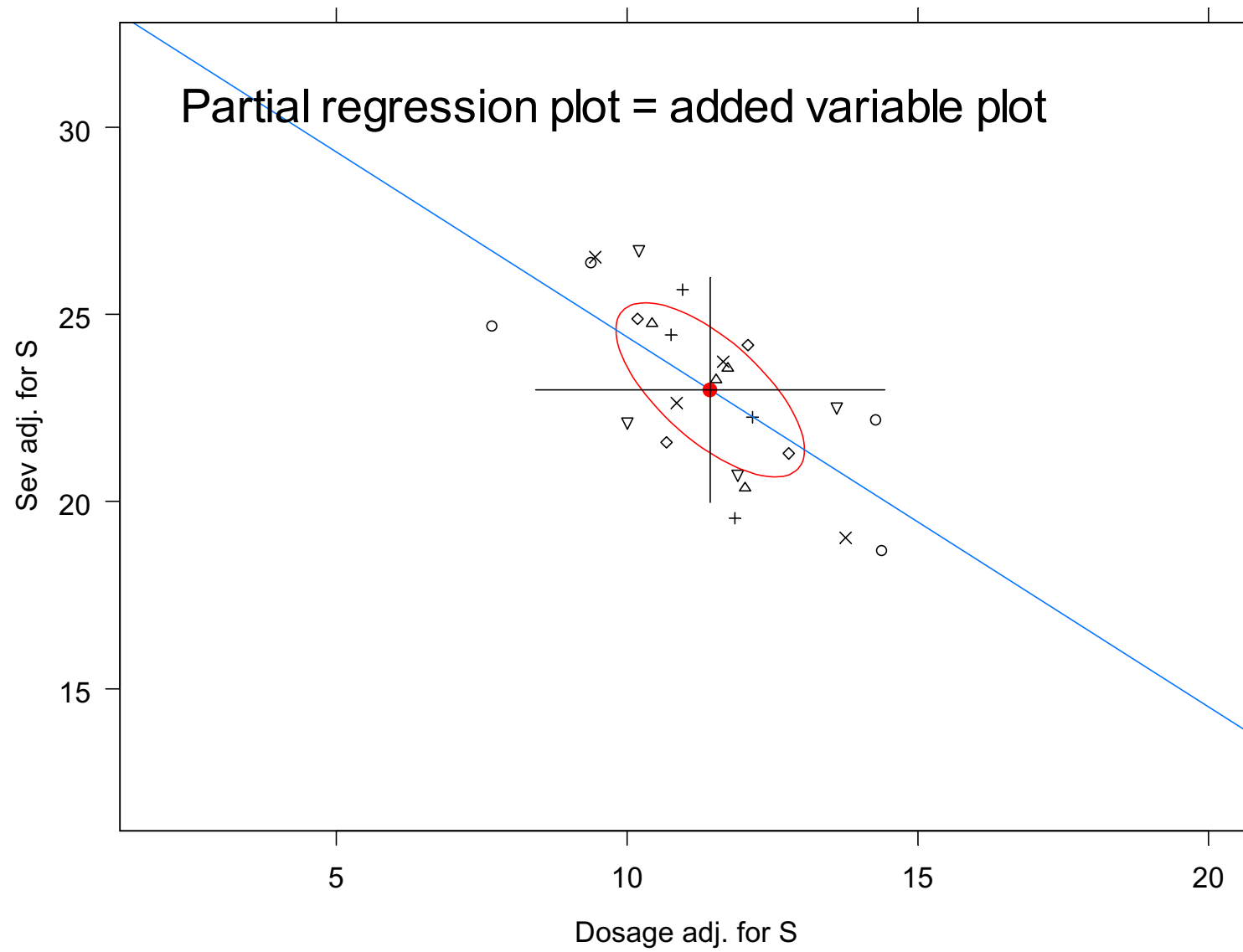


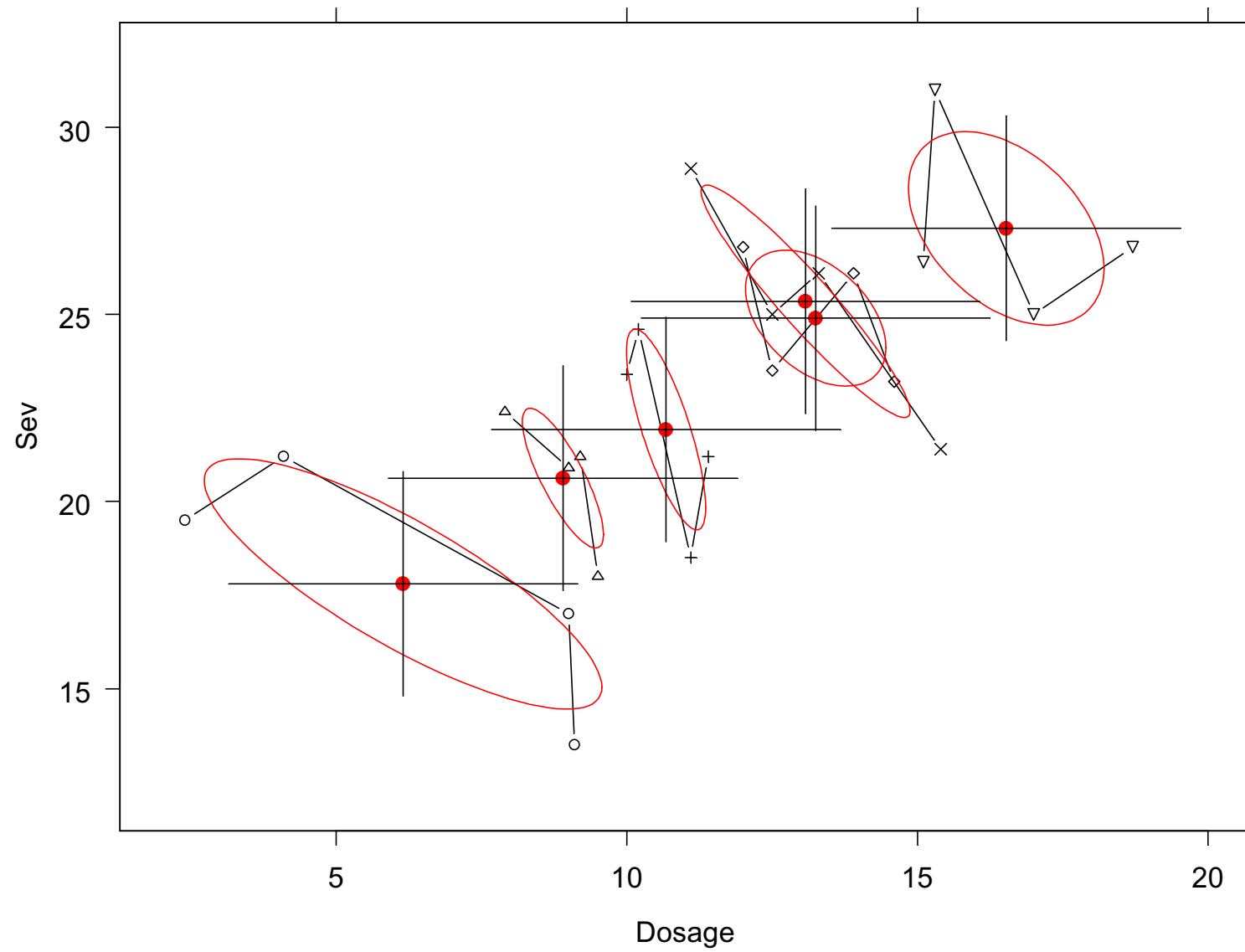
AVP

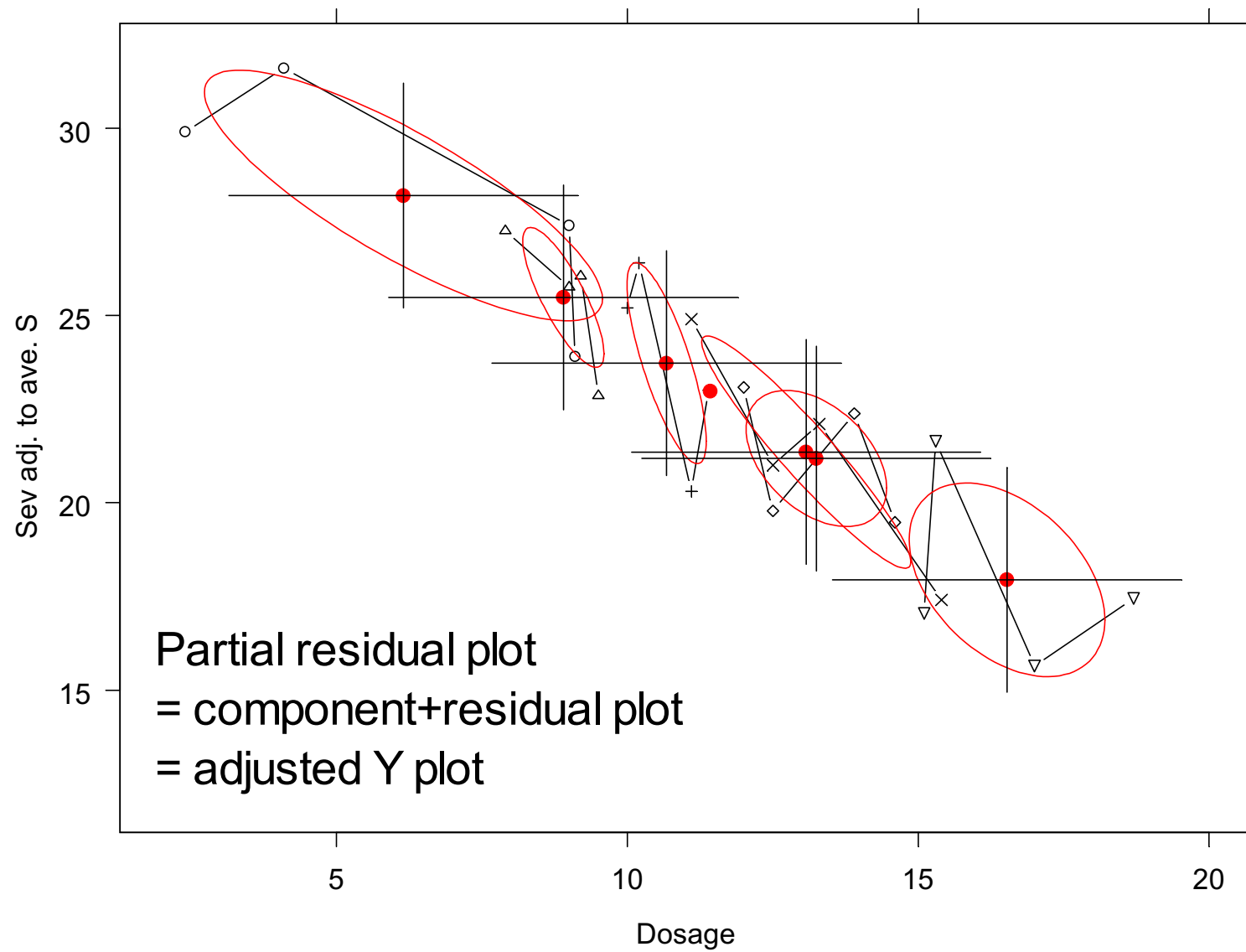
Frisch — 30?
Wangh — 40?
Lovell — 60?

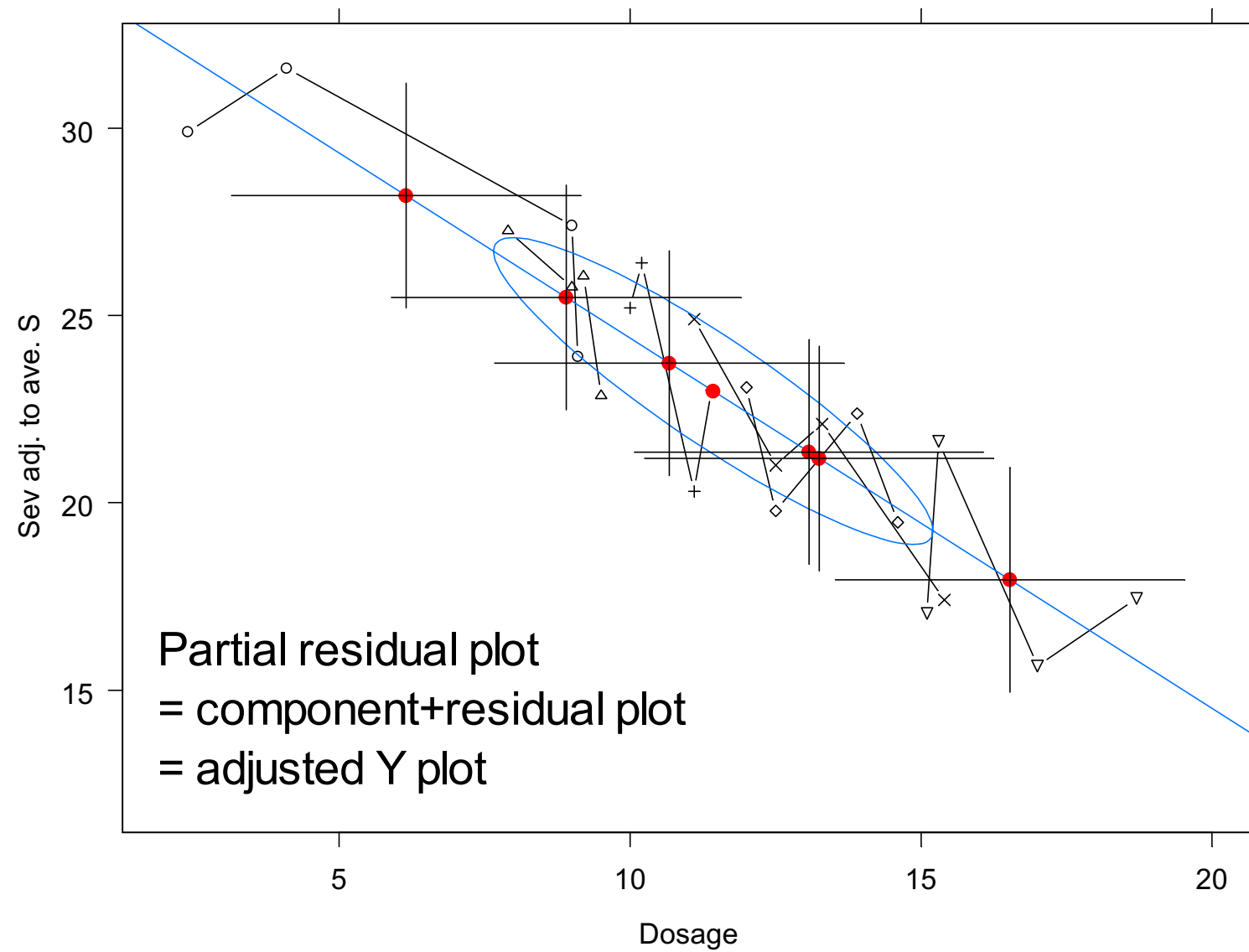
FWL

Dosage adj. for S









10 Added variable plot

10.1

Ideal tool for statistical sleuthing



Pay equity in a large law firm

	Title	Knowledge	Experience	Communication	Gender
58	Secretary III	5.0	4.5	4.0	F
70	Admin Asst	6.5	5.5	7.0	M
36	Secretary II	3.5	3.5	2.5	F
74	Secretary III	4.5	5.0	3.5	F
78	Secretary III	3.0	5.5	5.0	F
25	Secretary II	3.0	3.0	2.5	F
19	Admin Asst	7.0	7.0	7.0	M
42	Admin Asst	5.0	6.0	6.5	M
87	Clerk I	1.5	1.0	2.5	F
32	Secretary I	2.5	3.5	2.5	F
66	Admin Asst	4.5	6.0	6.5	F
47	Admin Asst	7.0	7.0	7.0	M
34	Secretary III	4.5	6.5	5.5	F
69	Secretary II	3.0	4.0	4.0	F
21	Secretary III	5.0	5.0	4.5	F
52	Secretary II	3.5	3.5	3.0	F
95	Receptionist	2.5	2.0	4.0	F

89	Admin Asst	5.0	5.5	7.0	M
----	------------	-----	-----	-----	---

... 95 positions

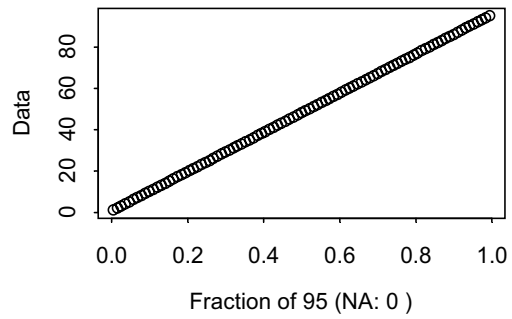
Summary

> summary(Law)

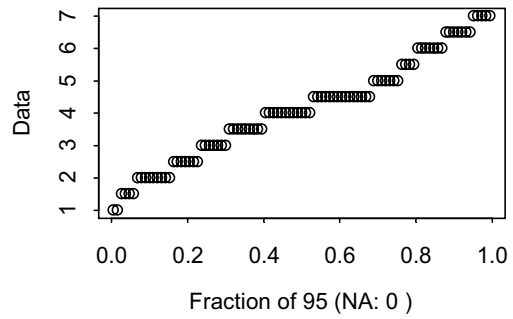
ID	Knowledge	Experience	Communication
Min.: 1.0	Min.:1.000	Min.:1.000	Min.:1.000
1st Qu.:24.5	1st Qu.:3.000	1st Qu.:3.000	1st Qu.:3.000
Median:48.0	Median:4.000	Median:4.500	Median:4.000
Mean:48.0	Mean:4.116	Mean:4.184	Mean:4.237
3rd Qu.:71.5	3rd Qu.:5.000	3rd Qu.:5.500	3rd Qu.:5.500
Max.:95.0	Max.:7.000	Max.:7.000	Max.:7.000

Gender	Female	Title	Wage
F:72	Min.:0.0000	Secretary III:24	Min.:19350
M:23	1st Qu.:1.0000	Secretary II:24	1st Qu.:39580
	Median:1.0000	Admin Asst:17	Median:50010
	Mean:0.7579	Clerk I:12	Mean:49180
	3rd Qu.:1.0000	Secretary IV: 6	3rd Qu.:58650
	Max.:1.0000	Secretary I: 5	Max.:84190
		(Other): 7	

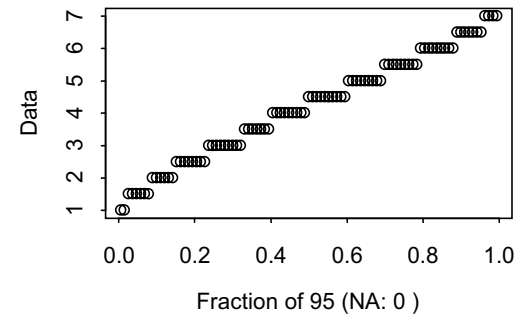
ID



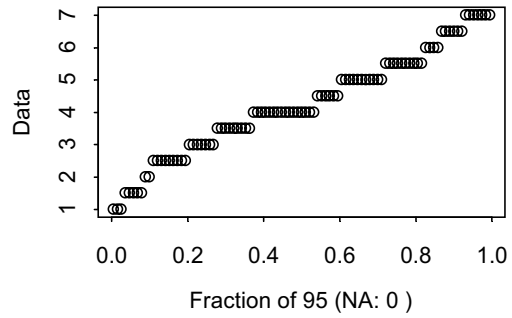
Knowledge



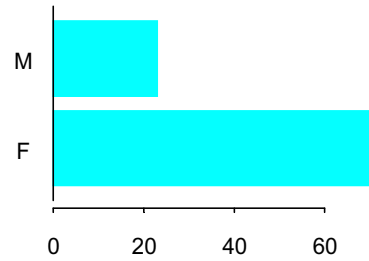
Experience



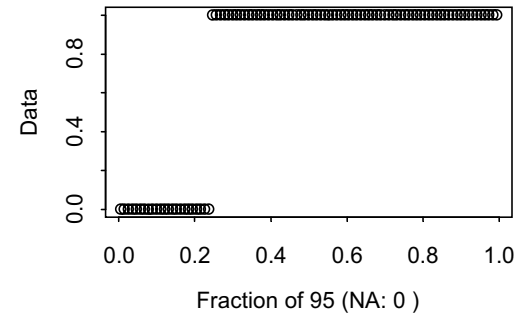
Communication



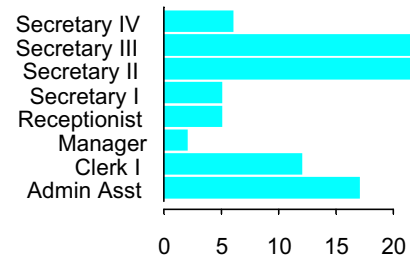
Gender



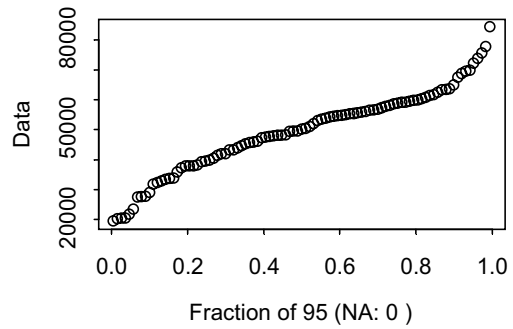
Female

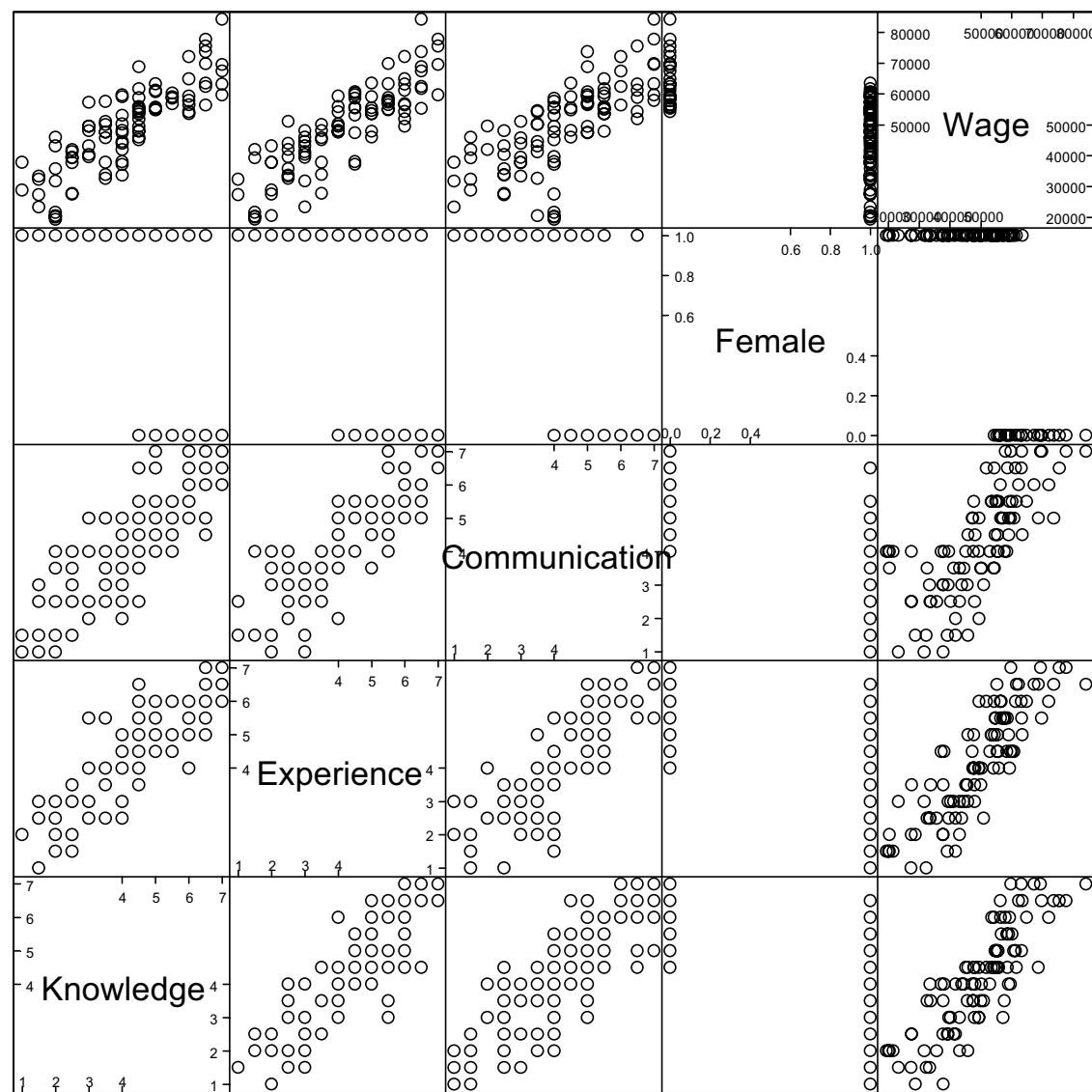


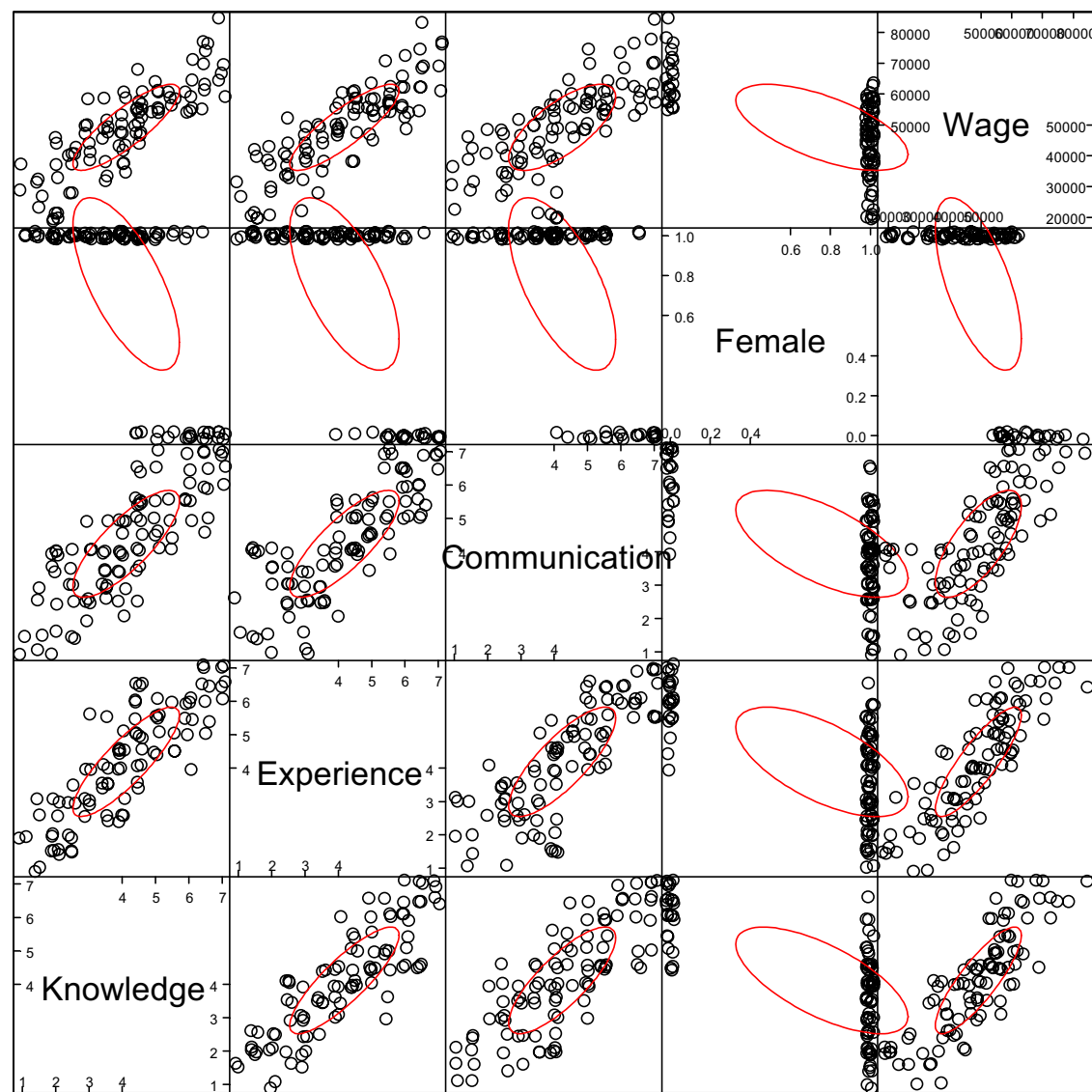
Title



Wage







Regression output:

```
> fit <- lm(Wage ~ Knowledge + Experience
            + Communication + Female, data = Law)
> summary(fit)
```

```
Call: lm(formula = Wage ~ Knowledge + Experience
         + Communication + Female, data = Law)
```

Residuals:

Min	1Q	Median	3Q	Max
-13279	-4824	972.8	5190	14498

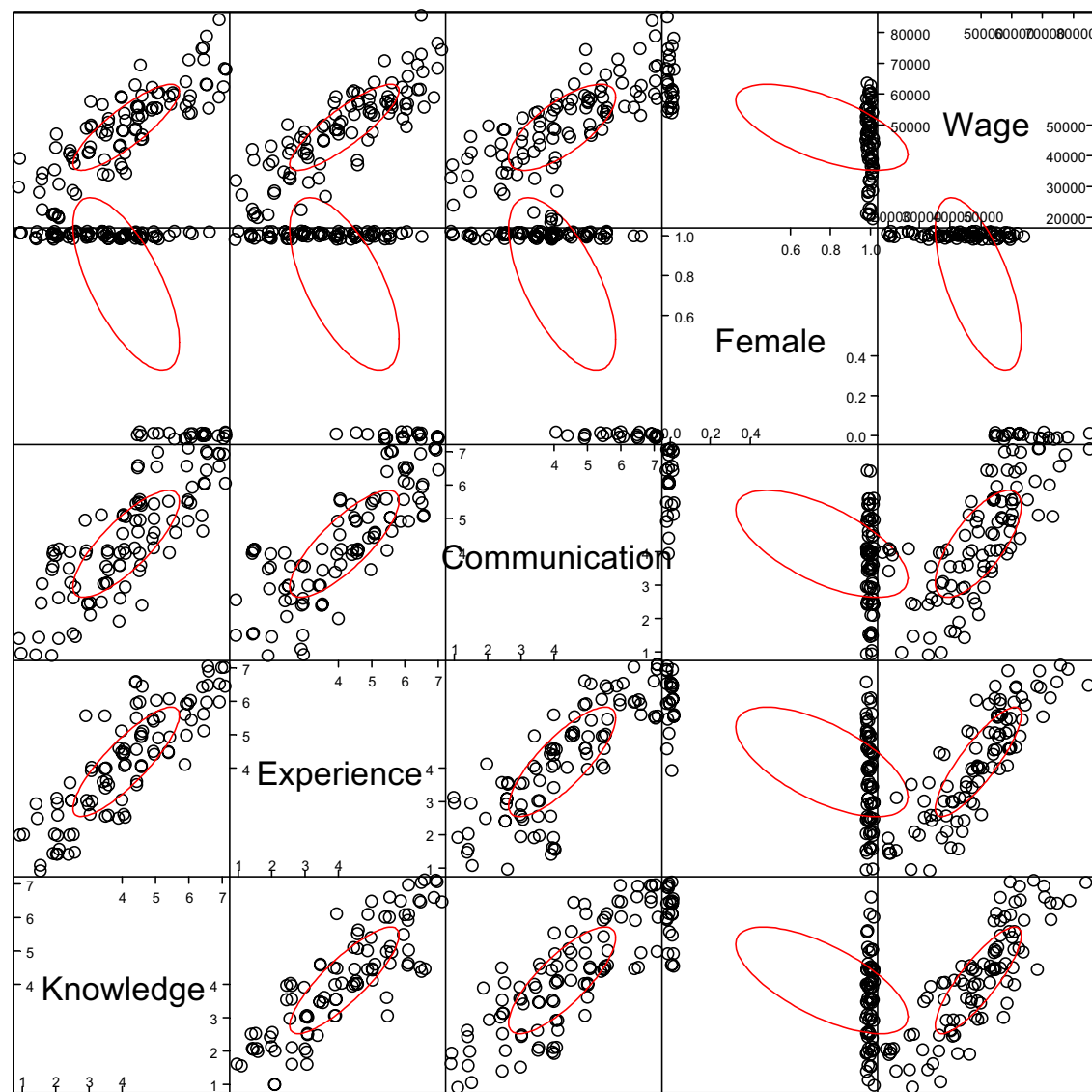
Coefficients:

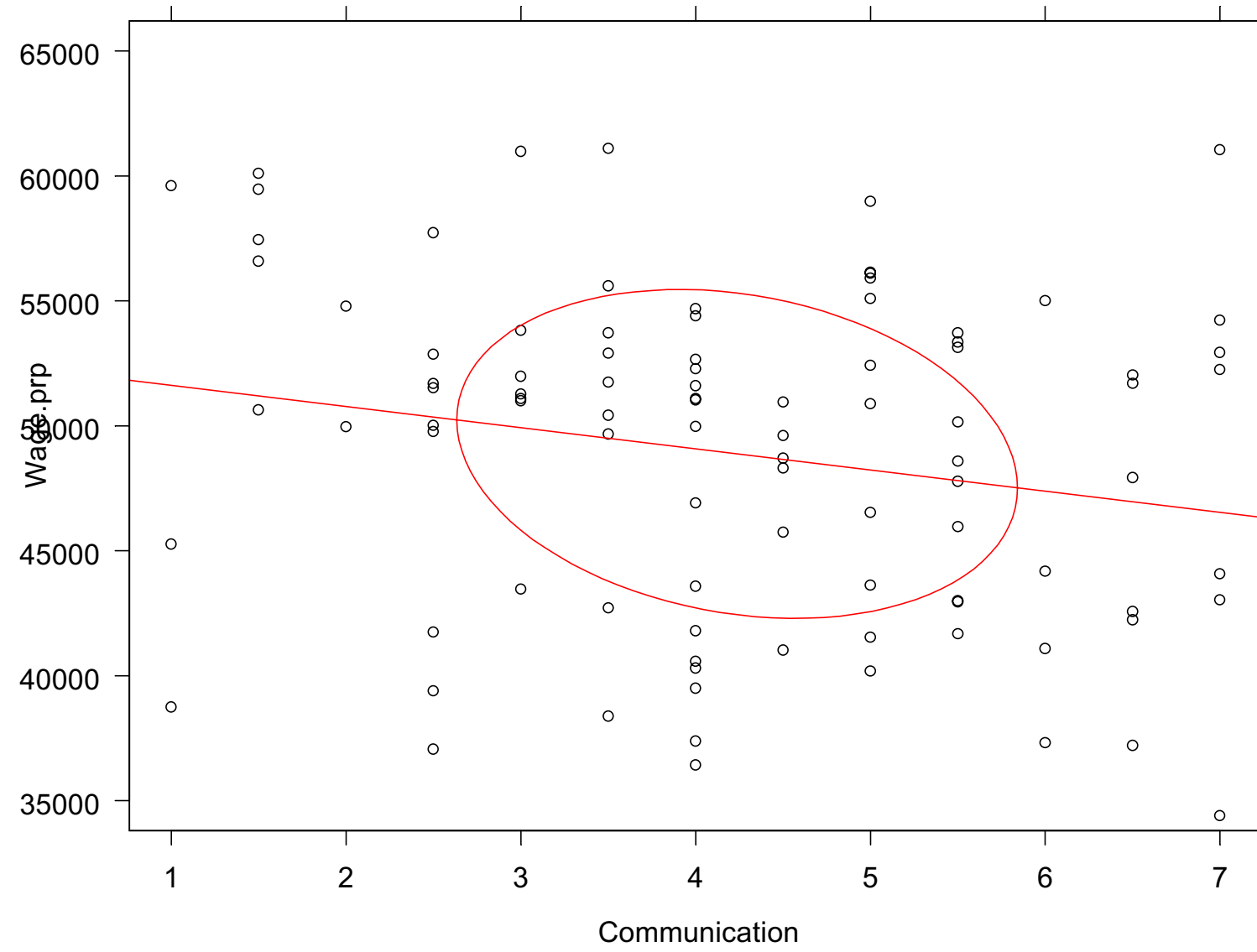
	Value	Std. Error	t value	Pr(> t)
(Intercept)	20153.7248	4031.6651	4.9989	0.0000
Knowledge	3825.2482	899.8848	4.2508	0.0001
Experience	4414.0819	854.7458	5.1642	0.0000
Communication	-846.8698	777.7750	-1.0888	0.2791
Female	-2113.9833	2206.0361	-0.9583	0.3405

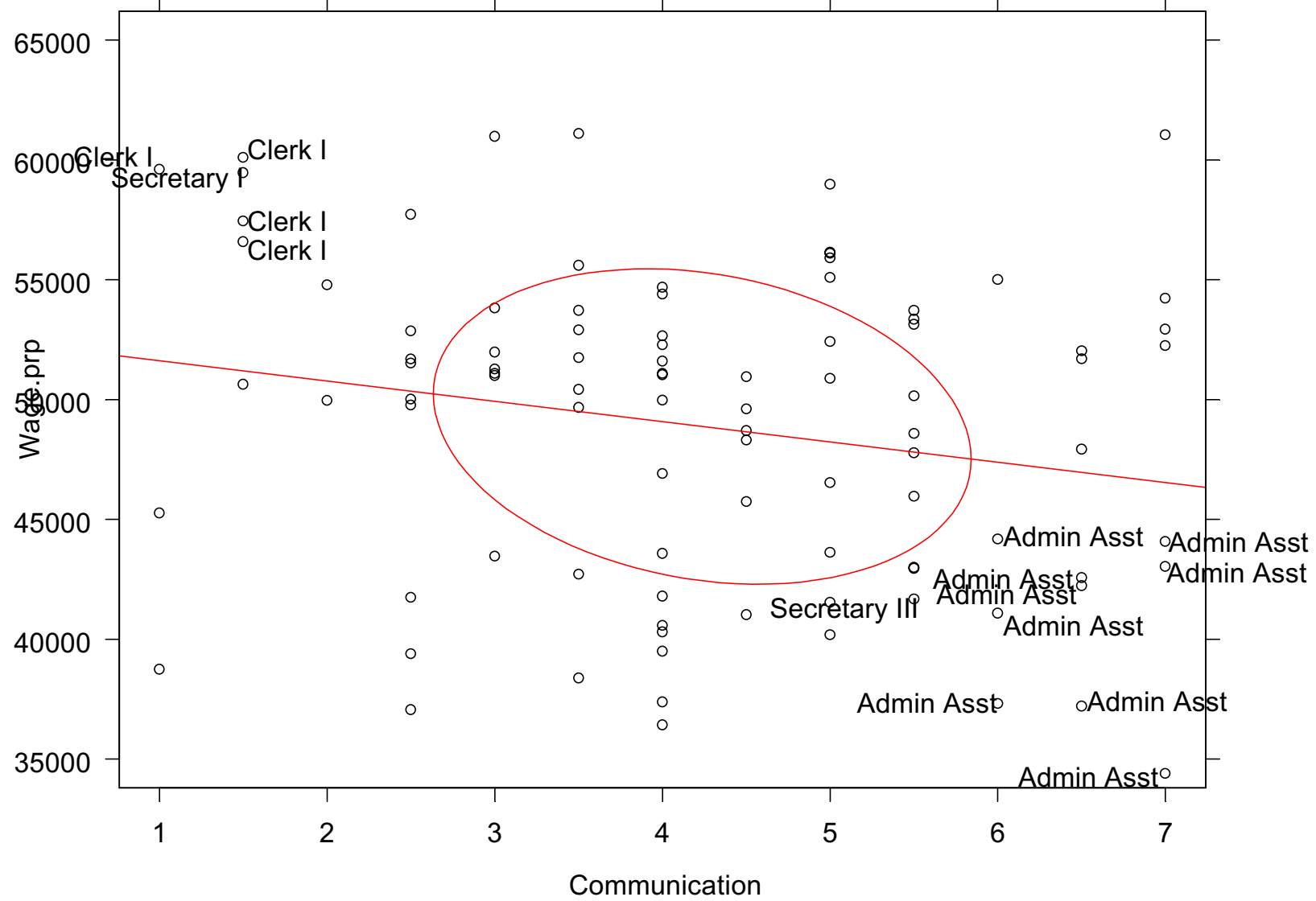
Residual standard error: 6577 on 90 degrees of freedom

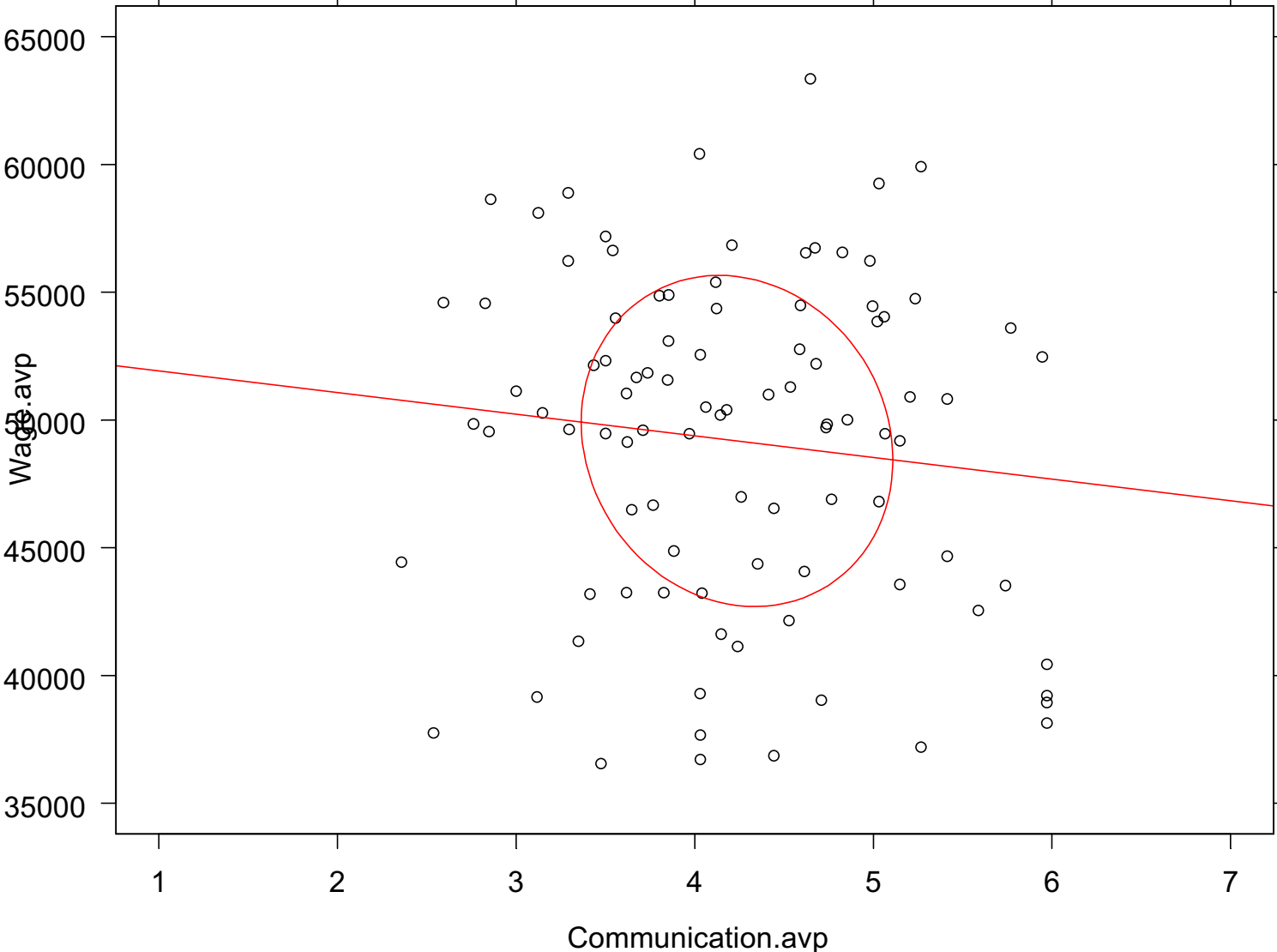
Multiple R-Squared: 0.7864

F-statistic: 82.86 on 4 and 90 degrees of freedom,
the p-value is 0









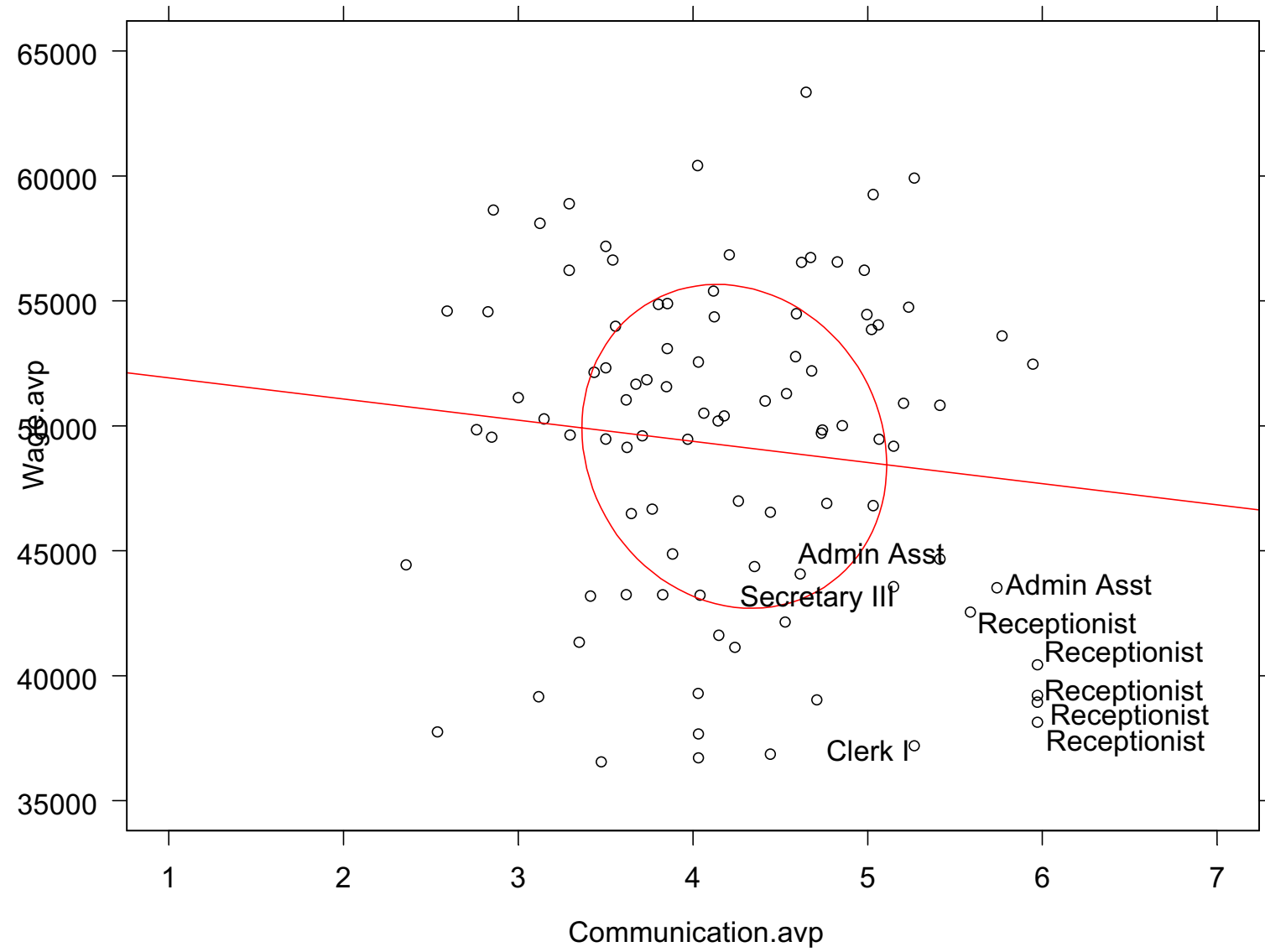
Note:

Horizontal width of ellipse in PRP = s_X

Horizontal width of ellipse in AVP = $s_{X \cdot (\text{other } Xs)}$

$$\sqrt{VIF} = \frac{s_X}{s_{X \cdot (\text{other } Xs)}}$$

So showing the PRP and the AVP side-by-side on the same horizontal scale can be used to convey information on multicollinearity.



Model with adjustment for Receptionists:

```
> fit <- lm( Wage ~Knowledge + Experience + Communication +
             Female + Recept, Law)
```

```
> summary(fit)
```

```
Call: lm(formula = Wage ~Knowledge + Experience + Communication +
          Female + Recept, data = Law)
```

Residuals:

Min	1Q	Median	3Q	Max
-15008	-3101	441.7	4261	14321

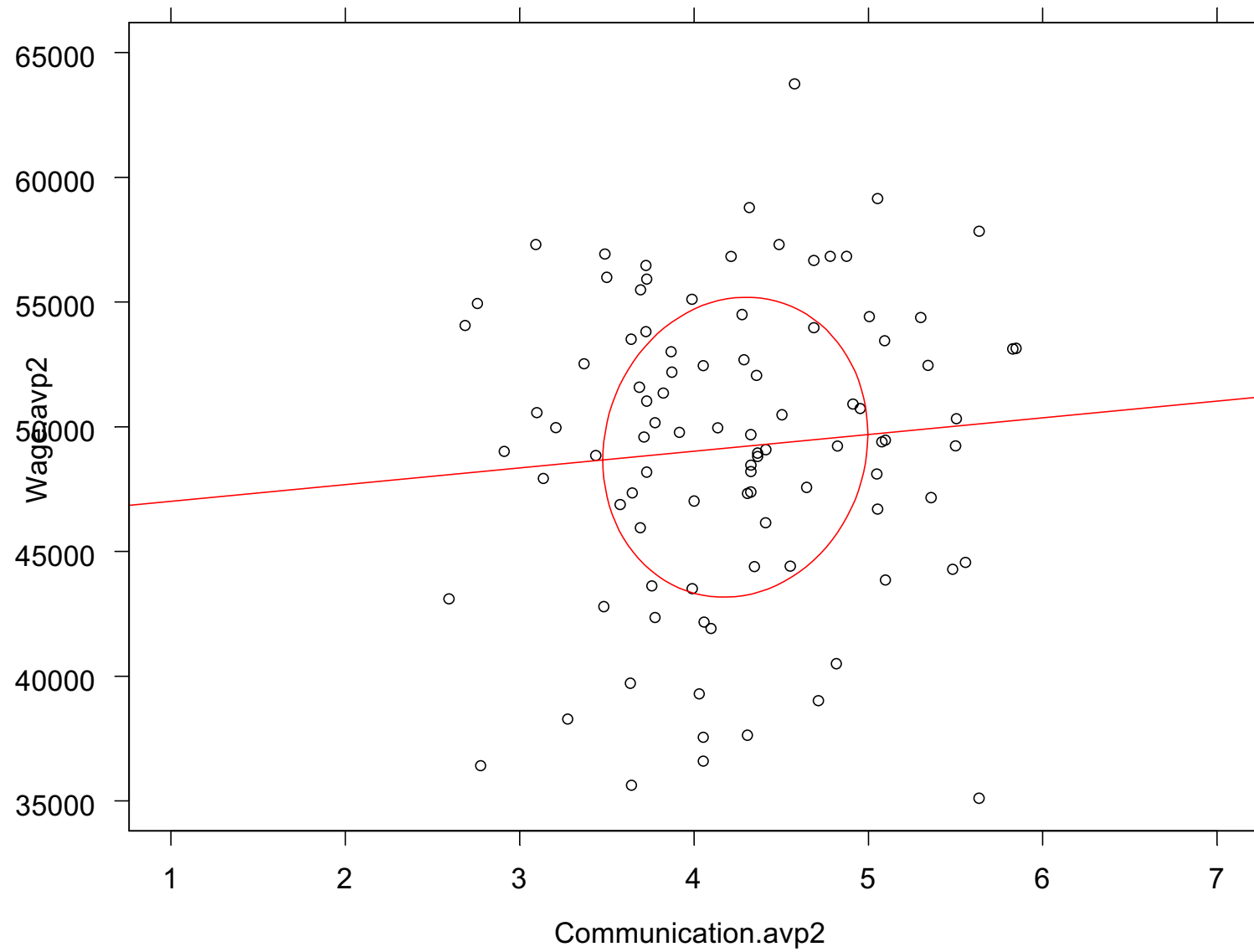
Coefficients:

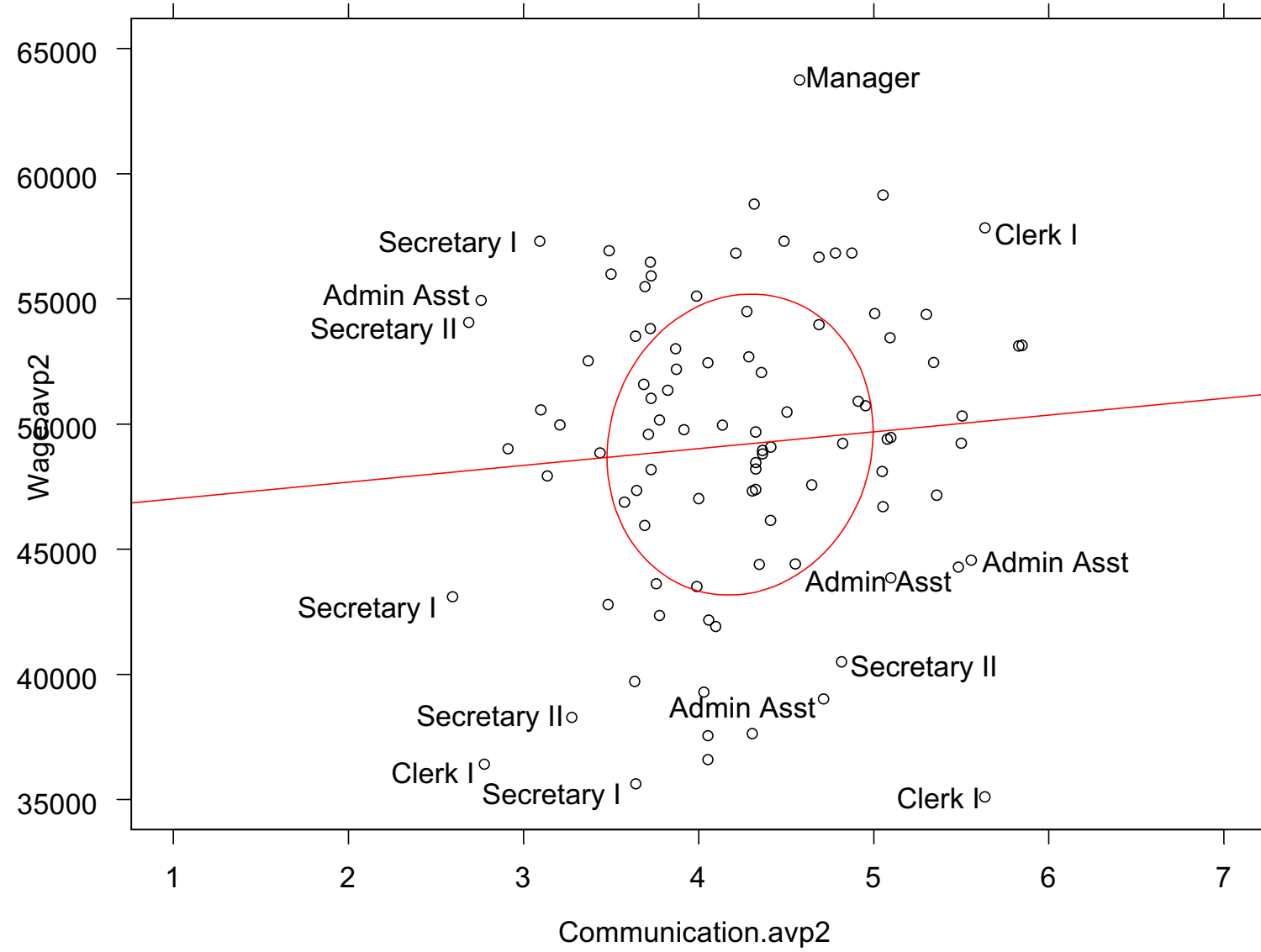
	Value	Std. Error	t value	Pr(> t)
(Intercept)	23317.3282	3869.9155	6.0253	0.0000
Knowledge	3234.6101	857.4269	3.7725	0.0003
Experience	2955.3080	892.0224	3.3130	0.0013
Communication	671.6964	835.8108	0.8036	0.4237
Female	-2606.6538	2069.4971	-1.2596	0.2111
Recept	-13098.1539	3539.6990	-3.7004	0.0004

Residual standard error: 6158 on 89 degrees of freedom

Multiple R-Squared: 0.8149

F-statistic: 78.37 on 5 and 89 degrees of freedom, the p-value is





17 References:

- Berk K.N. (1998) “Regression diagnostic plots in 3-D,” *Technometrics*, 40 (1): pp. 39-47.
- Fox, John (1997) *Applied Regression Analysis, Linear Models, and Related Methods*. Sage.
- Freedman, Pisani and Purves, (1997), *Statistics*, (3rd ed.) Norton.
- Monette, G. (1990). “The Geometry of Multiple Regression and Interactive 3D Graphics,” *Modern Methods of Data Analysis*, (J. Fox and J.S. Long, eds.), Sage, Newbury Park, Ca., pp. 209-256.
- Stone, Mervyn (1987), “Coordinate-free multivariable statistics: An illustrated geometric progression from Halmos to Gauss and Bayes”, Oxford University Press (Oxford)

