Linear regression

Grzegorz Chrupała

Tilburg University

Pearson's correlation coefficient

Pearson's correlation coefficient

- Magnitude of the covariance is not easy to interpret

Pearson's correlation coefficient

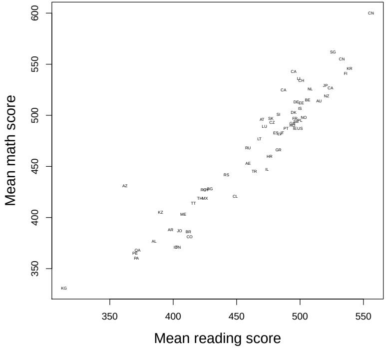
- Magnitude of the covariance is not easy to interpret
- Correlation coefficient, is **normalized** and corresponds to strength of the linear relation

Pearson's correlation coefficient

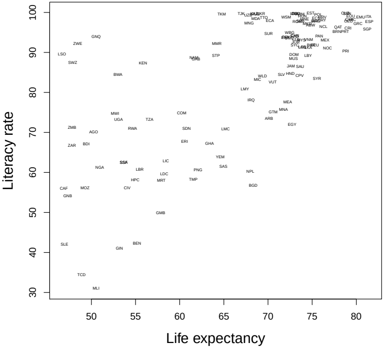
- Magnitude of the covariance is not easy to interpret
- Correlation coefficient, is **normalized** and corresponds to strength of the linear relation
- Divide variance by the product of the variables standard deviations

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

$$\rho = 0.95$$



$$\rho = 0.73$$



- As ice cream sales increase, the rate of drowning deaths increases sharply.
- Therefore, ice cream consumption causes drowning.

Correlation vs causation

Possible causal relationships between two events A and B measured by correlated random variables

Correlation vs causation

Possible causal relationships between two events A and B measured by correlated random variables

- A causes B

Correlation vs causation

Possible causal relationships between two events A and B measured by correlated random variables

- A causes B
- B causes A

Correlation vs causation

Possible causal relationships between two events A and B measured by correlated random variables

- A causes B
- B causes A
- C causes both A and B

Correlation vs causation

Possible causal relationships between two events A and B measured by correlated random variables

- A causes B
- B causes A
- C causes both A and B
- The correlation is a coincidence

Correlation vs causation

Possible causal relationships between two events A and B measured by correlated random variables

- A causes B
- B causes A
- C causes both A and B
- The correlation is a coincidence
- Some combination of the above

- Discovery of correlation can suggest a causal relationship

- Discovery of correlation can suggest a causal relationship
- But it can only be fully elucidated by an experimental study

- Discovery of correlation can suggest a causal relationship
- But it can only be fully elucidated by an experimental study
  - Vary a single variable while keeping all else equal

- Discovery of correlation can suggest a causal relationship
- But it can only be fully elucidated by an experimental study
    - Vary a single variable while keeping all else equal
    - Does the other variable co-vary?

Characterizing the relationship between two random variables

Regression analysis

Regression analysis

- Model relationships between variables

Regression analysis

- Model relationships between variables
- Specifically: model the dependent (output) variable as a function of the independent (input) variables

Regression analysis

- Model relationships between variables
- Specifically: model the dependent (output) variable as a function of the independent (input) variables
- Example:
  - **Describe** how people's weight depends on their height

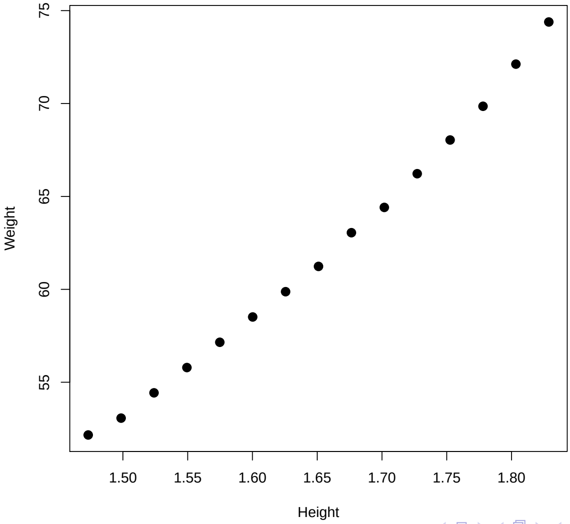
Regression analysis

- Model relationships between variables
- Specifically: model the dependent (output) variable as a function of the independent (input) variables
- Example:
  - **Describe** how people's weight depends on their height
  - **Predict** people's weight given their height

Sample data

	Height	Weight
1	1.47	52.2
2	1.50	53.1
3	1.52	54.4
4	1.55	55.8
5	1.57	57.2
6	1.60	58.5
7	1.63	59.9
8	1.65	61.2
9	1.68	63.0
10	1.70	64.4
11	1.73	66.2
12	1.75	68.0
13	1.78	69.9
14	1.80	72.1
15	1.83	74.4

Scatter plot



Model

- Single independent variable x
- Dependent variable y
- Model the relationship as a parametrized function $y = f(x)$:
  - $f(x) = ax^2 + bx + c$

Model

- Single independent variable x
- Dependent variable y
- Model the relationship as a parametrized function $y = f(x)$:
  - $f(x) = ax^2 + bx + c$
  - $f(x) = a\sin(x) + b$

Model

- Single independent variable x
- Dependent variable y
- Model the relationship as a parametrized function $y = f(x)$:
  - $f(x) = ax^2 + bx + c$
  - $f(x) = a\sin(x) + b$
  - $f(x) = ax + b$
- We focus on **linear** regression

Linear Regression

- Training data: observations paired with outcomes
- Observations are described by independent variables
- The model is a regression line $y = f(x) = ax + b$ which best fits the observations
  - a is the slope
  - b is the intercept (bias)
  - This model has two parameters (weigths, coefficients)
  - There is only one independent variable $= x$

Best fit

- Residual: difference between true value y and predicted value $f(x)$

Best fit

- Residual: difference between true value y and predicted value $f(x)$
- Find a line which minimizes Mean Squared Error:

$$MSE(f) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - f(x^{(i)}))^2$$

Best fit

- Residual: difference between true value y and predicted value $f(x)$
- Find a line which minimizes Mean Squared Error:

$$MSE(f) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - f(x^{(i)}))^2$$

- MSE: how much values of a variable deviate from regression line.

Best fit

- Residual: difference between true value y and predicted value $f(x)$
- Find a line which minimizes Mean Squared Error:

$$MSE(f) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - f(x^{(i)}))^2$$

- MSE: how much values of a variable deviate from regression line.
- Or how much of total variance is left unexplained by regression

R^2 - coefficient of determination

$$R^2 = \frac{Var[Y] - MSE(f)}{Var[Y]} = \frac{Var[f(X)]}{Var[Y]}$$

R^2 - coefficient of determination

$$R^2 = \frac{Var[Y] - MSE(f)}{Var[Y]} = \frac{Var[f(X)]}{Var[Y]}$$

Ratio of variance explained by f to total variance

The line of best fit can be calculated from the training data $X = x^{(1)} \dots x^{(n)}$ and $Y = y^{(1)} \dots y^{(n)}$

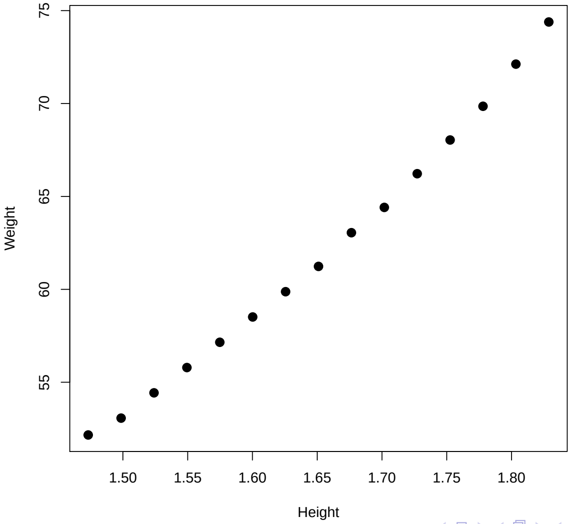
$$a = \frac{Cov(X,Y)}{Var(X)}$$
$$b = E[Y] - a$$

The line of best fit can be calculated from the training data $X = x^{(1)} \dots x^{(n)}$ and $Y = y^{(1)} \dots y^{(n)}$

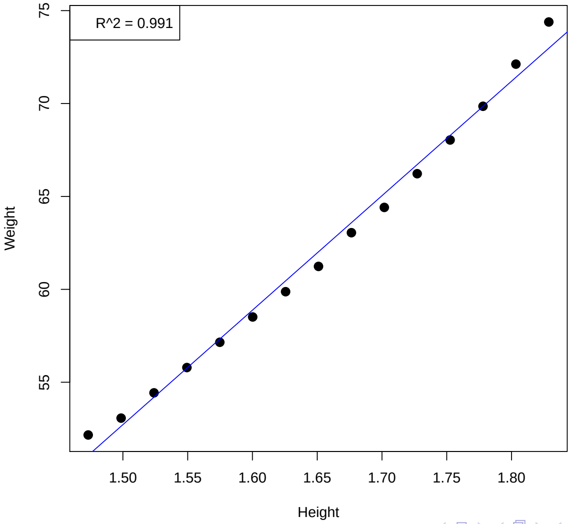
$$a = \frac{Cov(X, Y)}{Var(X)}$$
$$b = E[Y] - a$$

Use software to compute the estimates.

Scatter plot



Prediction of weight from height



```
   highsch score
1     TRUE    65
2     TRUE    98
3     TRUE    85
4     TRUE    83
5     TRUE   115
6    FALSE    98
7     TRUE    69
8     TRUE   106
9     TRUE   102
10    TRUE    95
11    TRUE    91
12    TRUE    58
13    TRUE    84
14    TRUE    78
15   FALSE   102
```

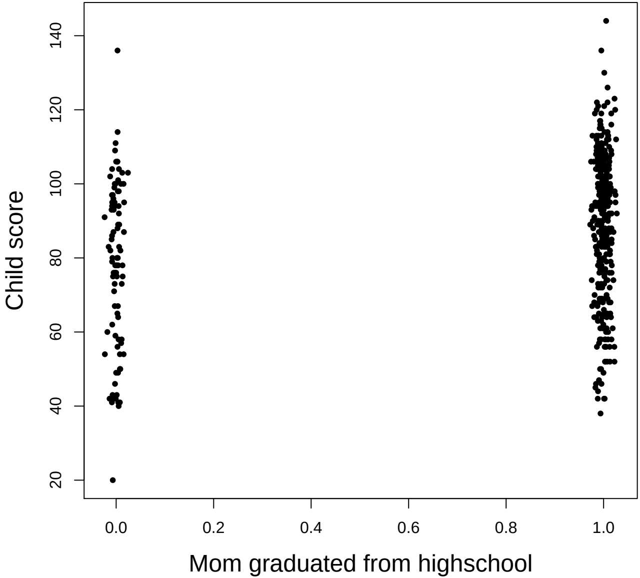
Binary predictor

Binary predictor

- Simplest categorical variable: binary value

Binary predictor

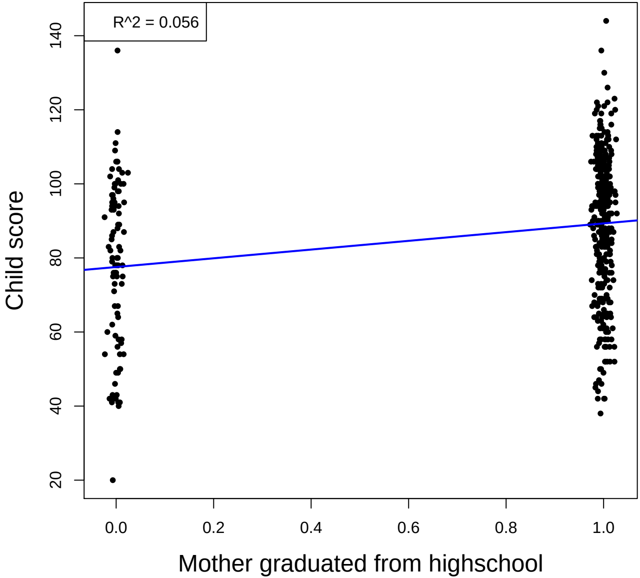
- Simplest categorical variable: binary value
- Code False $\rightarrow$ 0 and True $\rightarrow$ 1


Child score
20
40
60
80
100
120
140
0.0
0.2
0.4
0.6
0.8
1.0
Mom graduated from highschool

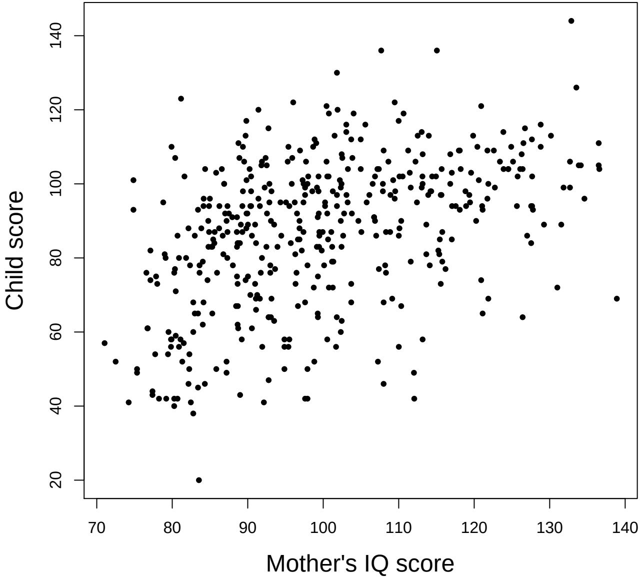
Equation

$$\text{child score} = a \times \text{highschool} + b$$
$$\text{child score} = 12 \times \text{highschool} + 78$$

- $12 \times 0 + 78$ - mean score of children whose mothers have **no highschool**
- $12 \times 1 + 78$ - mean score of children whose mothers **do have highschool**


R^2 = 0.056
Child score
Mother graduated from highschool
20
40
60
80
100
120
140
0.0
0.2
0.4
0.6
0.8
1.0

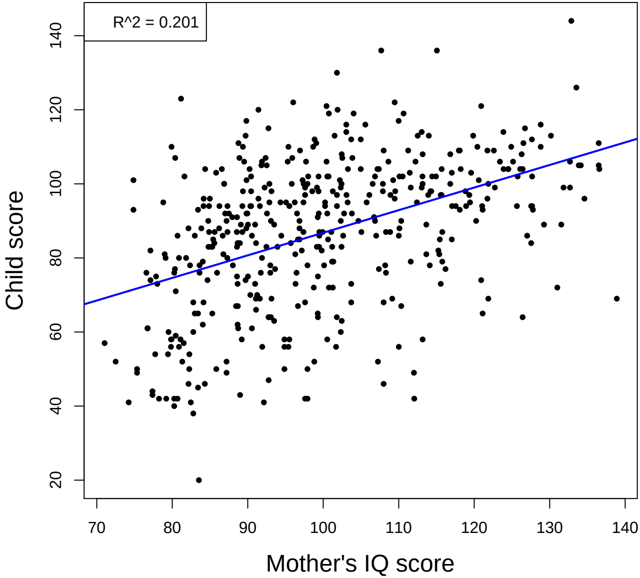
Child score as a function of mother's IQ


Child score
Mother's IQ score
20
40
60
80
100
120
140
70
80
90
100
110
120
130
140

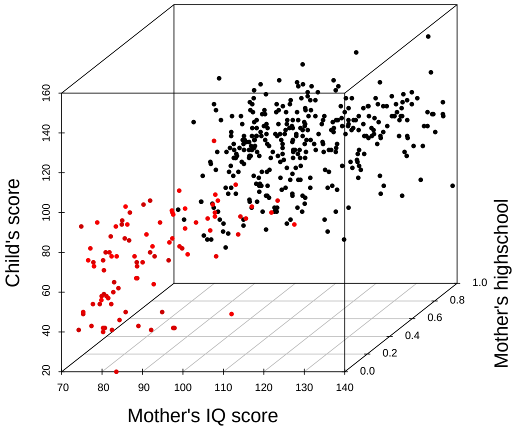
Child score as a function of mother's IQ

$$\text{child score} = a \times \text{mother's IQ} + b$$
$$\text{child score} = 0.6 \times \text{mother's IQ} + 26$$

- $a = 0.6$ – for each additional 10 points of mother's IQ, child's score goes up by 6
- What is the interpretation of $b = 26$?


R^2 = 0.201
Child score
Mother's IQ score
20
40
60
80
100
120
140
70
80
90
100
110
120
130
140

Multiple predictors



Multiple predictors

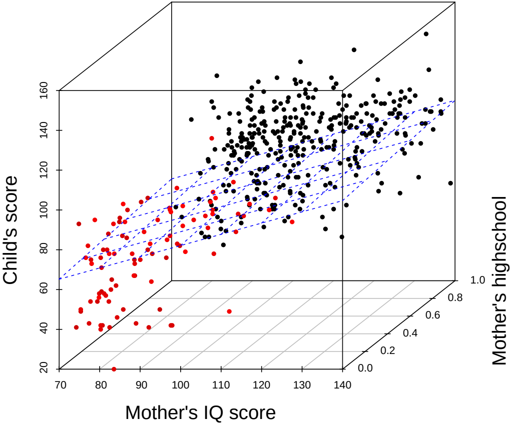
$$y = f(\mathbf{x}) = b + \sum_{i=1}^{d} a_i x_i,$$

where

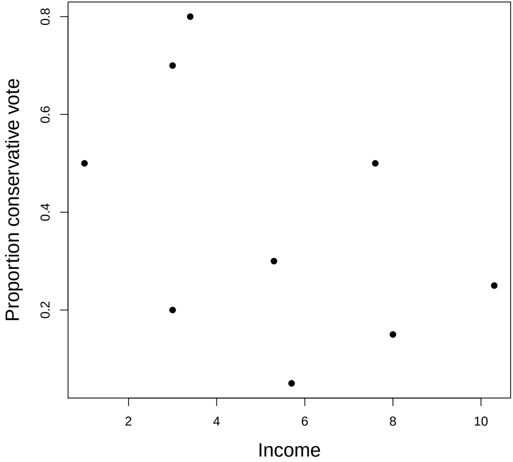
- y = outcome
- b = intercept
- $x_1..x_d$ = independent variables
- $a_1..a_d$ = coefficients

$$\text{child score} = b + a_1 \times \text{highschool} + a_2 \times \text{mother's IQ}$$

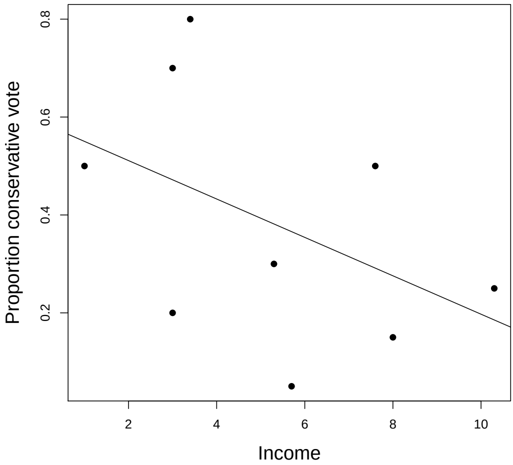
$$\text{child score} = 26 + 6 \times \text{highschool} + 0.6 \times \text{mother's IQ}$$


Child's score
Mother's IQ score
Mother's highschool
20
40
60
80
100
120
140
160
70
80
90
100
110
120
130
140
0.0
0.2
0.4
0.6
0.8
1.0

How does vote depend on income?

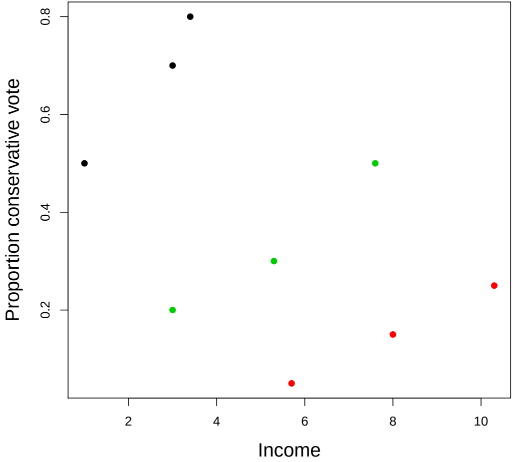


How does vote depend on income?



$$\text{cons} = 0.59 - 0.04 \times \text{income}$$

Add another predictor: region



$$\text{cons} = 0.51 - 0.51 \times \text{green} - 0.86 \times \text{red} + 0.06 \times \text{income}$$

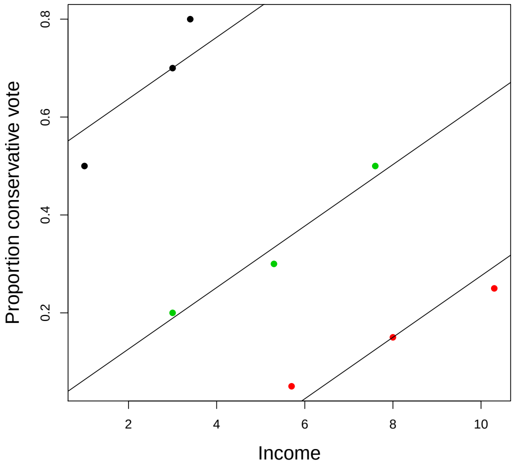
$$\text{cons} = 0.51 - 0.51 \times \text{green} - 0.86 \times \text{red} + 0.06 \times \text{income}$$

- Second model controls for the effect of the region variable

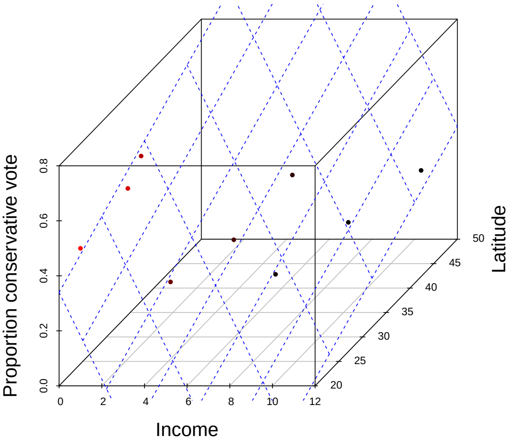
$$\text{cons} = 0.51 - 0.51 \times \text{green} - 0.86 \times \text{red} + 0.06 \times \text{income}$$

- Second model controls for the effect of the region variable
- **When holding region constant**, higher income predicts more conservative vote
- This type of effect is sometimes called **Simpson's paradox**

Simpson's paradox



Controling for latitude



$$\text{cons} = 1.4 + 0.14 \times \text{income} - 0.05 \times \text{latitude}$$

Understanding regression coefficients

Understanding regression coefficients

- Do not indicate any intrinsic effect

Understanding regression coefficients

- Do not indicate any intrinsic effect
- Should not be interpreted in isolation

Understanding regression coefficients

- Do not indicate any intrinsic effect
- Should not be interpreted in isolation
- Only make sense in the context of the whole model

Understanding regression coefficients

- Do not indicate any intrinsic effect
- Should not be interpreted in isolation
- Only make sense in the context of the whole model
- Multiple regression can help clarify confounds

Summary

Summary

- Pearson's correlation coefficient: strength of linear relation

Summary

- Pearson's correlation coefficient: strength of linear relation
  - Symmetric

Summary

- Pearson's correlation coefficient: strength of linear relation
  - Symmetric
- **Correlation often mistaken for causation**
- Linear regression

Summary

- Pearson's correlation coefficient: strength of linear relation
    - Symmetric
- **Correlation often mistaken for causation**
- Linear regression
    - Asymmetric: dependent and independent variables

Summary

- Pearson's correlation coefficient: strength of linear relation
  - Symmetric
- **Correlation often mistaken for causation**
- Linear regression
  - Asymmetric: dependent and independent variables
  - Functional form of linear relation

Summary

- Pearson's correlation coefficient: strength of linear relation
  - Symmetric
- **Correlation often mistaken for causation**
- Linear regression
  - Asymmetric: dependent and independent variables
  - Functional form of linear relation
  - Can be used for prediction