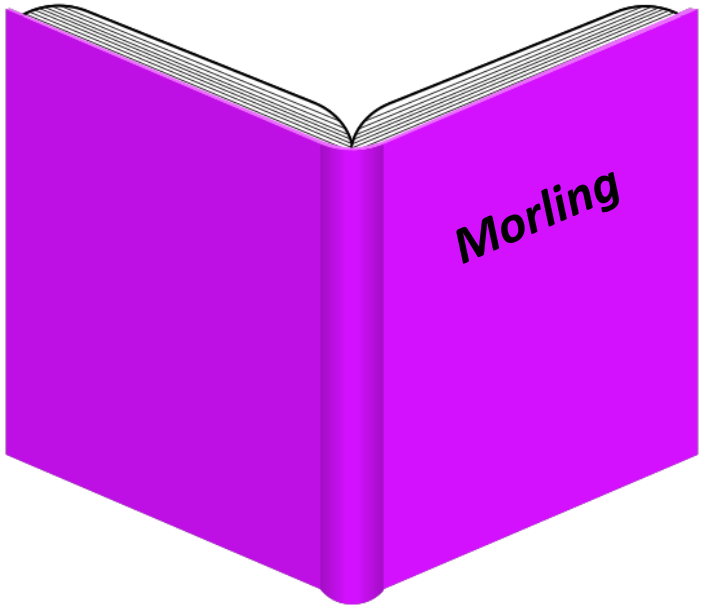




# Research Methods and Statistics

# Lecture 21: Regression analysis

Riet van Bork

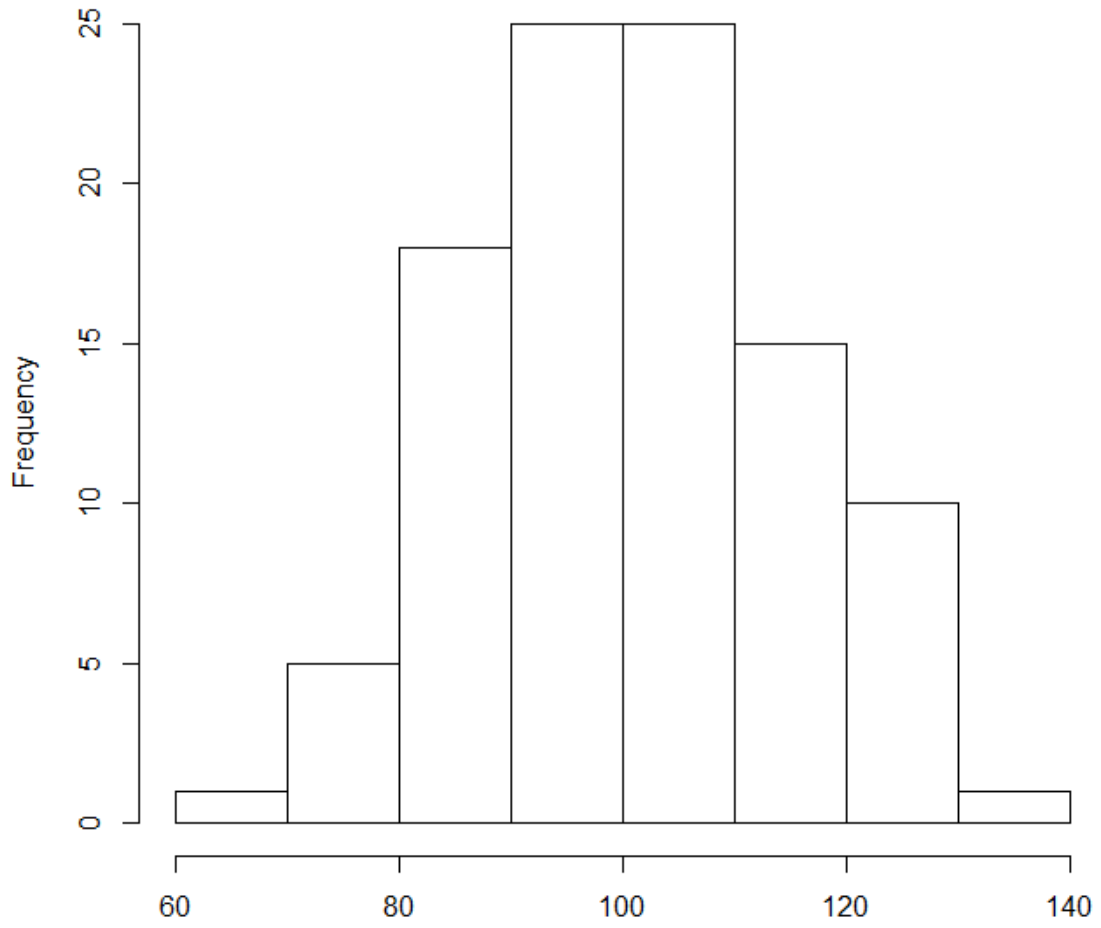
Pictures source: pixabay.org

Explanatory variable/independent variable

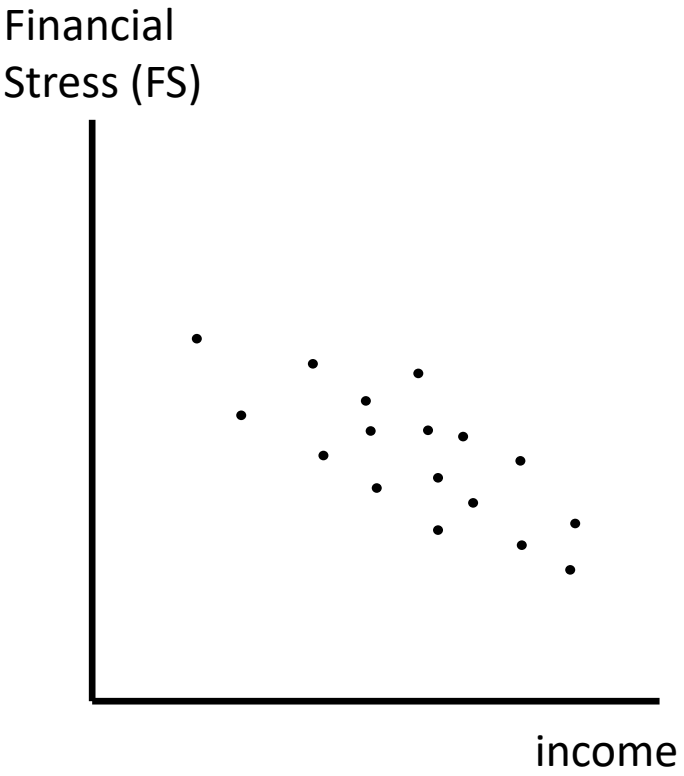
Response variable/ dependent variable

| | **Quantitative** | **Categorical** |
|---|---|---|
| **Quantitative** | Correlation<br>Regression | t-test<br>ANOVA |
| **Categorical** | Logistic regression | Contingency table |

# Today’s example: Financial stress test



# Predicting financial stress from income




Regression: Find a formula to predict ‘fincancial stress’ (FS) from ‘income’
In linear regression the regression line is a straight line:

$$\hat{y} = a + bx \qquad \widehat{\text{FS}} = \text{a} + \text{b} * \text{income}$$

# Today

**Correlation vs regression**

Constructing a regression line

How well does the regression line predict?

Population inferences

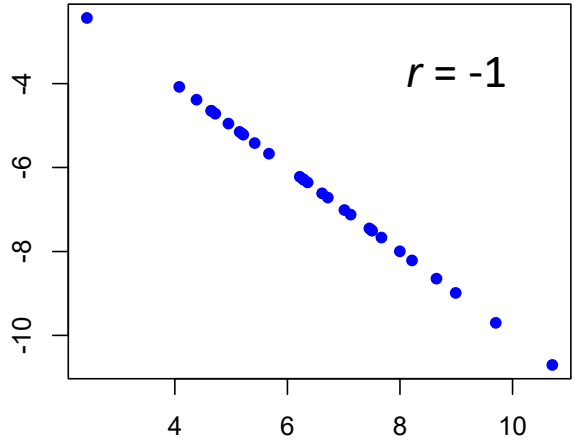
# Correlation

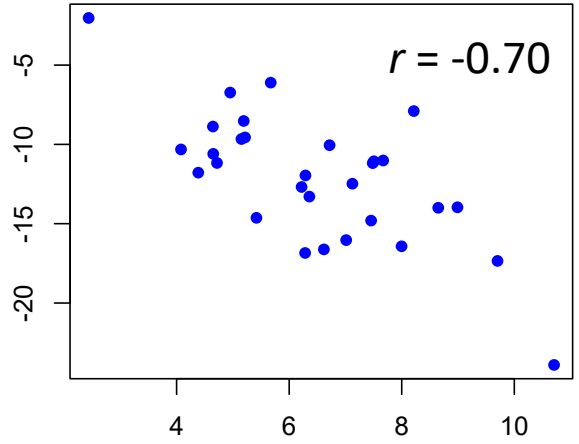
- Correlation coefficient, *r*
  - Number between -1 and 1 indicating ***the strength*** of the linear relation
    - E.g., *r* = -.54
  - But to predict an outcome value you do not only need to know the strength of the relation, you also need to take into account the scale on which the two variables are measured

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

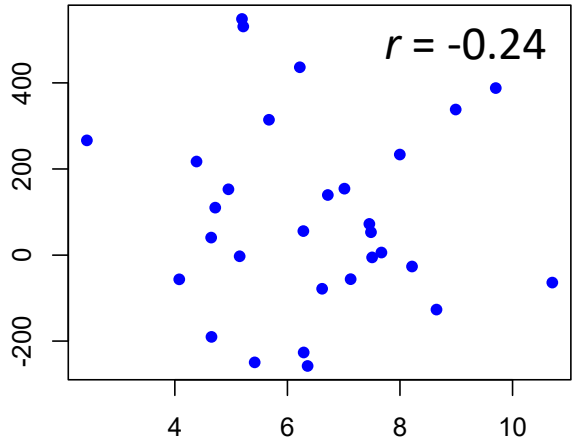
- Correlation is standardized: it does <u>not</u> depend on scale of X and Y

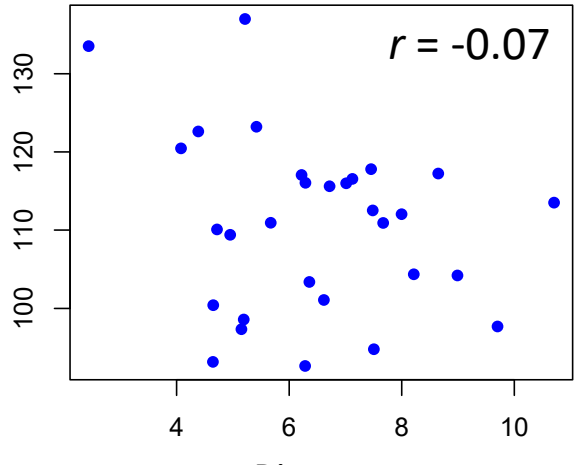
# Correlation strength




← Strong




← Weak

# Today

Correlation versus regression

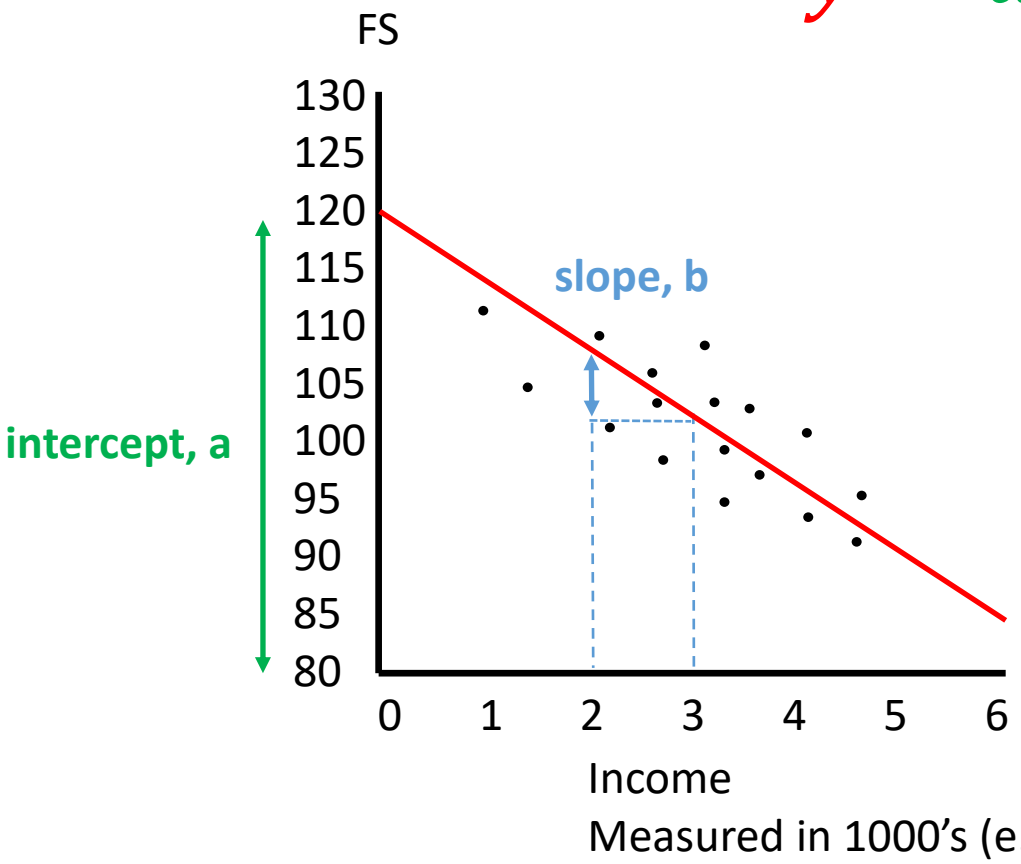
**Constructing a regression line**

How well does the regression line predict?

Population inferences

# Regression line

$$\hat{y} = a + bx \qquad \widehat{FS} = a + b \times \text{income}$$




**intercept, a: the value for $\hat{y}$ (FS) if x (income) is 0**

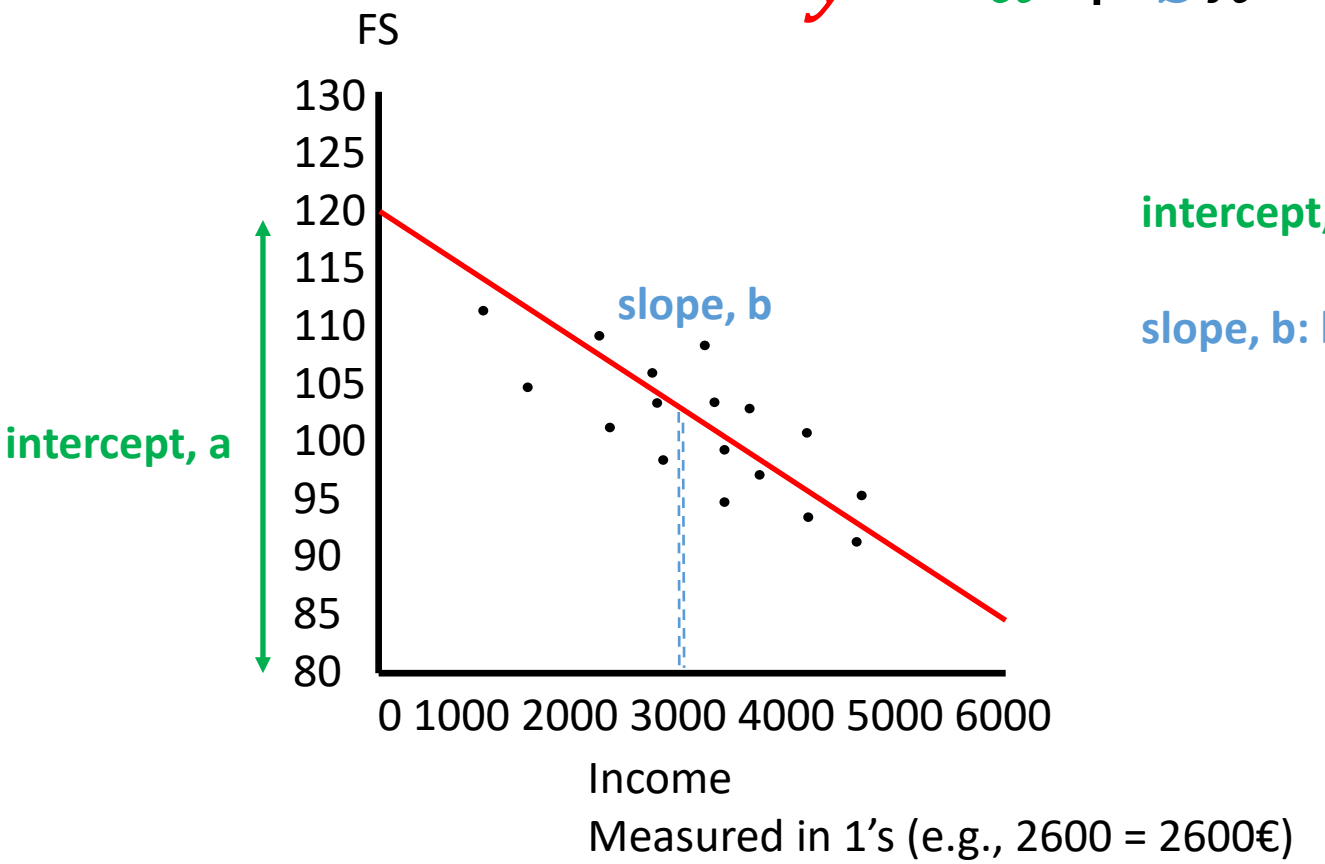
**slope, b: how much $\hat{y}$ (FS) changes if x (income) increases by 1 unit**

positive b (b>0), possitive association

negative b (b<0), negative association

# Regression line

$$\hat{y} = a + bx \qquad \widehat{FS} = a + b \times \text{income}$$




**intercept, a: the value for $\hat{y}$ (FS) if x (income) is 0**

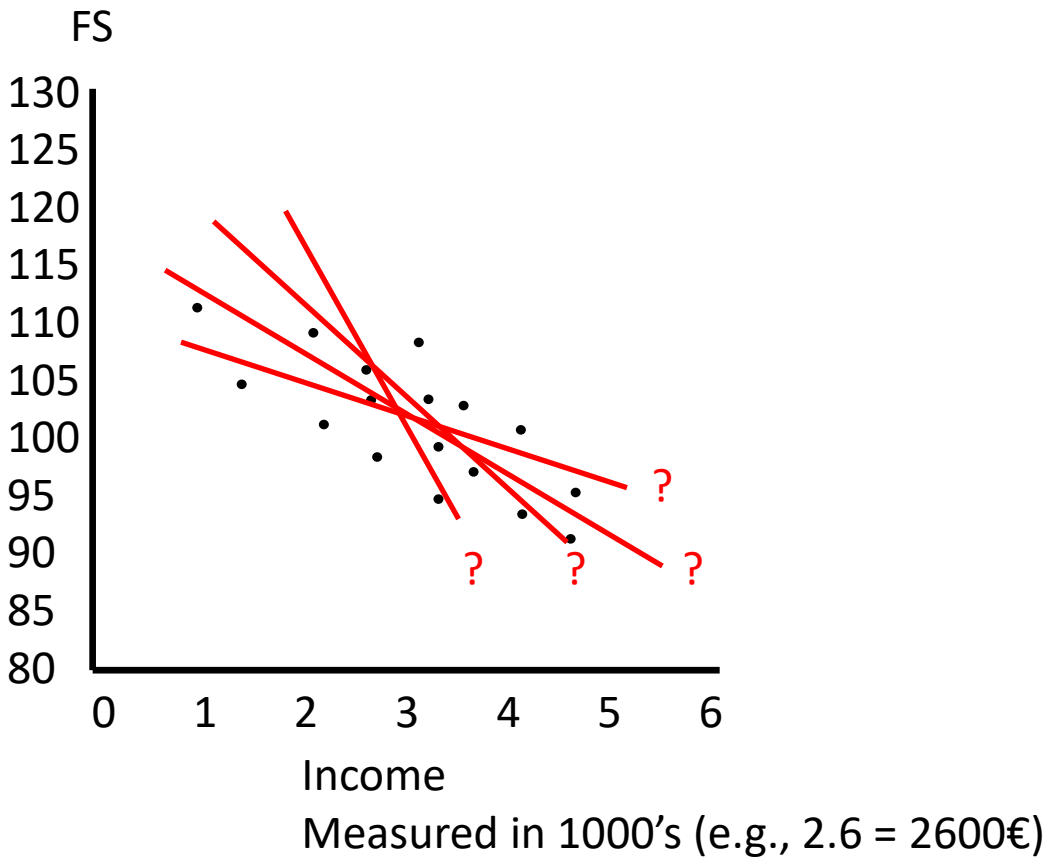
**slope, b: how much $\hat{y}$ (FS) changes if x (income) increases by 1 unit**

positive b (b>0), possitive association

negative b (b<0), negative association

Choosing a different scale influences the slope! Now b is much much smaller because b represents the change in FS for one unit change in income, which is here a change of only 1 euro! (instead of 1000 euro)

# Estimating the parameters



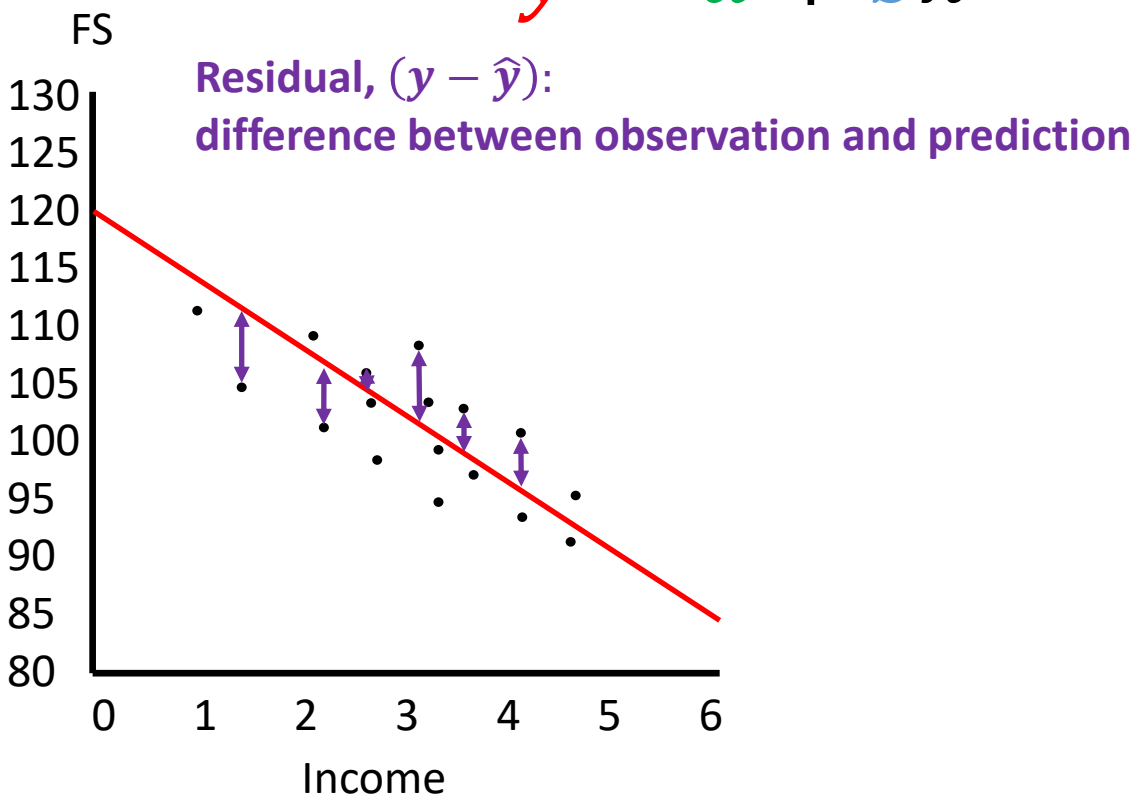

**Key problem**: How to draw a line through the points?

**(hopefully) intuitive idea:**
Draw the line in such a way that the distances to the line from points above the line are as much as the distances to the line from points below the line.
So some points are above and some below the line.

# Regression line

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

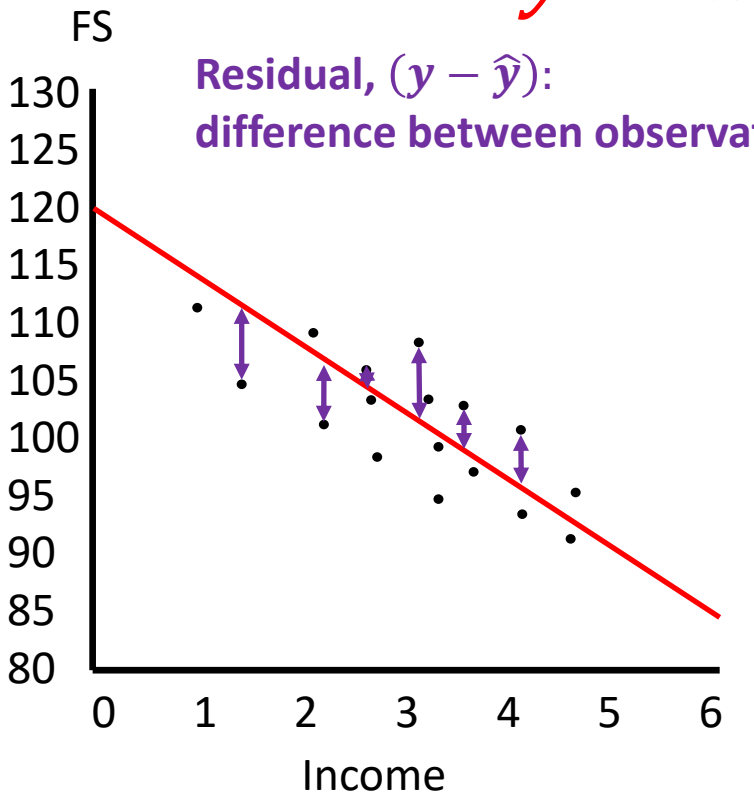
$$\widehat{FS} = a + b \times \text{income}$$



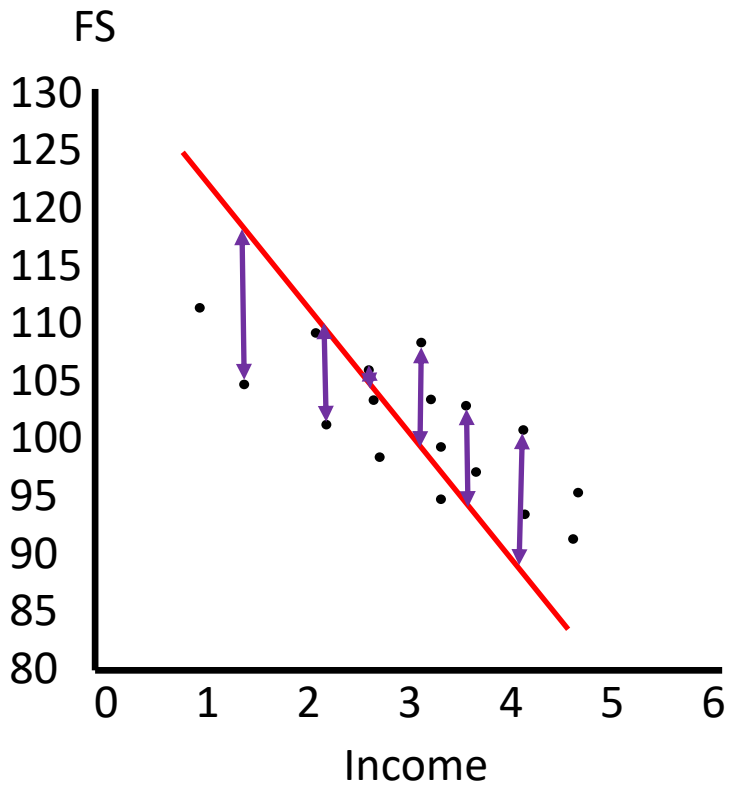
# Regression line

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

**Residual, $(y - \hat{y})$:**
**difference between observation and prediction**



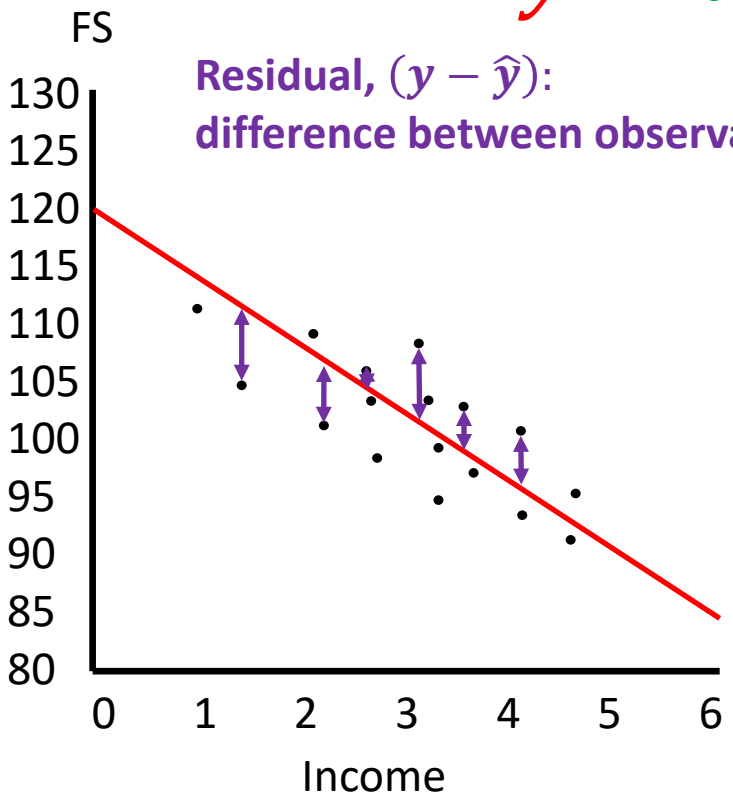

$$\widehat{FS} = a + b \times \text{income}$$




**Idea:** Choose the line that has the smallest possible residuals

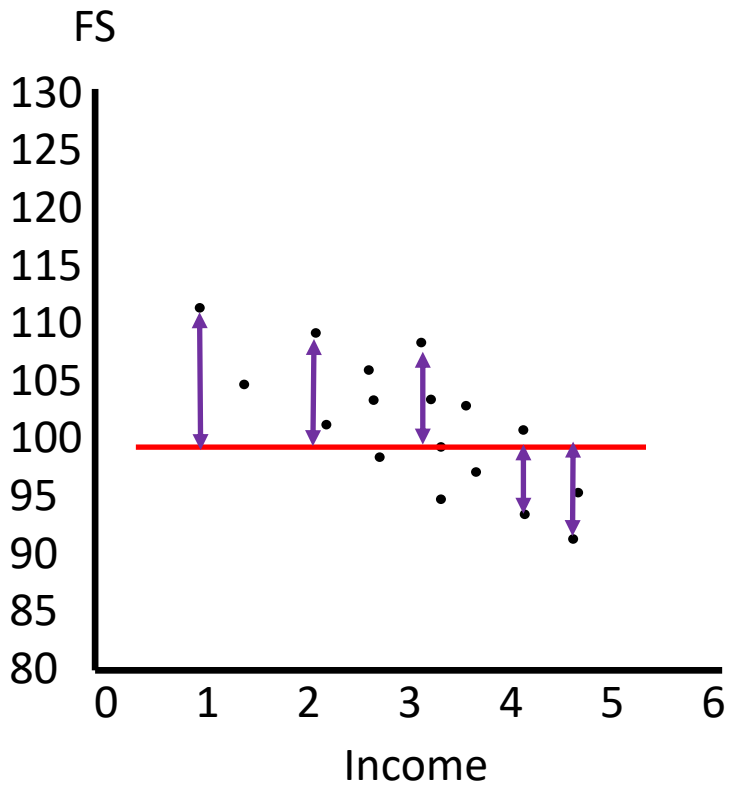
# Regression line

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




**Residual, $(y - \hat{y})$:**
**difference between observation and prediction**

$$\widehat{FS} = a + b \times \text{income}$$

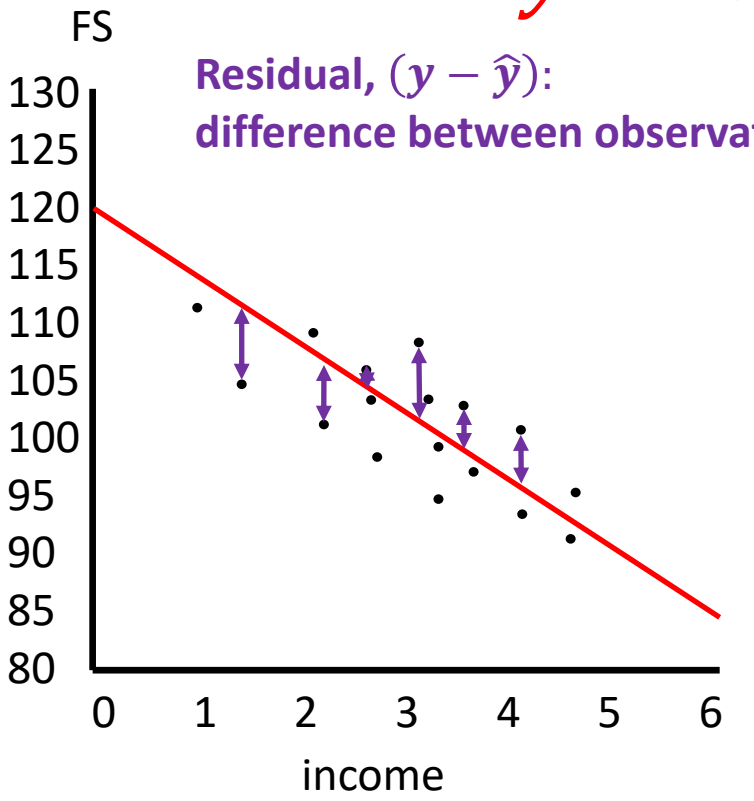



**Idea:** Choose the line that has the smallest possible residuals

# Regression line

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

$$\widehat{FS} = a + b \times \text{income}$$




**Residual, $(y - \hat{y})$:**
**difference between observation and prediction**

- Summing residuals $(y - \hat{y})$ does not work
  - positive residuals will cancel out negative residuals
  - So, instead, first square them
- →Choose the line with the smallest sum of squared residuals

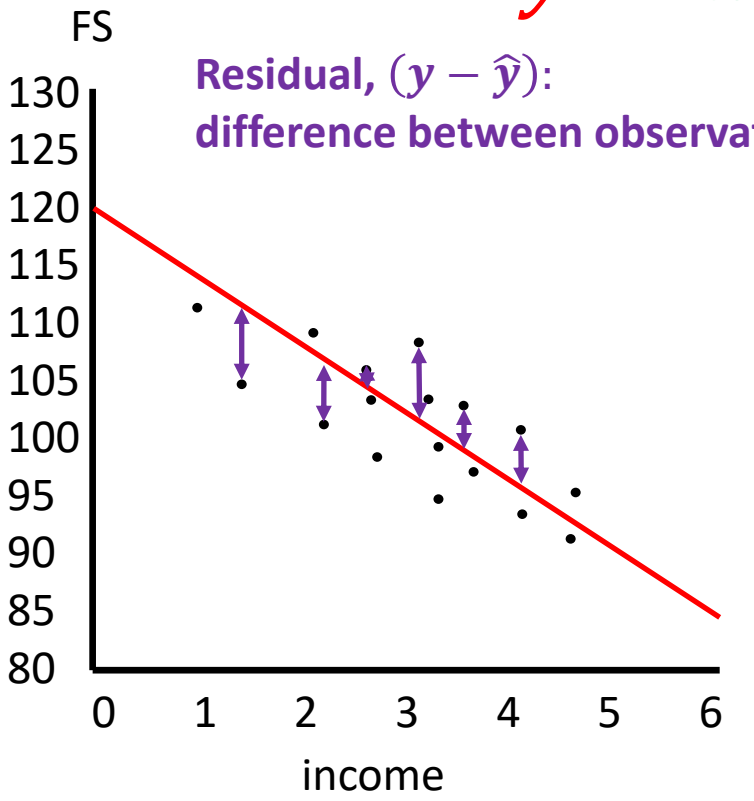
$$\Sigma(y - \hat{y})^2$$

  - Called a *least squares line*

**Idea:** Choose the line that has the smallest possible residuals

# Regression line

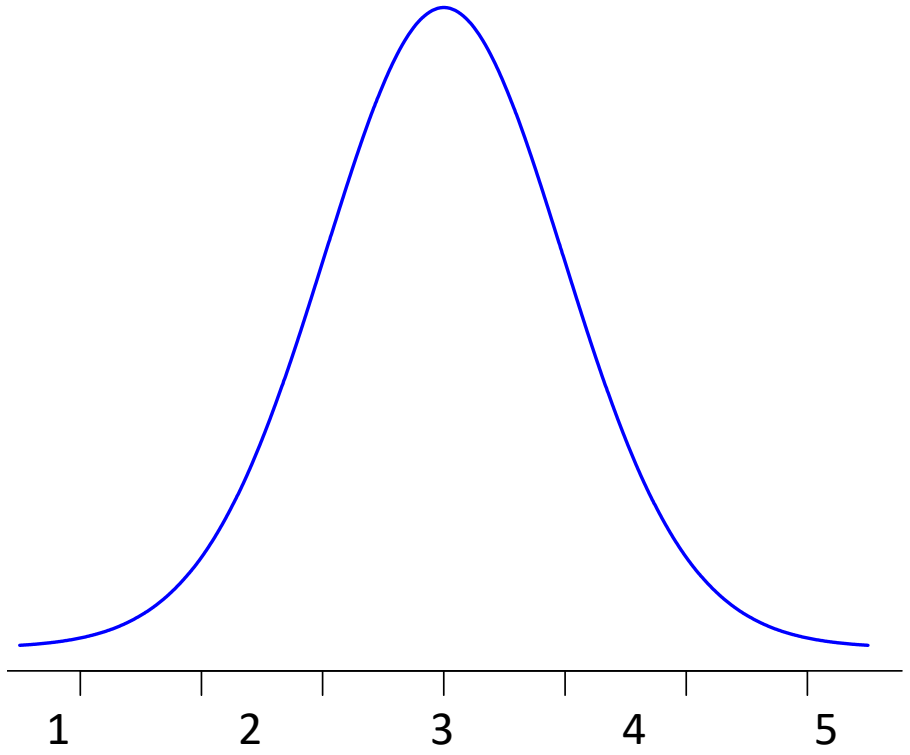
$$\hat{y} = a + bx$$



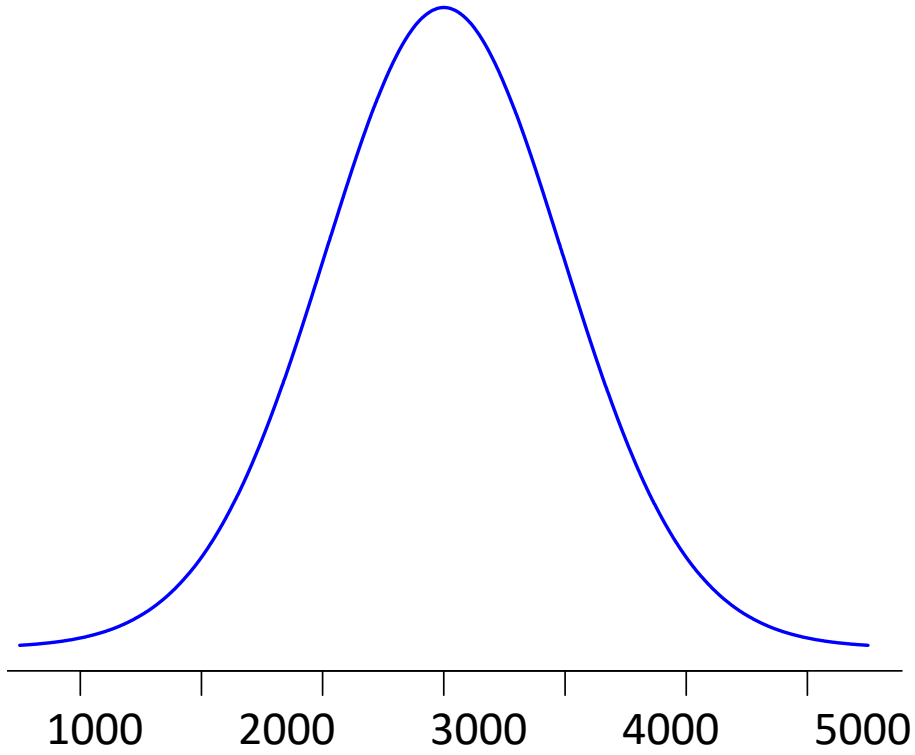

$$\widehat{FS} = a + b \times \text{income}$$

- This slope and intercept give the least squares line:
  - $b = r\left(\frac{s_y}{s_x}\right)$
  - $a = \bar{y} - b\bar{x}$
  - (section 3.3)

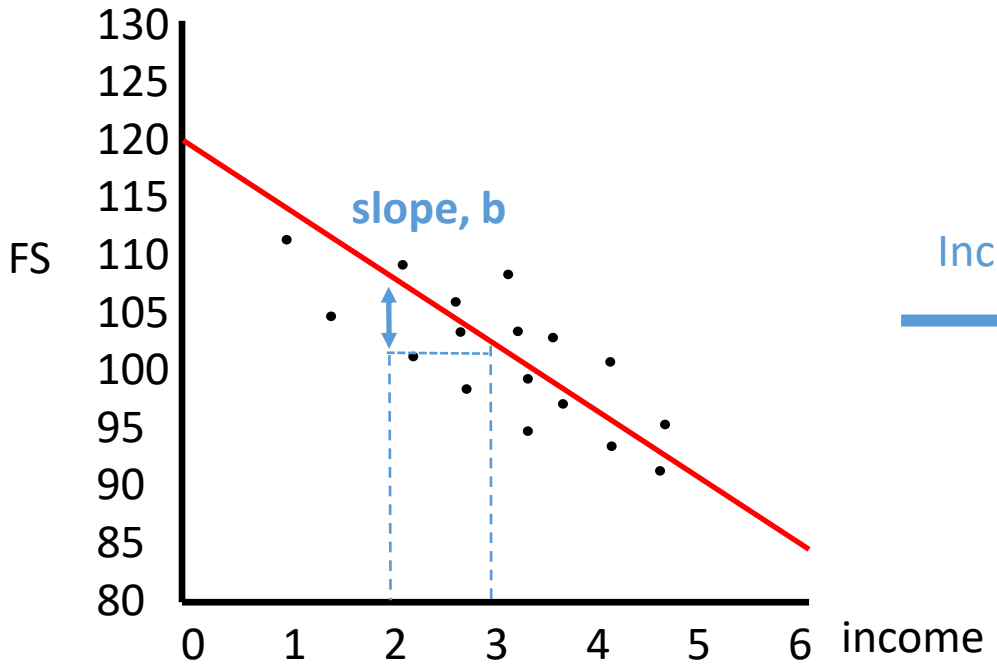
**Idea:** Choose the line that has the smallest possible residuals



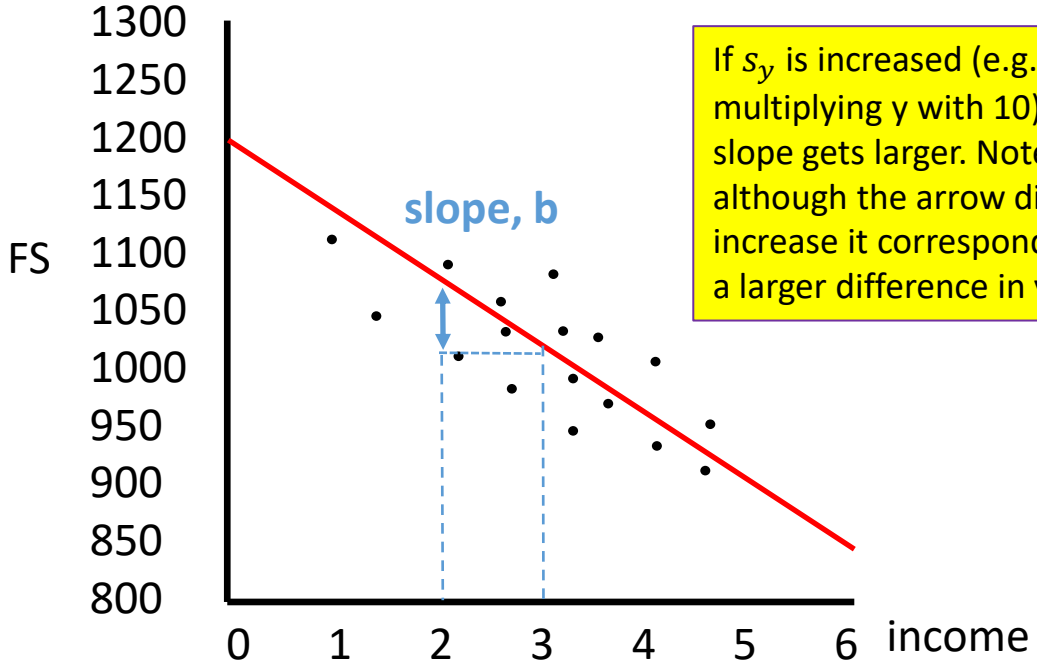

Measured in 1000’s of euros (1 = 1000 euro)



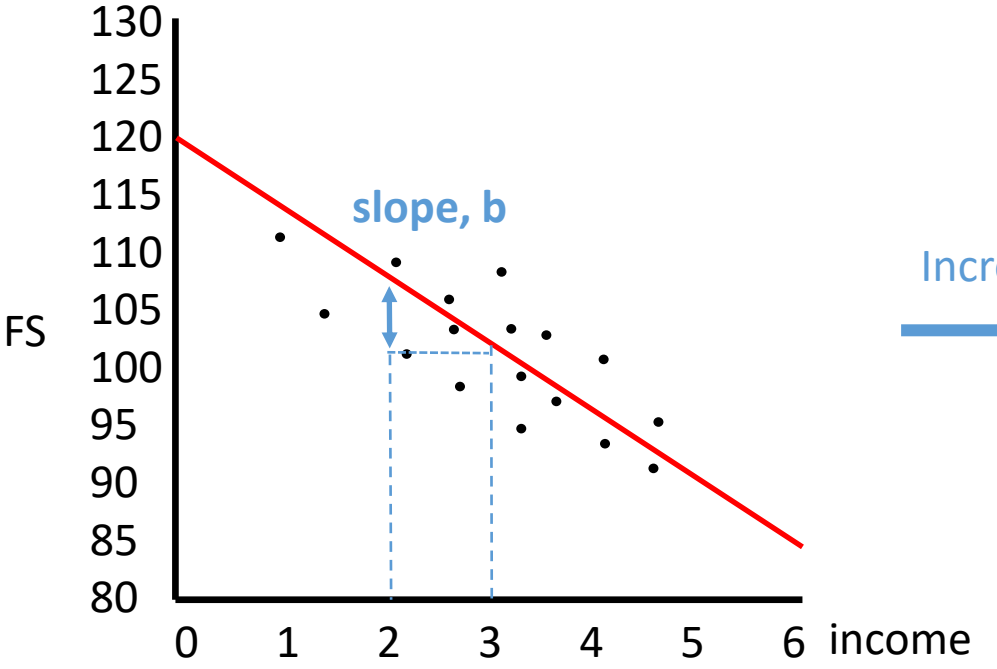

Measured in 1’s (1 = 1 euro)
Variable is multiplied by 1000, so is it’s SD!



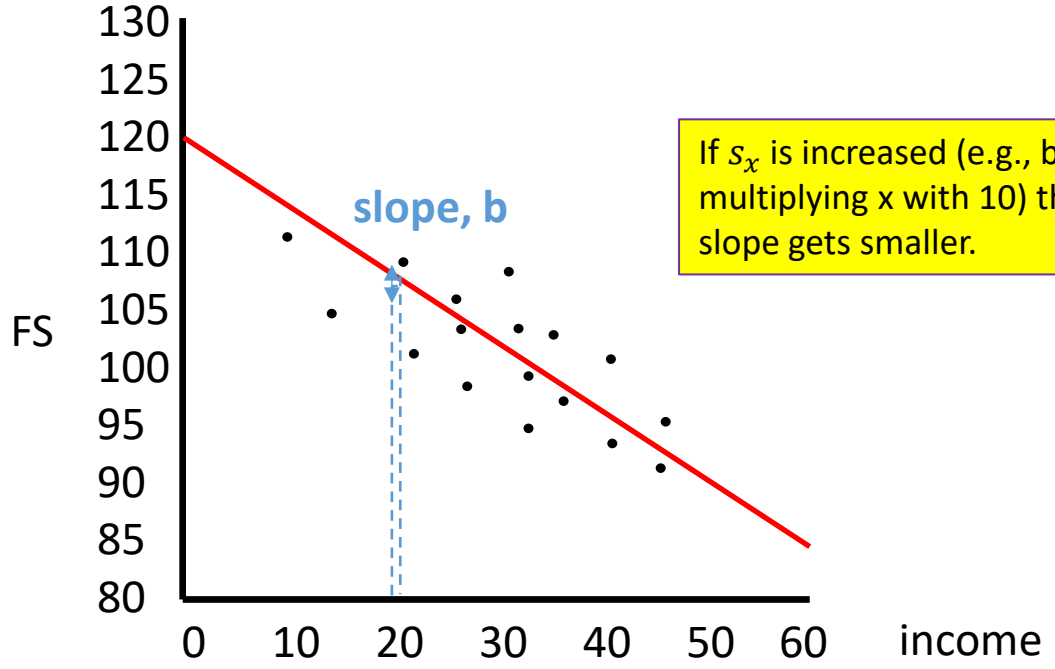

Increased $s_y$




If $s_y$ is increased (e.g., by multiplying y with 10) the slope gets larger. Note that although the arrow did not increase it corresponds with a larger difference in y (FS).



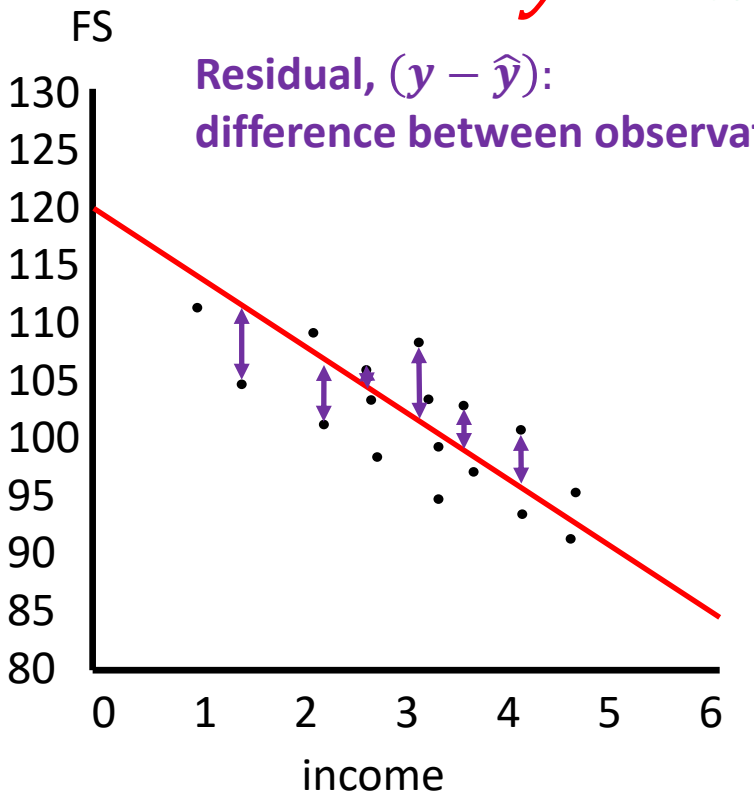

Increased $s_x$




If $s_x$ is increased (e.g., by multiplying x with 10) the slope gets smaller.

# Regression line

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




**Residual, $(y - \hat{y})$:**
**difference between observation and prediction**

$$\widehat{FS} = a + b \times \text{income}$$

- This slope and intercept give the least squares line:
  - $b = r\left(\frac{s_y}{s_x}\right)$
  - $a = \bar{y} - b\bar{x}$
  - (section 3.3)

**Idea:** Choose the line that has the smallest possible residuals

# Estimating parameters

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

- $b = r\frac{s_y}{s_x}$
- $a = \bar{y} - b\bar{x}$

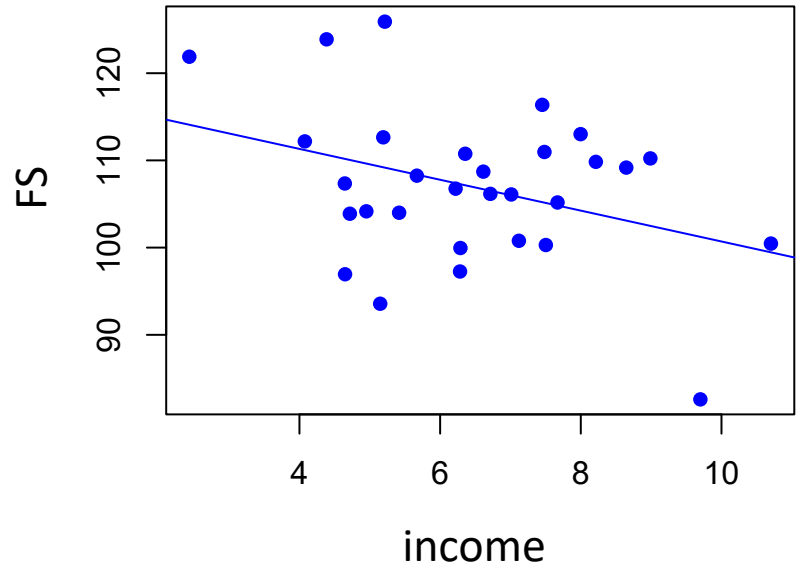
Use can use excel for these:

| | |
|---|---|
| correlation: | =CORREL(..) |
| SD: | =STDEV.S(..) |
| mean: | =AVERAGE(..) |

$\bar{x} = 6.45$ $\quad \bar{y} = 106.98$

$s_x = 1.81$ $\quad s_y = 8.87$

$r = -0.36$




$$b = -0.36 \times \frac{8.87}{1.81} = -1.77$$

Note that slope will have same sign as correlation!

$$a = 106.98 - (-1.77 \times 6.45) = 118.38$$

$$\hat{y} = 118.38 - 1.77x$$

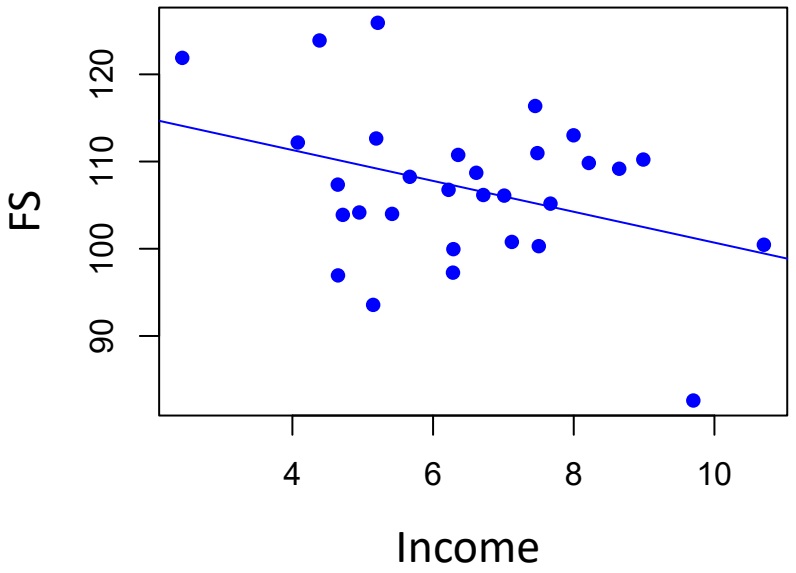
So, someone with a salary of 3000 per month, is expected to have a financial stress score of
118.38 – 1.77(3) = 113.07

# How can we interpret these numbers?

$\hat{y} = 118.38 - 1.77x$

- For a (hypothetical) income of 0, the predicted FS is 118.38

- If the income increases by 1 unit, then FS is predicted to decrease by 1.77



# Today

Correlation versus regression

Constructing a regression line

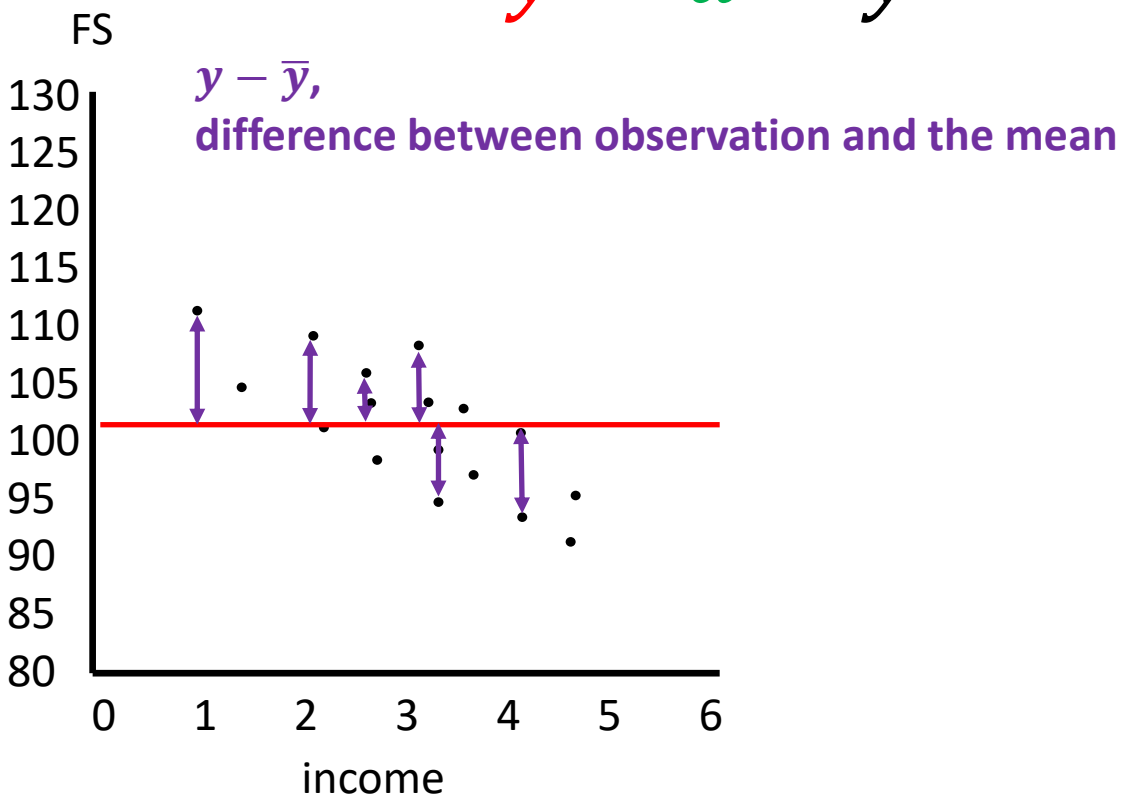
**How well does the regression line predict?**

Population inferences

# Using $\bar{y}$ to predict $y$

$$\hat{y} = a = \bar{y}$$

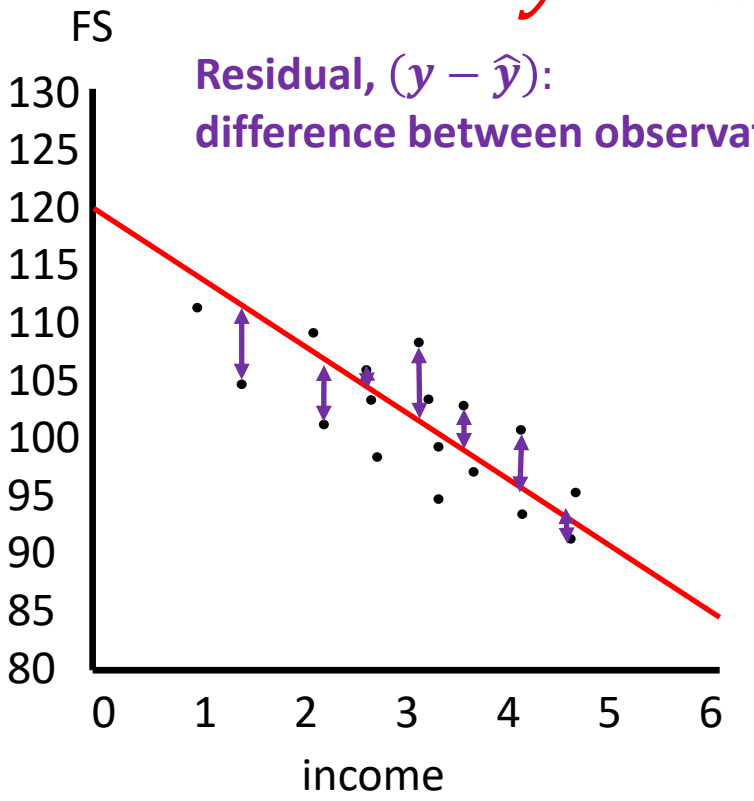
$$\widehat{FS} = a$$




- If we neglect x (income), the best prediction we can do is predicting an average FS, i.e., $\bar{y}$
- The difference is then: $y - \bar{y}$
  - Sum of squared differences
    $$\Sigma(y - \bar{y})^2$$
  - "*total sum of squares*" (total SS)
  - This is also a measure of 'total variance' (the numerator of the sample variance formula for y)

# Using the regression line to predict $y$

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

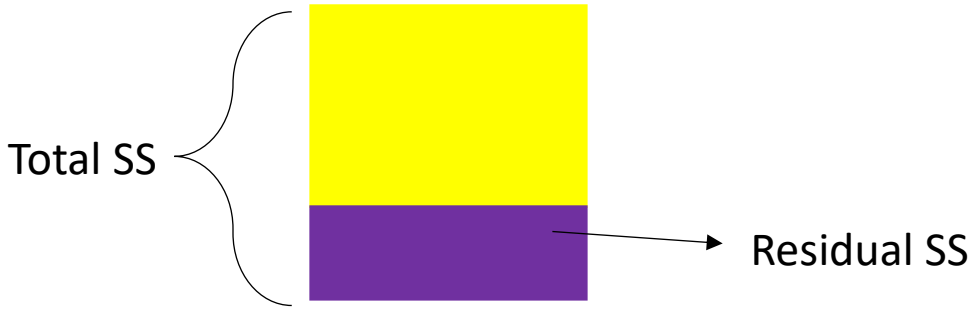


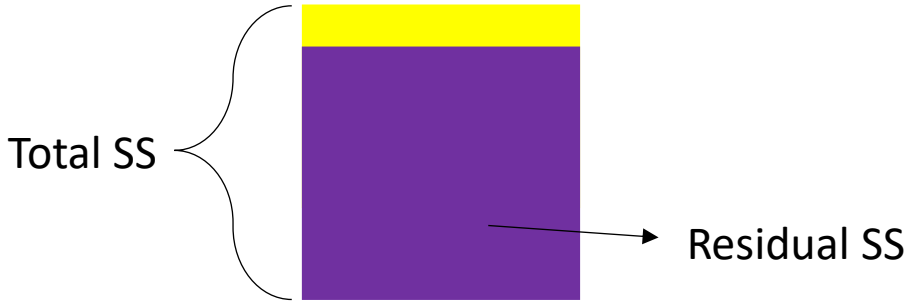

$$\widehat{FS} = a + b \times \text{income}$$

- If we include x (income), the prediction is the line $\hat{y}$
- The difference is then: $y - \hat{y}$
  - i.e., "residual"
  - With sum of squared residuals $\Sigma(y - \hat{y})^2$
  - "*residual sum of squares*" (RSS)

# How well does the regression line predict?

- The proportional decrease in the prediction error is $r^2 = \frac{\text{total SS} - \text{RSS}}{\text{total SS}}$
  - How much is the error decreased proportionally by adding the predictor




*$r^2$ is the proportion of variance in y (FS) explained by x (income), i.e., the proportion of yellow in the total SS*

Regression model **good**:

- RSS is small → $r^2$ large
- Regression model does much better than simply predicting the mean

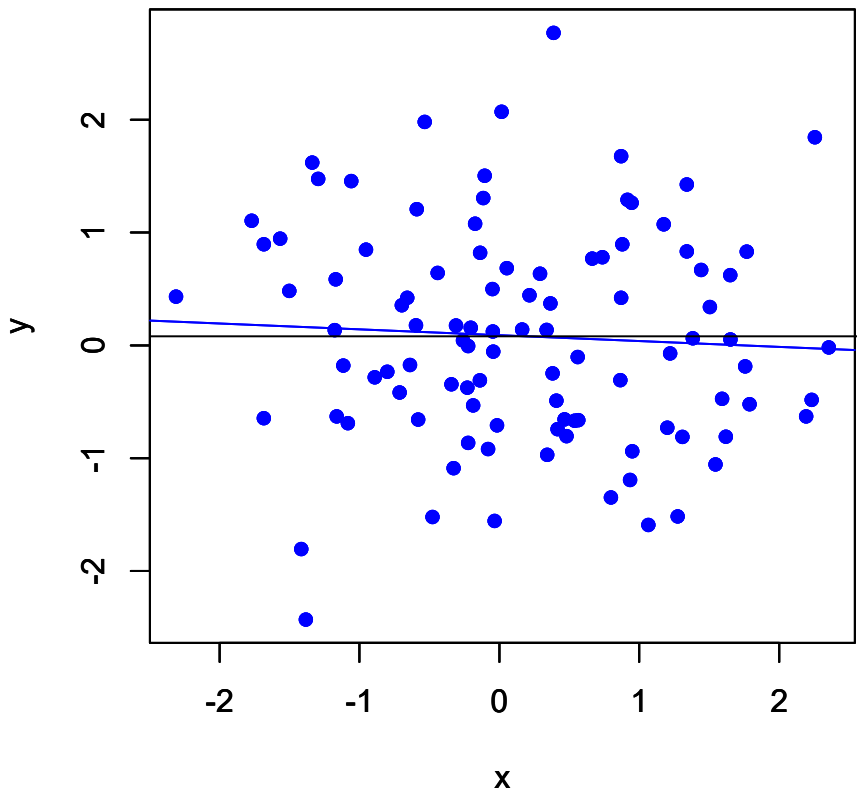
Regression model **bad**:

- RSS is large → $r^2$ small
- Regression model doesn't do much better than simply predicting the mean

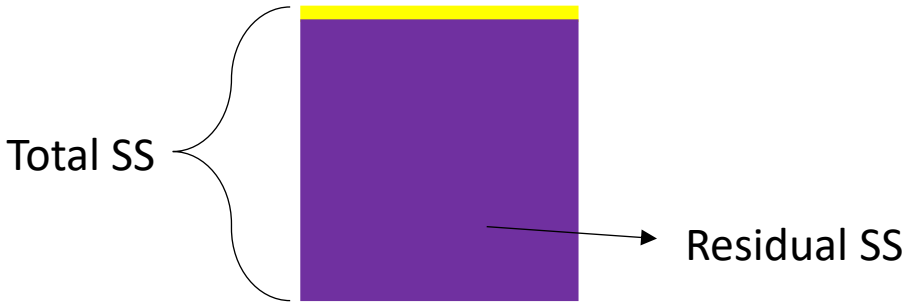
Compute using *r* (i.e., square the correlation coefficient)

# Examples

$\hat{y}$ =regression line (blue line)
$\bar{y}$ = mean (black line)
RSS= $\sum(y - \hat{y})$
Total SS = $\sum(y - \bar{y})$

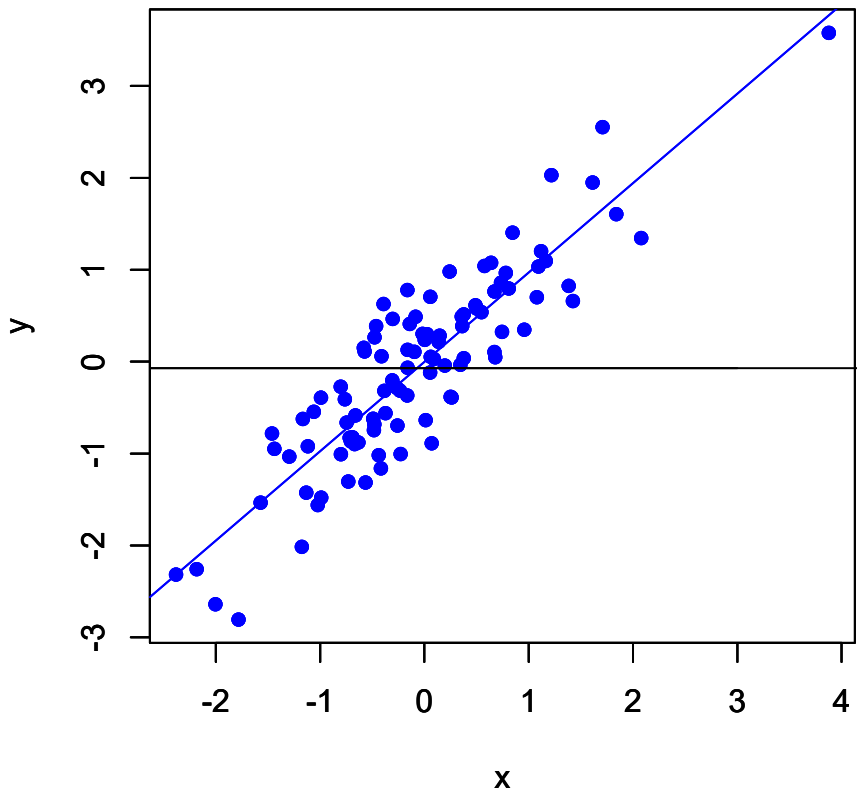



1. Compute regression line that has lowest RSS
2. Compute Total SS
3. Because Total SS ≈ RSS, $r^2 \approx 0$

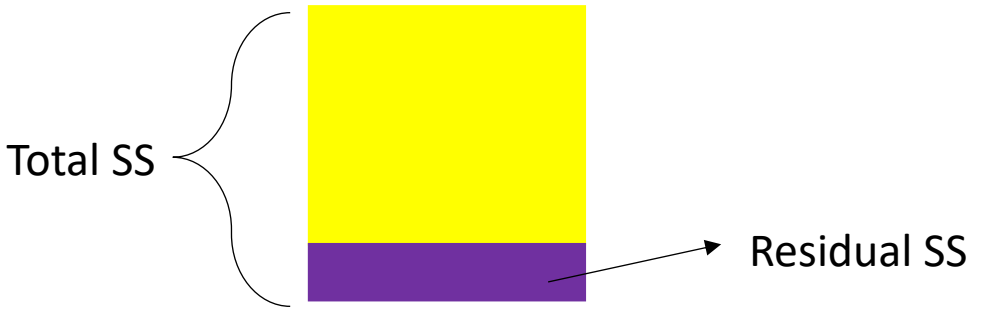


# Examples

$\hat{y}$ =regression line (blue line)
$\bar{y}$ = mean (black line)
RSS= $\sum(y - \hat{y})$
Total SS = $\sum(y - \bar{y})$




1. Compute regression line that has lowest RSS
2. Compute Total SS
3. Because Total SS > RSS, $r^2 > 0$

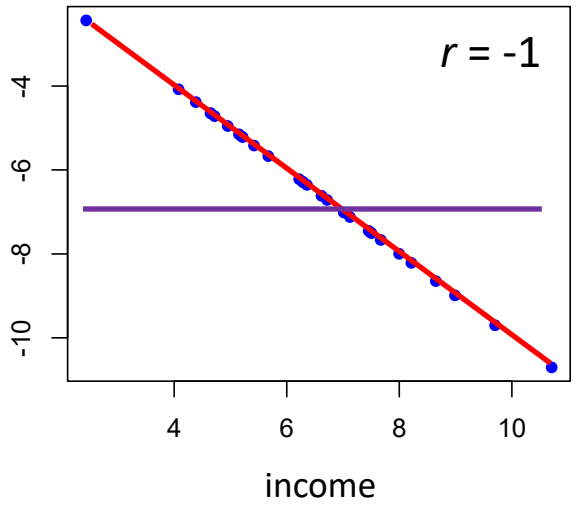


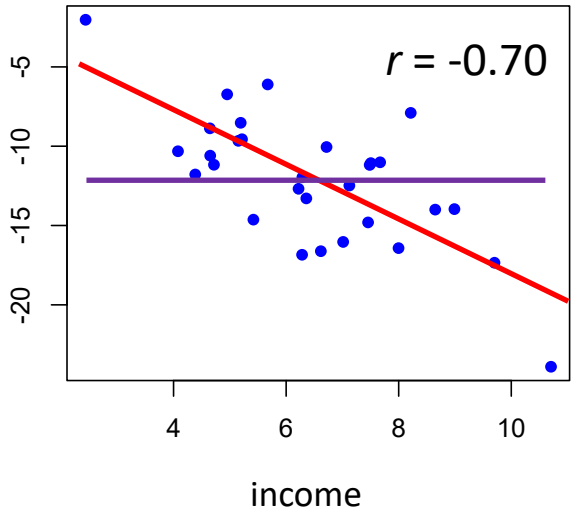
RSS is residuals to red line
TSS is residuals to purple line

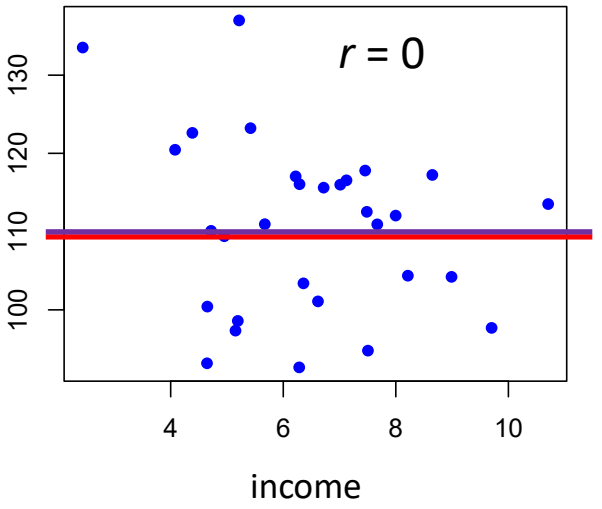
$$r^2 = \frac{\text{total SS} - \text{RSS}}{\text{total SS}}$$

$$\widehat{FS} = a + b \times \text{income}$$

If r = 0, b = 0 so $\widehat{FS} = a = \bar{y}$




RSS = 0

So $r^2 = \frac{\text{total SS} - 0}{\text{total SS}} = 1$

$0 < r^2 < 1$

RSS = total SS

So $r^2 = \frac{0}{\text{total SS}} = 0$

# How well does the regression line predict?

Predictor: mean

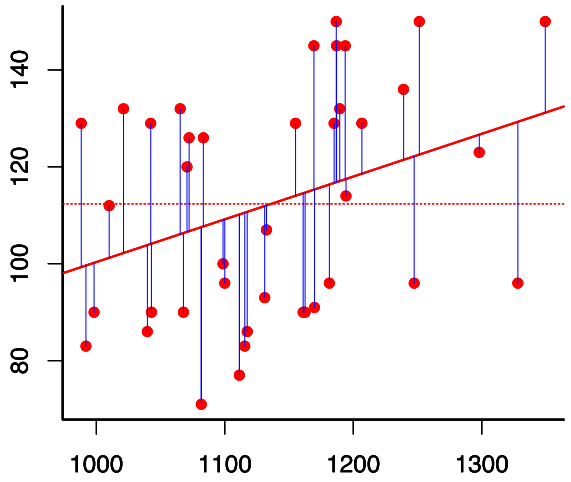
Total Sum of Squares (total SS):

$$\sum (y - \bar{y})^2 = 21751$$

Predictor: regression line

Residual Sum of Squares (RSS):

$$\sum (y - \hat{y})^2 = 19273$$



# How well does the regression line predict?

Predictor: mean

Total Sum of Squares (total SS):

$$\sum (y - \bar{y})^2 = 21751$$

Predictor: regression line

Residual Sum of Squares (RSS):

$$\sum (y - \hat{y})^2 = 19273$$

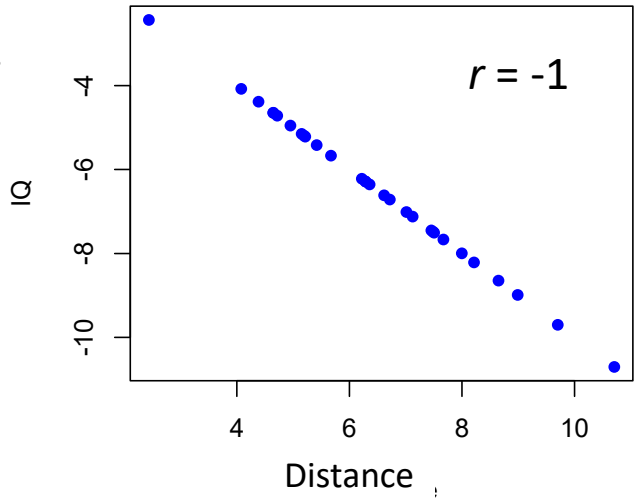
$$r^2 = \frac{\text{total SS} - \text{RSS}}{\text{total SS}} = \frac{21751 - 19273}{21751} = 0.114$$

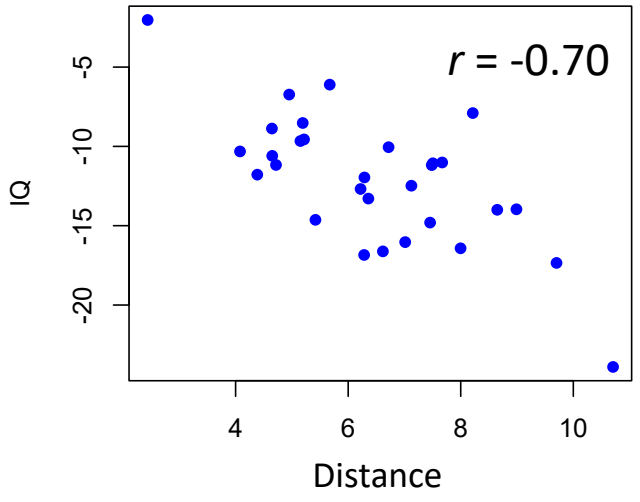
Thus: the variance around the regression line is 11.4% less than the total variance
Put differently: the error using $\hat{y}$ to predict $y$ is 11.4% smaller than the error using $\bar{y}$ to predict $y$
$r^2$ is thus the *proportional reduction in error*

→11.4% of the total variance in *y* is explained by *x*

# Correlation strength and explained variance

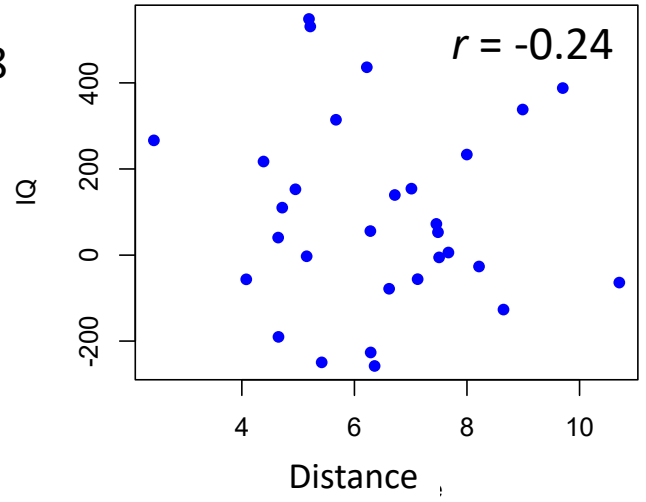
$r^2 = 1$
100% variance explained

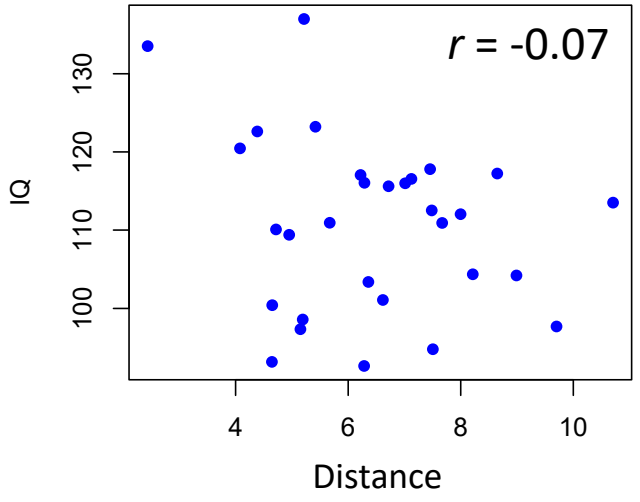



$r^2 = .49$
49% variance explained

$r^2 = 0.058$
~ 6% variance explained




$r^2 = 0.0049$
~ 0.5% variance explained

# Today

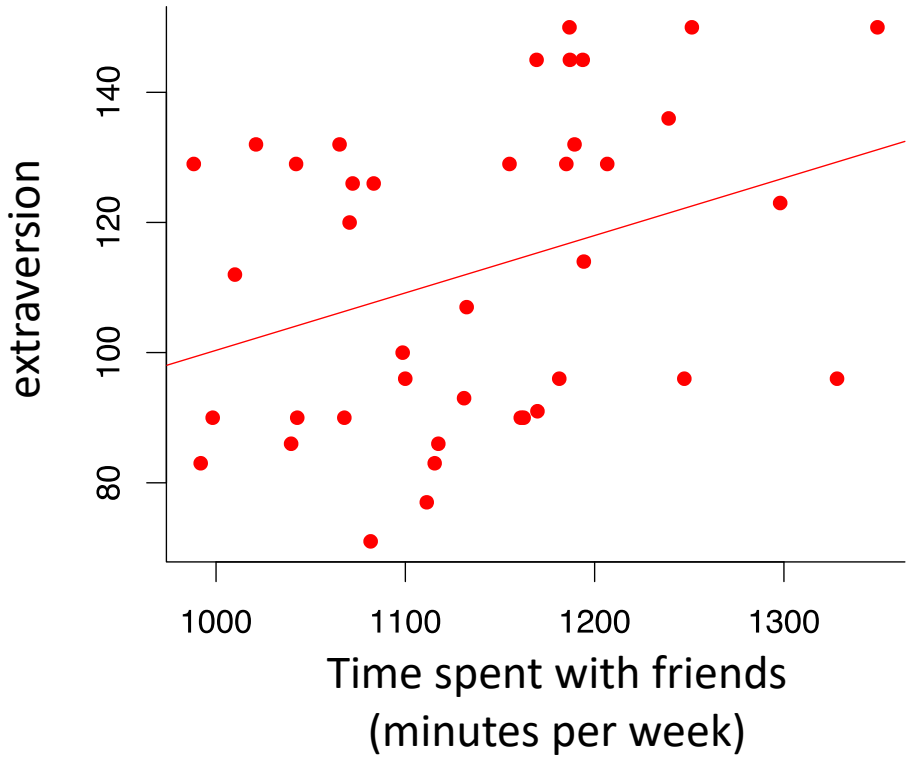
Correlation versus regression

Constructing a regression line

How well does the regression line predict?

**Population inferences**

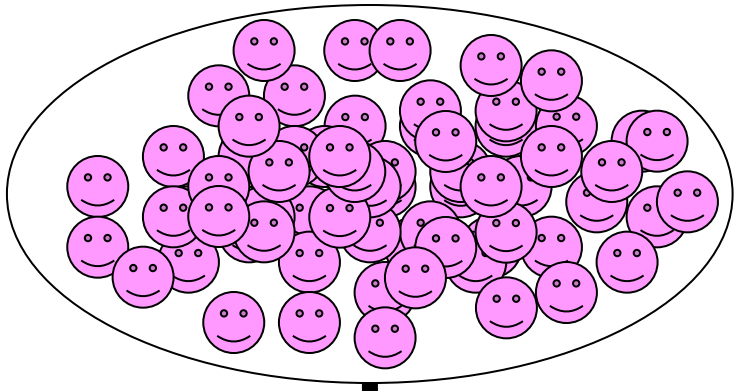
# Can we predict someone's extraversion score from the time they spend with friends?

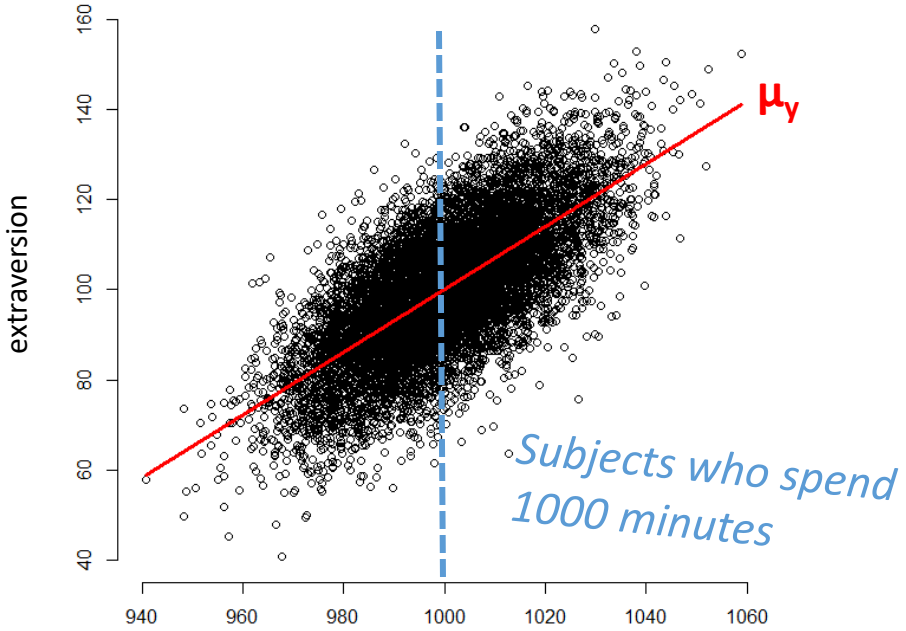



*n = 40*

**Population**

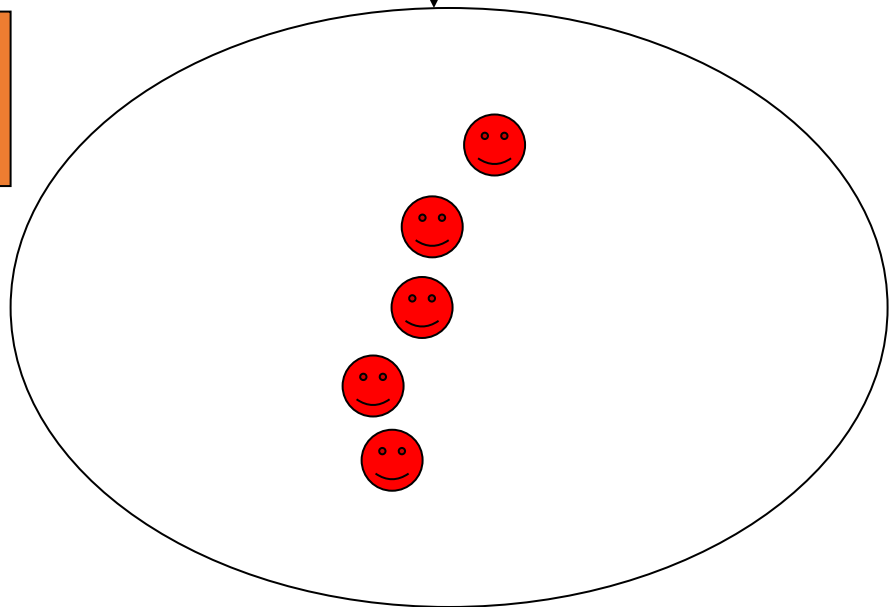
$$\mu_y = \alpha + \beta x$$

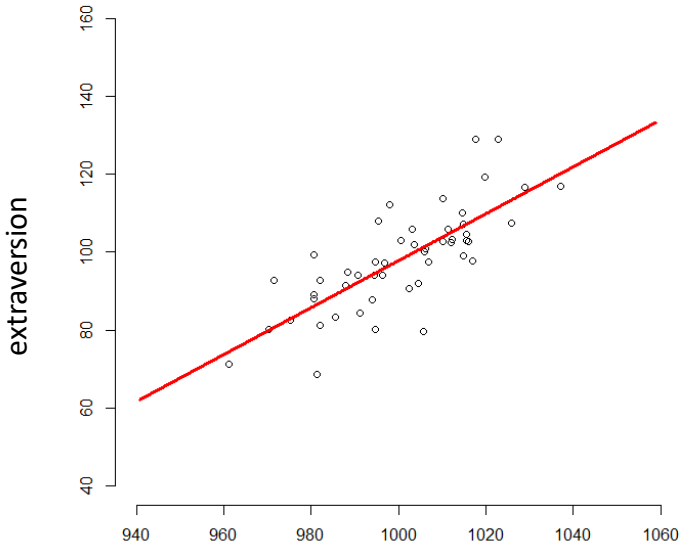





**Sample**

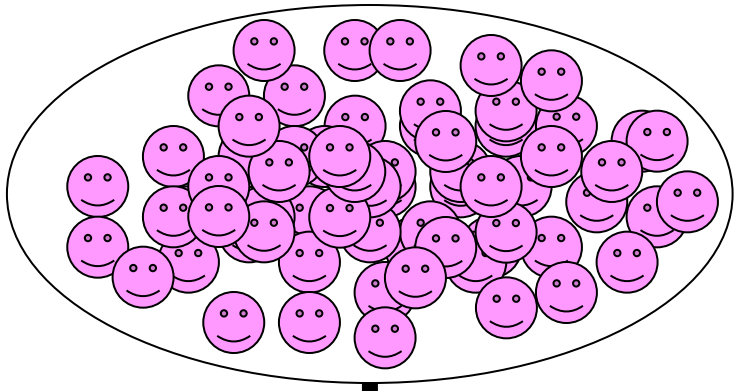
$$\hat{y} = a + bx$$

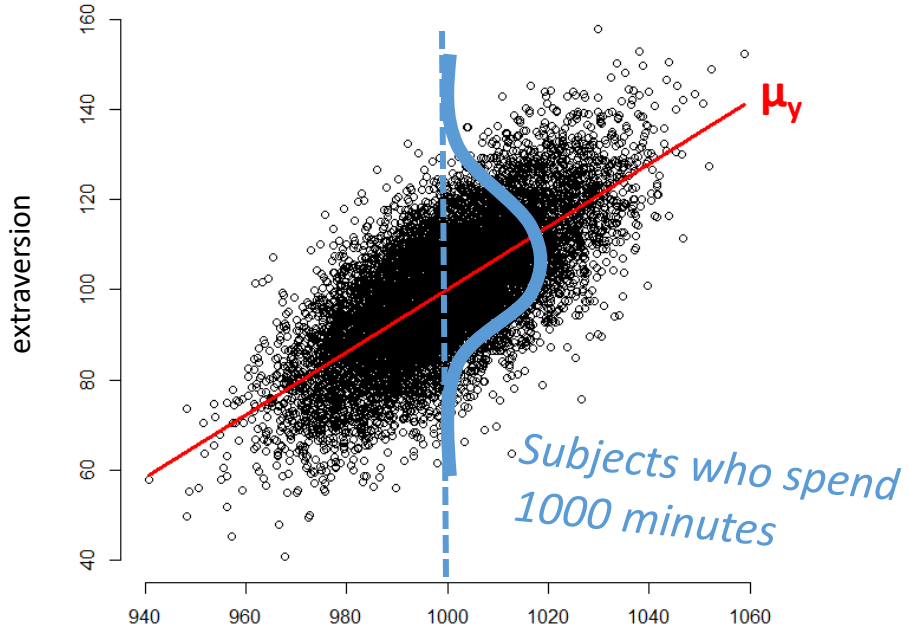




**Population**

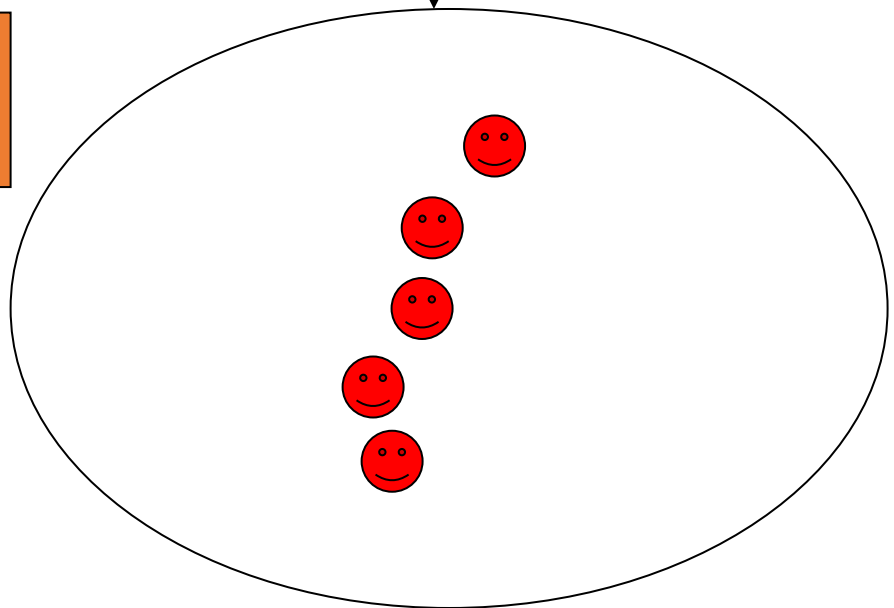
$\mu_y = \alpha + \beta x$

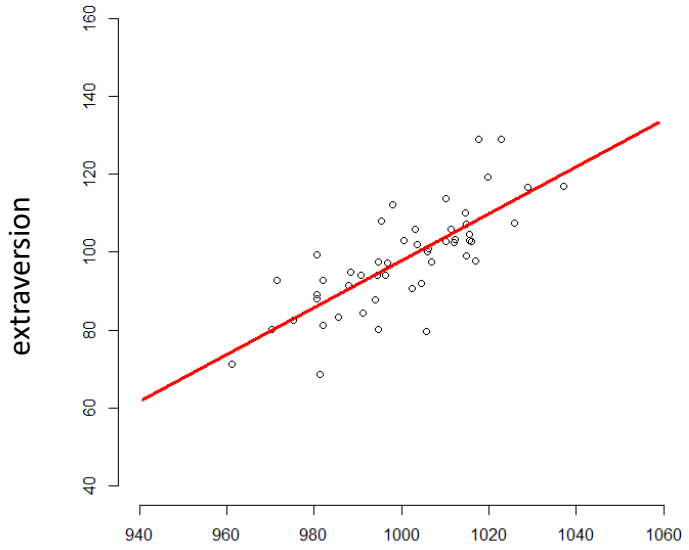




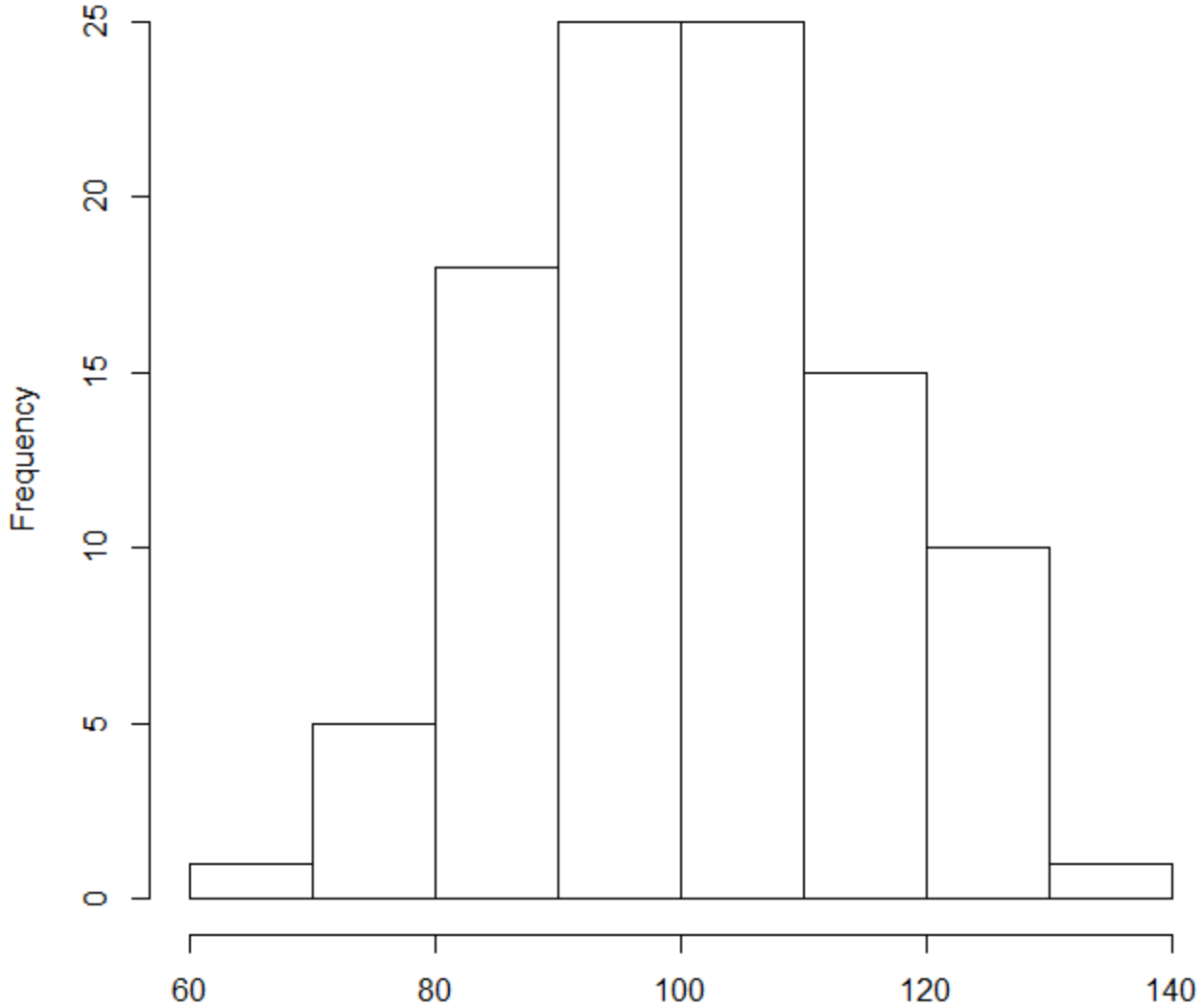

**Sample**

$\hat{y} = a + bx$



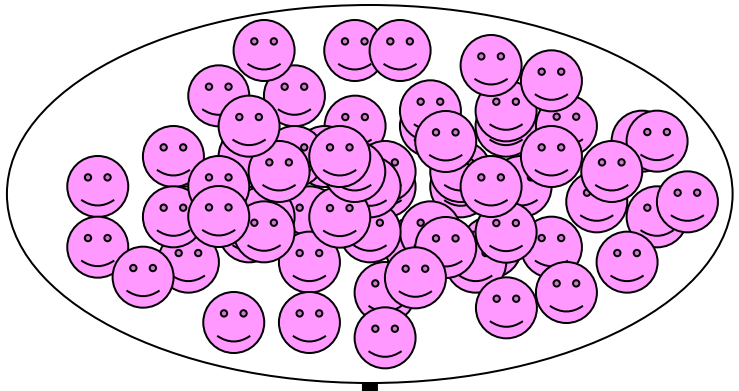


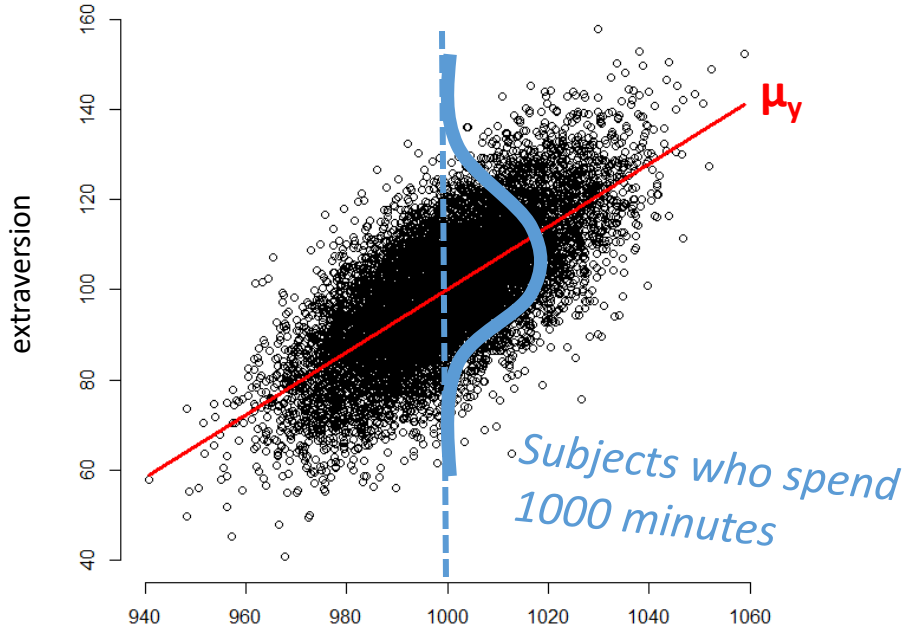
Extraversion histogram for subjects who spend
1000 minutes per week with friends



**Population**

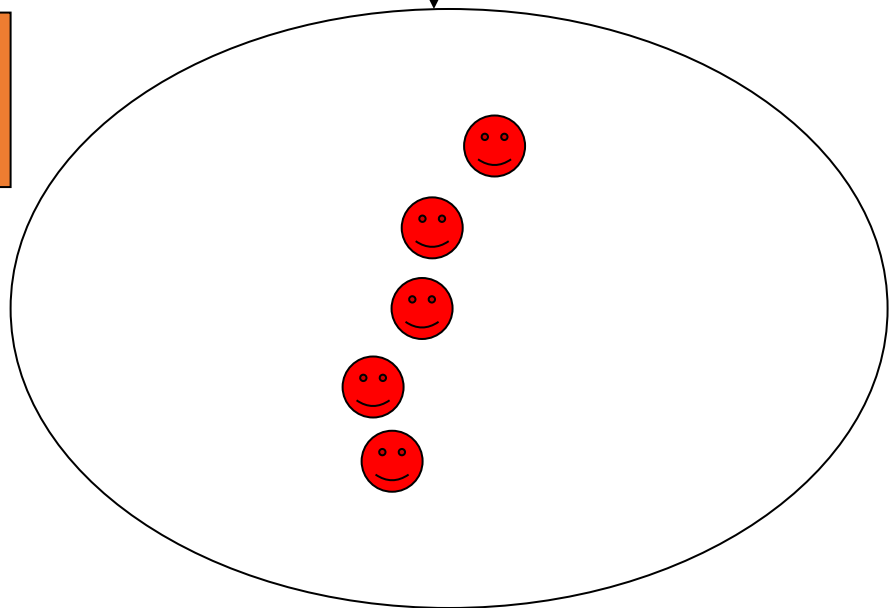
$\mu_y = \alpha + \beta x$

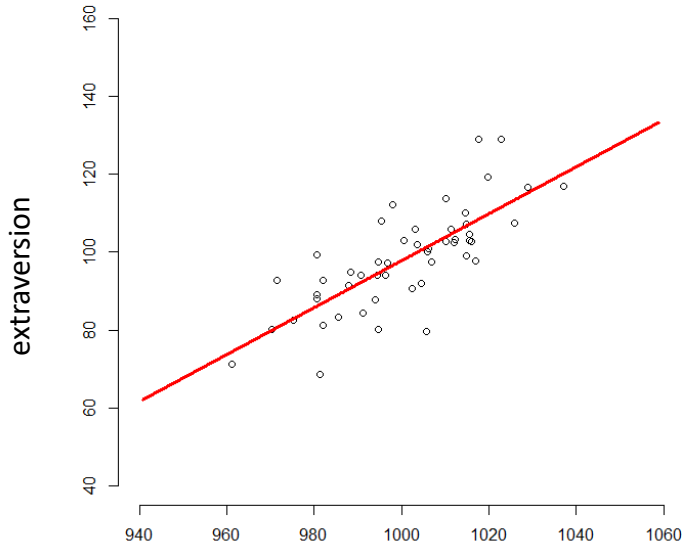





**Sample**

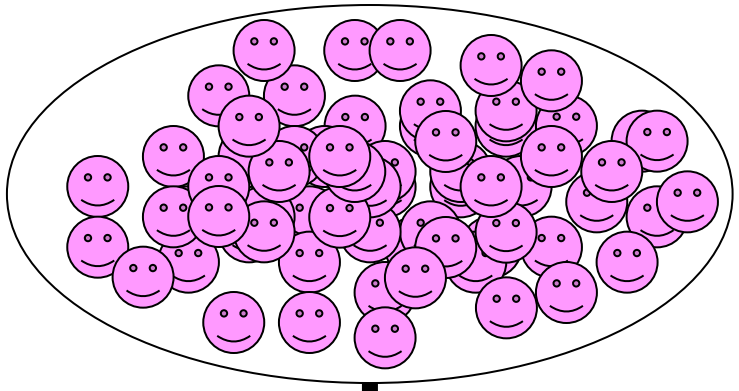
$\hat{y} = a + bx$

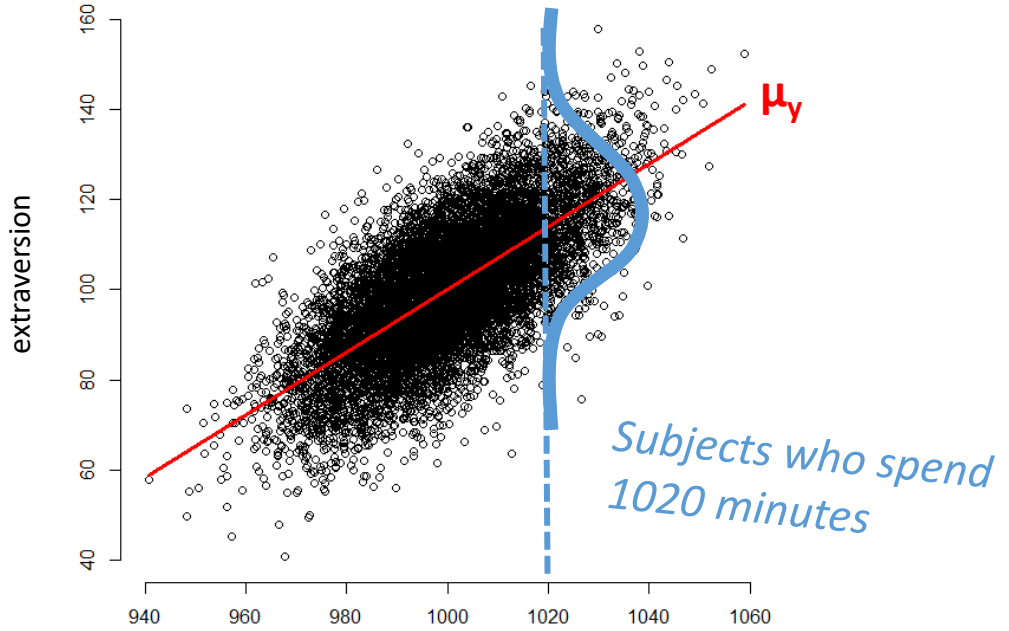




**Population**

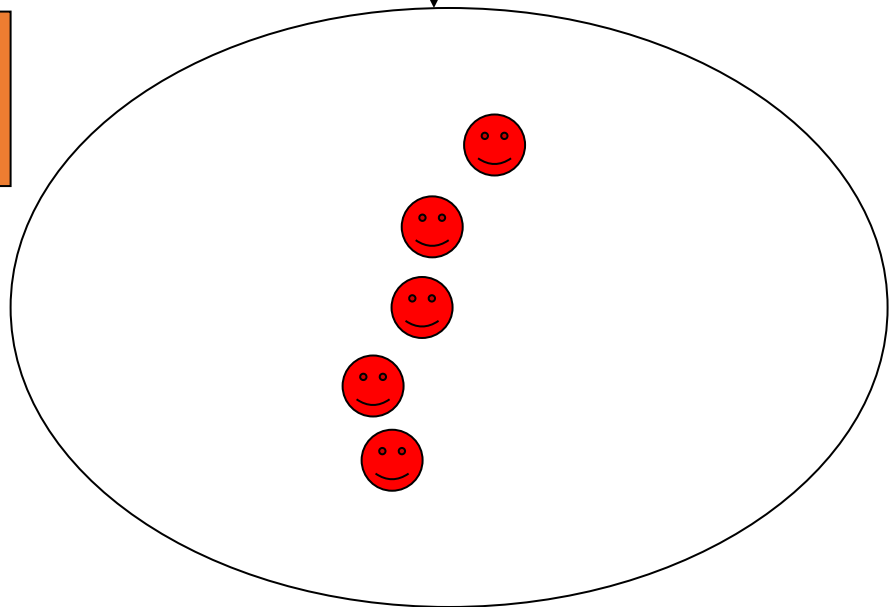
$\mu_y = \alpha + \beta x$

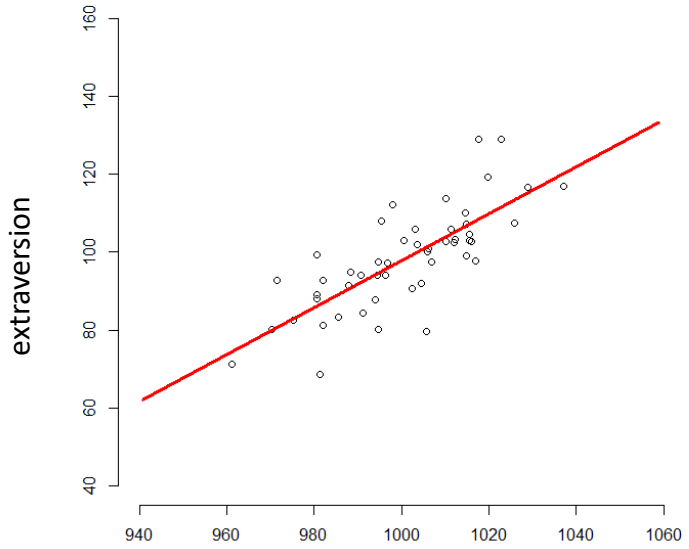




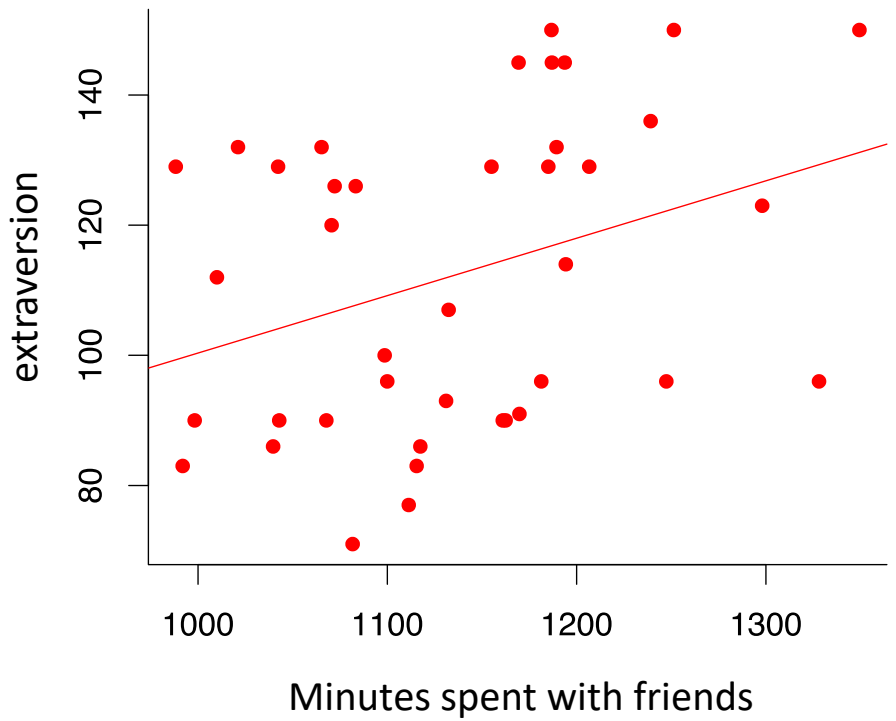

**Sample**

$\hat{y} = a + bx$





# Hypothesis test for the slope




- $\hat{y} = 12.15 + 0.08821x$
- b = 0.08821 → this sample
- What is $\beta$ in $\mu_y = \alpha + \beta x$? → population?

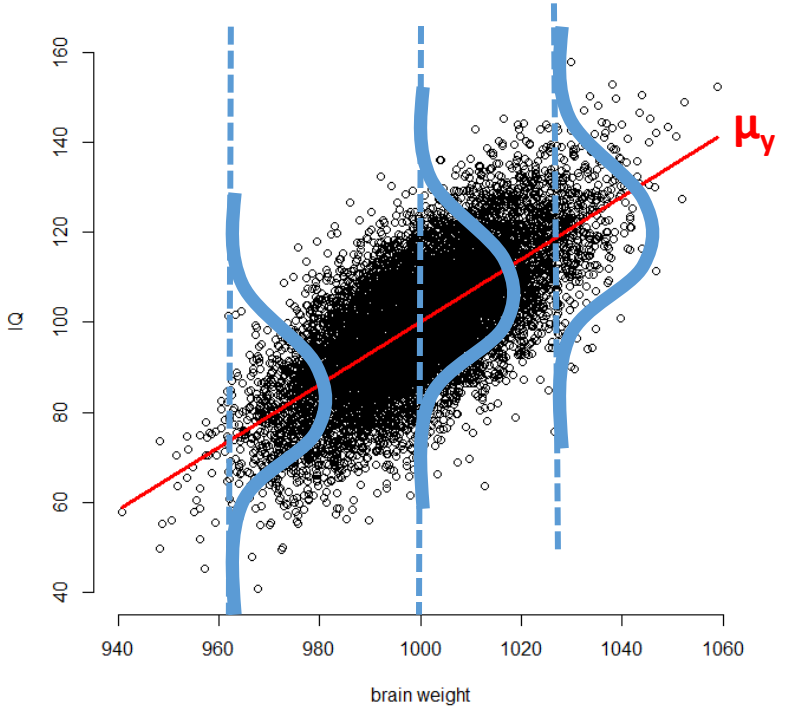
- We can for example test

$H_0$: $\beta = 0$ vs

$H_A$: $\beta \neq 0$

# 1. Assumptions

- Random sample
- x and y are ***linearly*** related (population satisfies $\mu_y = \alpha + \beta x$)
- For every value of x, y is normally distributed, with the ***same standard deviation***



# 2. Hypothesis

- $H_0$: $\beta = 0$
- $H_A$: $\beta \neq 0$

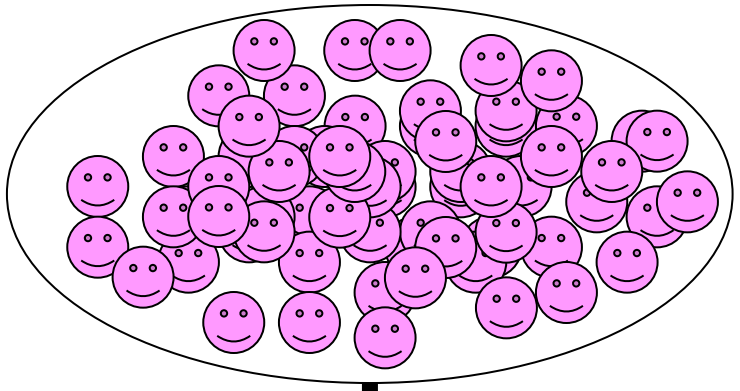
- Note: you can also test one-sided ($H_A$: $\beta > 0$ or $H_A$: $\beta < 0$ )

Thus: what is the probability that the statistic *b* takes the value we found in the sample (*b* = 0.08821) or more extreme given that $H_0$ is true?

- → What does the sampling distribution of *b* look like?

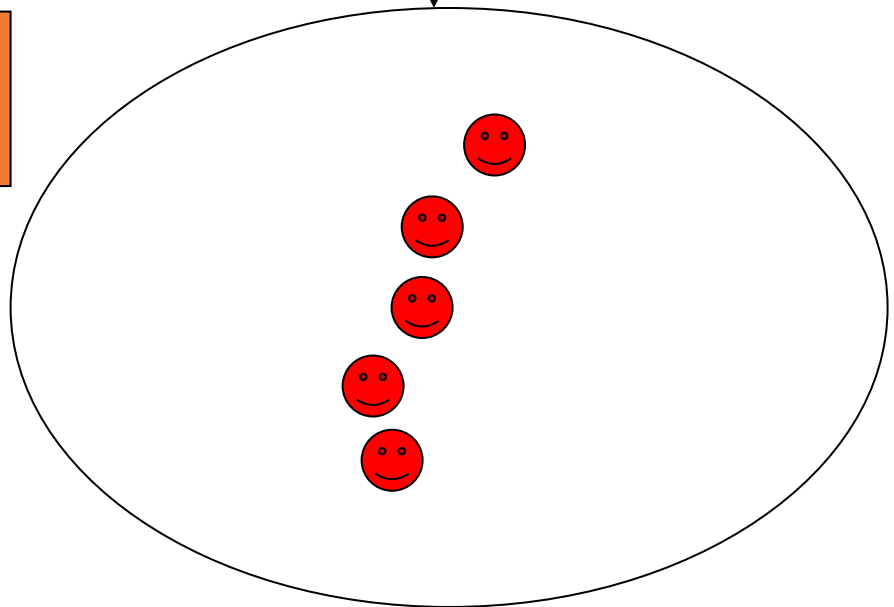
Population

$\mu_y = \alpha + \beta x$



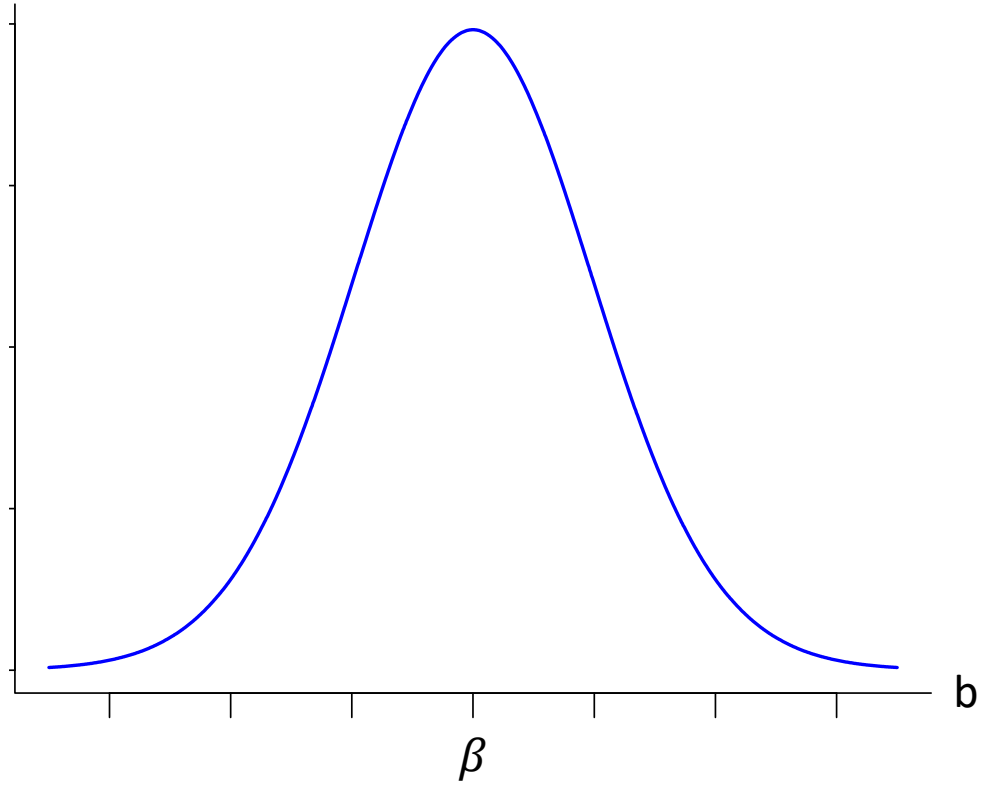
Sample

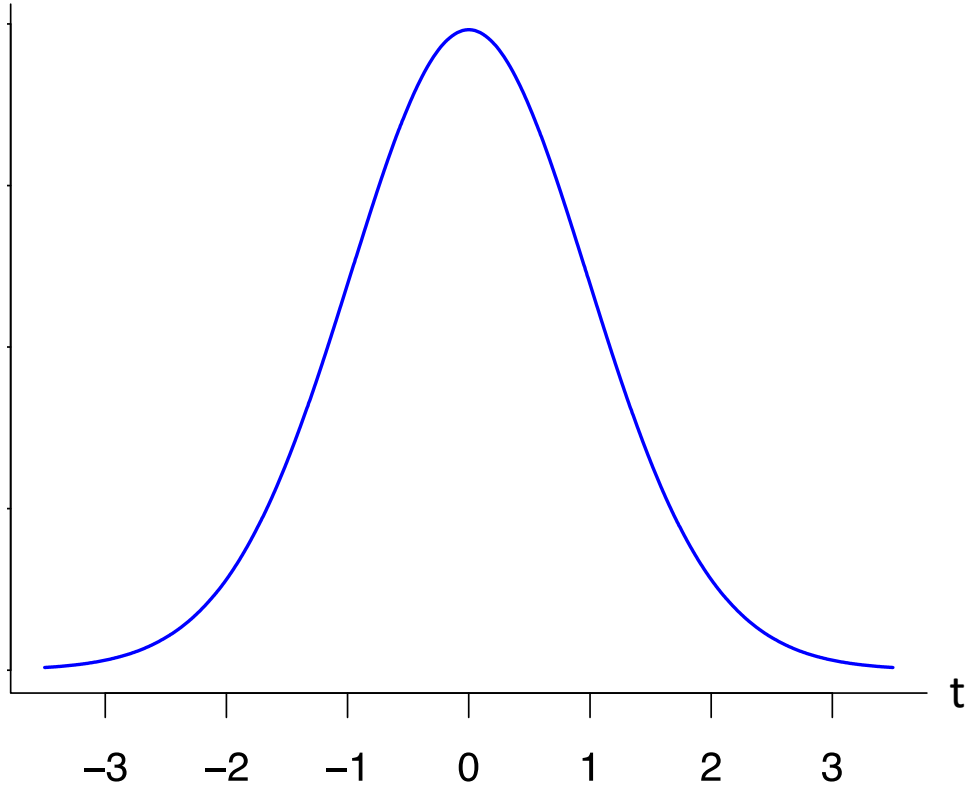
$\hat{y} = a + bx$



Sampling distribution of b

$$t = \frac{b - 0}{se} \qquad \text{df=n-2}$$




The standard error (which is the estimate of the SD of the sampling distribution) is something you don't need to calculate, it will be given

# 3. Test statistic

$t = \frac{b-0}{se}$ (the zero is because $H_0$: $\beta = 0$)

n - 2 degrees of freedom

$b$ = 0.0882 (calculated using the formula for slope)

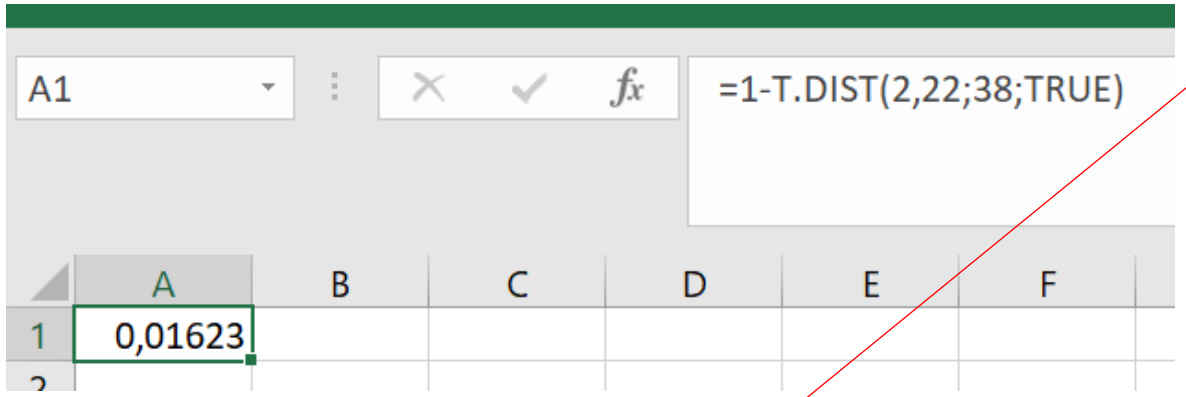
se = 0.03991 (given)

$$t = \frac{0.0882-0}{0.03991} = 2.22$$

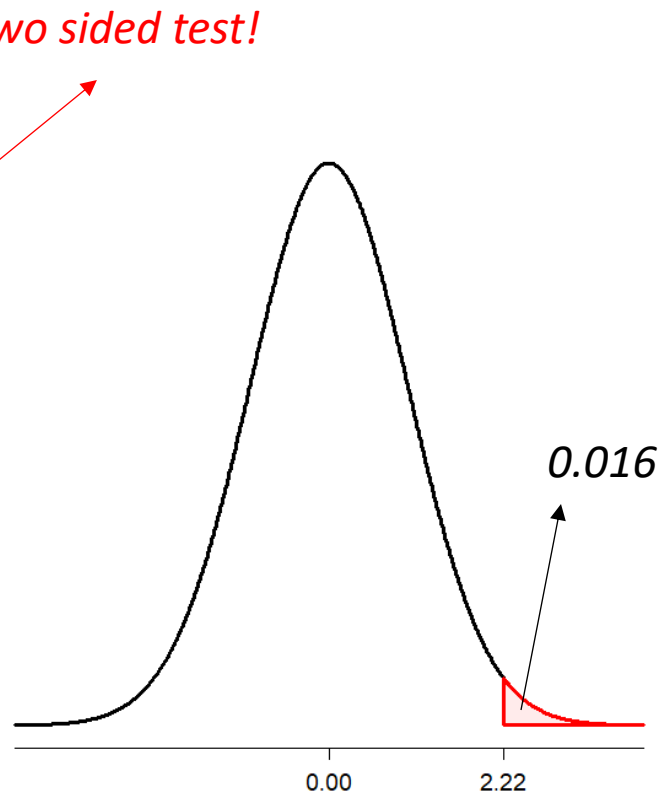
$$df = 40 - 2 = 38$$

# 4. P-value

- $t = 2.22$,
- $\mathrm{df} = \mathrm{n} - 2 = 40 - 2 = 38$
- From Excel:




- $p$ = 2 x 0.016= 0.032



# 5. Conclusion

- If we choose a level of significance $\alpha = 0.05$
- Because $p < .05$, we reject $H_0: \beta = 0$
- Thus the slope is significantly different from 0
- Time spent with friends predicts extraversion level

# Confidence Interval for $\beta$

- Instead of a hypothesis test, we can also compute a CI for $\beta$

- 95% CI: $b \pm t_{.025}(se)$
  - $b$ = 0.0882 (calculate using the formula for slope)
  - se = 0.03991 (given)

- df = n - 2 = 38
- $t_{.025} = 2.02$
  - Excel: `=T.INV(0,975;38)` → 2.02 (or `=T.INV(0,025;38)` → - 2.02)
- $b \pm t_{.025}(se) = 0.0882 \pm 2.02 \times 0.03991$ → 95%CI: (0.0076, 0.1688)

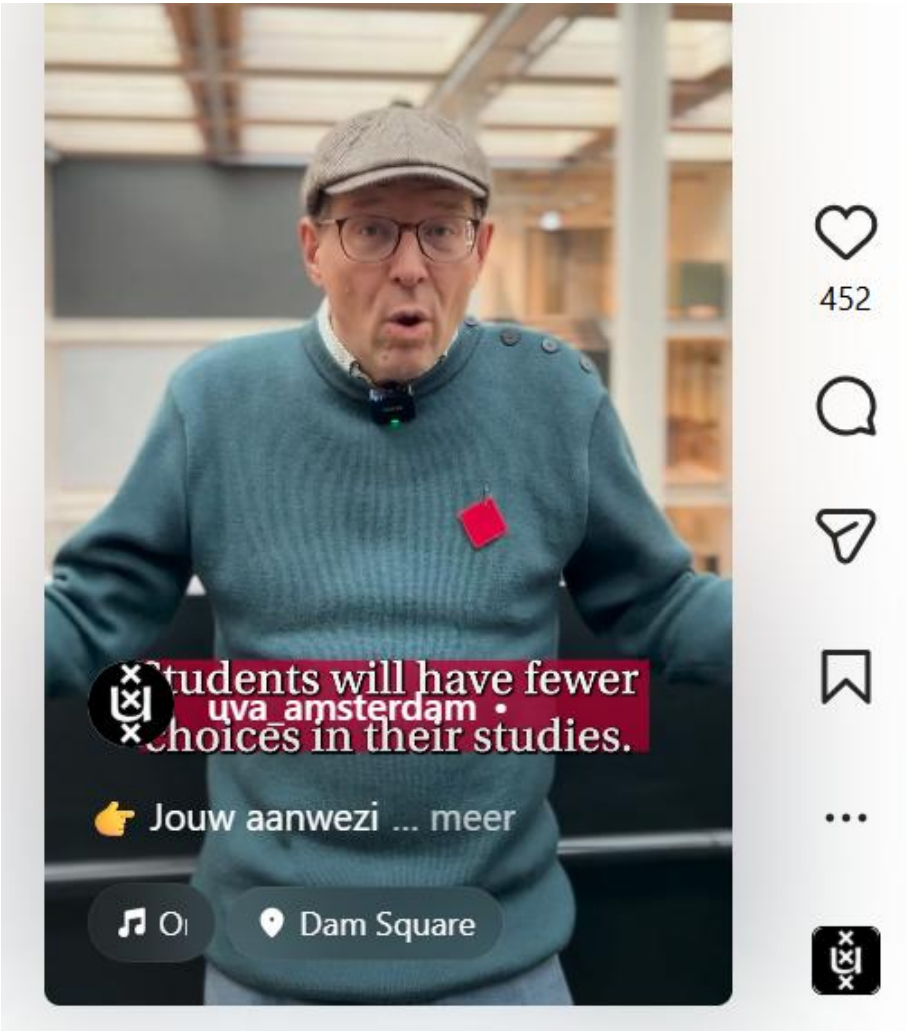
Because CI only includes values larger than 0, we can conclude that 0 is not a believable value: the slope is significantly larger than 0. Same conclusion as two-sided test: time spent with friends predicts extraversion level

# Summary

- A linear regression line is a straight line that predicts the value of a response variable *y* from the value of an explanatory variable *x*
- The correlation describes the strength of the relation and is a standardized slope
- The squared correlation $r^2$ measures how much better the regression line predicts *y* than the mean $\bar{y}$ predicts *y*. $r^2$ is the proportional reduction in error. It also has other interpretations such as “variance in y explained by x”
- Using a significance test, one can test whether the slope in the population differs from 0. Similarly, one can calculate a CI to calculate what the believable values are for $\beta$ (the slope in the population).

# Don’t forget:
# Next week, 9 dec: protest 12AM at Dam square

- No lecture!
- UvA professor Rens Bod (of WOinActie) explains why we should go to Dam square:
- https://www.instagram.com/reels/DRcAvEpjOlh/



# Example Exam Question

- Assume that the height of Dutch males is on average 184 cm, with a standard deviation of 7cm. Furthermore, assume that the correlation between the height of fathers and their adult sons is 0.4.
- Now consider Ben, the son of a father who is 186 cm tall. What is the predicted height of Ben?

a) 184
b) 186
c) 197

# Answer

- We need to estimate $\hat{y} = a + bx$
  - x: height of fathers
  - y: height of adult sons
  - Both are distributed equally
- Least squares method using the formula's:
  - $b = r\frac{s_y}{s_x}$
  - $a = \bar{y} - b\bar{x}$

- $b = r\frac{s_y}{s_x} = 0.4\frac{7}{7} = 0.4$
- $a = \bar{y} - b\bar{x} = 184 - 0.4 * 184 = 110$
- $\hat{y} = a + bx \rightarrow \hat{y} = 110 + 0.4x$
- $\hat{y} = 110 + 0.4 \times 186 \approx 184$

# Example Exam Question

- Assume that the height of Dutch males is on average 184 cm, with a standard deviation of 7cm. Furthermore, assume that the correlation between the height of fathers and their adult sons is 0.4.
- Now consider Ben, the son of a father who is 186 cm tall. What is the predicted height of Ben?

a) **184**
b) 186
c) 197

# Example exam question

Jennifer has a group of students do a difficult test and measures how long it takes students to finish the test. She wants to use the time it takes students to finish to predict their test score(TS), and constructs the following regression line:

$\widehat{TS} = a + b * time$

As a unit for time she can use a minute (i.e., time in minutes) or an hour (i.e., time in hours). Which of the following statements is true?

a) If time is on a scale of minutes, the slope will be *larger* than when the time is on a scale of hours.

b) If time is on a scale of minutes, the slope will be *smaller* than when the time is on a scale of hours.

c) The slope will not be affected by whether she chooses a minute or an hour as the unit of time.

# Example exam question

Jennifer has a group of students do a difficult test and measures how long it takes students to finish the test. She wants to use the time it takes students to finish to predict their test score(TS), and constructs the following regression line:

$$\widehat{TS} = a + b * time$$

As a unit for time she can use a minute (i.e., time in minutes) or an hour (i.e., time in hours). Which of the following statements is true?

a) If time is on a scale of minutes, the slope will be *larger* than when the time is on a scale of hours.

**b) If time is on a scale of minutes, the slope will be *smaller* than when the time is on a scale of hours.**

c) The slope will not be affected by whether she chooses a minute or an hour as the unit of time.

If she measures time in minutes the SD of time ($s_x$) is larger and so b will be smaller.

# Excercise from the book

**12.4 Higher income with experience** Suppose the regression line $\mu_y = -10{,}000 + 9500x$ models the relationship for the population of working adults in a country between $x$ = experience (in years) and the mean of $y$ = annual income (in euros). The conditional distribution of $y$ at each value of $x$ is modeled as normal with $\sigma = 6500$. Use this regression model to describe the mean and the variability around the mean for the conditional distribution at an experience of (a) 5 years and (b) 10 years.

# Solution to excercise 12.4

- The distribution of the annual incomes of working adults with experience of 5 years is normally distributed with mean = -10.000 + (9500*5) = 37500 and a standard deviation of 6500 (or variance of 6500^2 = 42250000).
- The distribution of the annual incomes of working adults with experience of 10 years is normally distributed with mean = -10.000 + (9500*10) = 85000 and a standard deviation of 6500 (or variance of 6500^2 = 42250000).

# Some other relevant excercises in the book:

- 3.25, 3.26, 3.27, 3.35, 3.40, 3.44, 3.49, 3.53, 12.7, 12.8, 12.11, 12.12, 12.16, 12.17, 12.18, 12.27, 12.35, 12.37
- Note: excercise 3.26 in Agresti 5th edition, is excercise 3.25 in an older edition of the book, but in this older edition it has a mistake: The question should read: "The regression equation is $\hat{y} = 6.71 - 0.024x$. Find the ..."

- If you want any help with excercises from the book: don't hesitate to ask questions on the discussion board!


We present to you this year's Christmas party...
The Naughty List
Dec. 3rd 2025
21:00 - 3:00 @Oliva
7€-members 9€-non-members
Tickets:
SES
AIM PPLE
STUDY ASSOCIATION
KALLIOPPE
VSPA
MACHIAVELLI