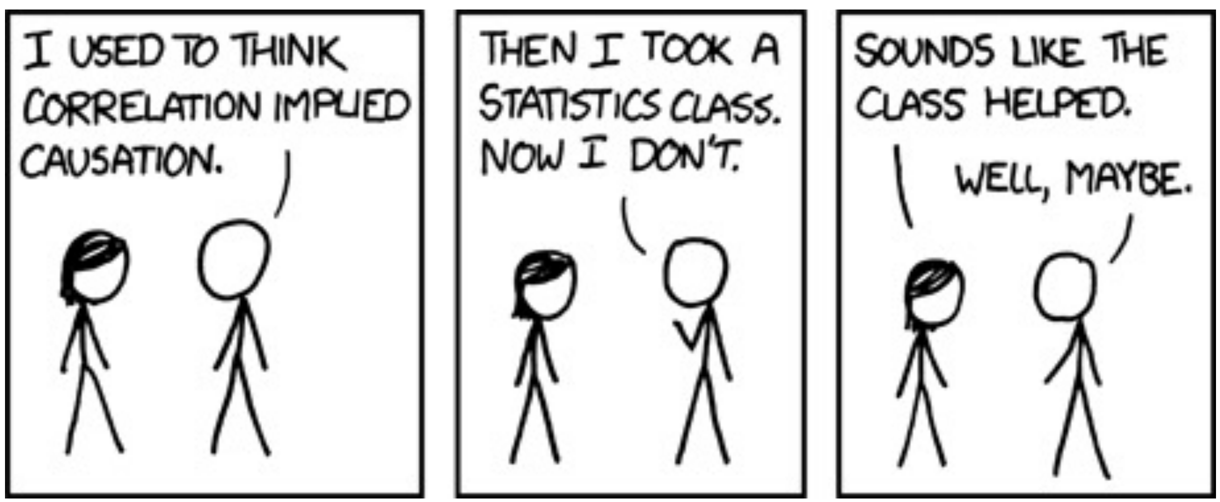
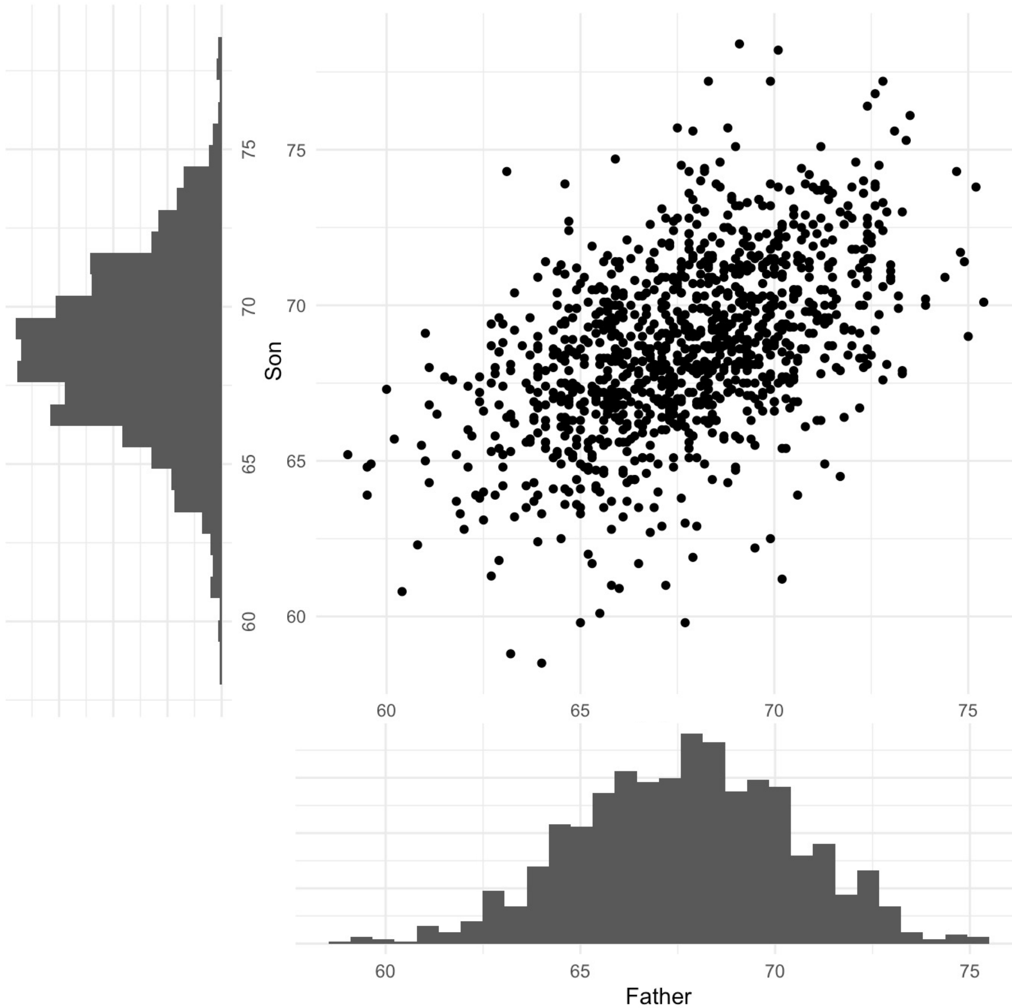
201ab Quantitative methods
L.08: Correlation, regression.



Alt-text:
Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

Projects!

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))
```



Questions we might want to ask:

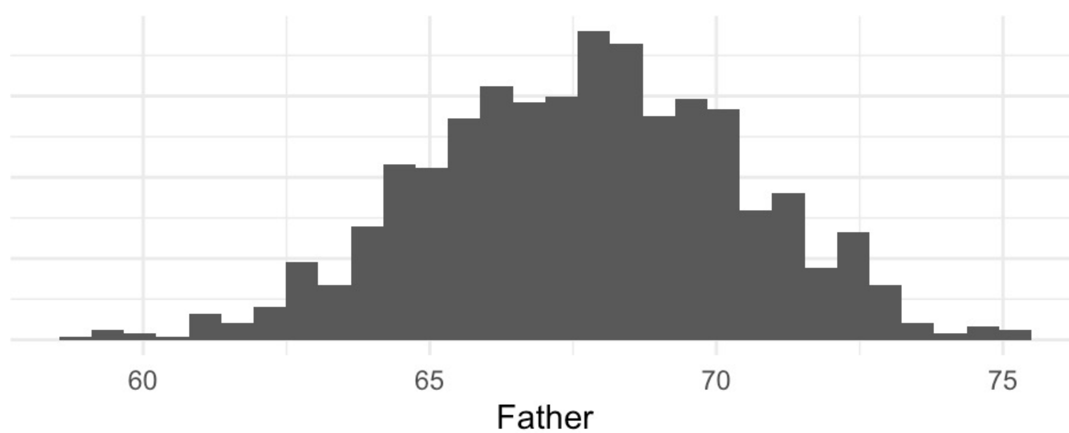
- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - What is our confidence interval on the mean of fathers' heights?
  - What is our prediction interval on the height of a new father?
- Are sons taller than their fathers?
  - Can we reject the null of mean=zero difference?
- What is the relationship between sons' and fathers' heights?

Questions we might want to ask:

- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - What is our confidence interval on the mean of fathers' heights?
  - What is our prediction interval on the height of a new father?
- Are sons taller than their fathers?
  - Can we reject the null of mean=zero difference?
- What is the relationship between sons and fathers heights?

- How do fathers' heights compare to the current UK male mean?
  - **Can we reject the null of the current UK mean?**
  - What is our confidence interval on the mean of fathers' heights?
  - What is our prediction interval on the height of a new father?

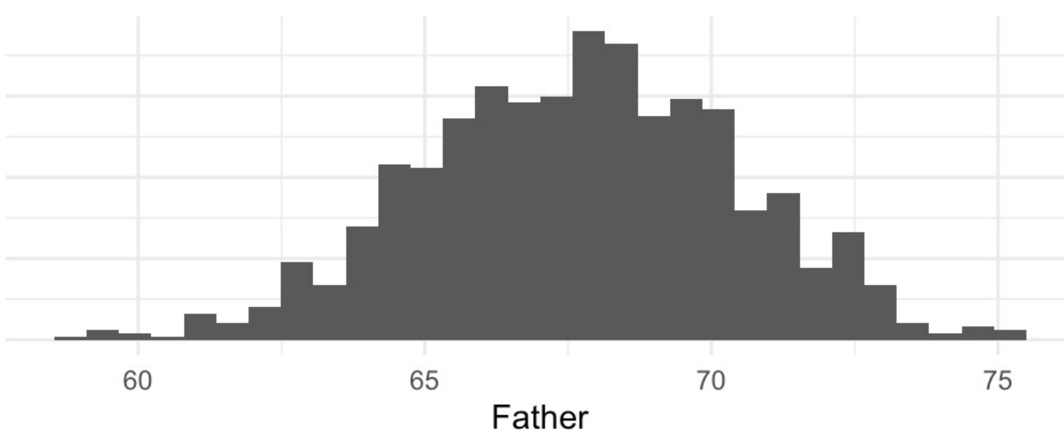
```
> f = fs$Father
> h0mean = 69
> (m = mean(f))
[1] 67.68683
> (n = length(f))
[1] 1078
> (s = sd(f))
[1] 2.745827
> (se_m = s/sqrt(n))
[1] 0.08363033
> (stat = (m-h0mean)/se_m)
[1] -15.70211
```



- How do fathers' heights compare to the current UK male mean?
  - **Can we reject the null of the current UK mean?**
  - What is our confidence interval on the mean of fathers' heights?
  - What is our prediction interval on the height of a new father?

```
> 2*pt(-abs(stat), df = n-1)
[1] 3.457638e-50
> 2*pnorm(-abs(stat))
[1] 1.462962e-55
```

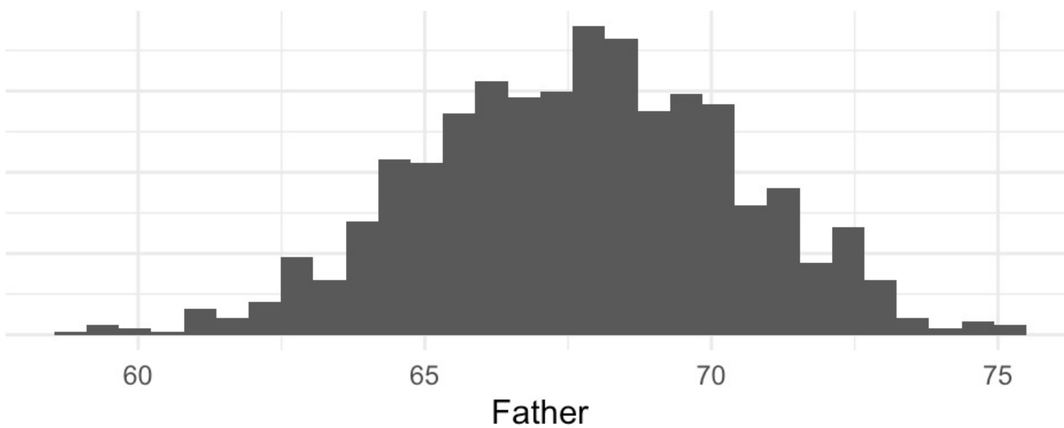
```
> f = fs$Father
> h0mean = 69
> (m = mean(f))
[1] 67.68683
> (n = length(f))
[1] 1078
> (s = sd(f))
[1] 2.745827
> (se_m = s/sqrt(n))
[1] 0.08363033
> (stat = (m-h0mean)/se_m)
[1] -15.70211
```



- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - **What is our confidence interval on the mean of fathers' heights?**
  - What is our prediction interval on the height of a new father?

```
> f = fs$Father
> h0mean = 69
> (m = mean(f))
[1] 67.68683
> (n = length(f))
[1] 1078
> (s = sd(f))
[1] 2.745827
> (se_m = s/sqrt(n))
[1] 0.08363033
> (stat = (m-h0mean)/se_m)
[1] -15.70211
```

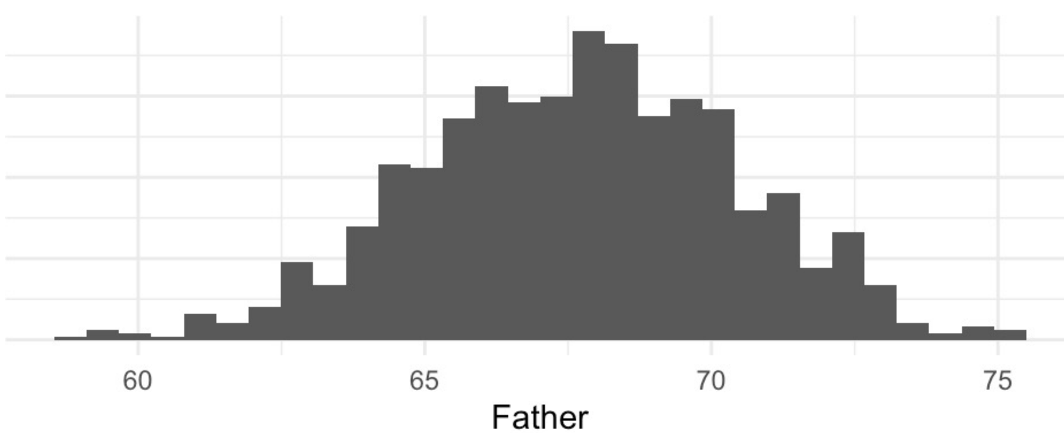
```
> (crit = abs(qt((1-0.95)/2, df = n-1)))
[1] 1.962169
> abs(qnorm((1-0.95)/2))
[1] 1.959964
> m + c(-1,1)*crit*se_m
[1] 67.52273 67.85092
```



- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - What is our confidence interval on the mean of fathers' heights?
  - **What is our prediction interval on the height of a new father?**

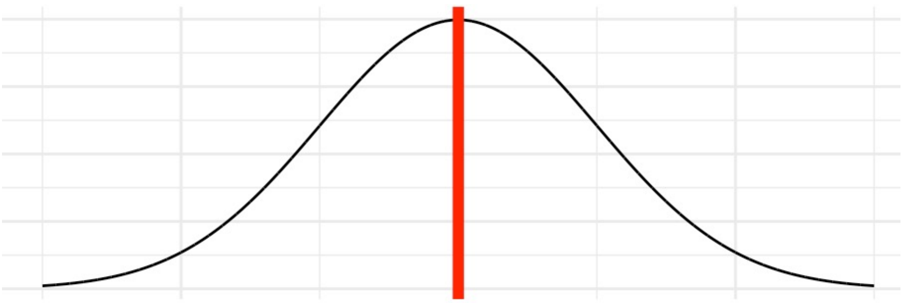
```
> f = fs$Father
> h0mean = 69
> (m = mean(f))
[1] 67.68683
> (n = length(f))
[1] 1078
> (s = sd(f))
[1] 2.745827
> (se_m = s/sqrt(n))
[1] 0.08363033
> (stat = (m-h0mean)/se_m)
[1] -15.70211
```

```
> (s_new = sqrt(s^2 + se_m^2))
[1] 2.7471
> m + c(-1,1)*crit*s_new
[1] 62.29655 73.07710
```



Linear model formulation

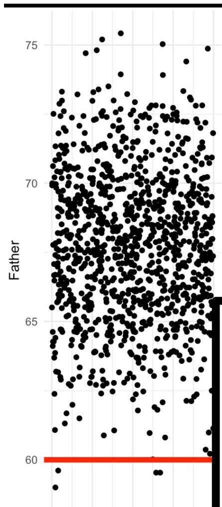
$$y_i = (1) \cdot \beta_0 + \epsilon_i$$
$$\epsilon_i \sim \mathbf{N}(0, \sigma_\epsilon)$$



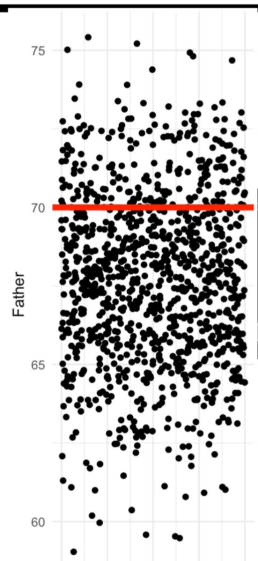
```
lm(f ~ 1)
```

Least squares fit.

```
tibble(n= n,
       SS= sum((f-b0)^2)) %>%
  mutate(MSE = SS/n,
         RMSE = sqrt(MSE))
```

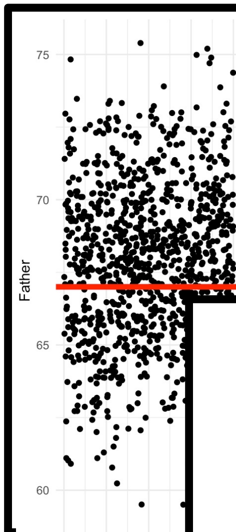


	n	SS	MSE	RMSE
	<int>	<dbl>	<dbl>	<dbl>
1	1078	71816.	66.6	8.16



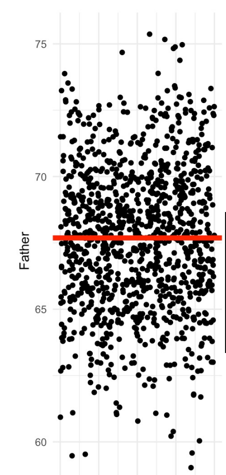
A tibble: 1 x 4

	n	SS	MSE	RMSE
	<int>	<dbl>	<dbl>	<dbl>
1	1078	13888.	12.9	3.59



A tibble: 1 x 4

	n	SS	MSE	RMSE
	<int>	<dbl>	<dbl>	<dbl>
1	1078	8629.	8.00	2.83



	n	SS	MSE	RMSE
	<int>	<dbl>	<dbl>	<dbl>
1	1078	8120.	7.53	2.74

- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - What is our confidence interval on the mean of fathers' heights?
  - What is our prediction interval on the height of a new father?

```
> lm(f~1) %>% summary()

Call:
lm(formula = f ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6868 -1.8868  0.1132  1.9132  7.7132 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 67.68683    0.08363   809.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.746 on 1077 degrees of freedom
```

```
> f = fs$Father
> h0mean = 69
> (m = mean(f))
[1] 67.68683
> (n = length(f))
[1] 1078
> (s = sd(f))
[1] 2.745827
> (se_m = s/sqrt(n))
[1] 0.08363033
> (stat = (m-h0mean)/se_m)
[1] -15.70211
```

```
> 2*pt(-abs(stat), df = n-1)
[1] 3.457638e-50
```

- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - What is our confidence interval on the mean of fathers' heights?
  - **What is our prediction interval on the height of a new father?**

```
> lm(f~1) %>%
+   predict.lm(newdata= data.frame(x=1),
+              interval = 'prediction',
+              level = 0.95)
       fit      lwr      upr
1 67.68683 62.29655 73.0771
```

```
> (s_new = sqrt(s^2 + se_m^2))
[1] 2.7471
> m + c(-1,1)*crit*s_new
[1] 62.29655 73.07710
```

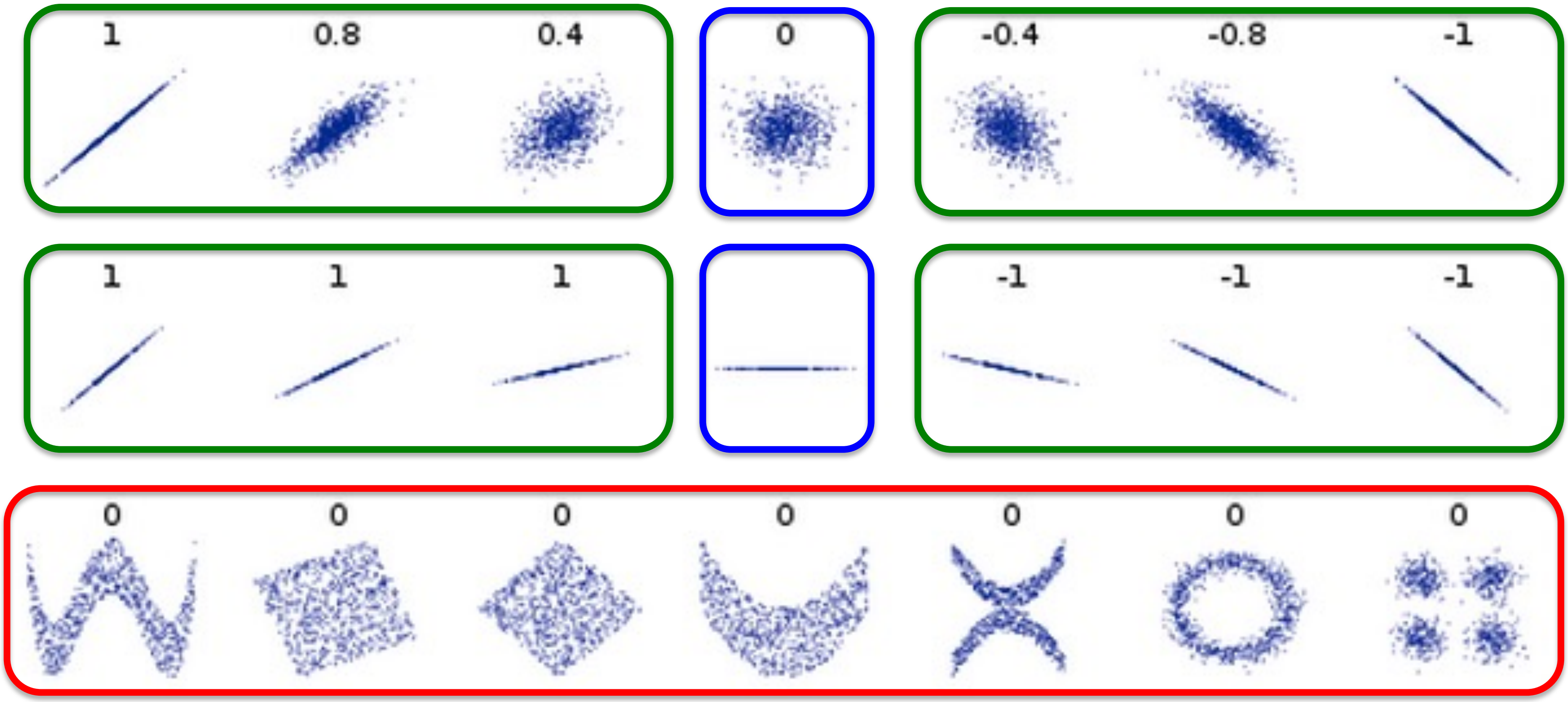
Evaluating a mean

- Fitting a mean, on the assumption of gaussian variability…
  - Requires that we use a t-distribution to respect the uncertainty of our standard deviation estimate.
  - Is the simplest/smallest ”linear model”: (just an intercept term)

Questions we might want to ask:

- How do fathers' heights compare to the current UK male mean?
  - Can we reject the null of the current UK mean?
  - What is our confidence interval on the mean of fathers' heights?
  - What is our prediction interval on the height of a new father?
- Are sons taller than their fathers?
  - Can we reject the null of mean=zero difference?
- What is the relationship between sons and fathers heights?

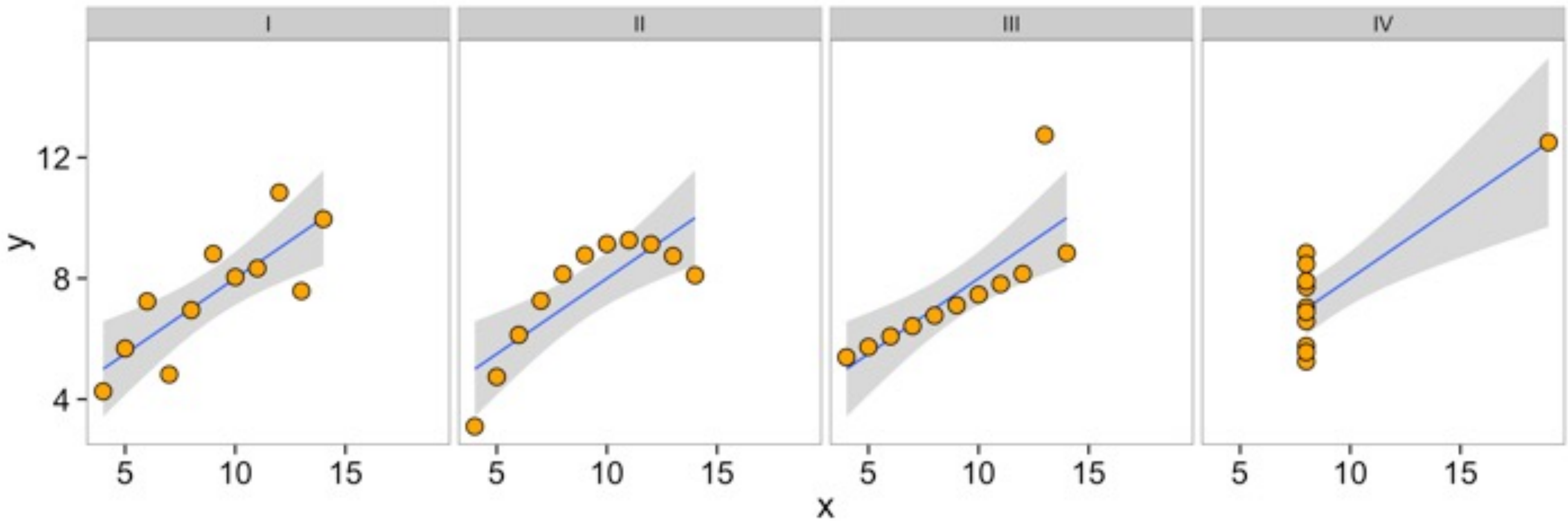
Relationship between two variables



X and Y can be…

- Independent.
- Dependent, but not linearly (tricky to measure in general)
- Linearly dependent (this is what we are going to measure)

Anscombe's quartet



Property	Value
Mean of *x* in each case	9 (exact)
Sample variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Sample variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

You can always fit a line; doesn't mean it's a good idea.

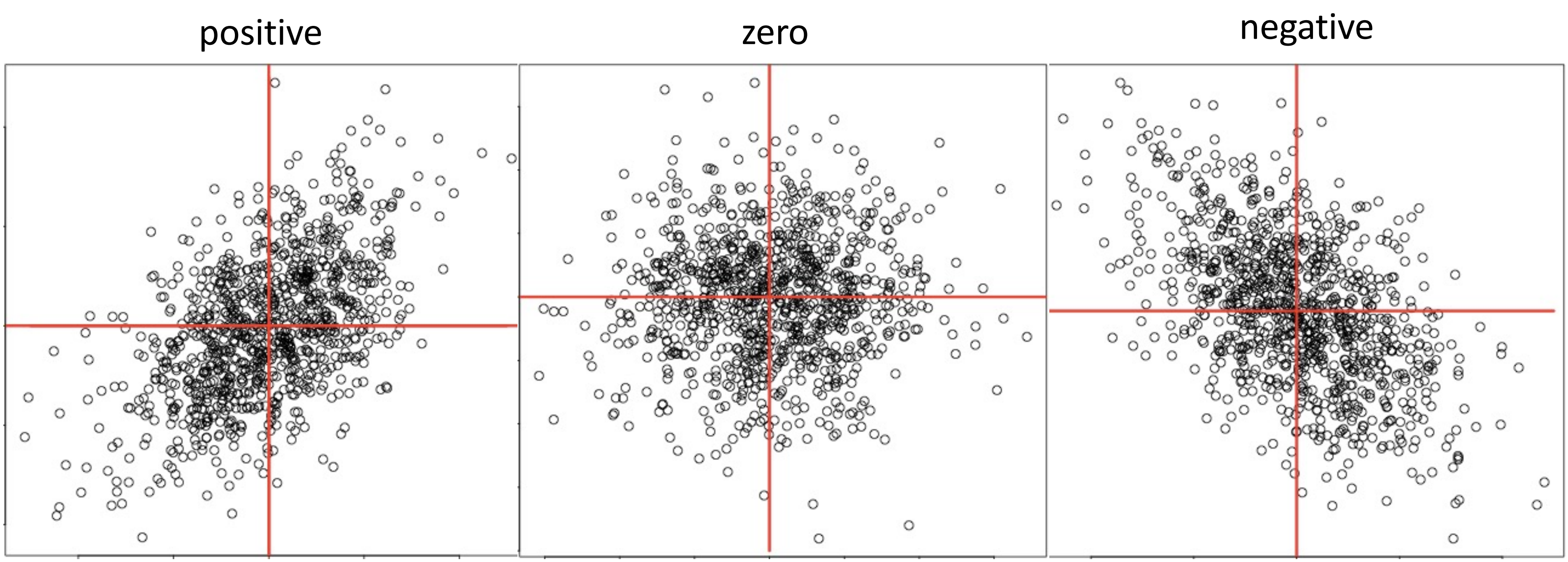
Measures of linear relationship

- *Covariance*: shared variance between x and y
- *Correlation*: standardized covariance
- *Coefficient of determination*: how much variance is captured by linear relationship.

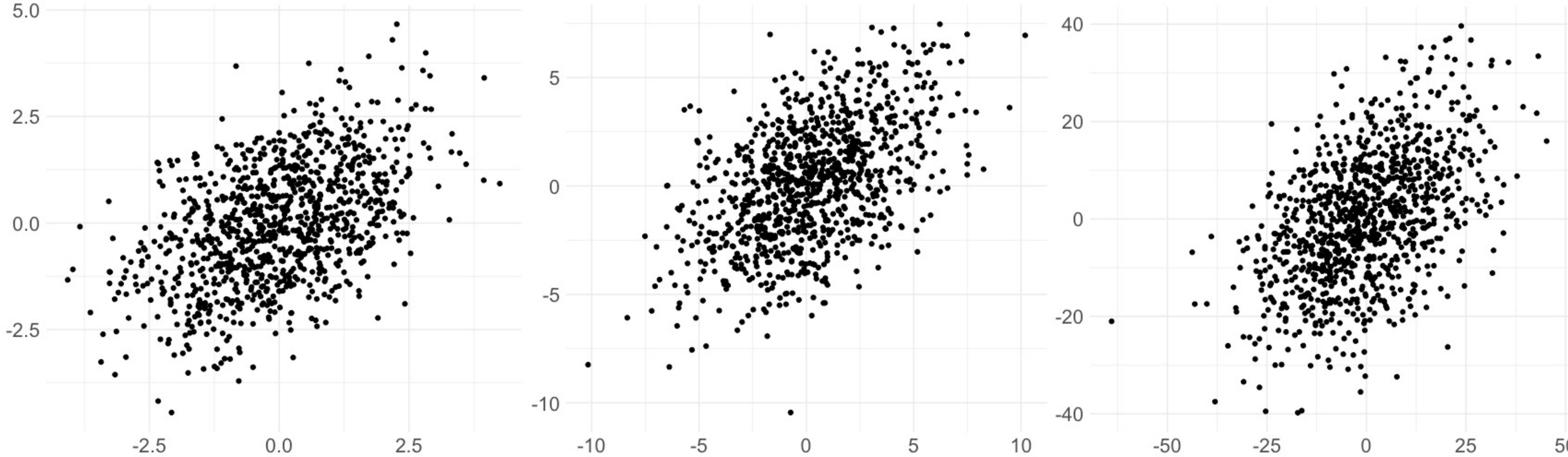
- *Regression slope of y~x*: predict y for given x (minimizing squared deviation of y from prediction)
- *Regression slope of x~y*: predict x for given y (minimizing squared deviation of x from prediction)
- *Principle component line*: (minimize squared deviation of (x,y) from line.)

Covariance: varying together.

When X deviates from the mean, does Y deviate from its mean. What is the size and direction of these shared deviations?



Covariance and correlation



Covariance (s_{xy})	1	4	100
Correlation (r)	0.5	0.5	0.5

Covariance: magnitude of shared variance.

Covariance will change with unit rescaling (heights in cm vs in)

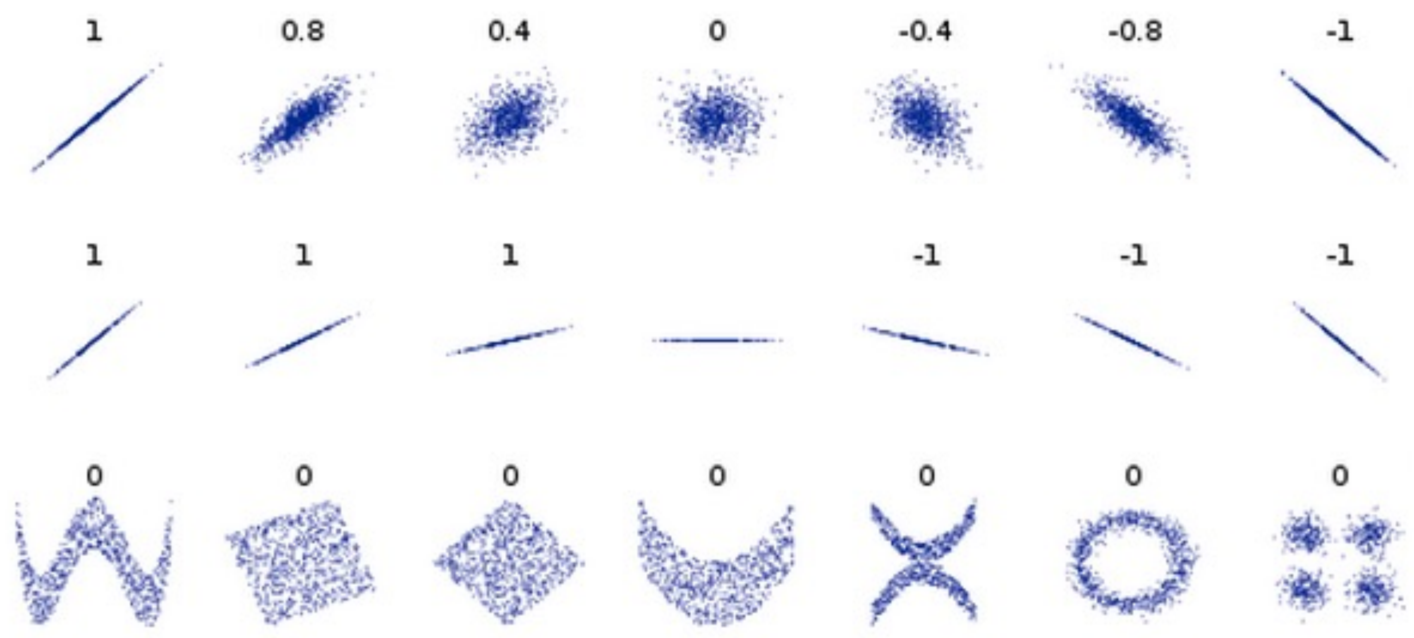
Correlation: Covariance scaled by the (marginal) variances of x and y

Correlation will not change with rescaling.

Correlation

Covariance scaled to the overall variances.

Between -1 and 1.
Measures direction, strength of linear relationship



Closer to 0 when variables are more independent.

Only sign of slope matters.

Non-linear relationships don't count.

Calculation correlation, covariance

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))        f = fs$Father;  s = fs$Son
```

```
cov(f,s)
```

3.8733

```
cov(f,f)
```

7.539566

```
var(f)
```

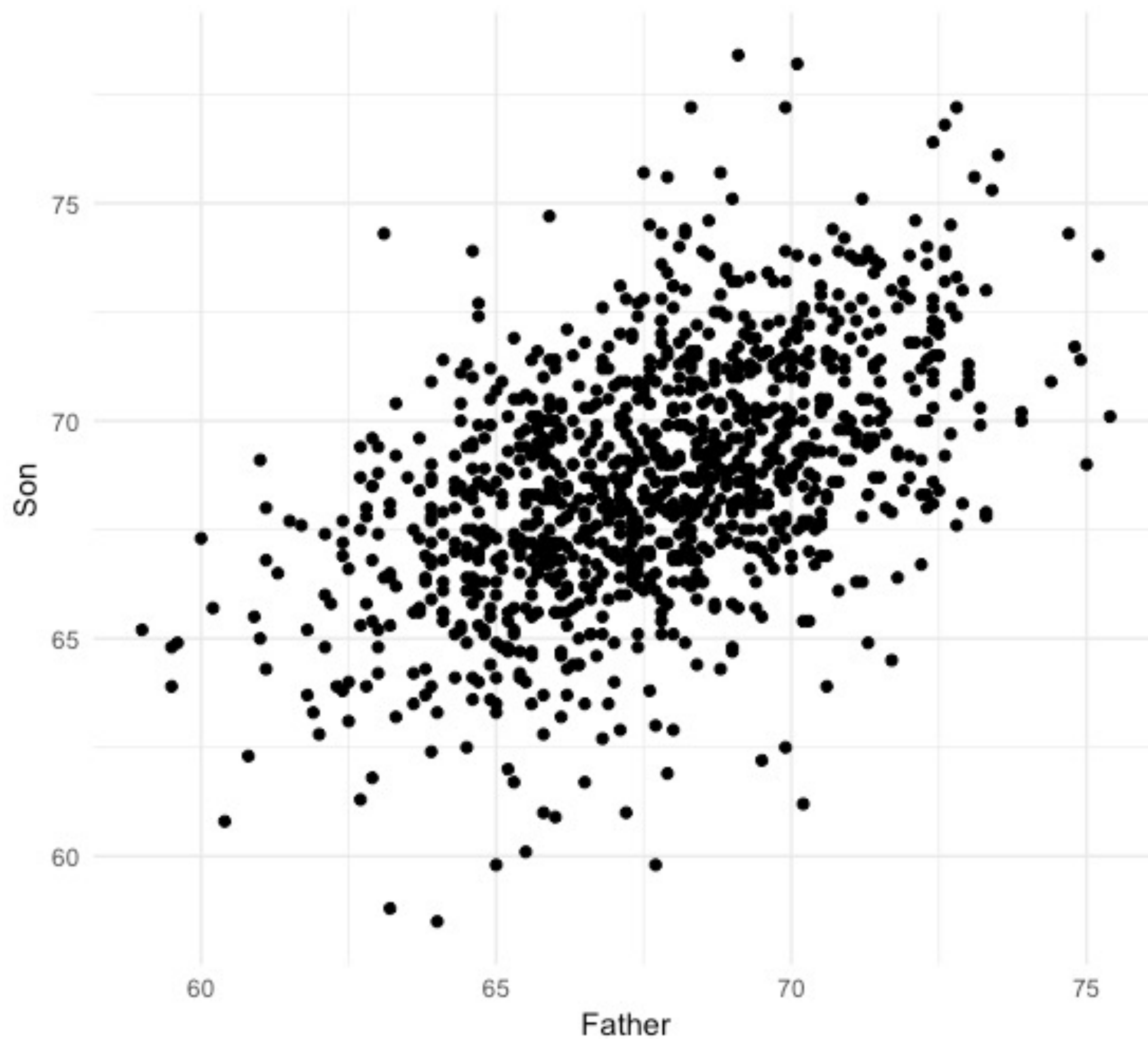
7.539566

```
cor(f,s)
```

0.5011627

```
cov(f,s)/(sd(f)*sd(s))
```

0.5011627



Calculation correlation, covariance

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
cov(f,s)
                                   3.8733
```

$$s_{xy} = \frac{1}{n-1} SP[x,y] = \frac{1}{n-1} \sum_{i=1}^{n} \left[(x_i - \bar{x})(y_i - \bar{y}) \right]$$

Sample covariance

"sum of products"

```
cov(f,f)
                                 7.539566
var(f)
                                 7.539566
```

$$s_x^2 = \frac{1}{n-1} SS[x] = \frac{1}{n-1} \sum_{i=1}^{n} \left[(x_i - \bar{x})(x_i - \bar{x}) \right]$$

Sample variance

"sum of squares"

```
cor(f,s)
                                0.5011627
cov(f,s)/(sd(f)*sd(s))
                                0.5011627
```

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{SP[x,y]}{\sqrt{SS[x] * SS[y]}}$$

Sample correlation

The covariance matrix

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
cov(fs)
```

```
         Father      Son
Father 7.539566 3.875382
Son    3.875382 7.930949
```

var(f)	cov(f,s)
7.539566	3.875382
cov(f,s)	**var(s)**
3.875382	7.930949

Linear transformations

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))
```

Original variables

```
f = fs$Father
s = fs$Son
```

mean(f)	67.68683
mean(s)	68.68423
sd(f)	2.745827
sd(s)	2.816194
cov(f,s)	3.875382
cor(f,s)	0.5011627

Shifted variables

```
f = fs$Father + 2
s = fs$Son + 3
```

mean(f)	69.68683
mean(s)	71.68423
sd(f)	2.745827
sd(s)	2.816194
cov(f,s)	3.875382
cor(f,s)	0.5011627

Scaled variables

```
f = fs$Father * 2
s = fs$Son * 3
```

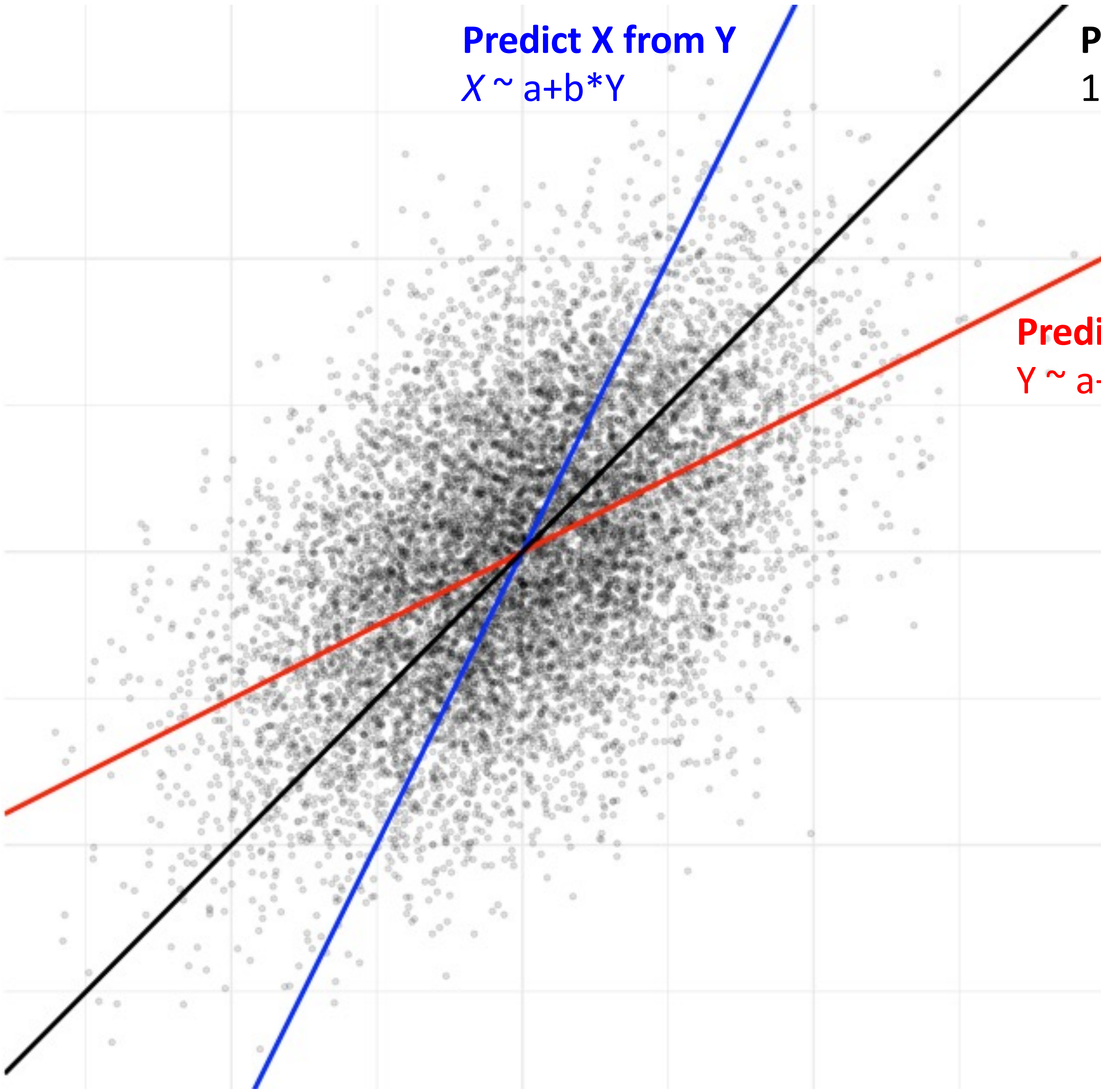
mean(f)	135.3737
mean(s)	206.0527
sd(f)	5.491654
sd(s)	8.448582
cov(f,s)	23.25229
cor(f,s)	0.5011627

Shifting influences the mean, nothing else.
Scaling changes mean, variance, sd, covariance, but not the *correlation*:

The correlation normalizes the covariance to the sd of x,y, so is constant.

What line would you draw?

Different regressions, lines



They all go through
[mean(x), mean(y)]

They differ in their definition
of error to be minimized.

OLS regression model.

Y is is a line (w.r.t. X) plus "error"

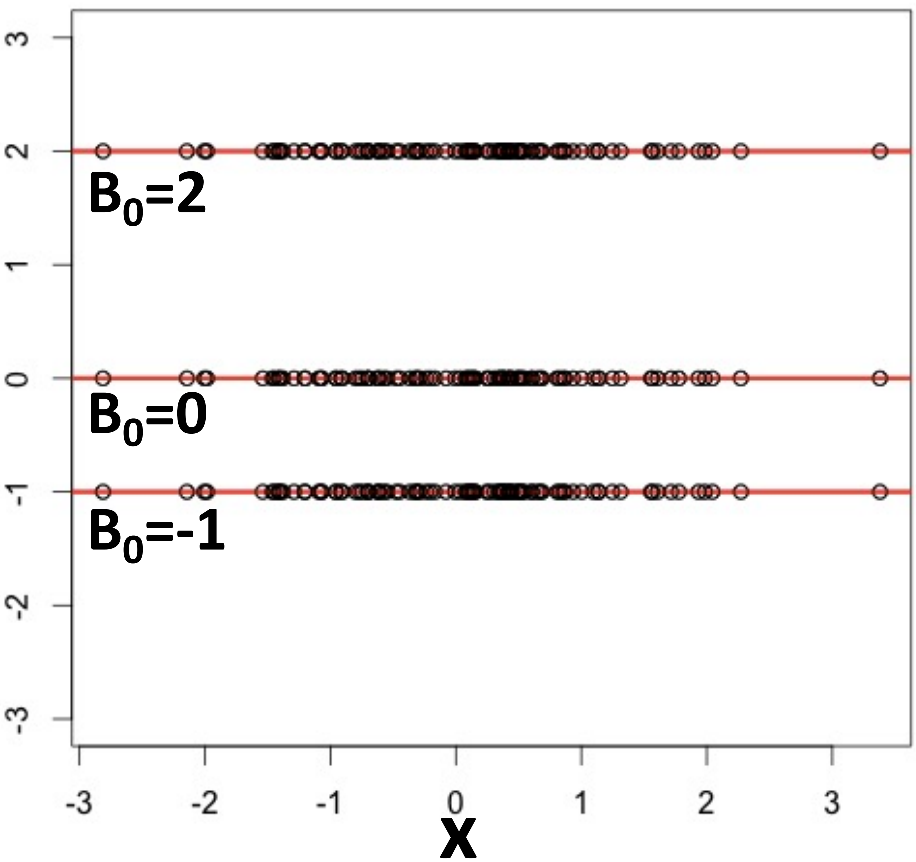
$$\boxed{Y_i} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{X_i} + \boxed{\varepsilon_i}$$

$$\boxed{\text{Score on Y for the ith individual}} = \boxed{\text{Y Intercept}} + \left(\boxed{\text{Slope } (\textit{Effect})} \times \boxed{\text{Score on X for the ith individual}}\right) + \boxed{\text{Error}}$$

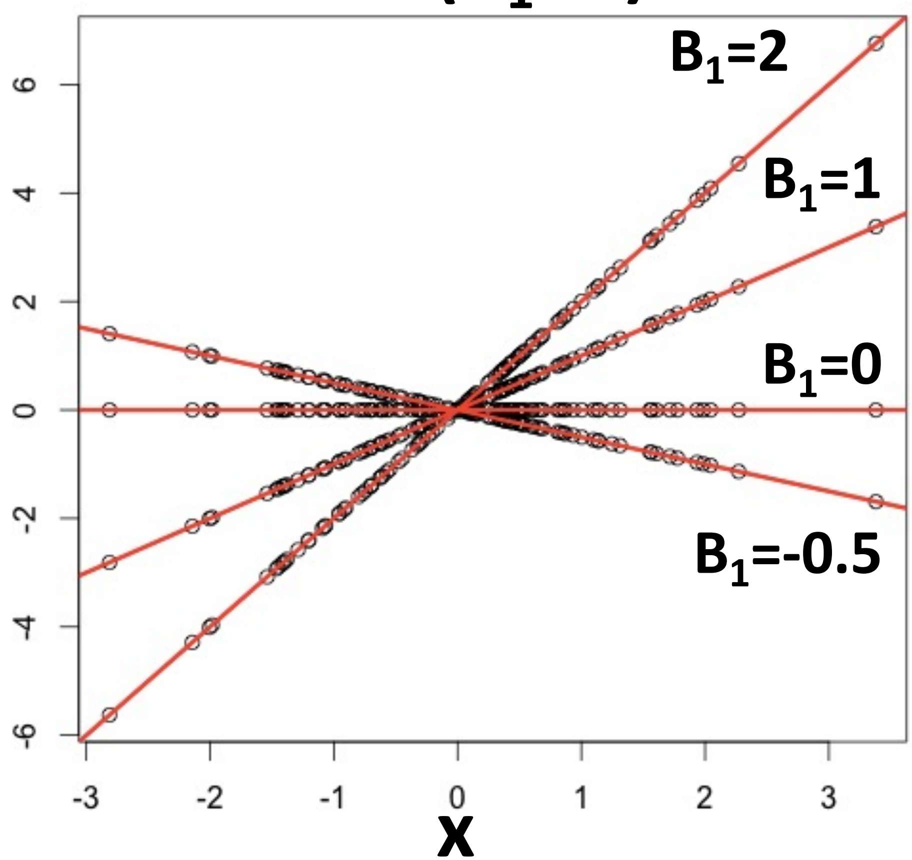
Error assumed to be independent, identically distributed, Gaussian noise.

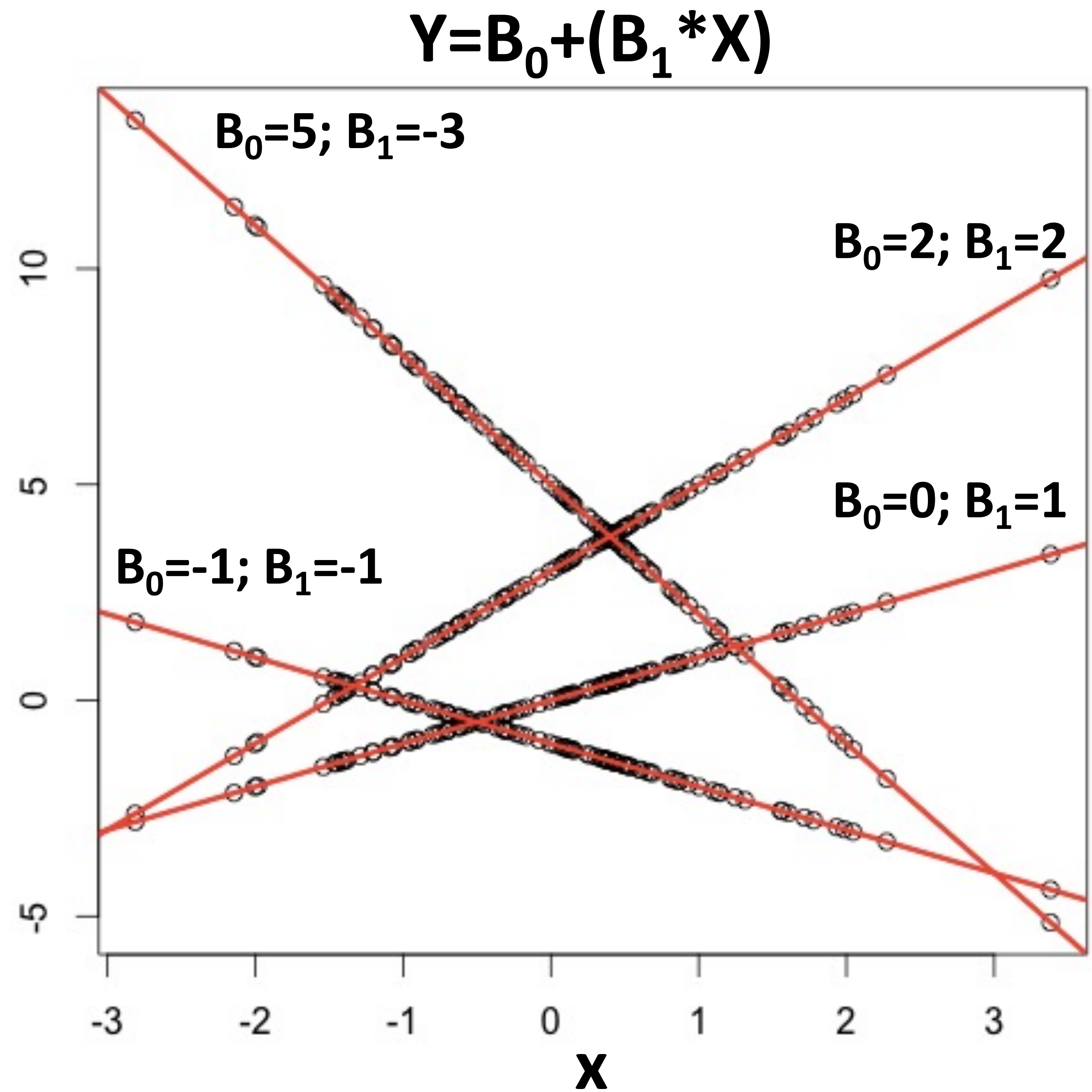
$$\varepsilon_i \sim N(0, \sigma_\varepsilon)$$

$Y=B_0+(0*X)$

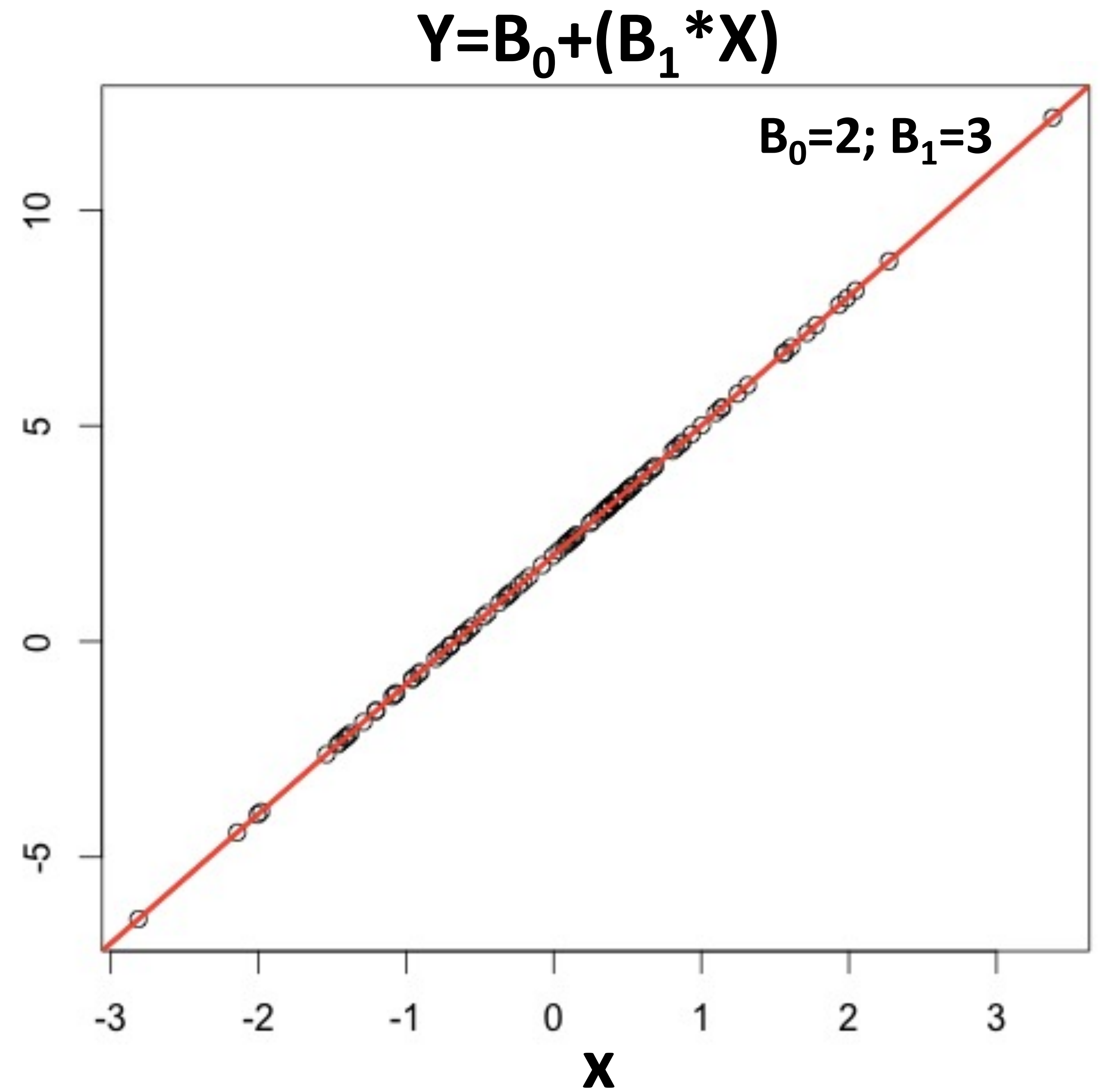


$Y=0+(B_1*X)$



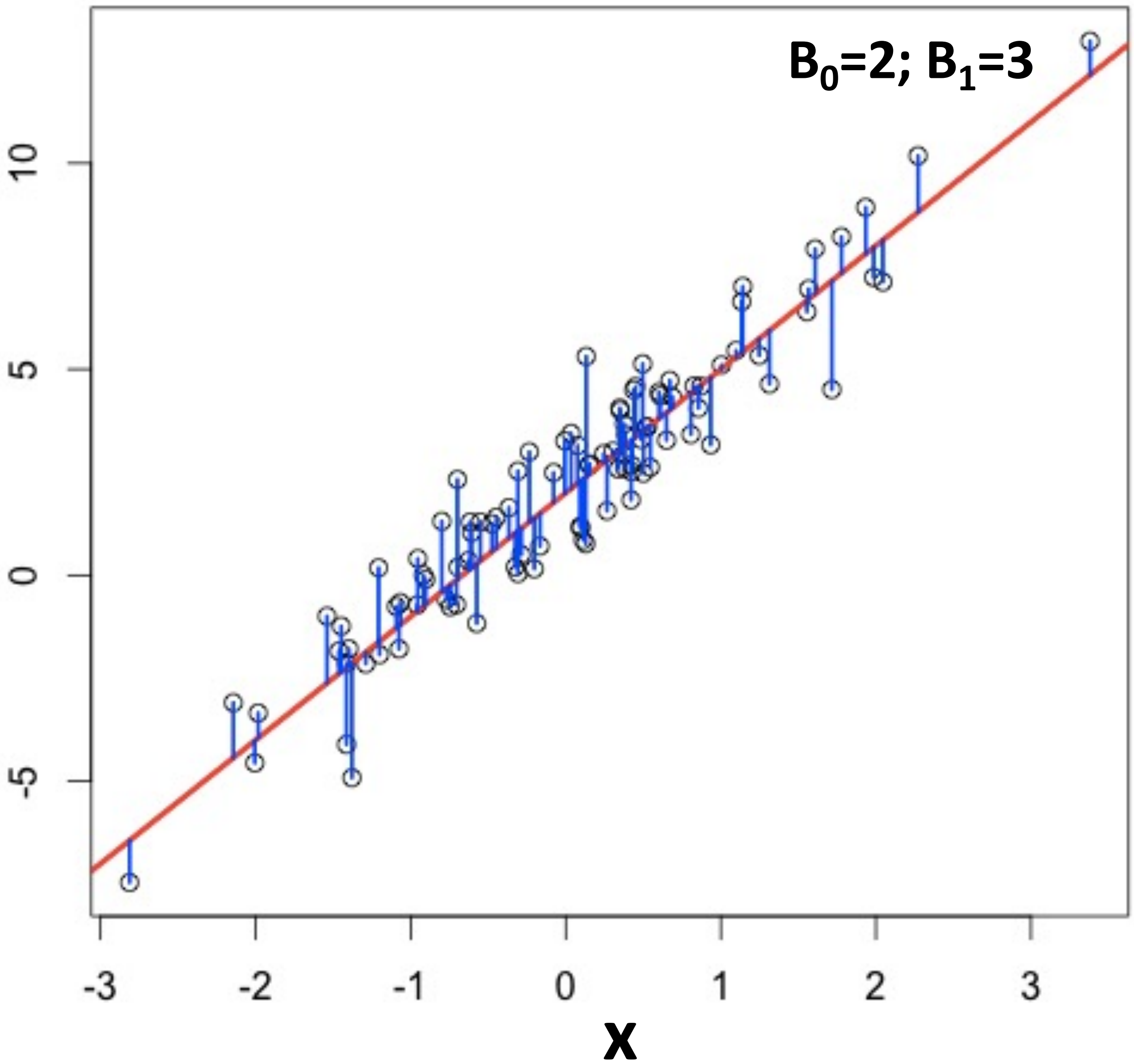


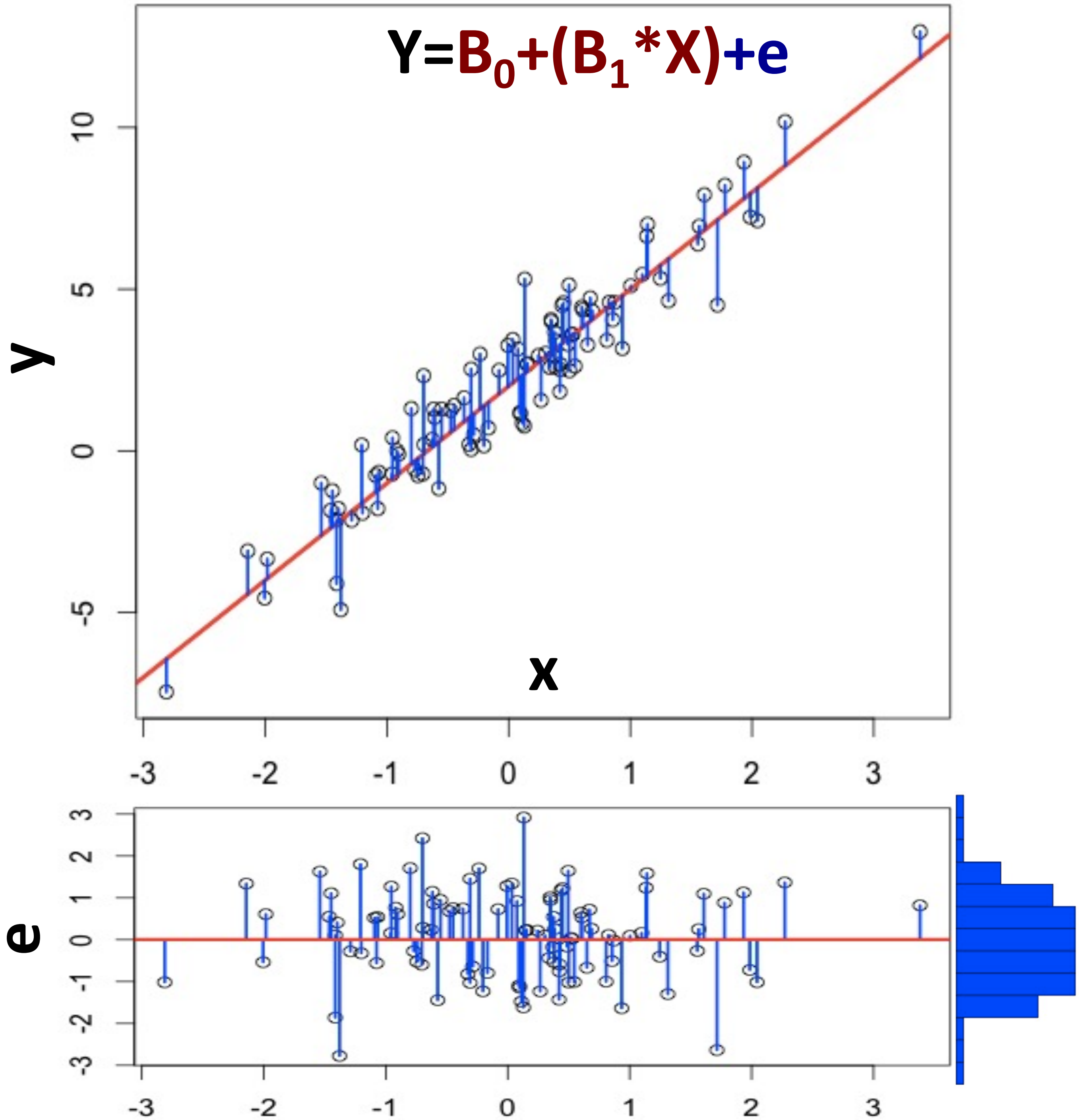
Y=B0+(B1*X)
B0=5; B1=-3
B0=2; B1=2
B0=0; B1=1
B0=-1; B1=-1
10
5
0
-5
-3
-2
-1
0
1
2
3
X



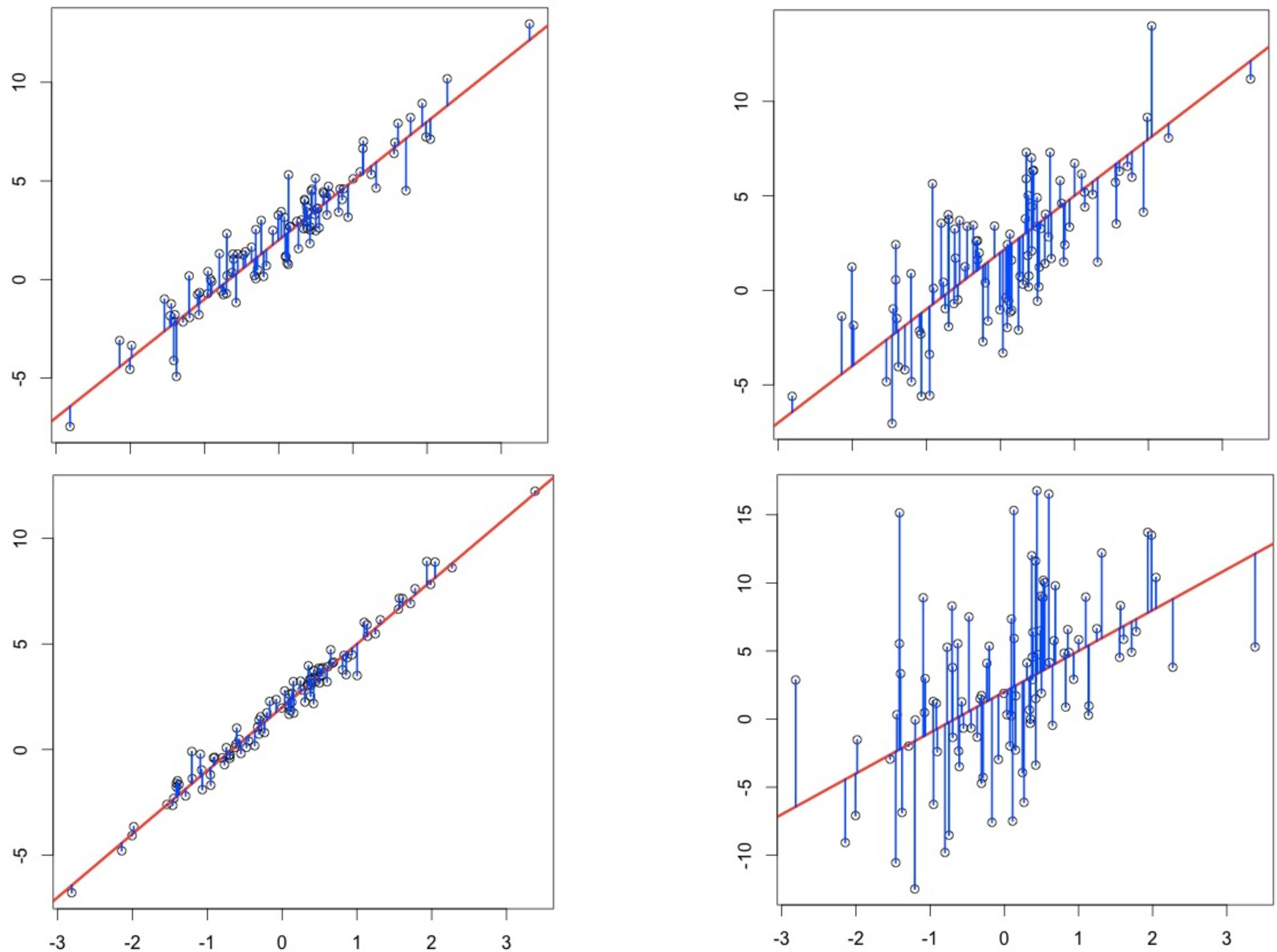
Y=B0+(B1*X)
B0=2; B1=3
10
5
0
-5
-3
-2
-1
0
1
2
3
X

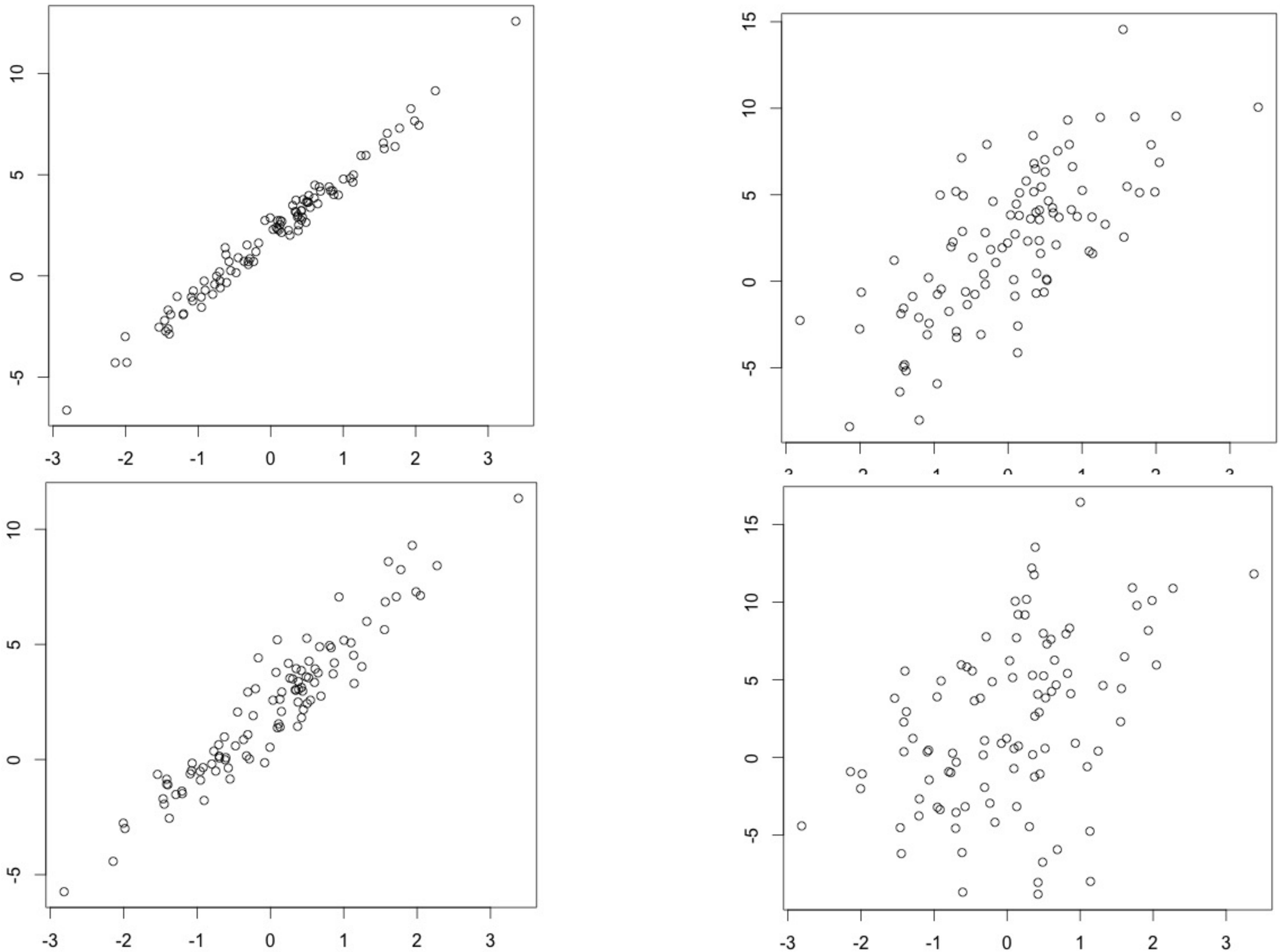
$Y = B_0 + (B_1 * X) + e$





Y=B0+(B1*X)+e
y
x
e
10
5
0
-5
-3
-2
-1
0
1
2
3


10
5
0
-5
-3
-2
-1
0
1
2
3
15
-10



OLS regression model.

Y is is a line (w.r.t. X) plus “error”

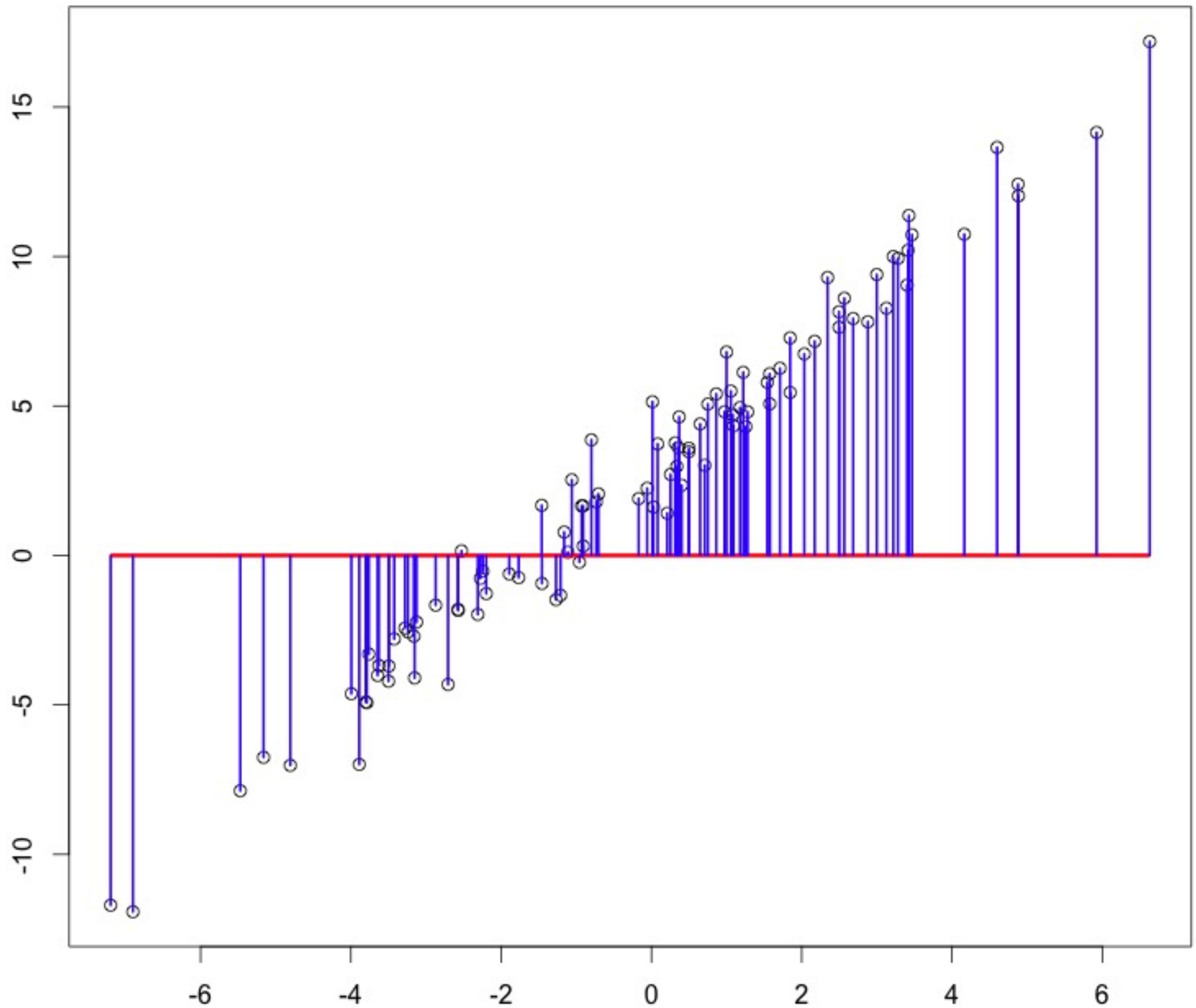
$$\boxed{Y_i} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{X_i} + \boxed{\varepsilon_i}$$

$$\boxed{\begin{array}{c}\text{Score on Y}\\ \text{for the ith}\\ \text{individual}\end{array}} = \boxed{\begin{array}{c}\text{Y}\\ \text{Intercept}\end{array}} + \left(\boxed{\begin{array}{c}\text{Slope}\\ \textit{(Effect)}\end{array}} \times \boxed{\begin{array}{c}\text{Score on X}\\ \text{for the ith}\\ \text{individual}\end{array}}\right) + \boxed{\text{Error}}$$

Inference goal is to estimate B0, B1, error.

This is harder when there is more error.

Minimize squared error *in y*



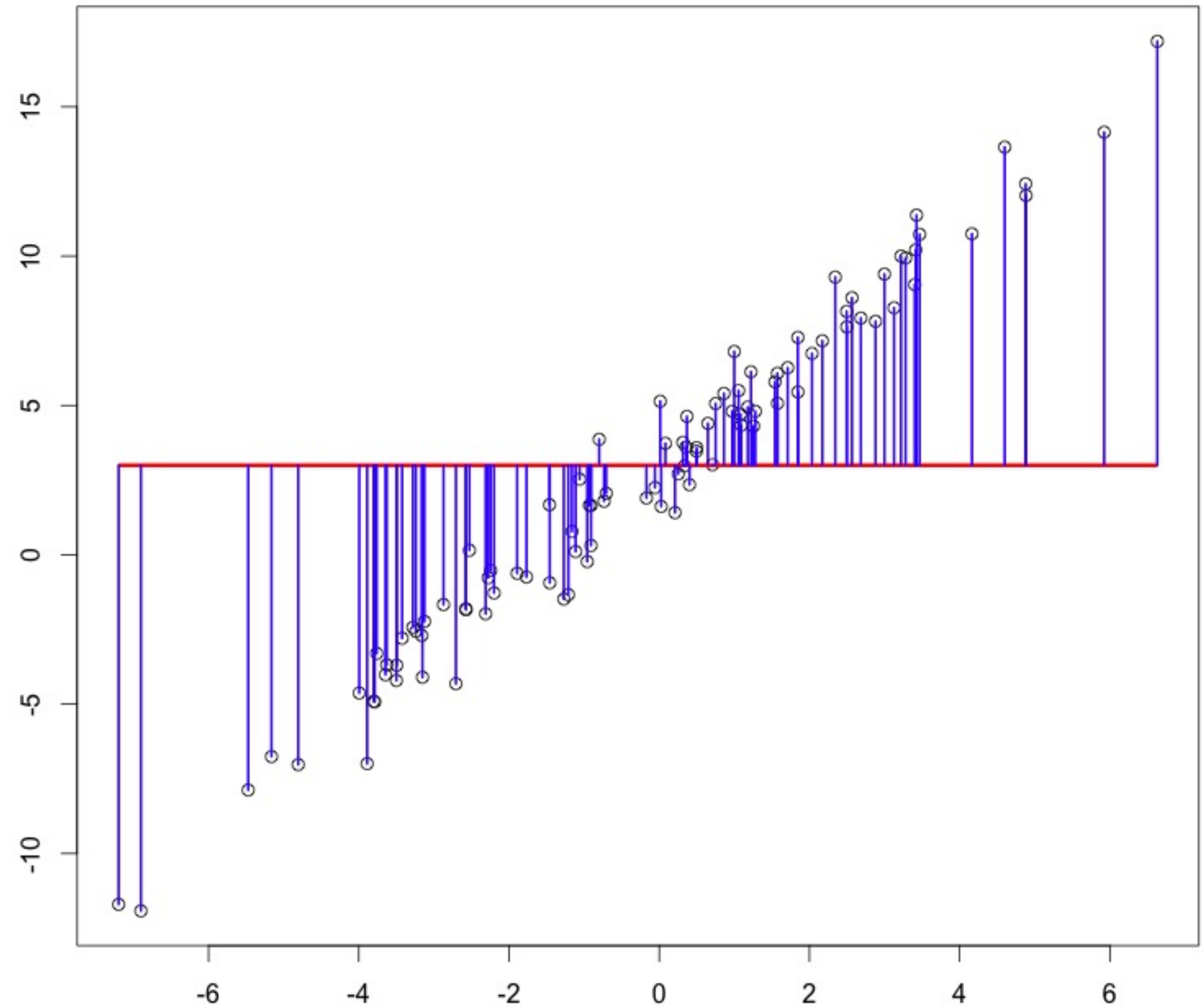
$$SS[error] = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i \right)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sum of squared error = 3837

Minimize squared error *in y*



$$SS[error] = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

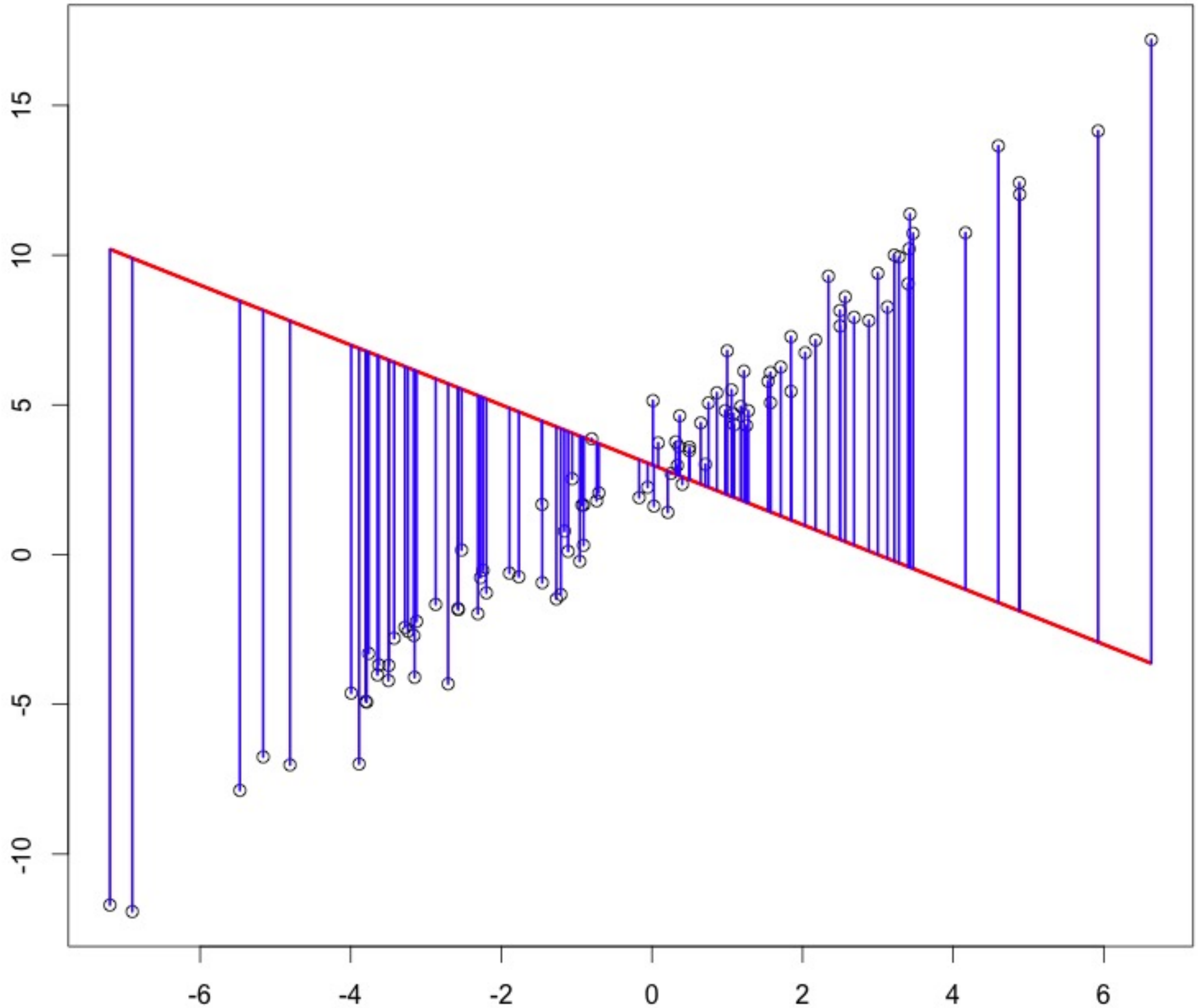
$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i\right)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sum of squared error = 3174

If we don't get to vary slope from 0, our squared error minimizing line is the horizontal that passes through the mean of y.

Minimize squared error *in y*



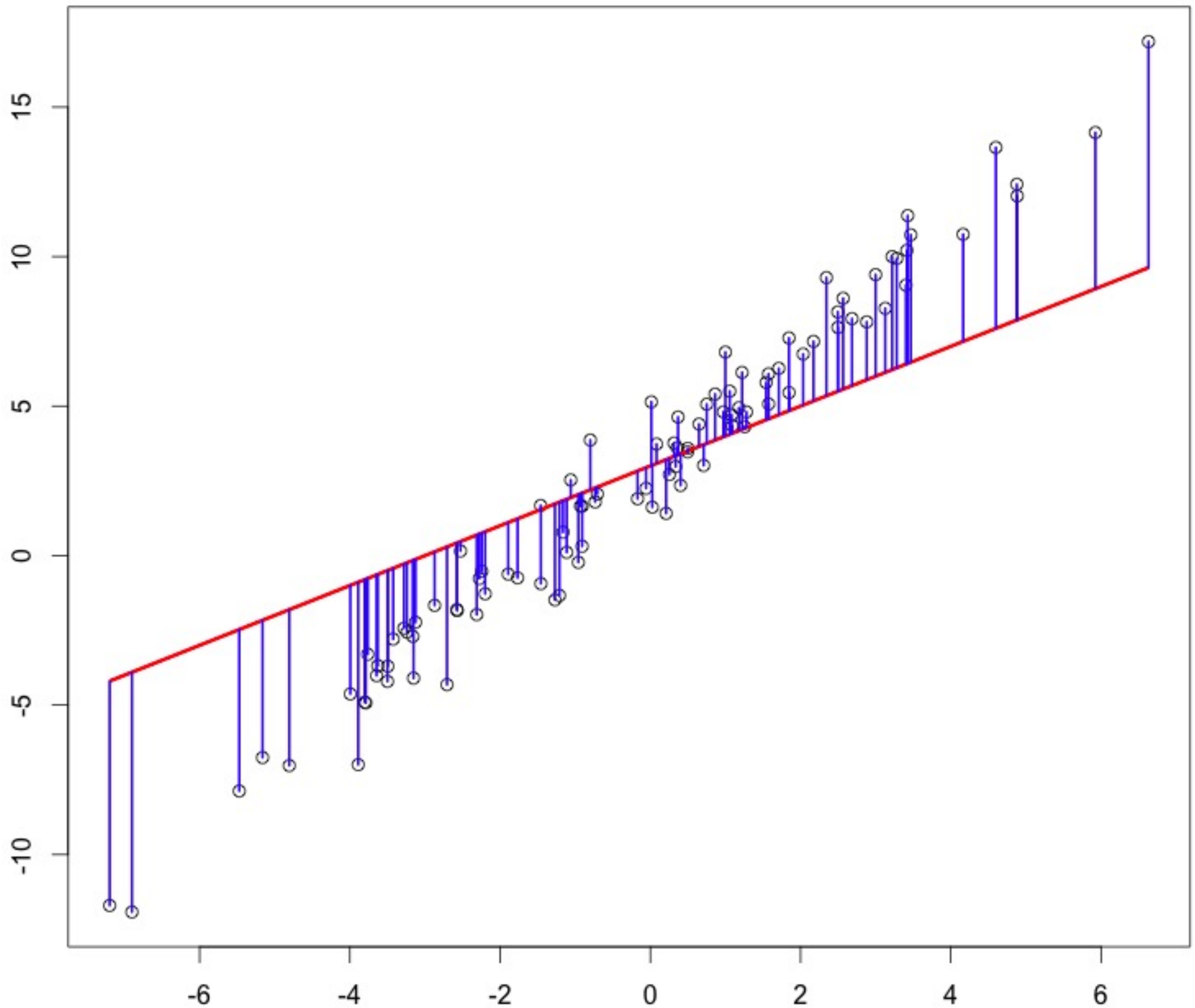
$$SS[error] = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i\right)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sum of squared error = 7050

Minimize squared error *in y*



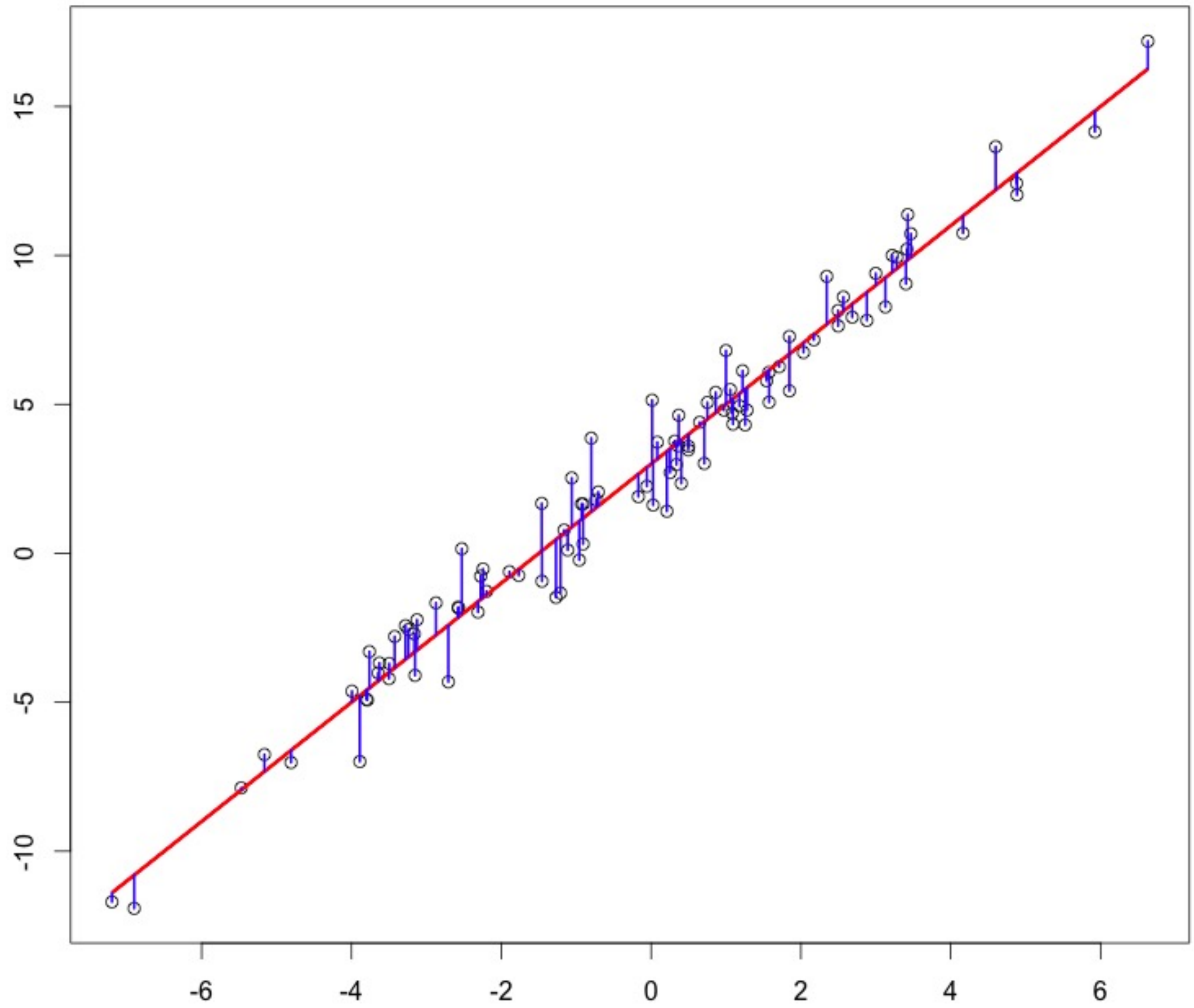
$$SS[error] = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i\right)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sum of squared error = 855

Minimize squared error *in y*



$$SS[error] = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i \right)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sum of squared error = 93

Regression in R via lm( )

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))        f = fs$Father;  s = fs$Son
```

```
lm(data = fs, Son~Father)
```

```
                Coefficients:
(Intercept)           Father
     33.893            0.514
```

Formula syntax:
response ~ explanatory variables

Regression in R via lm( )

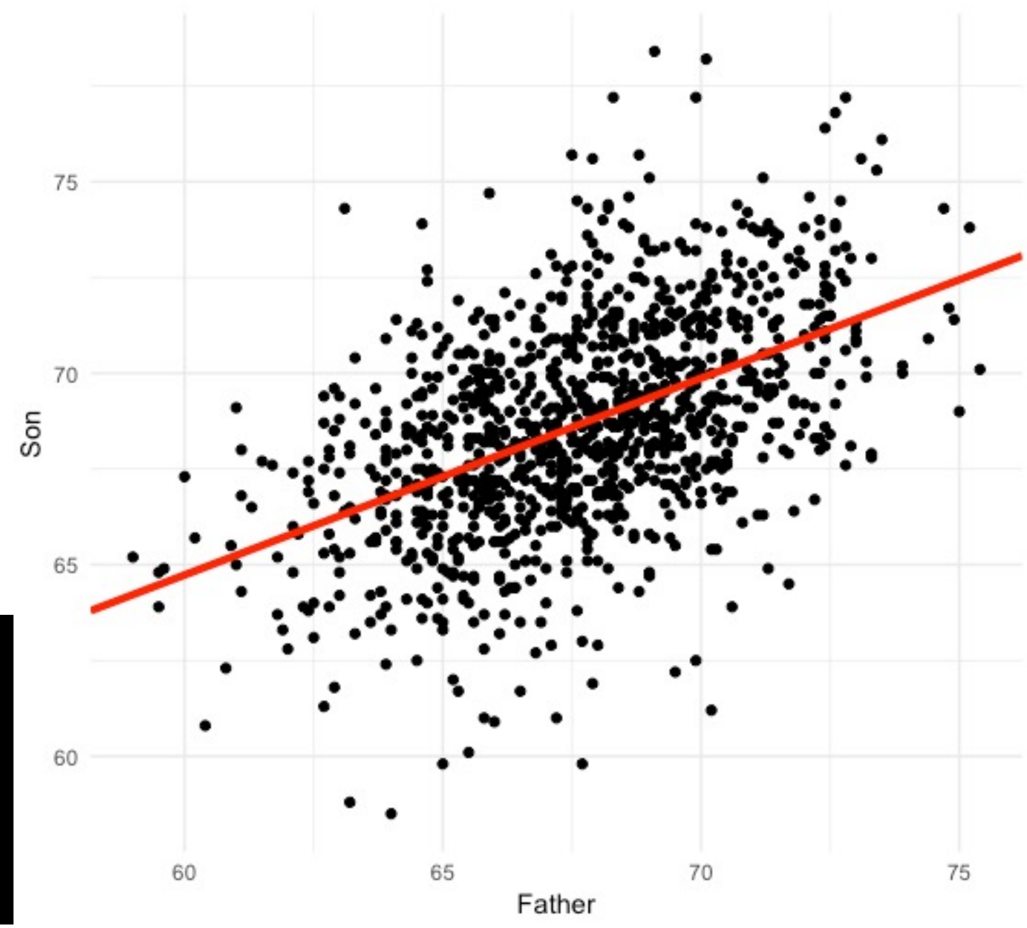
Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
lm(data = fs, Son~Father)
```

```
                     Coefficients:
         (Intercept)         Father
              33.893          0.514
```

```
ggplot(fs, aes(x=Father, y=Son))+
     geom_point()+
     geom_abline(intercept = 33.893,
          slope = 0.514,
          color="red",
          size=1.5)+
     theme_minimal()
```



Regression in R

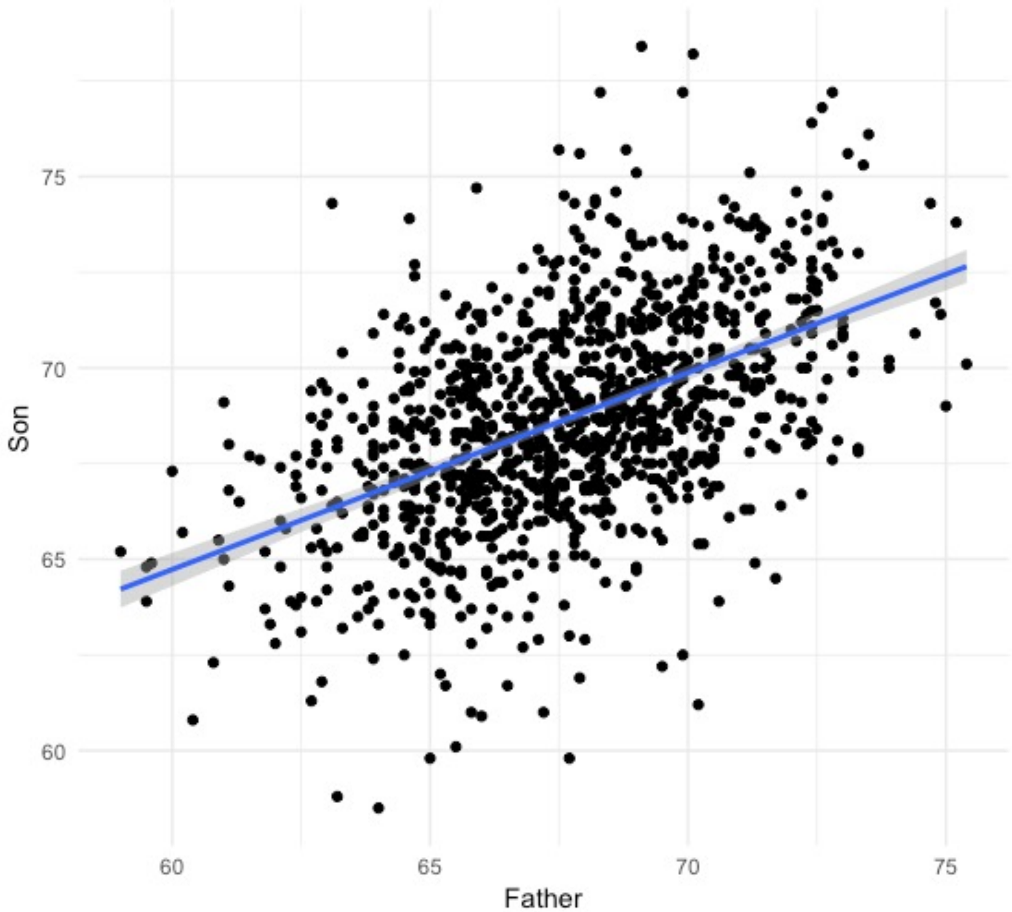
Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;   s = fs$Son
```

```
lm(data = fs, Son~Father)
```

```
                     Coefficients:
         (Intercept)        Father
              33.893         0.514
```

```
ggplot(fs, aes(x=Father, y=Son))+
     geom_point()+
     geom_smooth(method = "lm")+
     theme_minimal()
```



Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8910 -1.5361 -0.0092  1.6359  8.9894 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505 
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

OLS regression: estimate of slope

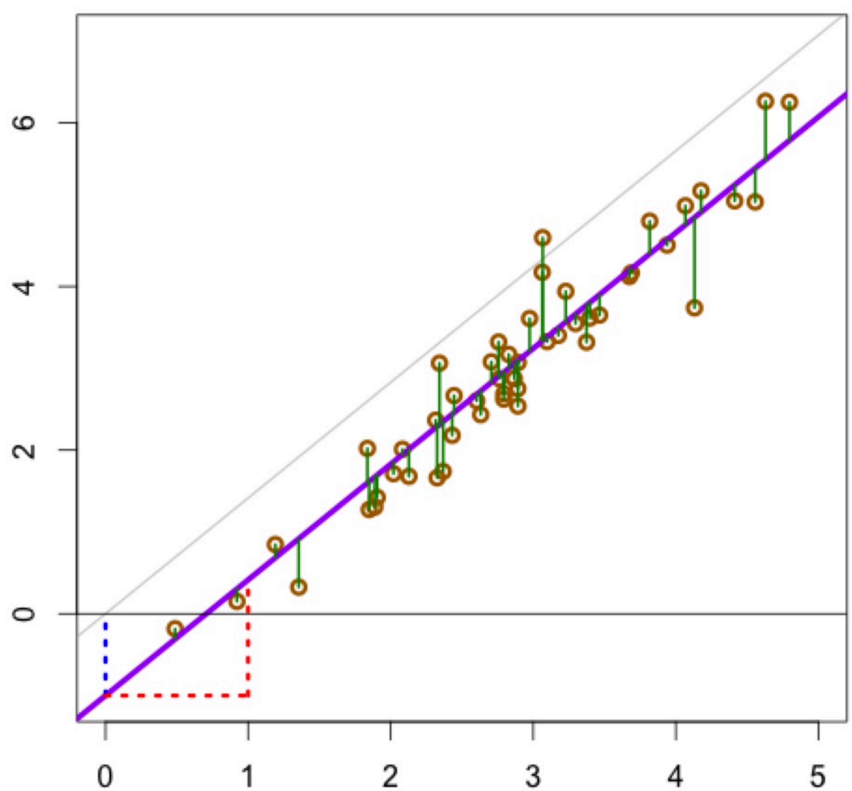
$$\boxed{Y_i} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{X_i} + \boxed{\varepsilon_i}$$

Score on Y for the ith individual = Y Intercept + (Slope *(Effect)* × Score on X for the ith individual) + Error

Least squares estimates

Line that minimizes sum of squared errors

This is the line that gives us E[Y|X]



$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x}$$

There are equivalent formulae using covariance, etc.

A few consequences of E[Y|X] (slope)

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x}$$

- Correlation is the slope of the z-scores.
- Regression to the mean.
- Asymmetry between y~x and x~y.

Correlation is the slope of z-scores.

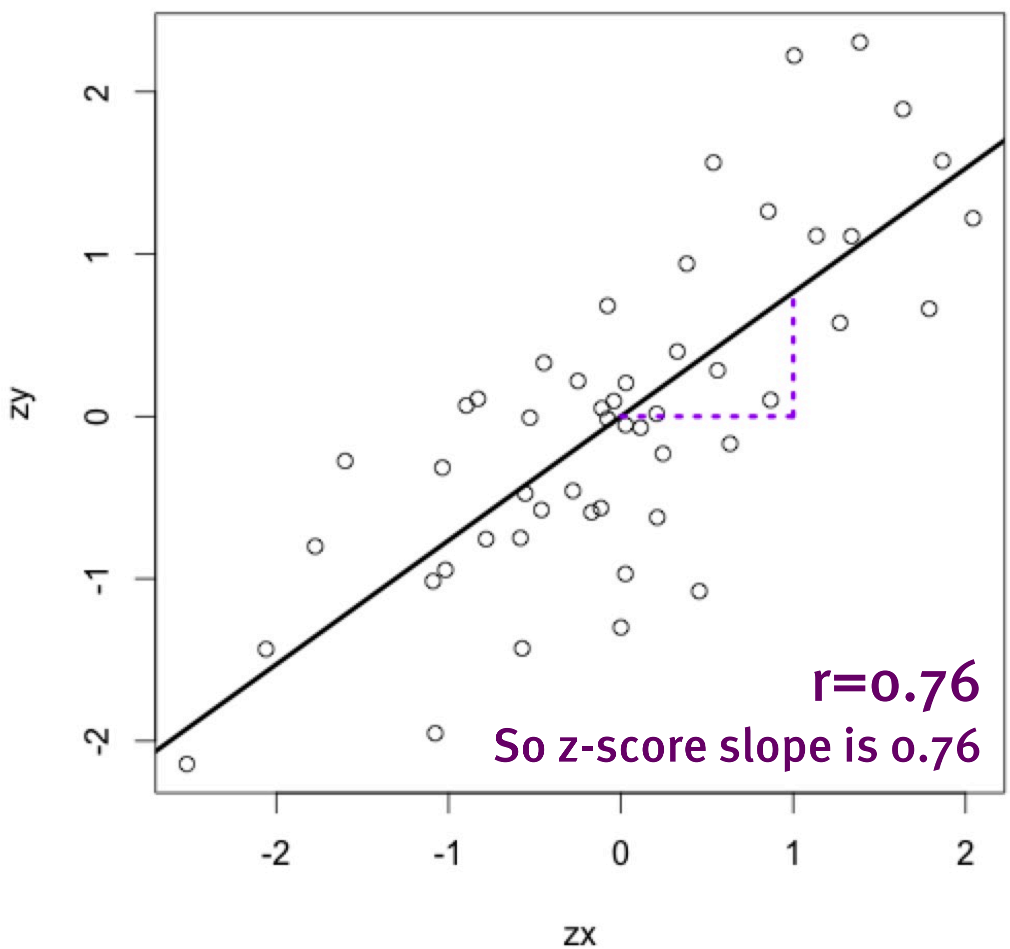
The correlation coefficient is the slope of the z-scores: how many standard deviations in y do you go up for every 1 s.d. increase in x?

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x}$$

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_i^{(x)} z_i^{(y)}$$

Regression to the mean

The correlation coefficient is the slope of the z-scores...



This means that (unless the correlation is perfect) the y value will not be as extreme as the x value.

E.g., test-retest reliability is never perfect. So people who do really well/badly (very big positive/negative z-score) on one test, will tend to be closer to the average on the retest

E.g., very tall/short parents will tend to have children closer to average.

E.g., very good performance by stock brokers in one quarter is likely to be followed by average performance.

$Y \sim b_0 + b_1(X) + e \neq X \sim b_0 + b_1(Y) + e$

Regression of Y as fx. of X gives different line than X as fx. of Y. Why?

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$(\hat{y}_i - \hat{\beta}_0) / \hat{\beta}_1 = x$$

$$a = 1 / \hat{\beta}_1$$

$$b = -\hat{\beta}_0 / \hat{\beta}_1$$

$$\hat{x} = a \cdot y + b$$

```
b0=33.89
b1=0.514
a = 1/b1
b = b0/b1
```

```
a [1] 1.94
b [1] 65.93
```

```
summary(lm(sons~fathers))
```

```
Coefficients:
            Estimate
(Intercept) 33.88660
fathers      0.51409
```

So, since

son.height ~ father.height*0.5 + 34

we might expect

father.height ~ son.height*2 + 66

And we would be very wrong!

```
summary(lm(fathers~sons))
```

```
Coefficients:
            Estimate
(Intercept) 34.10745
sons         0.48890
```

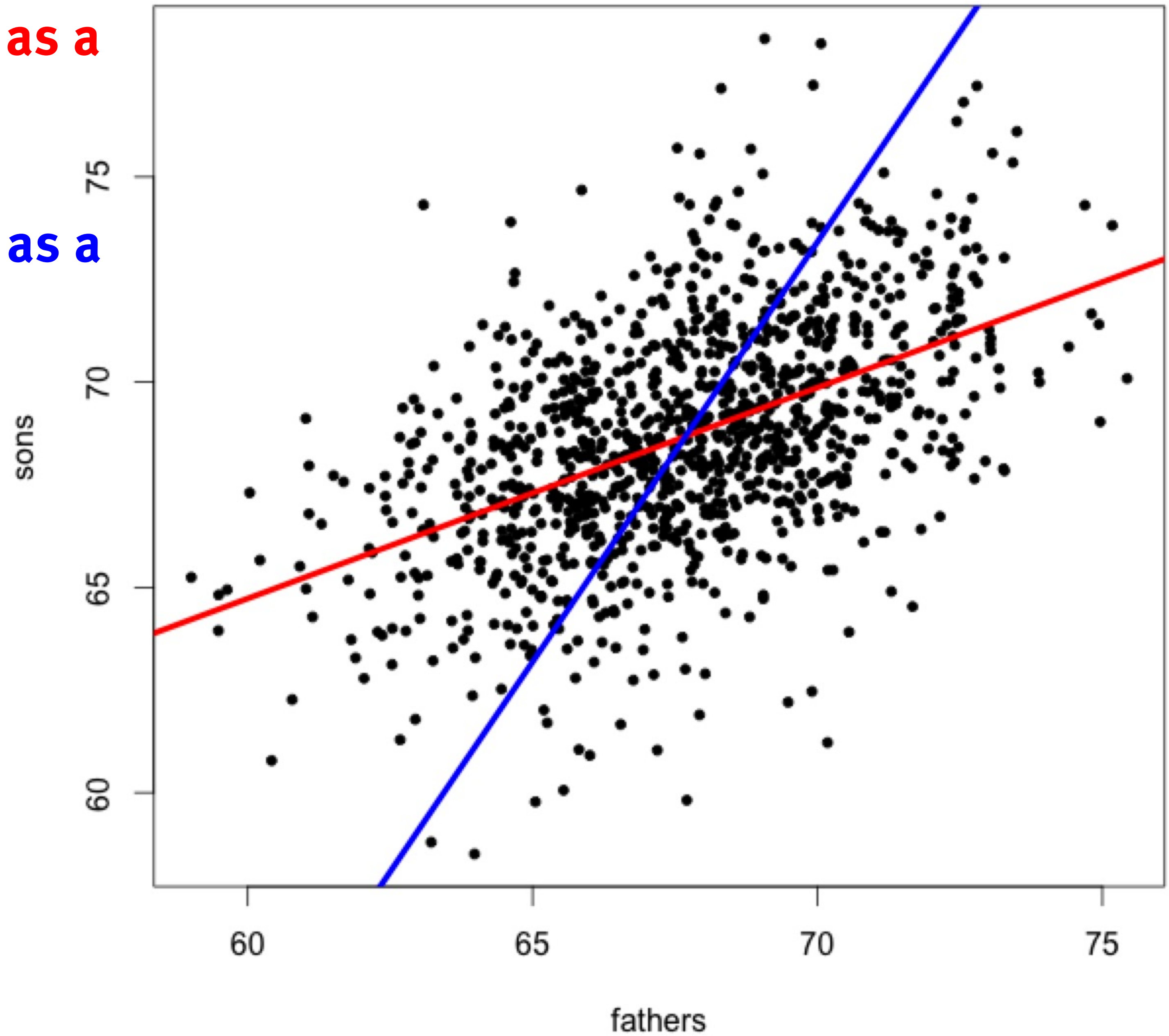
These coefficients are very different from what we get by using the same line and just algebraically shuffling to get x~y

Why?

$Y \sim b_0 + b_1(X) + e \neq X \sim b_0 + b_1(Y) + e$

Regression of y as a function of x

Regression of x as a function of y

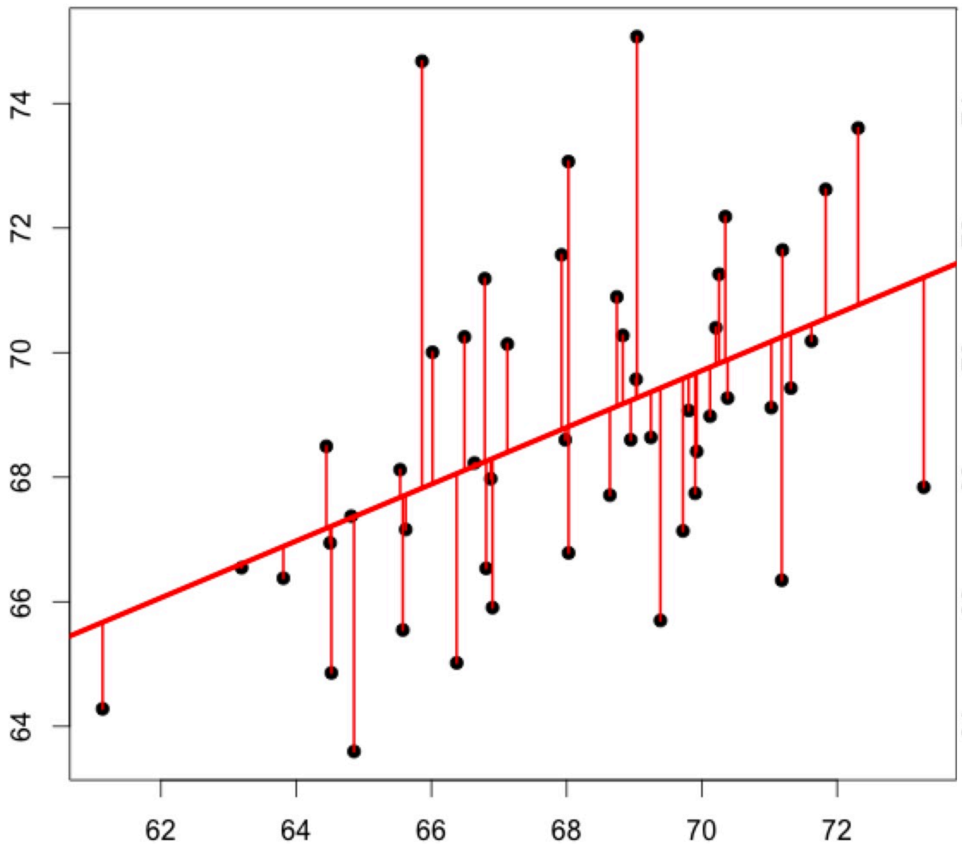


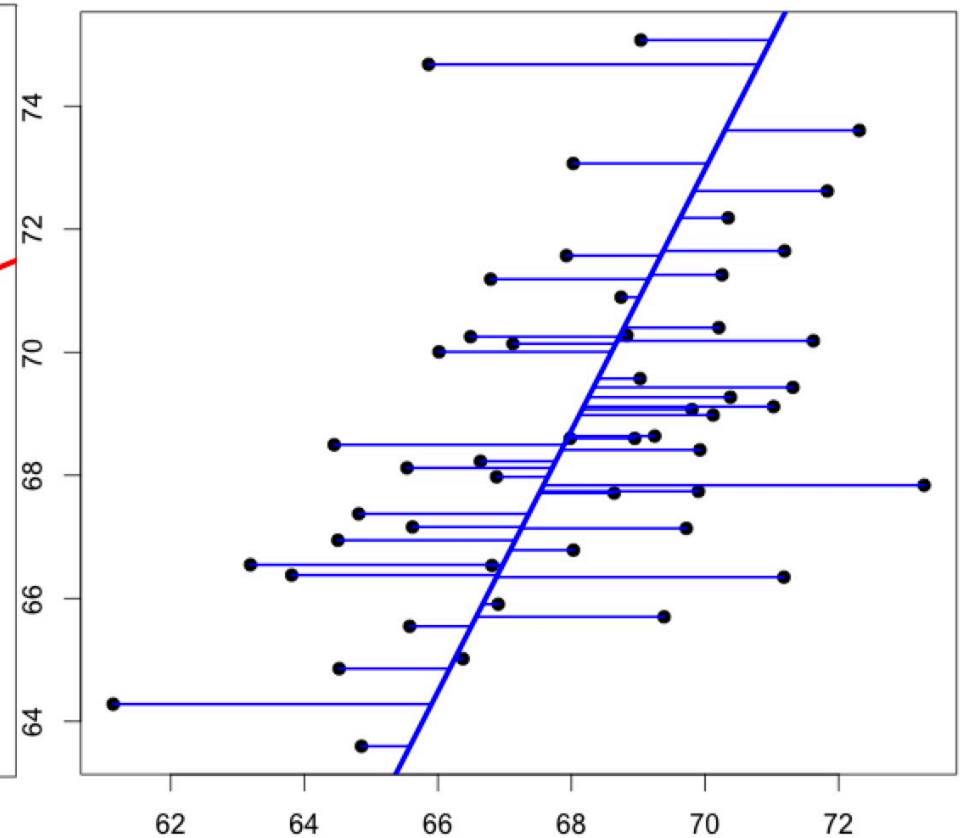
$Y \sim b_0 + b_1(X) + e \neq X \sim b_0 + b_1(Y) + e$

Why is regression of Y as fx. of X different than X as fx. of Y?

Regression of y as a function of x minimizes squared errors in y

Regression of x as a function of y minimizes squared errors in x



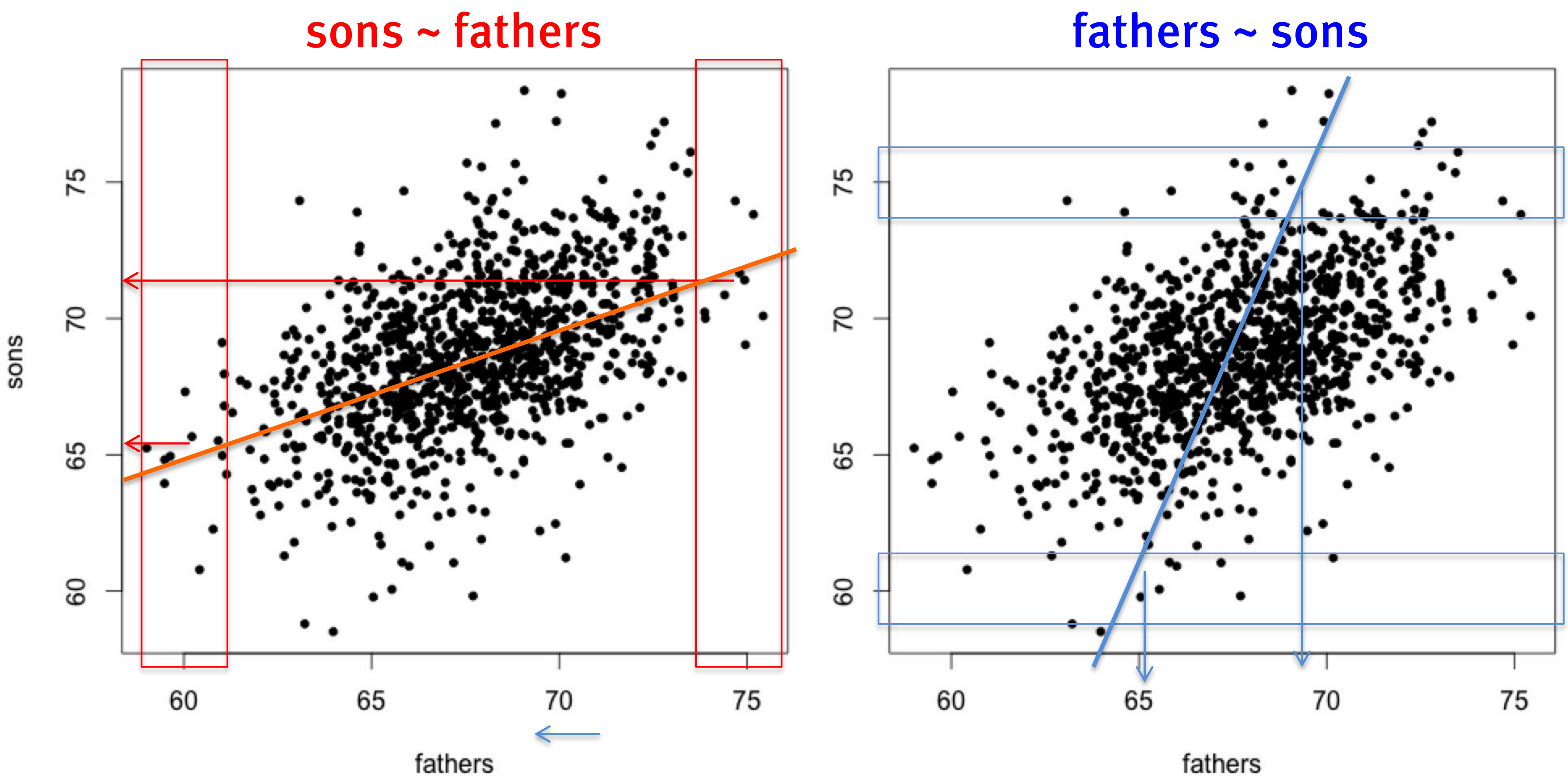


OLS regression assumes predictor is not uncertain…

Regression: conditional means

Regression estimates conditional means. E.g., y~x estimates mean(y | x)

Consequently we get a few weird phenomena:
- slope of y~x differs from inverse of x~y.
- Regression to the mean:
  - mean(x) for extreme y is less extreme, mean(y) for extreme x is less extreme.

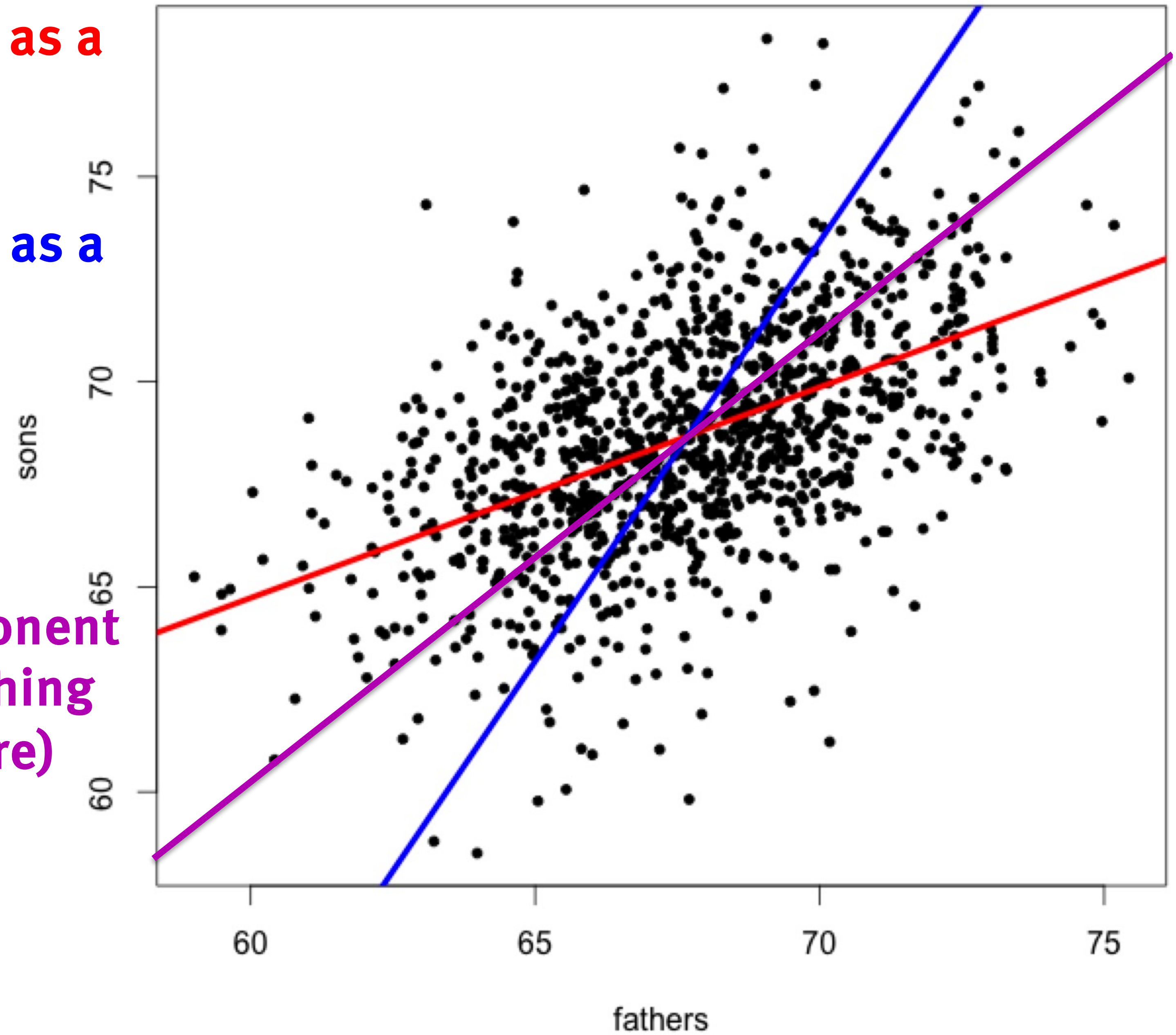


Y~bo+b1(X)+e ≠ X~bo+b1(Y)+e

Regression of y as a function of x

Regression of x as a function of y

Principle component line (not something we estimate here)



Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))        f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

OLS regression: estimate of intercept

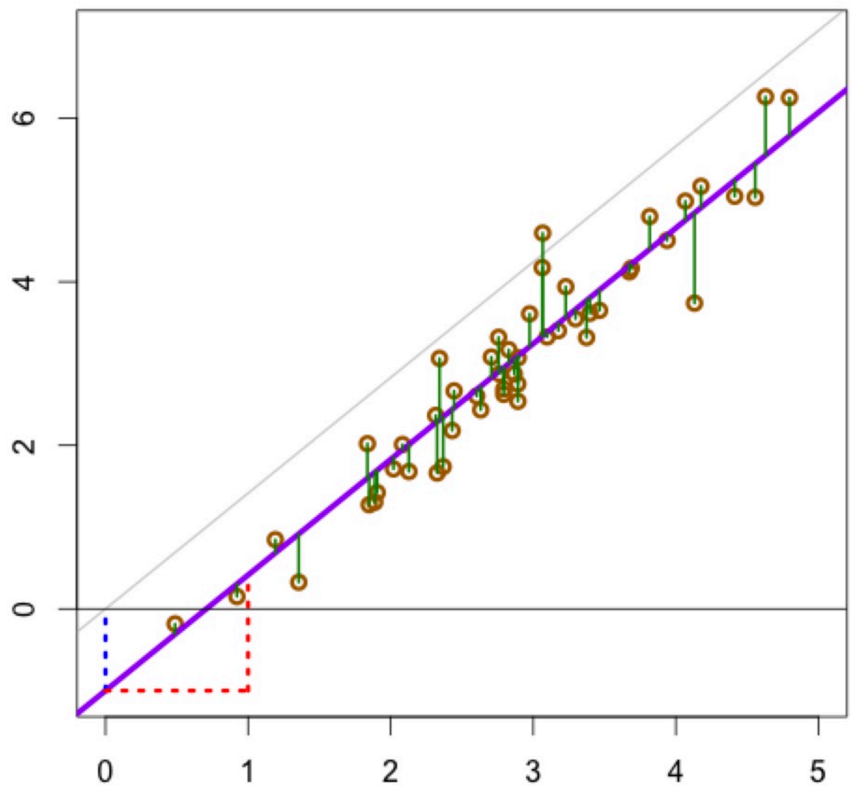
$$\boxed{Y_i} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{X_i} + \boxed{\varepsilon_i}$$

Score on Y for the ith individual = Y Intercept + (Slope *(Effect)* × Score on X for the ith individual) + Error

Least squares estimates

Line that minimizes sum of squared errors

This is the line that gives us E[Y|X]



$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

This comes from the constraint that the line must go through [mean(x), mean(y)].

OLS regression: estimate of intercept

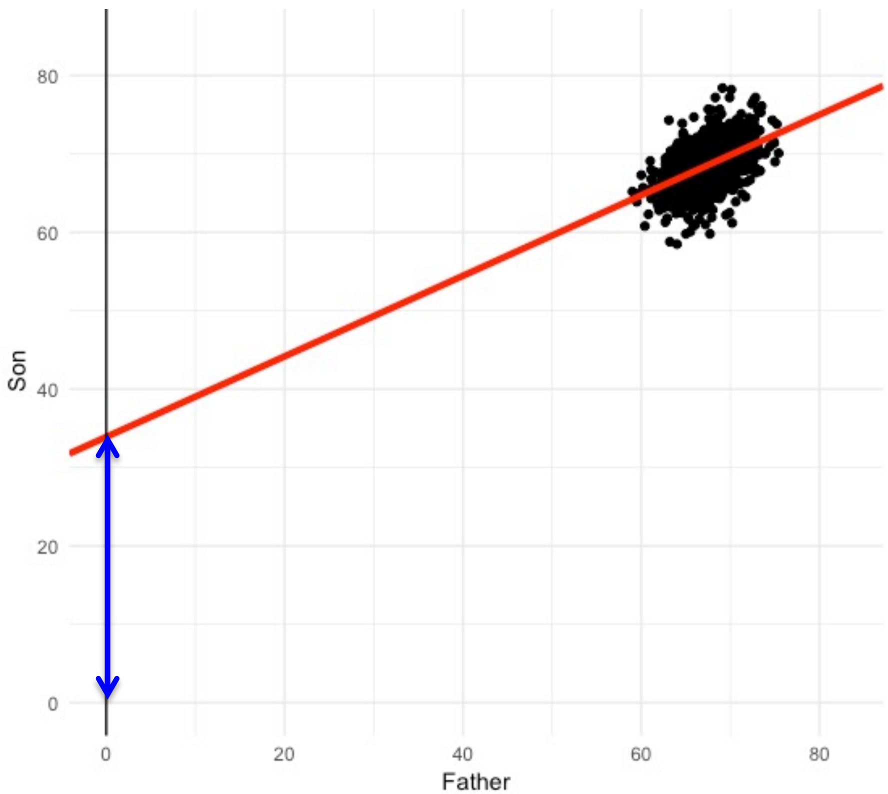
$$\boxed{Y_i} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{X_i} + \boxed{\varepsilon_i}$$

Score on Y for the ith individual = Y Intercept + (Slope *(Effect)* × Score on X for the ith individual) + Error

Least squares estimates

Line that minimizes sum of squared errors

This is the line that gives us E[Y|X]



$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Interpretation of intercept is rather challenging. It is the predicted y value at x=0. e.g., the height of a son whose father is 0 inches tall.

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

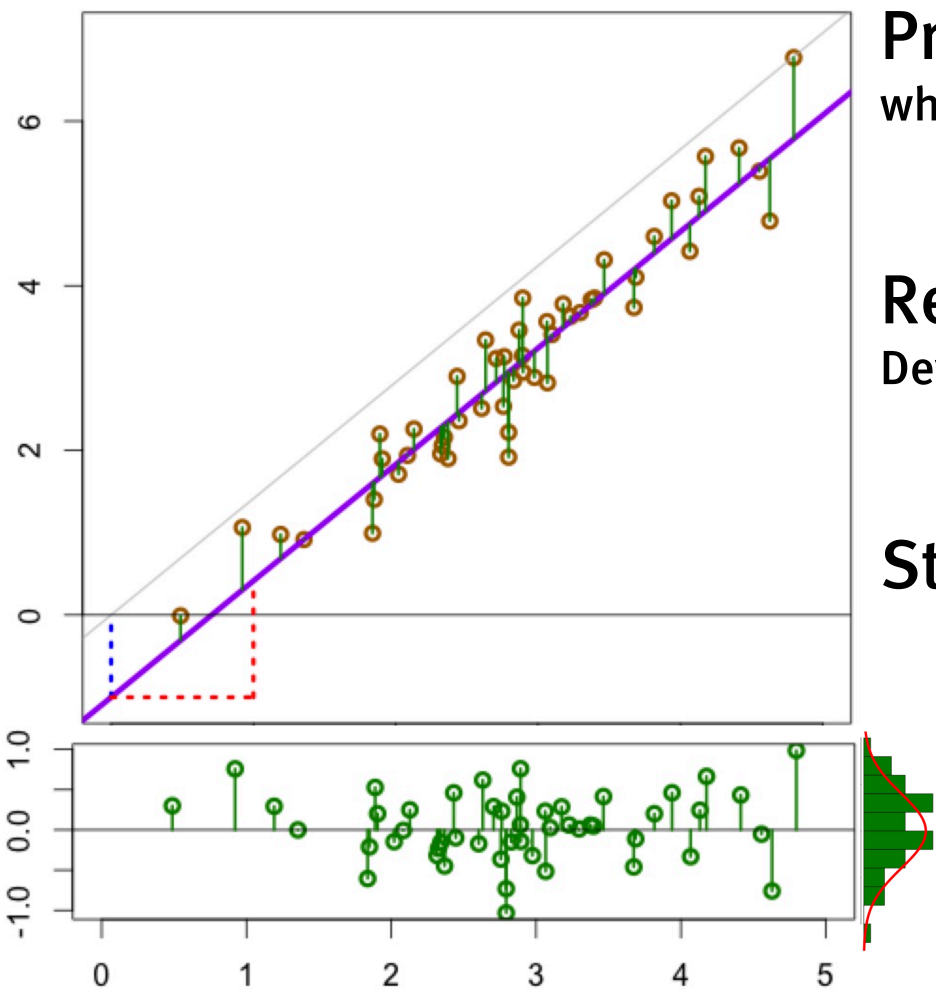
```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

OLS regression: estimate of residuals

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Score on Y for the ith individual = Y Intercept + (Slope (Effect) × Score on X for the ith individual) + Error



Least squares estimates

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Predicted y values

where the estimated line passes at each x value

$$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$$

Residuals (estimated error)

Deviation of real y value from line prediction

$$\boxed{\hat{\varepsilon}_i = \left(y_i - \hat{y}_i\right)}$$

Standard deviation of residuals

$$\hat{\sigma}_\varepsilon = s_r = \sqrt{\frac{1}{\boxed{n-2}} \boxed{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}}$$

The sum of squared errors: SS[e]

df=n-2, we fit two parameters (B0,B1)

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Standard Error of the Slope

Estimated slope

```
b1 = r.fs*sy/sx
```

```
[1] 0.5141
```

$$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x}$$

```
sr = sqrt(sum((sons-fathers*b1-b0)^2)/(n-2))
```

```
[1] 2.436556
```

Std. dev. of residuals

$$s_r = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Sampling s.d. of estimated slope (std. err. of slope)

$$\boxed{s\{\hat{\beta}_1\}} = \boxed{\boxed{s_r}\sqrt{\frac{1}{\boxed{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}} = \boxed{\boxed{s_r}\sqrt{\frac{1}{\boxed{\boxed{s_x^2}(n-1)}}}} = \boxed{\frac{\boxed{s_r}}{\boxed{s_x}}/\sqrt{n-1}}$$

SD / variance of x

Standard error of the slope

Sum of squares of X

Standard deviation of the residuals

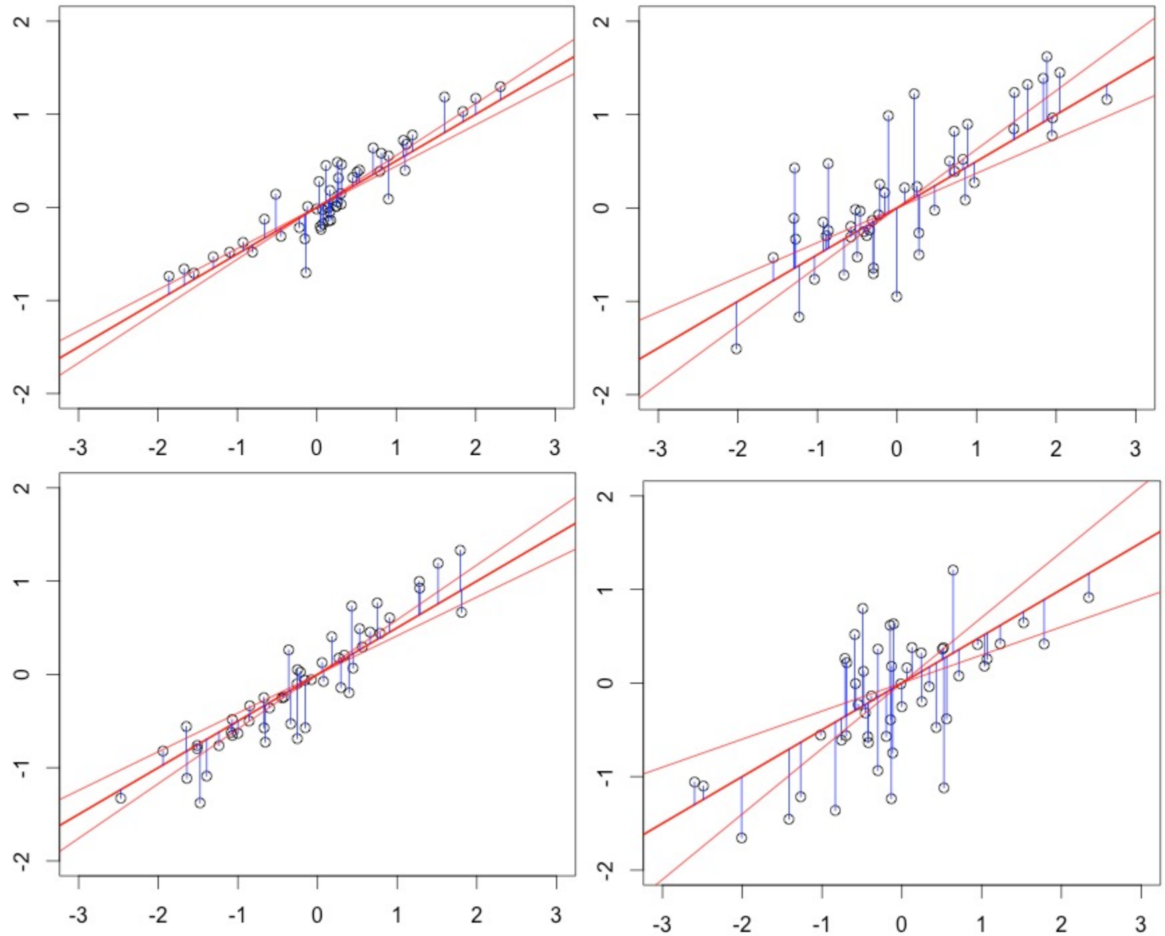
$$SS[x] = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

What makes our slope estimate better?

$$s\{\hat{\beta}_1\} = \frac{\boxed{s_r}}{s_x} / \sqrt{n-1}$$

Standard error of the slope is lower (and so slope estimate is better) when:

- Error around the line is smaller (lower sd of residuals)

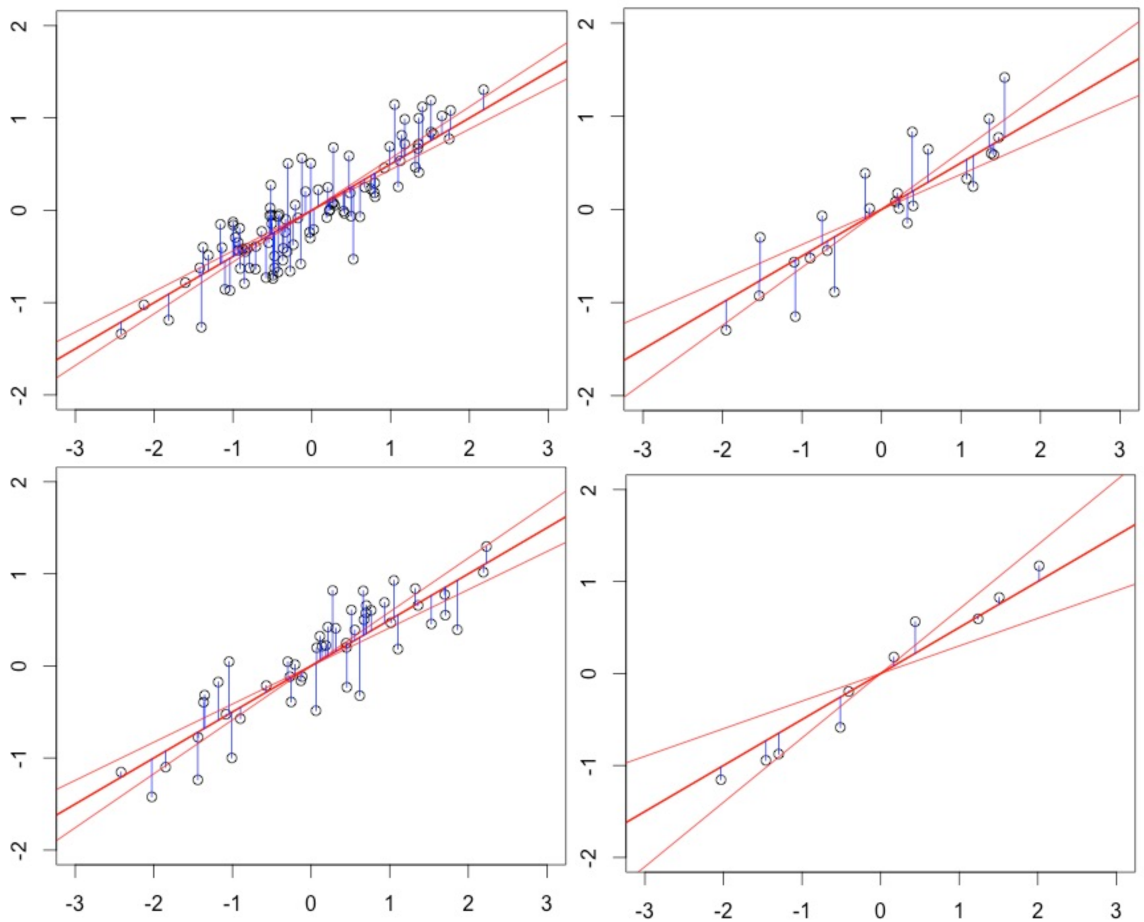


What makes our slope estimate better?

$$s\{\hat{\beta}_1\} = \frac{\boxed{s_r}}{s_x} / \sqrt{\boxed{n-1}}$$

Standard error of the slope is lower (and so slope estimate is better) when:

- Error around the line is smaller (lower sd of residuals)
- We have more data.

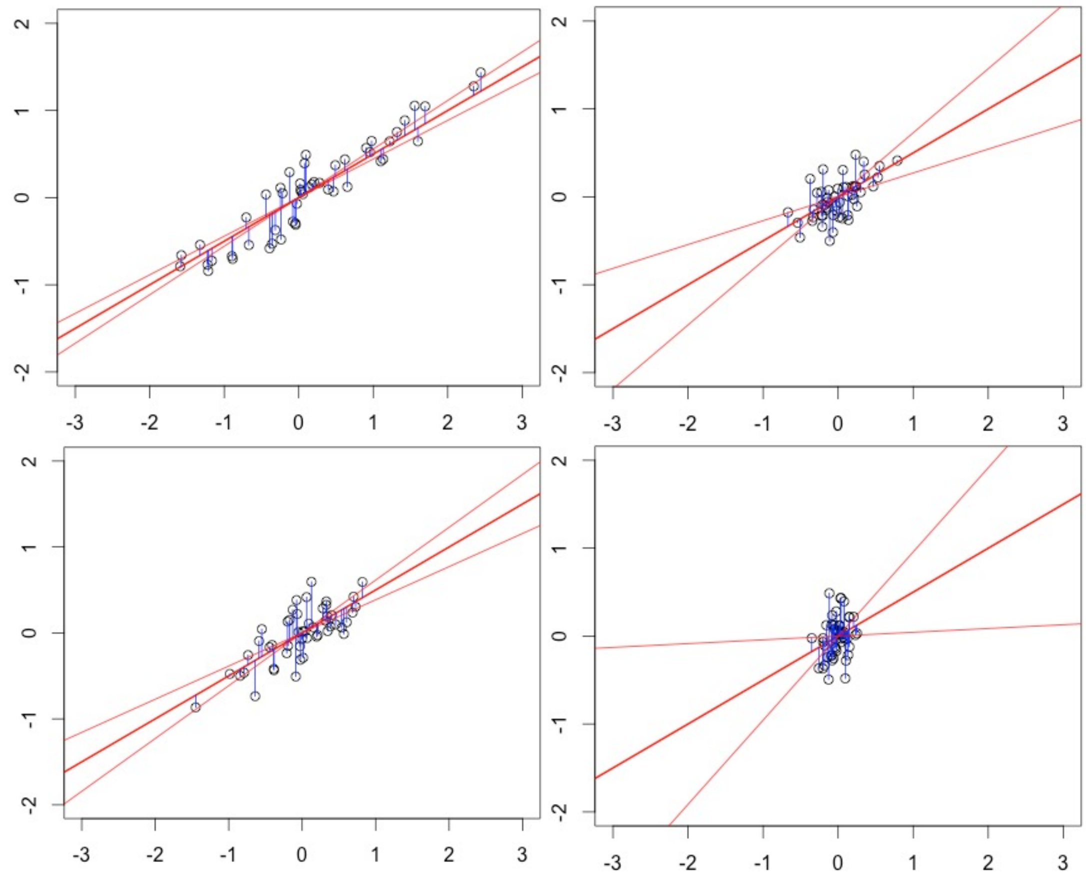


What makes our slope estimate better?

$$s\{\hat{\beta}_1\} = \frac{\boxed{s_r}}{\boxed{s_x}} / \sqrt{\boxed{n-1}}$$

Standard error of the slope is lower (and so slope estimate is better) when:

- Error around the line is smaller (lower sd of residuals)
- We have more data.
- X is more spread out (higher sd of x)



Why? SD of x determines the range of x, and the amount of variation in y due to variation in x. Thus, signal (var y due to x) to noise (var y due to error) ratio goes up.

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Standard error of the intercept

Estimated intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

This comes from the constraint that the line must go through [mean(x), mean(y)].

Sampling s.d. of estimated intercept (std. err. of intercept)

$$\boxed{s\{\hat{\beta}_0\}} = \boxed{\boxed{s_r}\sqrt{\boxed{\frac{1}{n}} + \boxed{\frac{(\bar{x}-0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}} = \boxed{\sqrt{\boxed{\frac{\boxed{s_r}^2}{n}} + \boxed{\left(\bar{x}\boxed{\left(\frac{\boxed{s_r}}{s_x\sqrt{n-1}}\right)}\right)^2}}}$$

Standard error of the intercept

Standard deviation of the residuals

Error in estimating the mean of y.

Error from extrapolating slope to x=0

The familiar std. error of the slope!

Standard error of the intercept

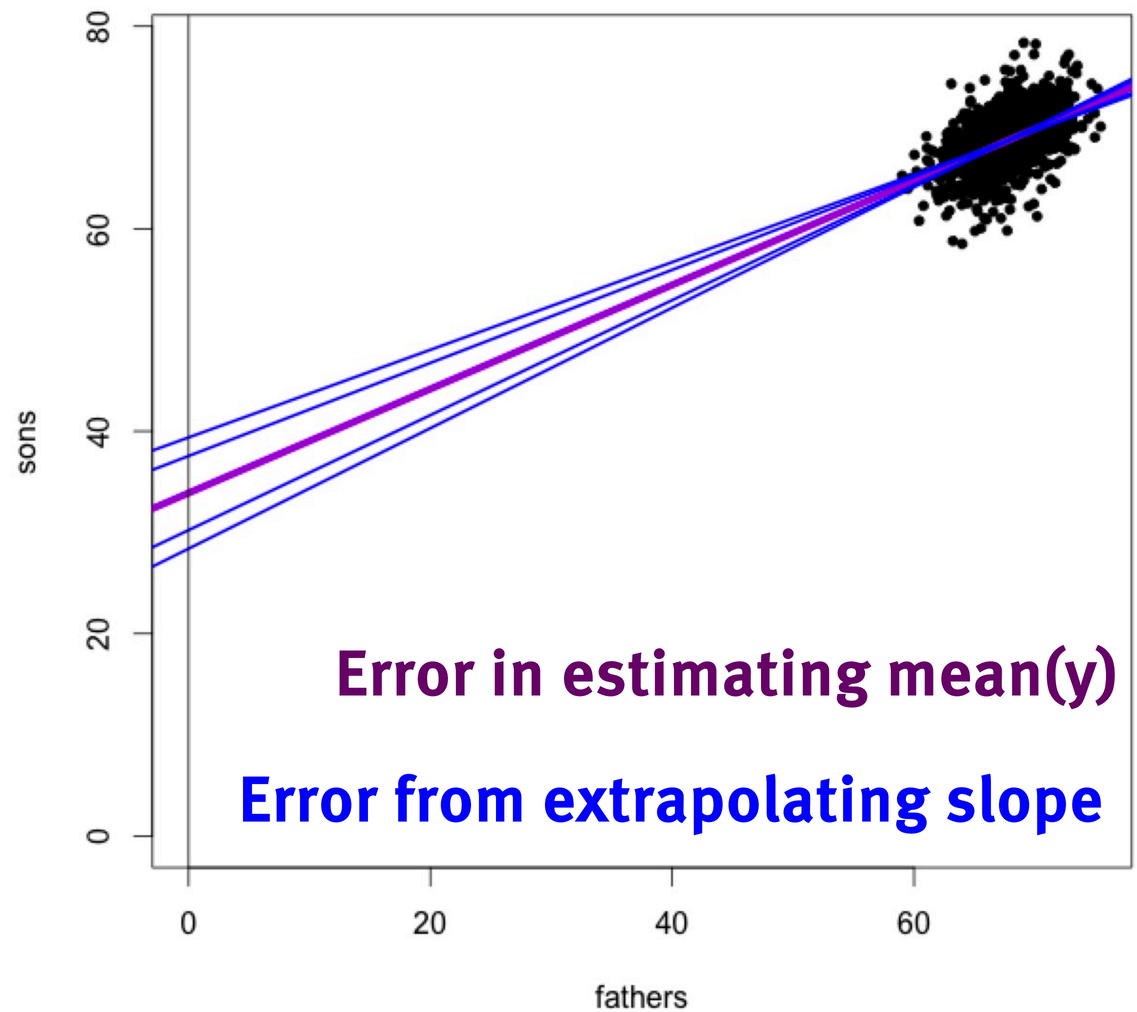
Estimated intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

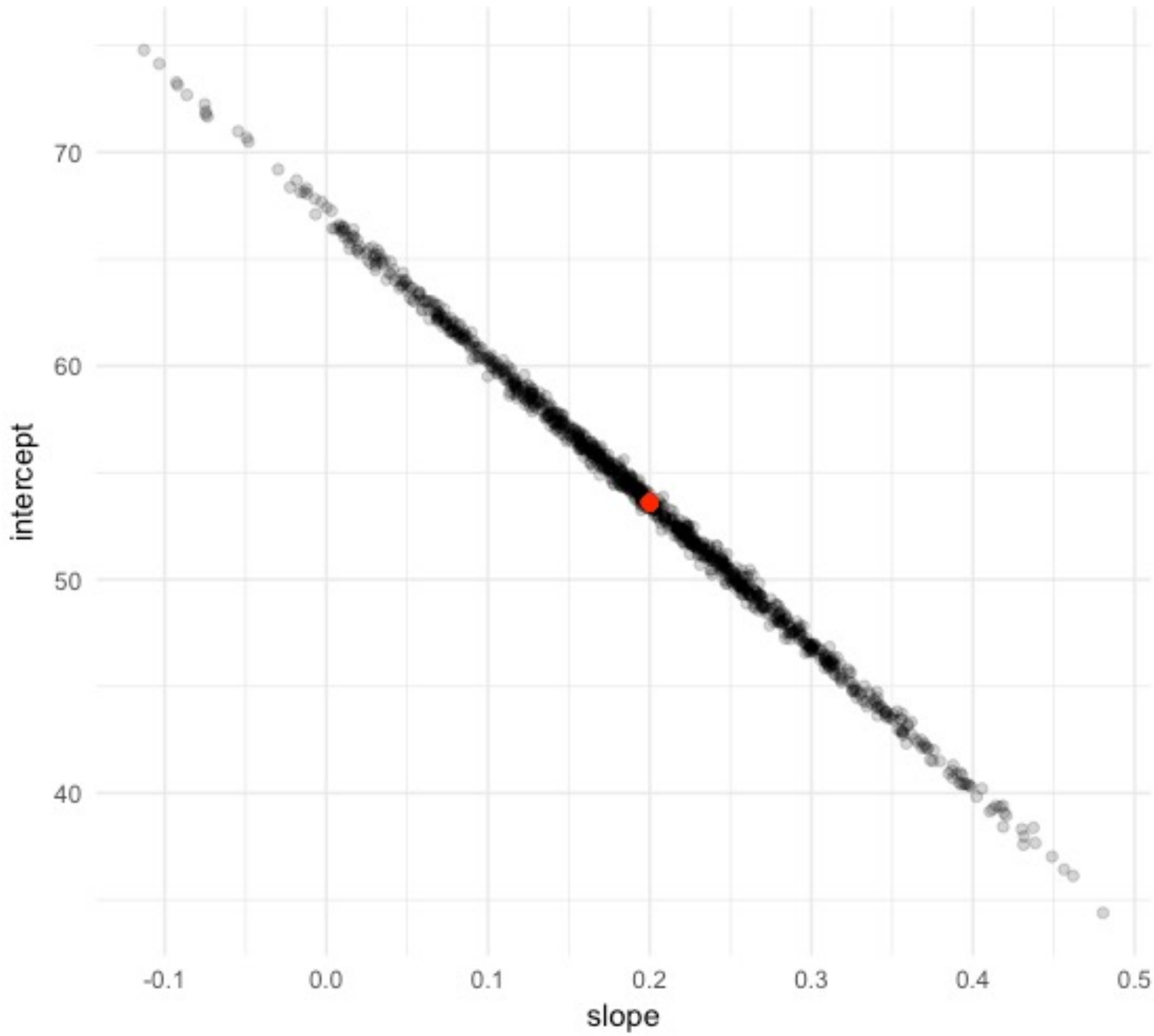
This comes from the constraint that the line must go through [mean(x), mean(y)].
So we have to extrapolate line to x=o to find intercept.

Sampling s.d. of estimated intercept (std. err. of intercept)

$$s\{\hat{\beta}_0\} = \sqrt{\boxed{\frac{s_r^2}{n}} + \boxed{\left(\bar{x}\left(\frac{s_r}{s_x\sqrt{n-1}}\right)\right)^2}}$$



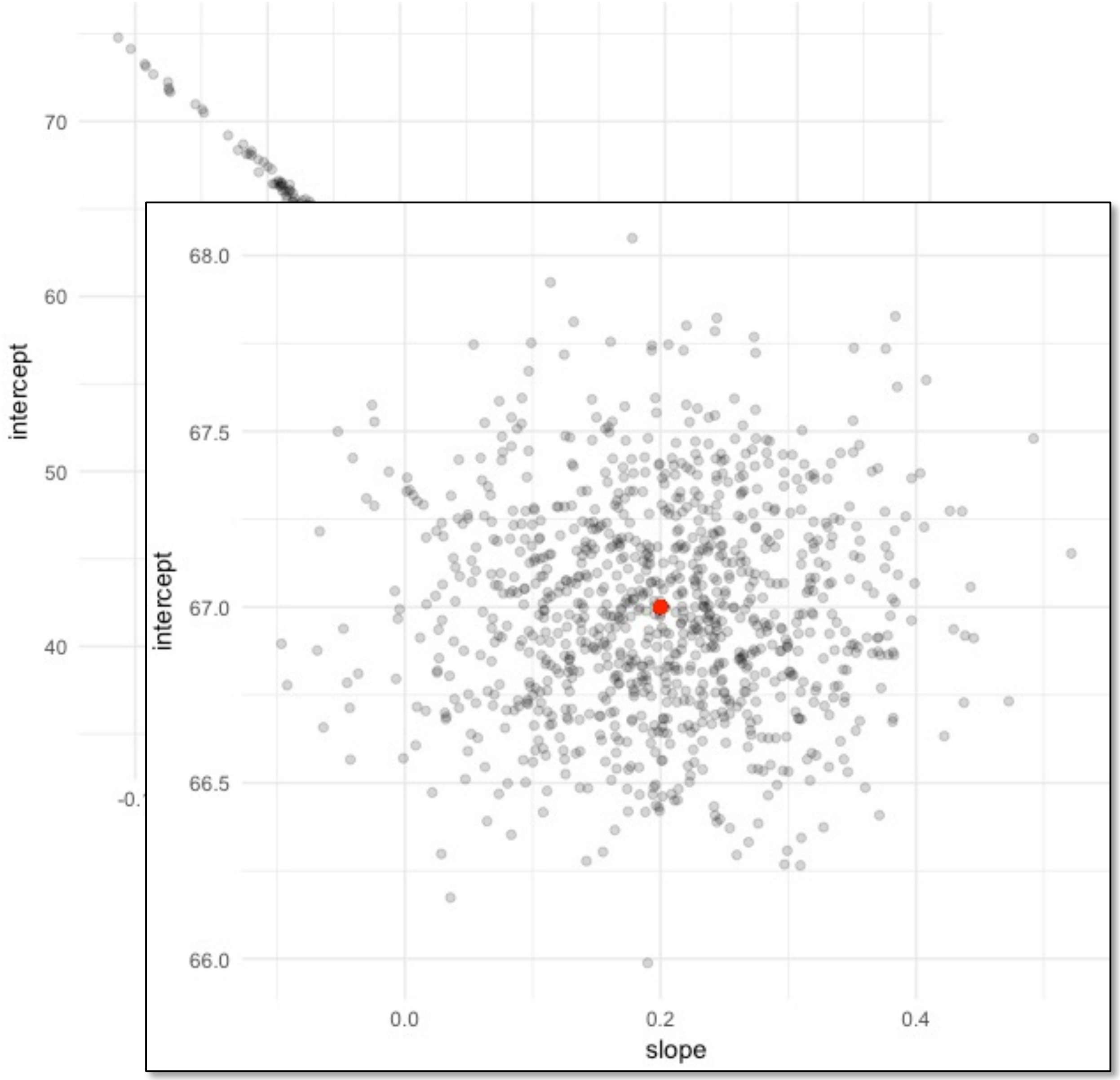
Correlation of estimation errors.



Error from extrapolating slope means:

Errors of slope and intercept will be very correlated (if we get the slope wrong, we will get the intercept wrong).
How bad this correlation is depends on how far we have to extrapolate: Mean(x)-o
The sign of this correlation depends on sign of mean(x).

Marginal std. error of intercept



Standard error of intercept is the *marginal* standard errors. So this very large correlation will look like a very large error in estimating intercept.

Centering x is generally a very good idea:

x’ = x-mean(x)

lm(y~x’)

Gets rid of huge errors in intercept, and also makes intercept interpretable as mean(y) at mean(x) (rather than mean(y) at x=0)

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))        f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Standard Errors of coefficients

Standard error of the slope *decreases* with:

- Smaller s.d. of residuals
- Larger sample size
- Larger spread of x values

$$s\{\hat{\beta}_1\} = \frac{s_r}{s_x} / \sqrt{n-1}$$

Standard error of the intercept *decreases* with:

- Smaller s.d. of residuals
- Larger sample size
- Smaller std. distance between 0 and mean(x)

$$s\{\hat{\beta}_0\} = s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

We do the usual t-test procedures to test null hypotheses and obtain confidence intervals

With df=n-2: degrees of freedom in estimating the s.d. of residuals.

$$t_{b1} = \frac{\hat{\beta}_1 - h0}{s\{\hat{\beta}_1\}} \qquad \hat{\beta}_1 \pm t_{\alpha/2} s\{\hat{\beta}_1\}$$

```
summary(lm(sons~fathers))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.88660    1.83235   18.49   <2e-16 ***
fathers      0.51409    0.02705   19.01   <2e-16 ***

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared: 0.2513,  Adjusted R-squared: 0.2506
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

s_e S.D. of residuals
SSE = se^2*df

d.f. of residuals

These t-statistics and p values are calculated just like all other t statistics:

t = (estimate – null.value)/se{estimate}

Default null.value=0

So t tests are asking if those parameter estimates differ from zero.

df: df for estimating sample variance (residual std. deviation/error)

Can define confidence intervals the usual way as well:

estimate +/- t.crit * se{estimate}

e.g., 95% C.I. on slope: 0.514 +/- (~)2 * 0.027 => (0.46, 0.57)

$\hat{B}_0$ Estimate of intercept

$s\{\hat{B}_0\}$ Std. err. of intercept

$$t_{(n-2)} = \frac{\hat{B}_0 - 0}{s\{\hat{B}_0\}}$$

T-test of intercept

```
summary(lm(sons~fathers))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.88660    1.83235   18.49   <2e-16 ***
fathers      0.51409    0.02705   19.01   <2e-16 ***

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared: 0.2513,    Adjusted R-squared: 0.2506
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

$$t_{(n-2)} = \frac{\hat{B}_1 - 0}{s\{\hat{B}_1\}}$$

T-test of slope

$s\{\hat{B}_1\}$ Std. err. of slope

$\hat{B}_1$ Estimate of slope

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))        f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

158.1928
157.0185
153.2481
156.1513
154.1769
155.1849
155.4694
155.9177
153.8620
158.7263
156.3841
156.9075
156.9597
155.8952
160.1060
159.2632
157.8709
156.5646
158.1436
154.6955
159.4184
159.5932
158.9586
156.9553
155.9073
156.1151
157.5840
155.2092
156.7197
156.1086
155.4311
154.4730
154.2109
157.4233
155.7556
157.1322
155.8327
156.0758

Variation and randomness

- Measure weight 168 times

Variation and randomness

153.2481
153.8620
154.1769
154.2109
154.2850
154.4140
154.4730
154.6955
154.7180
154.8091
154.9224
154.9990
154.9997
155.0386
155.1849
155.2092
155.3161
155.4191
155.4311
155.4667

⋮

161.5555
161.5896
162.0160
162.0885
162.0995
162.1995
163.1148

- Measure weight 168 times
- Sort measurements:

Variation and randomness

153.2481
153.8620
154.1769
154.2109
154.2850
154.4140
154.4730
154.6955
154.7180
154.8091
154.9224
154.9990
154.9997
155.0386
155.1849
155.2092
155.3161
155.4191
155.4311
155.4667
⋮
161.5555
161.5896
162.0160
162.0885
162.0995
162.1995
163.1148

- Measure weight 168 times
- Bin the measurements

153.2481
153.8620

152-154

154.1769
154.2109
154.2850
154.4140
154.4730
154.6955
154.7180
154.8091
154.9224
154.9990
154.9997
155.0386
155.1849
155.2092
155.3161
155.4191
155.4311
155.4667
155.4694
155.5219
155.5740
155.5786
155.6990
155.7556
155.8230
155.8327
155.8952
155.9073
155.9177
155.9982

154-156

156.0251
156.0758
156.1086
156.1151
156.1513
156.2832
156.3841
156.3873
156.4799
156.5246
156.5405
156.5520
156.5634
156.5646
156.6763
156.6920
156.6960
156.7093
156.7197
156.7343
156.8443
156.8820
156.9075
156.9169
156.9553
156.9597
156.9831
157.0185
157.0686
157.0901
157.0917
157.1322
157.1376
157.1692
157.1886
157.2173
157.2534
157.2818
157.4020
157.4057
157.4233
157.4354
157.5128
157.5840
157.6364
157.6622
157.6892
157.7325
157.7732
157.7927
157.8709
157.9105
157.9139
157.9295
157.9991

156-158

158.0551
158.0623
158.0717
158.1436
158.1813
158.1928
158.2152
158.2264
158.2900
158.3271
158.3519
158.3566
158.3953
158.4081
158.4175
158.4654
158.4779
158.4850
158.6486
158.6562
158.6797
158.7263
158.7267
158.7663
158.7801
158.7813
158.7818
158.8892
158.9586
159.0148
159.0371
159.0489
159.0561
159.0561
159.0801
159.1355
159.1790
159.2632
159.3555
159.3869
159.4184
159.5593
159.5843
159.5880
159.5932
159.6446
159.6699
159.7500
159.7729
159.8272
159.8597
159.9212
159.9562
159.9878

158-160

160.0779
160.1060
160.1280
160.2301
160.2902
160.3108
160.3463
160.3494
160.3508
160.3592
160.3994
160.4515
160.4914
160.6324
160.6519
160.6937
161.0825
161.1887
161.2539
161.3951
161.5555
161.5896

160-162

162.0160
162.0885
162.0995
162.1995
163.1148

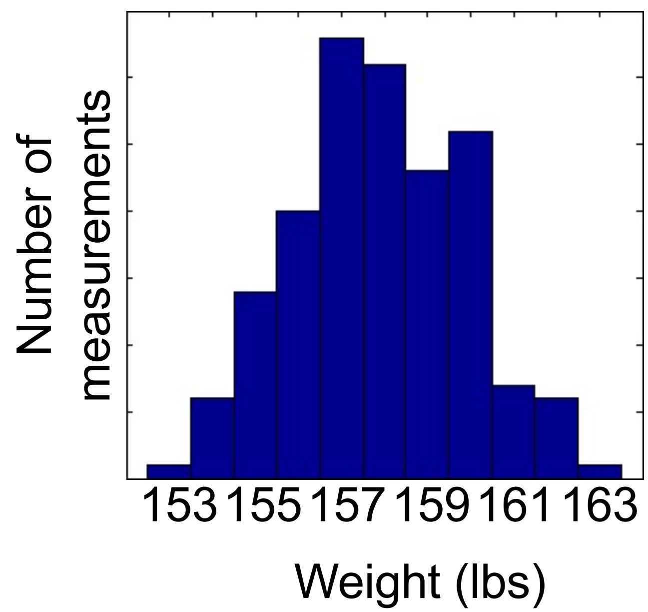
160-162

Variation and randomness

153.2481
153.8620
154.1769
154.2109
154.2850
154.4140
154.4730
154.6955
154.7180
154.8091
154.9224
154.9990
154.9997
155.0386
155.1849
155.2092
155.3161
155.4191
155.4311
155.4667

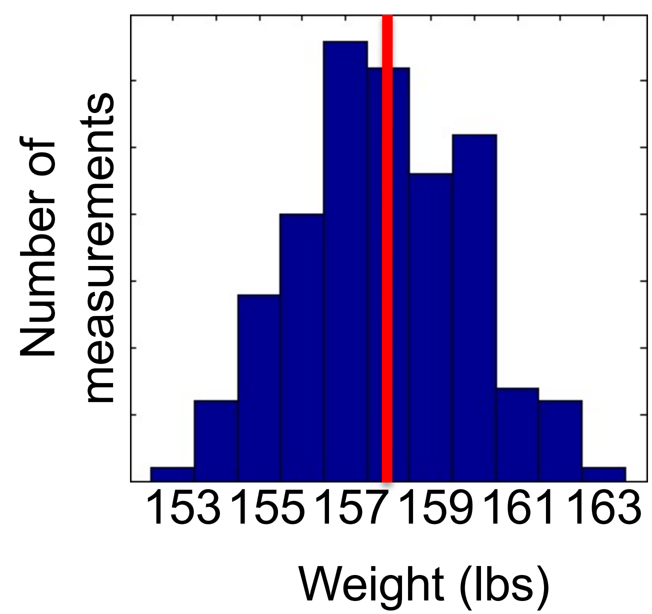
161.5555
161.5896
162.0160
162.0885
162.0995
162.1995
163.1148

- Measure weight 168 times
- Make a histogram



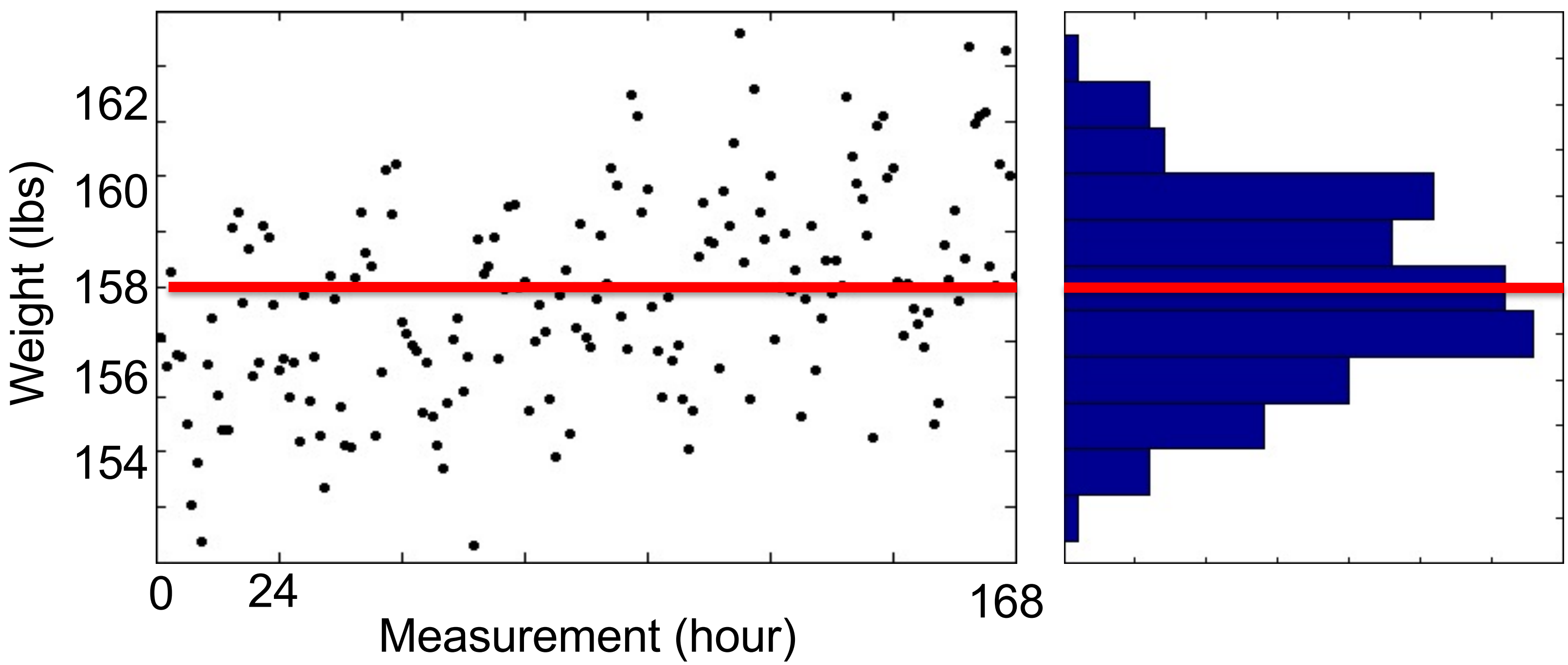
Variation and randomness

- Measure weight 168 times
- What is my weight?
- Different each time I measure it.
- Mean is 157.9
- Variation around the mean (157.9) is "random", as far as I know.



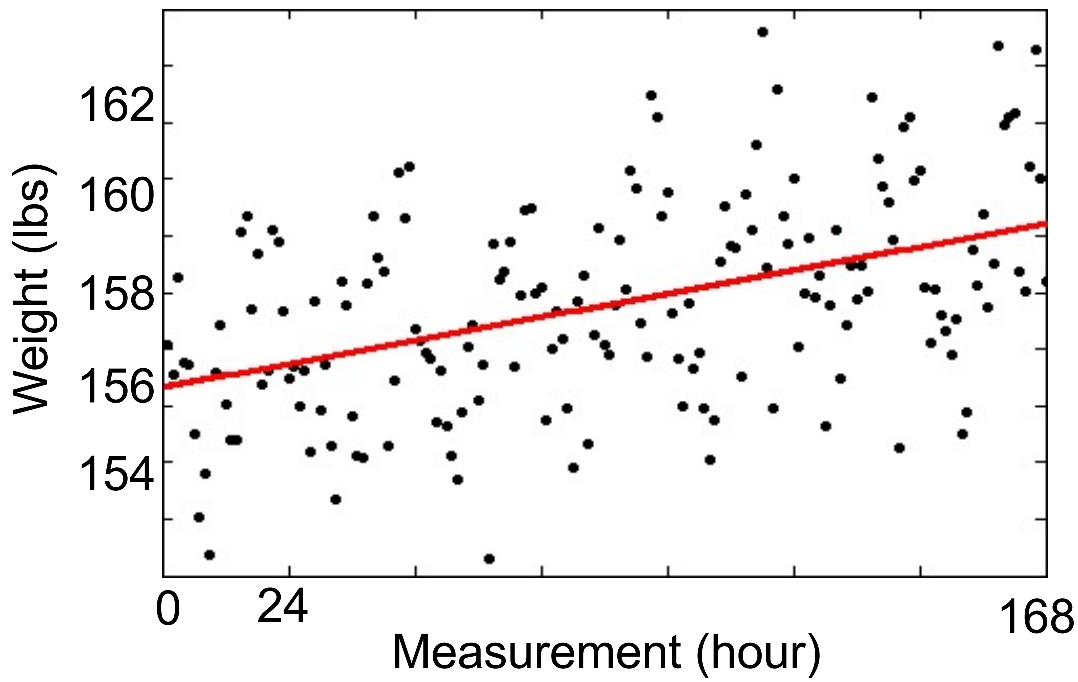
Variation and randomness

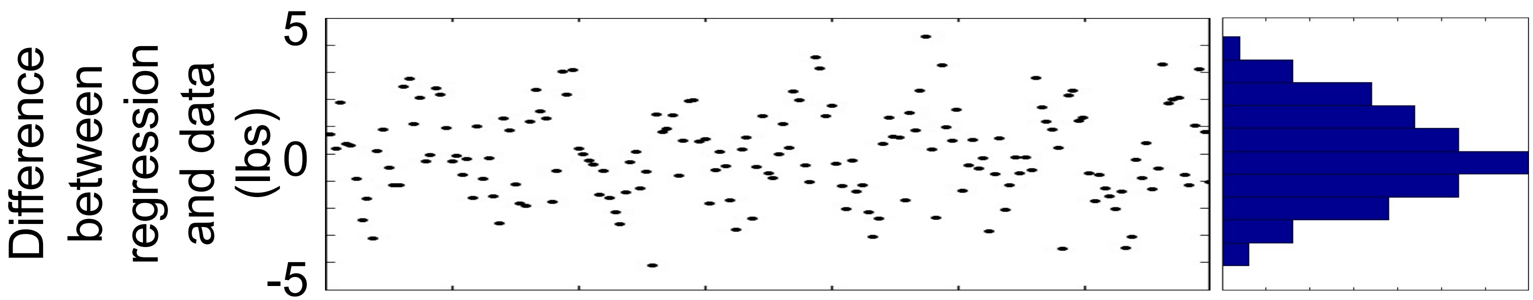
- Oh yeah:
- 168 measurements are hourly for 7 consecutive days



Variation and randomness

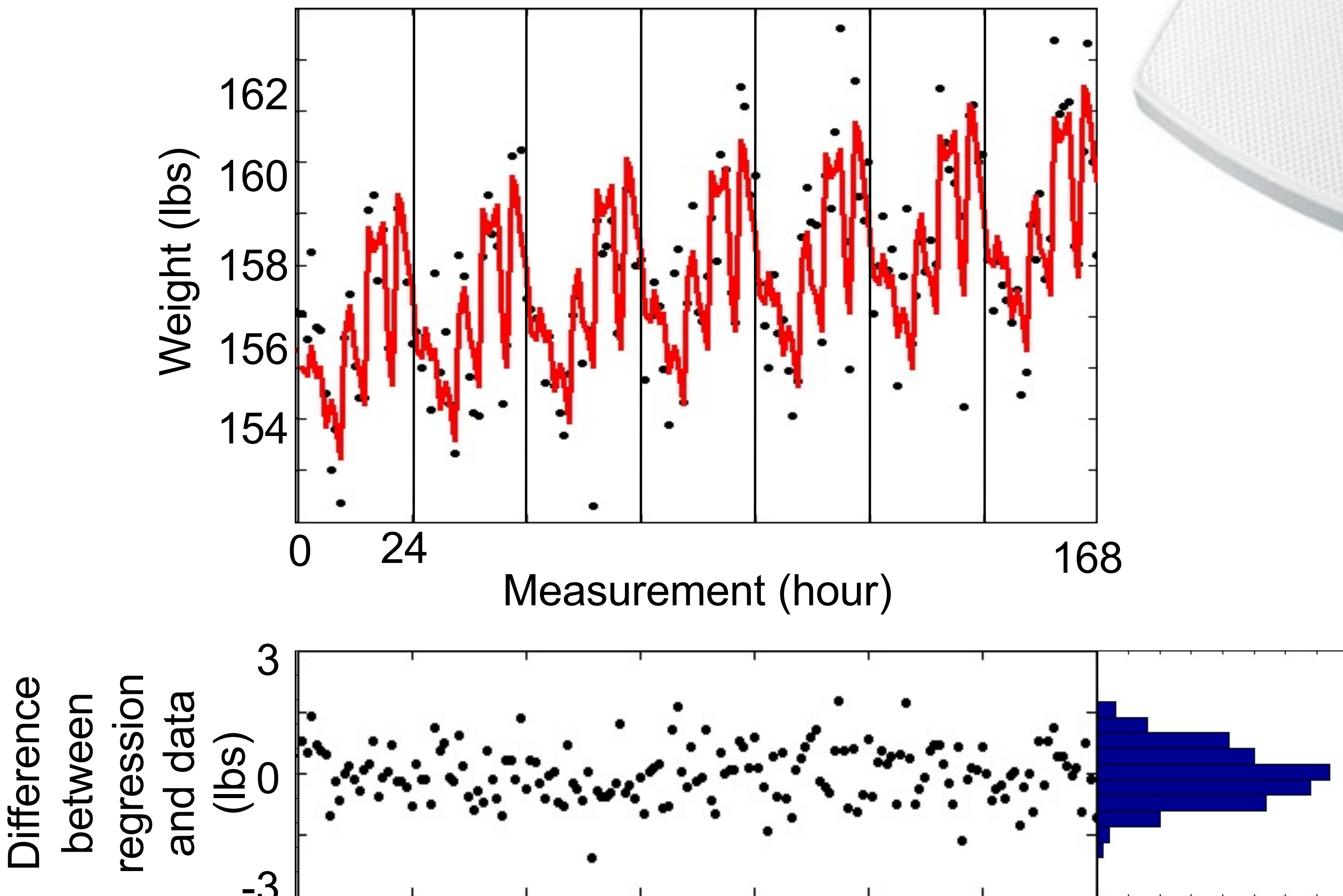
- Taking trend into account reduces the apparent randomness





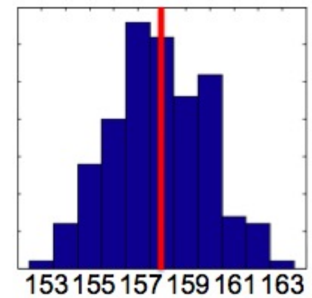
Variation and randomness

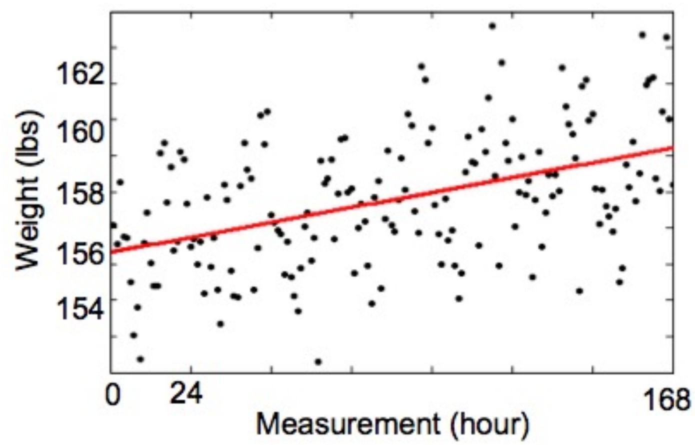
- Taking cyclical (hourly, daily) patterns further reduces error

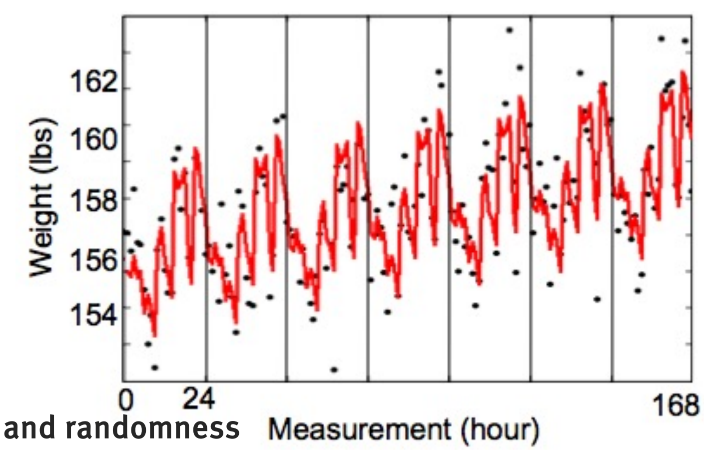


Variation and randomness

- What is my weight?
  - It was ~155 lbs a week ago
  - I am gaining ~0.015 lbs/hr
  - Weight systematically fluctuates over a range of 5 lbs in a 24 hr cycle.
  - After taking all that into account, there is still some unexplained variation of +/- 2 lbs
    (perhaps random error? More likely systematic deviations from regular trends in daily cycle, or systematic variation in how the scale operates)

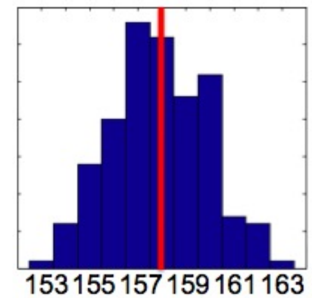




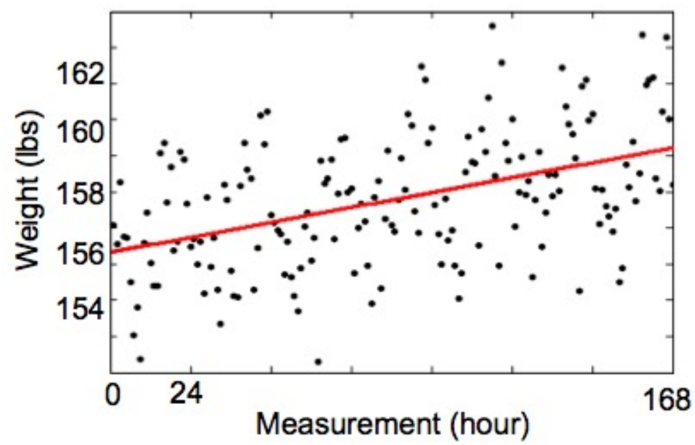


Variation and randomness

- Unaccounted-for variation is considered “random”



- This can be called:
  - “noise”
  - “random error”
  - “sampling variability”



- Someone’s “noise” may be another’s “signal”, depending on what you know about the data and what analytical tools you have at your disposal.