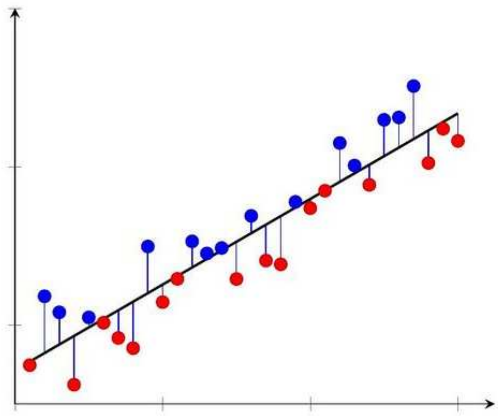
# Examining associations (Regression) ‘taking measurements on people’

Dr Hakim-Moulay Dehbi
June, 2017



## What can be done with a linear regression

Questions

- Is blood pressure related to age ?
- Is lung function related to height ?
- Is inflammation related to air pollution ?
- Is atherosclerosis related to noise exposure at night ?

- An approach consists of using **linear regression**
  - “If we increase the explanatory variable by one unit, by how much does the outcome change ?”

# Typical form of a linear regression

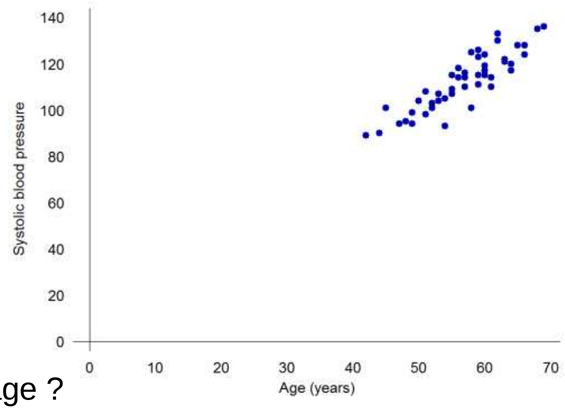
**Y = a + b * X + error**

- Y is the outcome ; X is the explanatory variable
- Error (a.k.a. residual): the difference between the value estimated by the regression and the true value
- Assumptions:
  - All observations must be independent to each other (e.g. different people, samples, animals etc.)
  - The errors (residuals=observed minus estimated values) are normally distributed

*If these assumptions are not satisfied, linear regression may not be appropriate*

# Example: systolic blood pressure and age

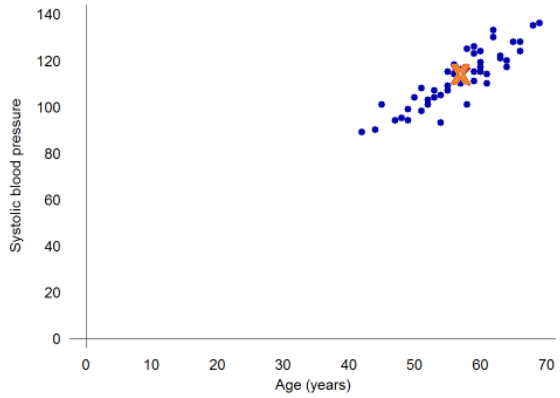
- 50 subjects
- Data on age in years and systolic blood pressure (SBP [mmHg]) at entry
- Age is normally distributed
- Mean age (range): 57 years (42 to 69)
- Mean SBP (range): 113 mmHg (89 to 136)
- Is there an association between SBP and age ?



## Drawing a straight line between the points

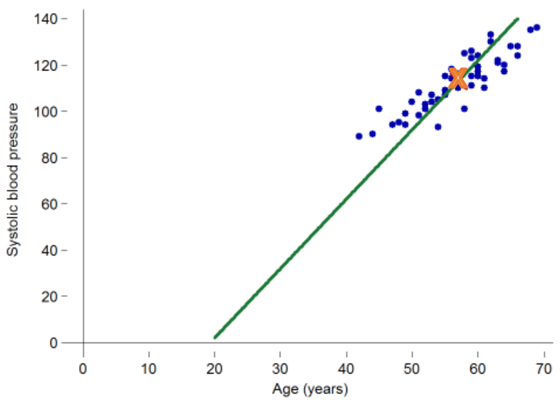
- How can we draw a straight line between these points
- Where do we centre the line ?
- Most appropriate choice:
  - at the mean value of BP
  - at the mean value of Age

X

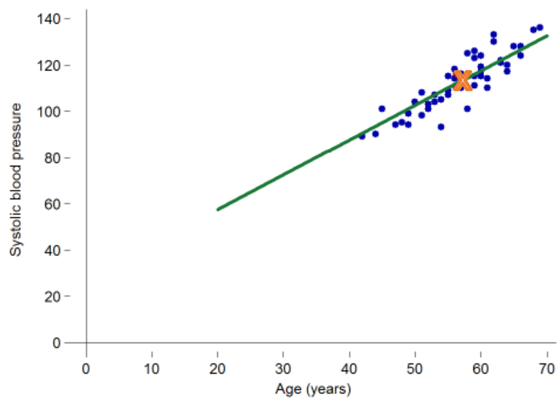



## Many possible lines

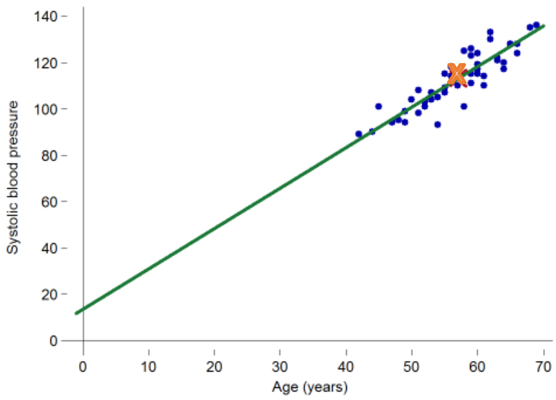
This is one possibility for the line :




This is another possibility :



# The best fit

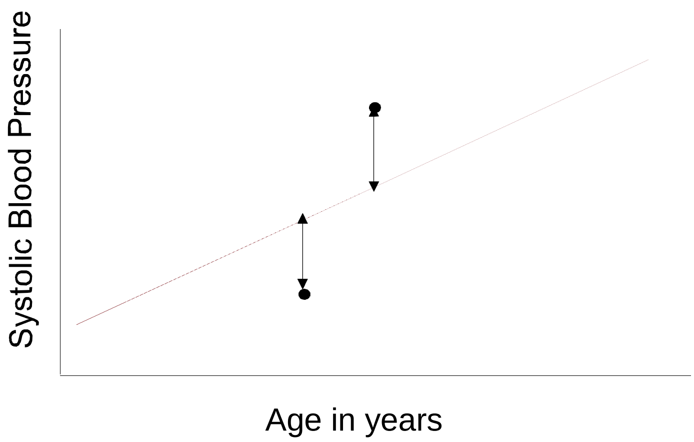



- Remember: Y = a + b * X
  - a is where the green line crosses the y-axis - about 13
  - b is the slope of the line - about 1.7

The best fit always go through the mean value of X and Y

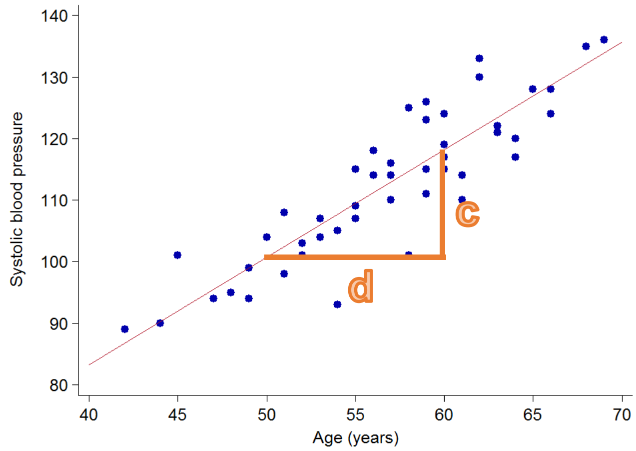
# How to find the best fit

By minimising the square of the errors !




- The red line is the line that fits the points best
- The sum of the errors will always be 0 if the line goes through the means of Y and X
- The best line is the one for which the sum of all the squared differences is the smallest out of all possible lines
  - Almost all points are away from the line of best fit

# How to work out the slope




- b = c / d = slope (regression coefficient)
  - At age = 50, BP ~ 102 mmHg
  - At age = 60, BP ~ 119 mmHg
- hence c=17, d=10 => when x increases by 10, y increases by 17
- slope: b ~ 17 / 10 = 1.7 mmHg/year

# Interpretation

**SBP = 13.3 + 1.7 * age**

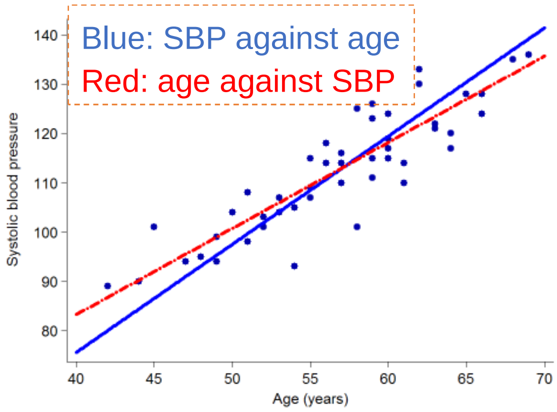
- for a 1 year increase in age , blood pressure increases, on average, by 1.7 mmHg

OR, because it is a *linear* association:

- for a 10 year increase in age , blood pressure increases, on average, by 17 mmHg

=> Older people have a higher blood pressure than younger people, on average

# Regressing Y against X is *not the same* as X against Y




- The question asked is different
  - Y against X (SBP against age)
    - "What is the change in SBP with an increase in age of 1 year"
  - X against Y (age against SBP)
    - "What is the change in age with an increase of 1 mmHg in SBP"

# Confidence interval (CI) for the slope

**What is the uncertainty in the estimate of the slope ?**

- Based on our sample: for a 1 year increase in age, SBP increases by 1.7 mmHg

What would be the result in another sample ?

- We use the Standard Error (SE) to construct a 95% CI around the slope, using

$$CI = b \pm 1.96 * SE$$

Example from regression of SBP on age:
**b** = 1.75 **SE of b** = 0.13 **(1.96 x SE of b)**= 0.25
95% CI = **b ± 0.25** = 1.5 to 2.0 mmHg per 1 year increase in age

# Quick quiz

**What will the line look like if there is no association between the variables ?**

It will be a **horizontal line**. The slope coefficient will be **0**

**The 95% CI is from 1.5 to 2.0. What does this tell us about the p-value for 'b' ?**

The interval **does not contain 0** => the p-value will be **less than 0.05**

# SPSS output for the SBP vs. age example

**Parameter Estimates**

Dependent Variable: Systolic_BP

| Parameter | B | Std. Error | t | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Intercept | 13.300 | 7.331 | 1.814 | .076 | -1.440 | 28.041 |
| Age | 1.749 | .128 | 13.695 | .000 | 1.492 | 2.005 |

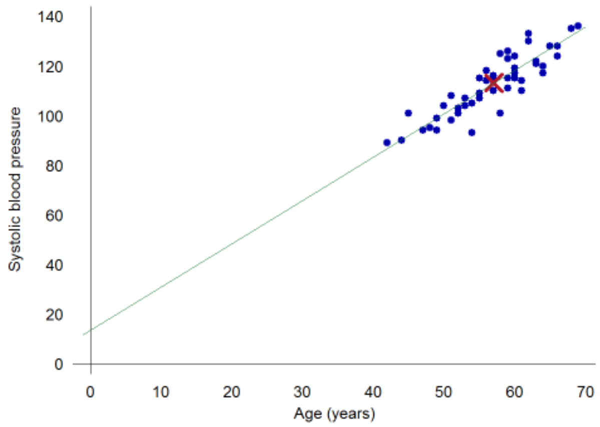
a

b

95% CI

# Correlation coefficient

$r^2$, is the correlation coefficient squared. It can be interpreted as the amount of the variability in the y-variable that is explained by the x-variable.




- Correlation of SBP and age:

r = 0.89

- $r^2 = 0.89 * 0.89 = 0.80$

= > **age explains 80% of the variation in SBP**

But remember, correlation does not imply causation (!)

# Interpretation – recap

**SBP = 13.3 + 1.7 * Age – 95% CI for "b": 1.5-2.0 – P-value < 0.001**

- Slope "b" = 1.7 mm Hg/year

On average, for an increase in age by 1 year, blood pressure increases by 1.7 mmHg

- 95% CI: 1.5 to 2

If we were to repeat this analysis on 100 samples of 50 people, 95 of the 100 Cis would include the true value. Our CI is one of them. And it may or may not include the true value

- P-value: <0.001

The 95% CI does not contain the 'no effect value' (0). The p-value reflects this.

# Predictions: using the estimated equation

**SBP = 13.3 + 1.7 * Age**

The regression equation can be used to make predictions about the value of the dependent variable based upon a subject's explanatory variable(s).

- What is the predicted SBP of a patient who is 53 years old ?

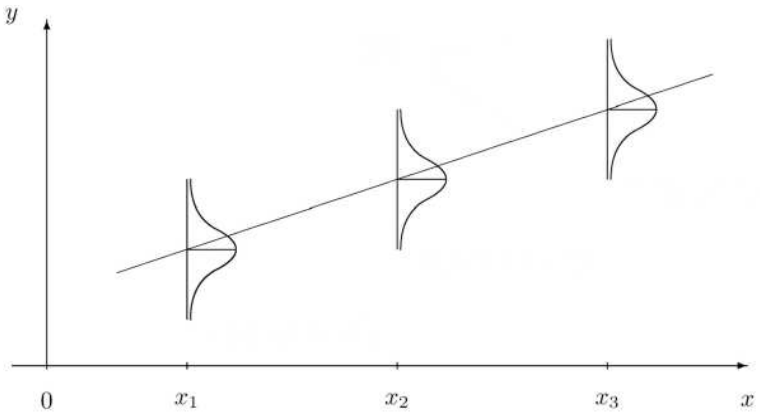
13.3 + 1.7x53 = 103.4 mmHg

- This is only an estimate of the average SBP for a 53-year-old. There is variation within people with the same explanatory variables.
  - Statistical software can give prediction intervals (a CI for the prediction)
- Do not use this equation to predict outside the range of values used to create it (!)
  - E.g., it is not valid to use this to predict the SBP for a six-year-old child

# Model checking: normality of residuals


- Residuals are the difference between each observed point, and the value predicted from the model
- A key assumption of linear regression:
  - residuals are normally distributed
    - with mean of 0
    - with the same variance at all levels of the explanatory variable(s)
- In our example, there is someone aged 42 who has SBP of 89 mmHg. The model predicts a SBS of Y=13.3 + (1.75 x 42) = 86.8 => the residual = 89 – 86.8 = 2.2

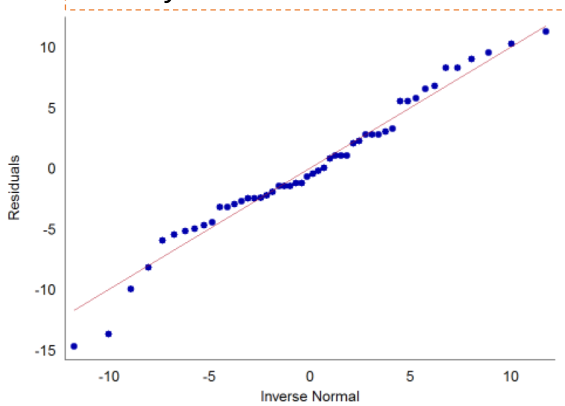
# Normality of residuals – visualizing it

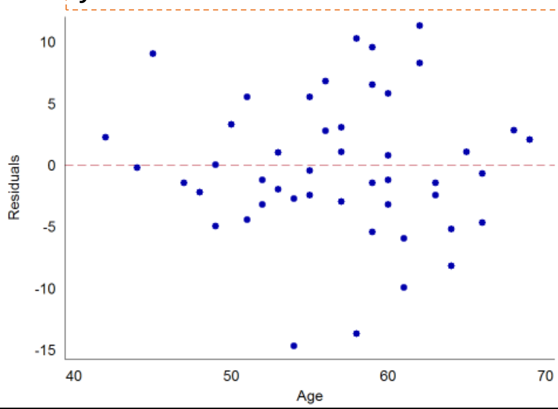



# Normality of residuals – examining plots

The quantile-quantile plot should look like a straight line along the identity line

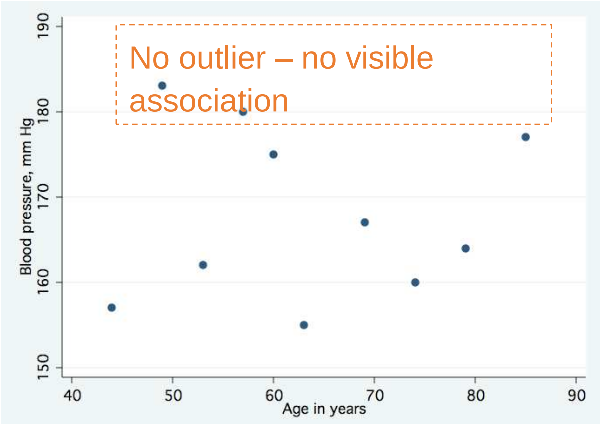
The scatter plot should look like a random scatter around the line y = 0

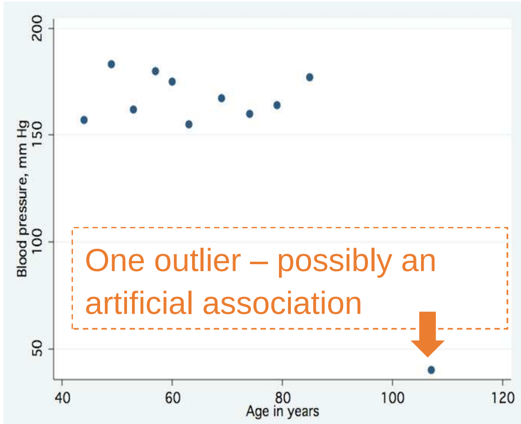


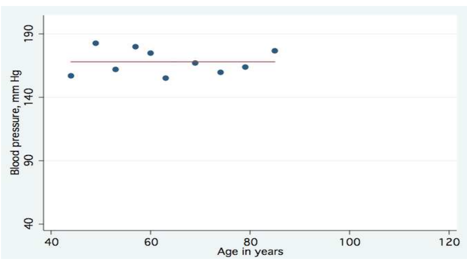
# Outliers

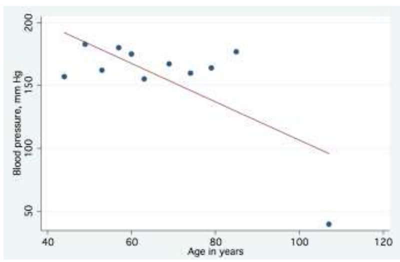
**An outlier is a real data value that happens to be much smaller or larger than expected**




# Outliers can influence the results markedly




| bp | Coef. | Std. Err. | P-value | [95% CI] | |
|---|---|---|---|---|---|
| age | .0037545 | .2662023 | 0.989 | -.610109 | .617618 |
| _cons | 167.7623 | 17.18335 | 0.000 | 128.1375 | 207.3872 |

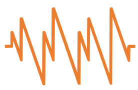
| | Coef. | SE | P-value | [95% CI] |
|---|---|---|---|---|
| age | -1.5 | 0.52 | 0.017 | -2.70 to -0.35 |
| cons | 258.84 | 36.02 | <0.001 | 177.36 to 340.33 |

No evidence of an association

A negative association

# How to deal with outliers

- First check whether the data value is correct
  - If unsure and the value is likely to be implausible, then ignore in the analysis
  - If the value is correct, we need to consider carefully how to deal with it
- Including outliers probably won't affect your conclusions, if there are very few compared to the overall number of observations
- You can run your analysis with / without outliers to check this



# Other kinds of explanatory variables

- In a linear regression
  - the outcome is always continuous
  - the explanatory variable(s) does not have to be continuous; it can be categorical => a linear regression can be used to compare measurements between two or more groups
- Categorical variable with 2 levels => results identical to t-test
  - *But*, a regression can include other factors at the same time, which the t-test cannot

# Example: regression of SBP in men vs. women

**Parameter Estimates**

Dependent Variable: Systolic_BP

| Parameter | B | Std. Error | t | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Intercept | 119.880 | 2.137 | 56.100 | .000 | 115.583 | 124.177 |
| [Gender=0] | -13.600 | 3.022 | -4.500 | .000 | -19.676 | -7.524 |
| [Gender=1] | 0[a] | . | . | . | . | . |

a. This parameter is set to zero because it is redundant.

- There is strong evidence (p<0.001) that gender is associated with SBP
- The slope of -13.6 is the difference between the means of the two groups
  - the 95% CI is -19.7 to -7.5
- Gender = 1 is the baseline in this case

  => the intercept 119.9 is the mean SBP in the baseline group

  => 119.8 – 13.6 = 106.3 is the mean SBP in the other group

# Example: t-test of SBP in men vs. women

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| Systolic_BP | Equal variances assumed | .550 | .462 | -4.500 | 48 | .000 | -13.600 | 3.022 | -19.676 | -7.524 |
| | Equal variances not assumed | | | -4.500 | 46.984 | .000 | -13.600 | 3.022 | -19.680 | -7.520 |

- The mean difference and the 95% CI are the same as with a linear regression
- A linear regression with two groups is identical to a t-test
  - the advantage of regression: one can adjust for additional factors

# Example: regression of SBP vs. social class

**Tests of Between-Subjects Effects**

Dependent Variable: Systolic_BP

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 48.513[a] | 3 | 16.171 | .096 | .962 |
| Intercept | 638767.102 | 1 | 638767.102 | 3794.738 | .000 |
| SE_class | 48.513 | 3 | 16.171 | .096 | .962 |
| Error | 7743.167 | 46 | 168.330 | | |
| Total | 647146.000 | 50 | | | |
| Corrected Total | 7791.680 | 49 | | | |

a. R Squared = .006 (Adjusted R Squared = -.059)

**Parameter Estimates**

Dependent Variable: Systolic_BP

| Parameter | B | Std. Error | t | Sig. | 95% Confidence Interval Lower Bound | 95% Confidence Interval Upper Bound |
|---|---|---|---|---|---|---|
| Intercept | 111.923 | 3.598 | 31.104 | .000 | 104.680 | 119.166 |
| [SE_class=High] | .462 | 5.089 | .091 | .928 | -9.782 | 10.705 |
| [SE_class=Lower middle | 1.994 | 5.194 | .384 | .703 | -8.461 | 12.448 |
| [SE_class=Upper middle | 2.327 | 5.194 | .448 | .656 | -8.128 | 12.782 |
| [SE_class=Low] | 0[a] | . | . | . | . | . |

a. This parameter is set to zero because it is redundant.

- Social class: 4 levels
- P-value = 0.96
- Intercept = 111.9 is the SBP in the baseline group (SES = low)
- Parameters: difference between mean SBP in low SES and other categories

# Example: regression of SBP vs. social class

**Parameter Estimates**

Dependent Variable: Systolic_BP

| Parameter | B | Std. Error | t | Sig. | 95% Confidence Interval Lower Bound | 95% Confidence Interval Upper Bound |
|---|---|---|---|---|---|---|
| Intercept | 99.062 | 1.729 | 57.300 | .000 | 95.585 | 102.540 |
| [Age_group=>65] | 26.471 | 2.485 | 10.651 | .000 | 21.471 | 31.471 |
| [Age_group=60-65] | 15.990 | 2.346 | 6.815 | .000 | 11.270 | 20.711 |
| [Age_group=<60] | 0[a] | . | . | . | . | . |

- Age is a continuous variable, but it can be categorized
  - Group 1 (baseline): under 60 ; mean SBP = 99.06
  - Group 2: 60 – 65 ; mean SBP = 115.05
  - Group 3: > 65 ; mean SBP = 125.53
- The intercept is the mean SBP in the baseline group
- The coefficients are the differences in mean SBP between the categories

## Categorizing a continuous variable

Advantages
- Easier to report and present graphically
- May be more useful / clinically relevant
- Useful if there is a non-linear trend

Disadvantages
- Some information is lost
- Cases close to the cutoffs may be misclassified
- The cutoffs can be arbitrary

# Thank you for your attention

# Facebook experiment - questions

1. List 2 things that you like and 2 things that you don't like that about each of the following sections:
   - Introduction
   - Methods
   - Results
2. What are the main results ? Are there any additional strength(s) and/or limitation(s) that should be added to the Discussion
3. Assess the overall quality of the layout, format and presentation, and give it a score from 1 (very poor) to 5 (excellent)

# Facebook experiment – Cohen's D

- Cohen's D is a statistic used to report the standardised difference between two means
- It is the difference between the means (M2 – M1) divided by the pooled standard deviation (pooled SD) : (M2 – M1) / pooled SD
- In general, it is interpreted with cut-offs:
  - < 0.2 : small difference
  - 0.2 – 0.8 : medium difference
  - > 0.8 : large difference
- Excellent visual representation on: http://rpsychologist.com/d3/cohend/