Regression

Data Mining for Business and Governance*
29/8/2017

TILBURG UNIVERSITY

**Formerly known as Social Data Mining*

Course Schedule

`v05.09.2017` (subject to change – always check the latest version!)

#	Date	Lectures (Theory - Willem)	Date	Video Lectures (Applications - Chris)	Video Practicals & Notebooks
1	29-08	Introduction to Data Mining	31-08	Introduction to Data Science	Introduction to jupyter, pandas, and scikit-learn
2	05-09	Regression	07-09	Representing Data: Vectors, Types, Databases	Handling & Interpreting Data, Plotting
3	12-09	Classification	14-09	Working with Text Data Part 1 (17-09)	DIY Pandas + scikit-learn
4	19-09	Algorithm Fitting & Tuning	21-09	Working with Text Data Part 2	**No practical** -> time to prepare for midterm.
5	26-09	**Midterm**	28-09	Best Practices, Common Pitfalls & Research	Preprocessing + Pipelines, MNIST Challenge
6	03-10	Data Reduction & Decomposition	05-10	Mining Massive Data, Ensemble Methods	Online / Out-of-Core Learning
7	10-10	Time Series Analysis	12-10	Applications of Deep Learning	Social Media and Multi-modal Data
8	17-10	Clustering and Graphs	19-10	Explaining Models, Ethics, Privacy	Unsupervised Learning: Intuitions and Metrics

Overview: Regression

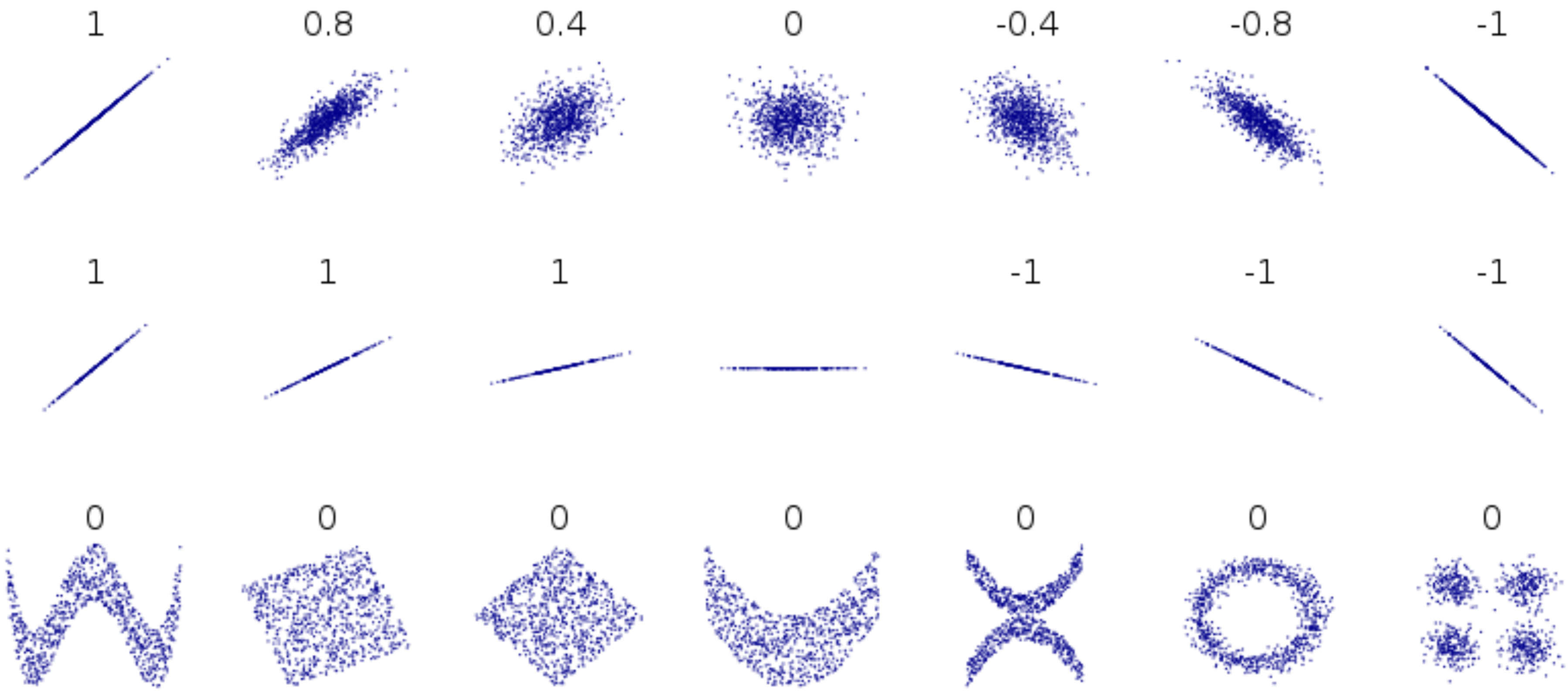
- **Covariance and Correlation**
- Correlation vs Causation
- Linear Regression with two variables
- Mean Squared Error
- Binary Predictor
- Multiple Predictors
- Simpson’s paradox
- Interpreting regression coefficients

Overview: Regression

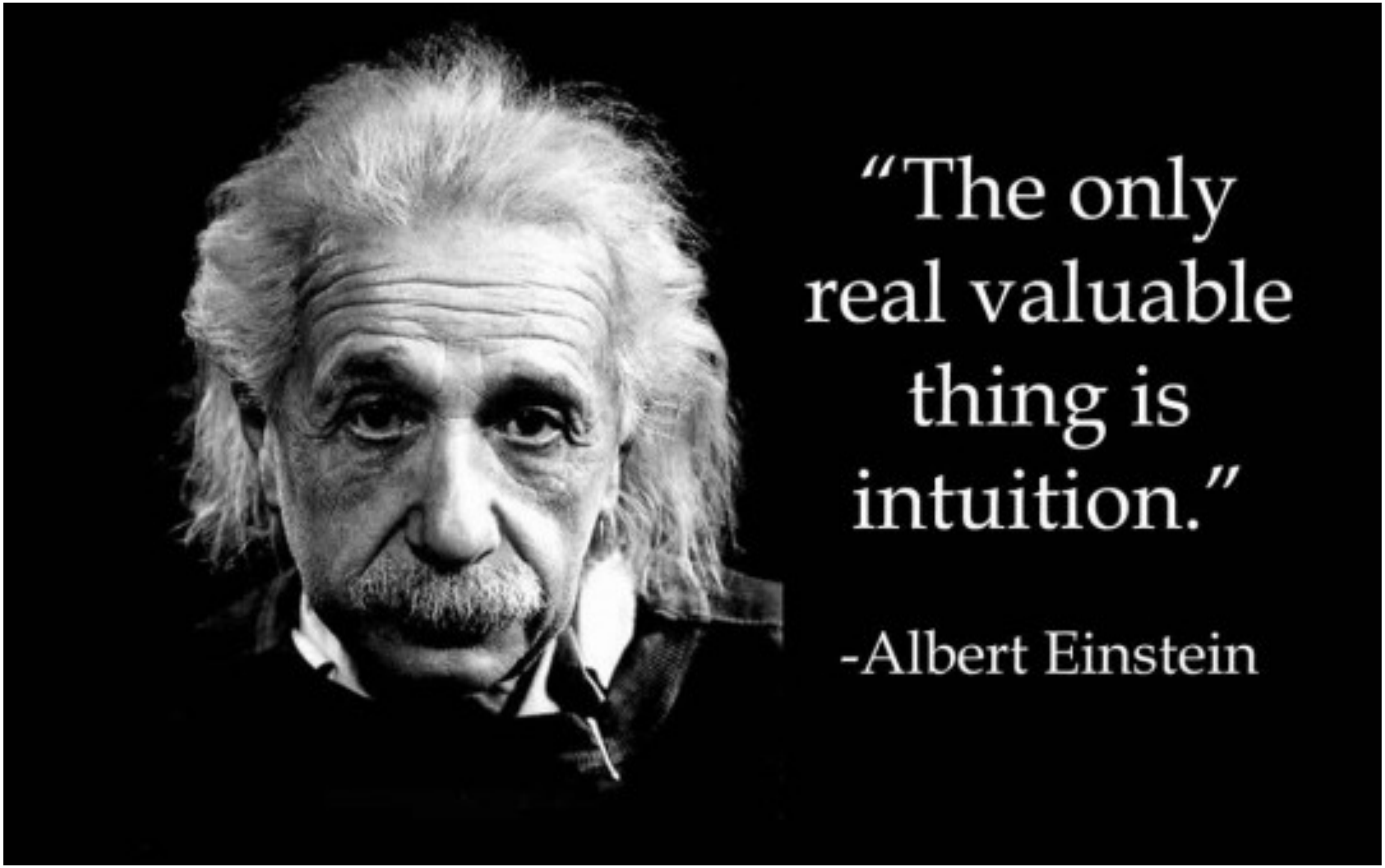
- **Covariance and Correlation**
- Correlation vs Causation
- Linear Regression with two variables
- Mean Squared Error
- Binary Predictor
- Multiple Predictors
- Simpson’s paradox
- Interpreting regression coefficients

Correlation Coefficient

- Pearson's *r* measures the strength of **linear** relationship (dependency)



Covariance and Correlation



Covariance and Correlation

- **Covariance** is a measure of the joint variability of two variables (X,Y)

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Relation between X and Y (requires a pair of inputs)

- Minus(X,Y)
- Mean(X)
- Cov = function

$$\text{Mean(Y)} = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$$

"For i to n → apply … and sum"

$$\sum_{i=1}^{4} 2X_i = 2*1 + 2*2 + 2*3 + 2*4 = 20 = \sum_{i=1}^{4} 2Xi$$

Covariance and Correlation

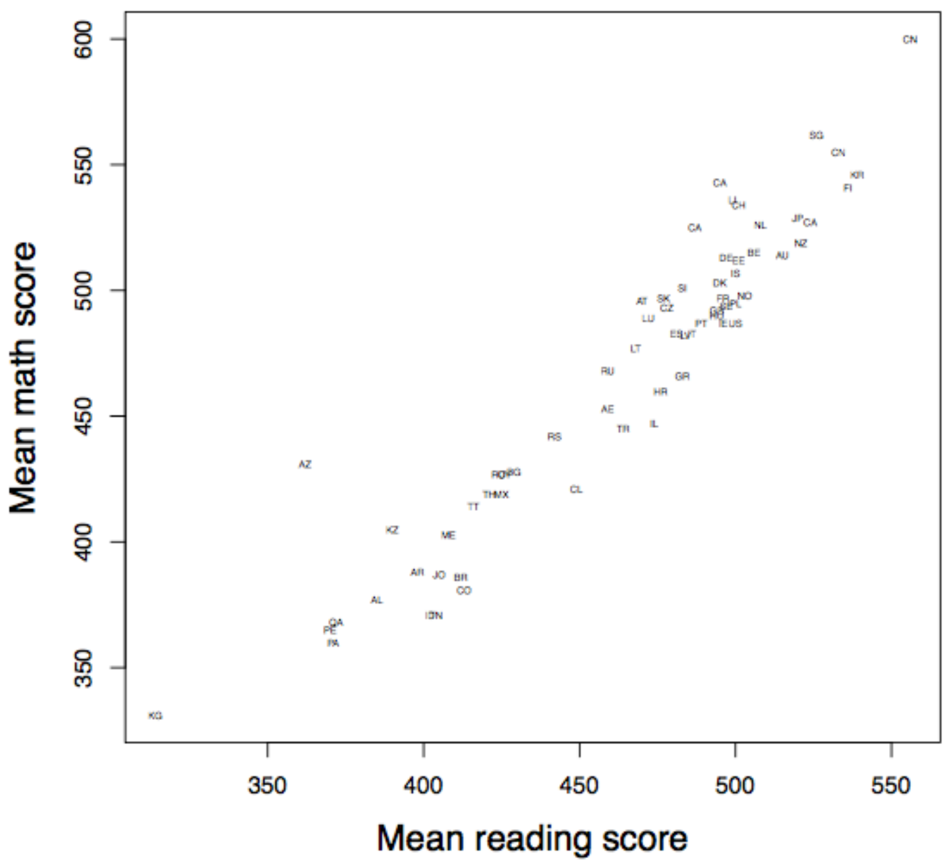
- **Covariance** is a measure of the joint variability of two variables (X,Y)
- Magnitude of the covariance is not easy to interpret

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Correlation coefficient, is normalized and corresponds to strength of the linear relation
- Divide variance by the product of the variables standard deviations

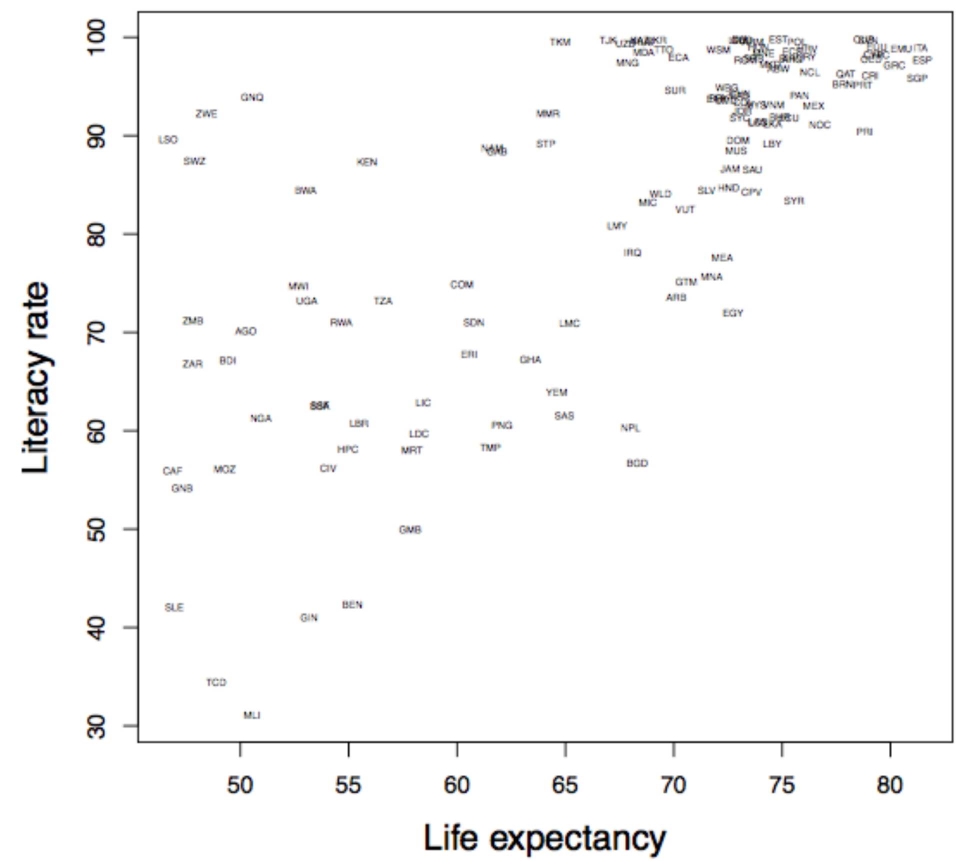
$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$R = 0.95$



Mean math score
Mean reading score

$R = 0.73$

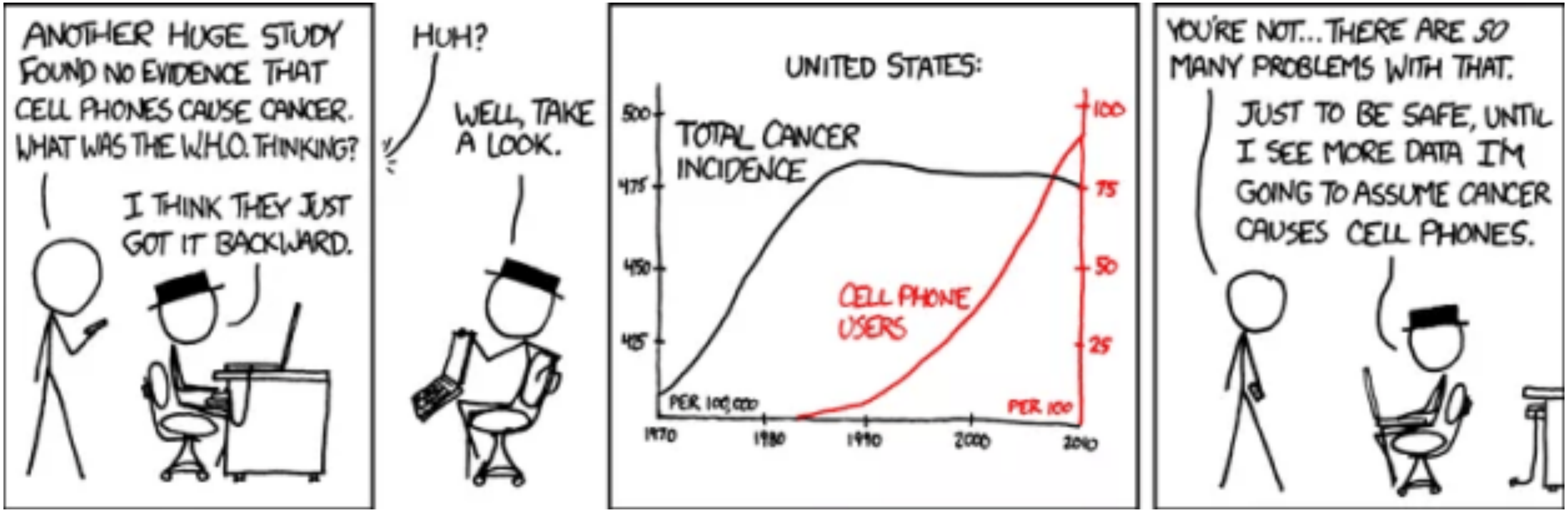


Literacy rate
Life expectancy

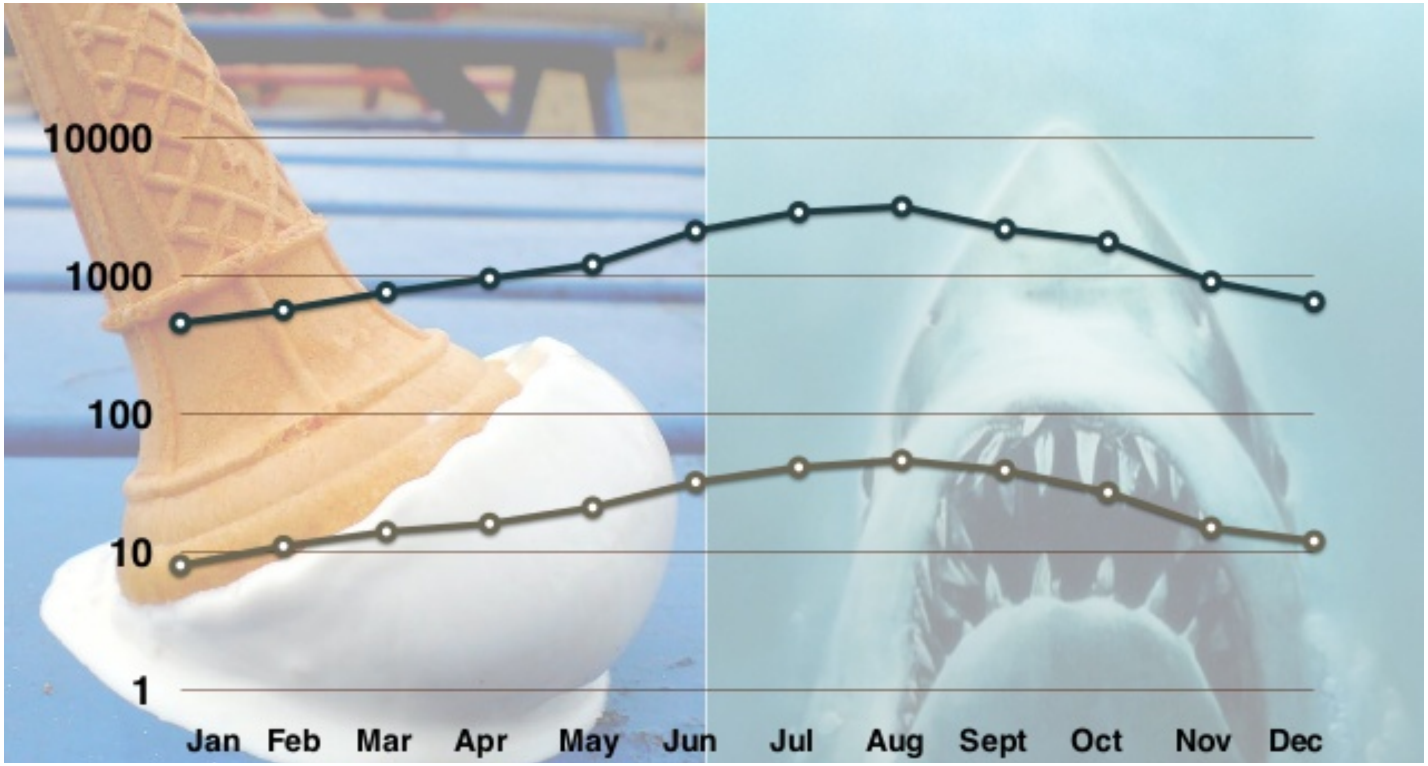
Overview: Regression

- Covariance and Correlation
- **Correlation vs Causation**
- Linear Regression with two variables
- Mean Squared Error
- Binary Predictor
- Multiple Predictors
- Simpson’s paradox
- Interpreting regression coefficients

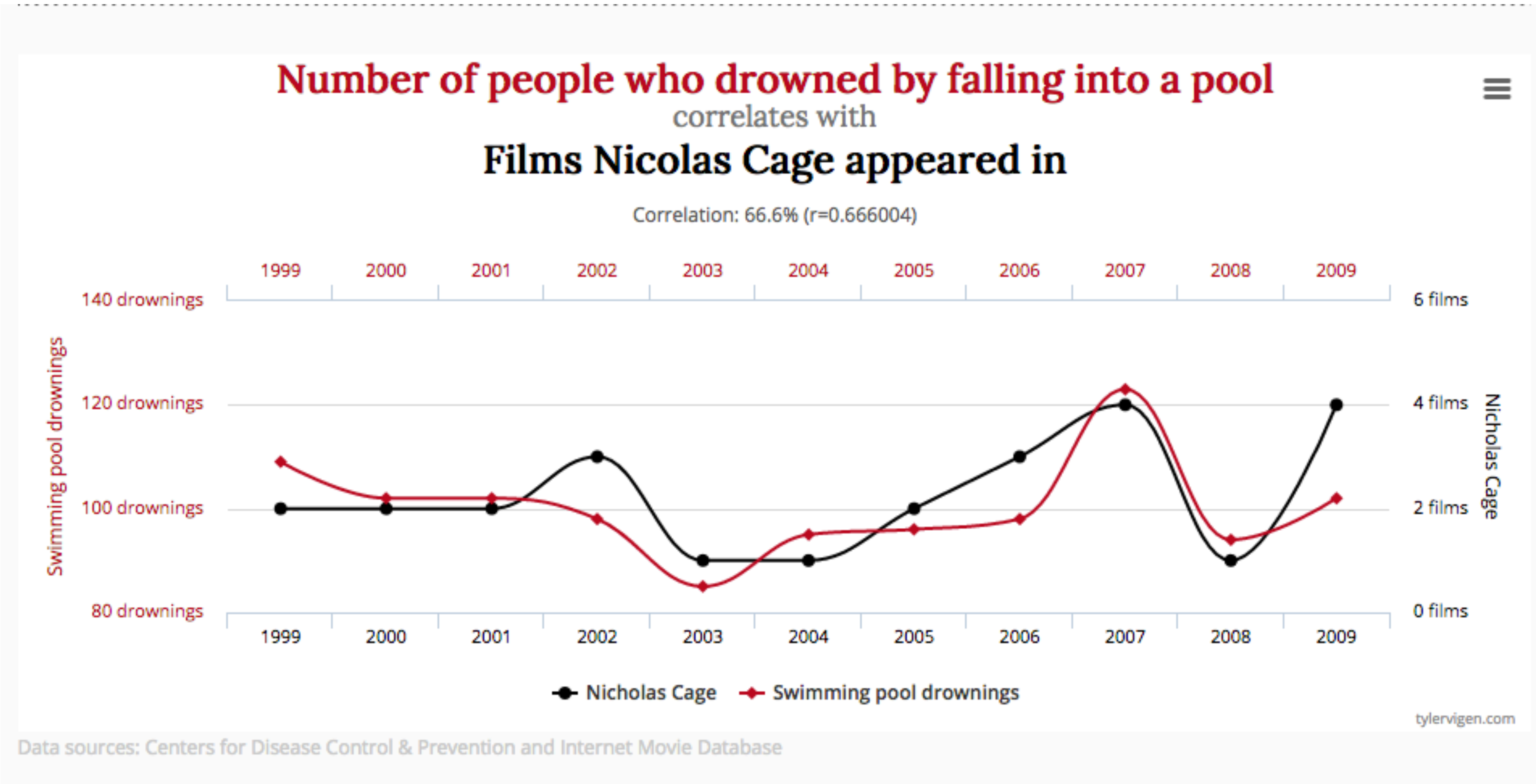
Correlation vs Causation



Ice cream sales vs Shark attacks



- As ice cream sales increase, the number of shark attacks increases sharply.
- Therefore, ice cream consumption causes shark attacks

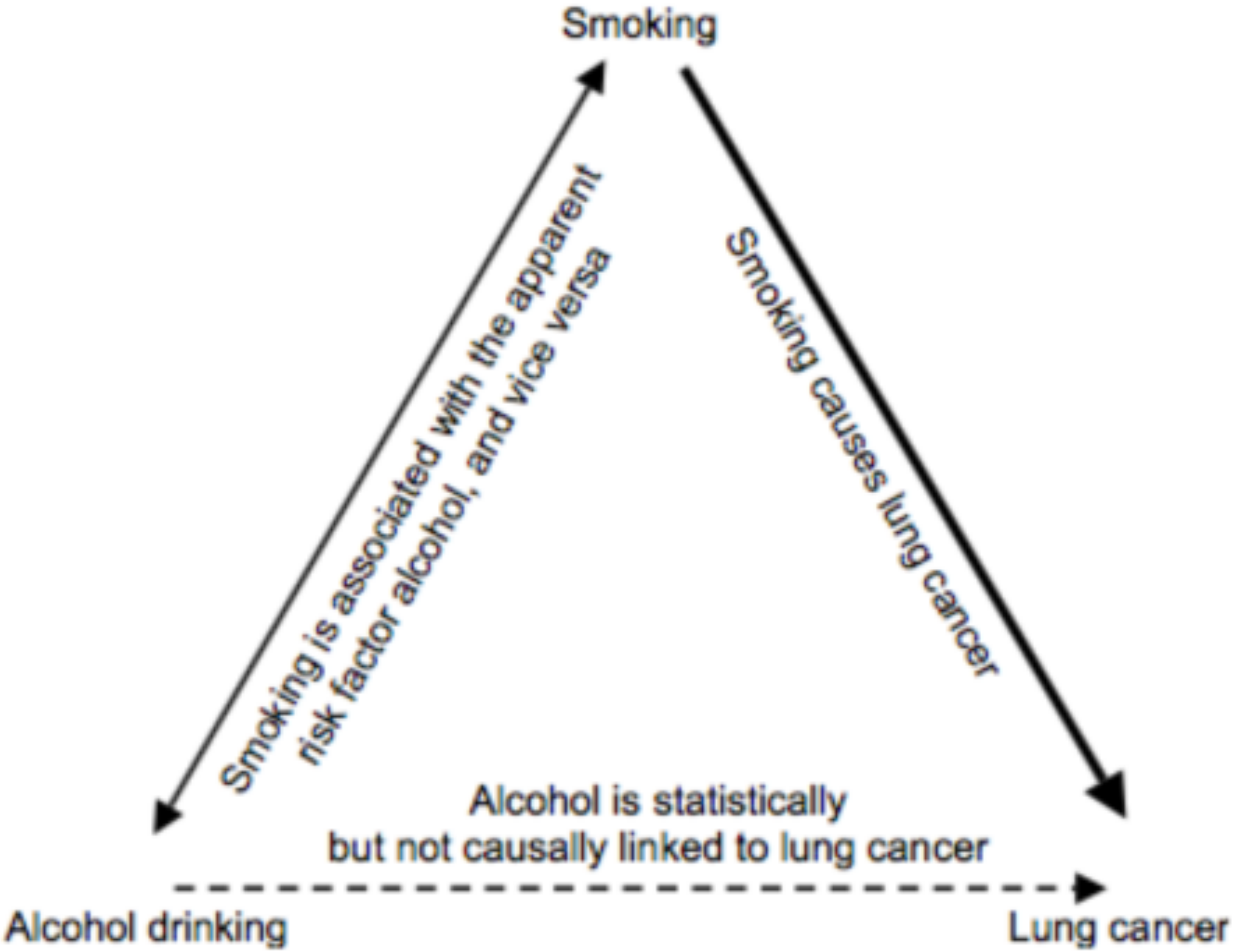

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

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
  (C might be known, hidden or a confounders)
- The correlation is a coincidence
- Some combination of the above

An example of different associations



Correlation vs Causation

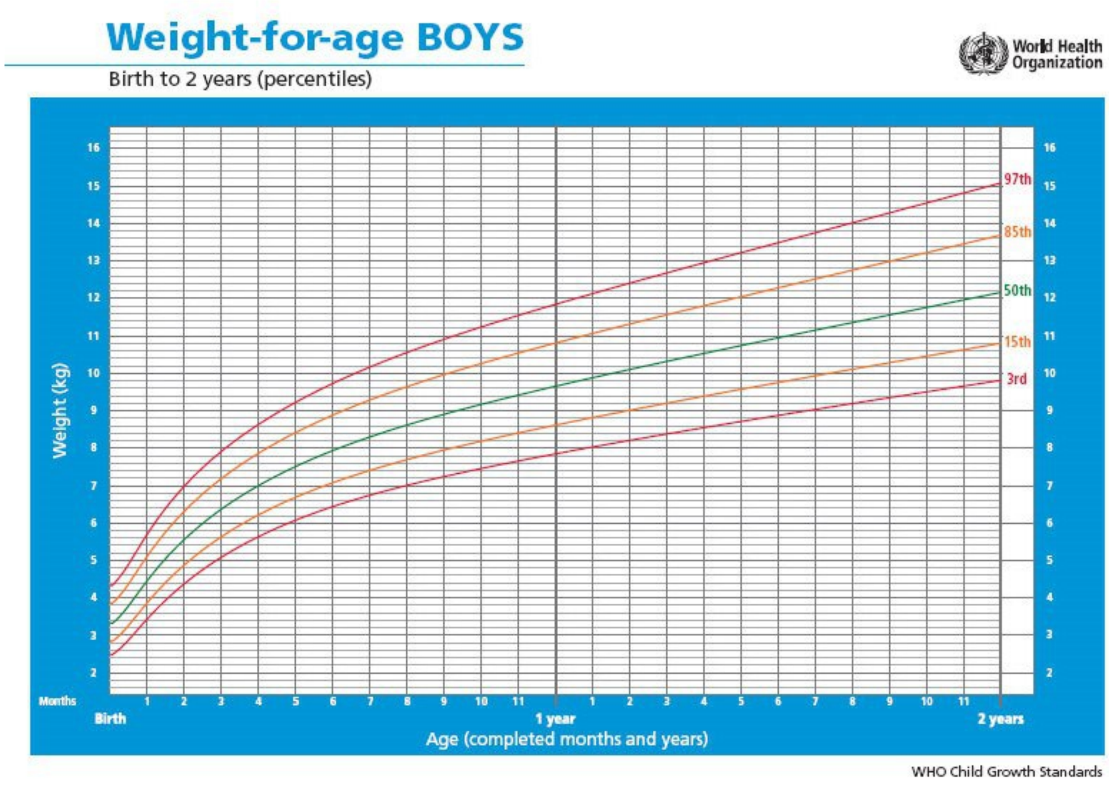
- Discovery of correlation can suggest a causal relationship
- Some argue that causation can only be fully elucidated by an experimental study
    - Vary a single variable while keeping all else equal
    - Does the other variable co-vary?

Overview: Regression

- Covariance and Correlation
- Correlation vs Causation
- **Linear Regression with two variables**
- Mean Squared Error
- Binary Predictor
- Multiple Predictors
- Simpson’s paradox
- Interpreting regression coefficients

Linear Regression

- A very valuable tool in (Data) Science
- Regression Analysis is used to:
  - **Describe** the relationship between random variables
  - **Predict** the value of one variable based on another variable



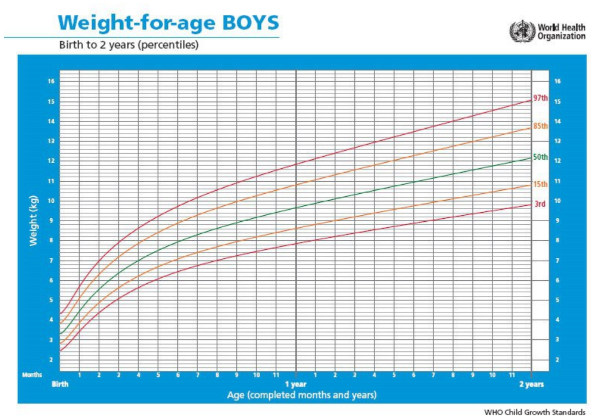
Regression

- A very valuable tool in (Data) Science
- Regression Analysis is used to:
  - **Describe** the relationship between random variables
  - **Predict** the value of one variable based on another variable

- Model the relationship between two variables
  - Dependent (output), or response, or target variable (Y axis)
  - Independent variables (input), predictors, or features, in the case of one variable (X axis)

Regression Model

- Single independent variable x (predictor)
- Dependent variable y (target)
- 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

How to perform a regression

“A construction company renovates old homes in the Netherlands. They have found that its earning on renovation work is dependent on the area payroll.”

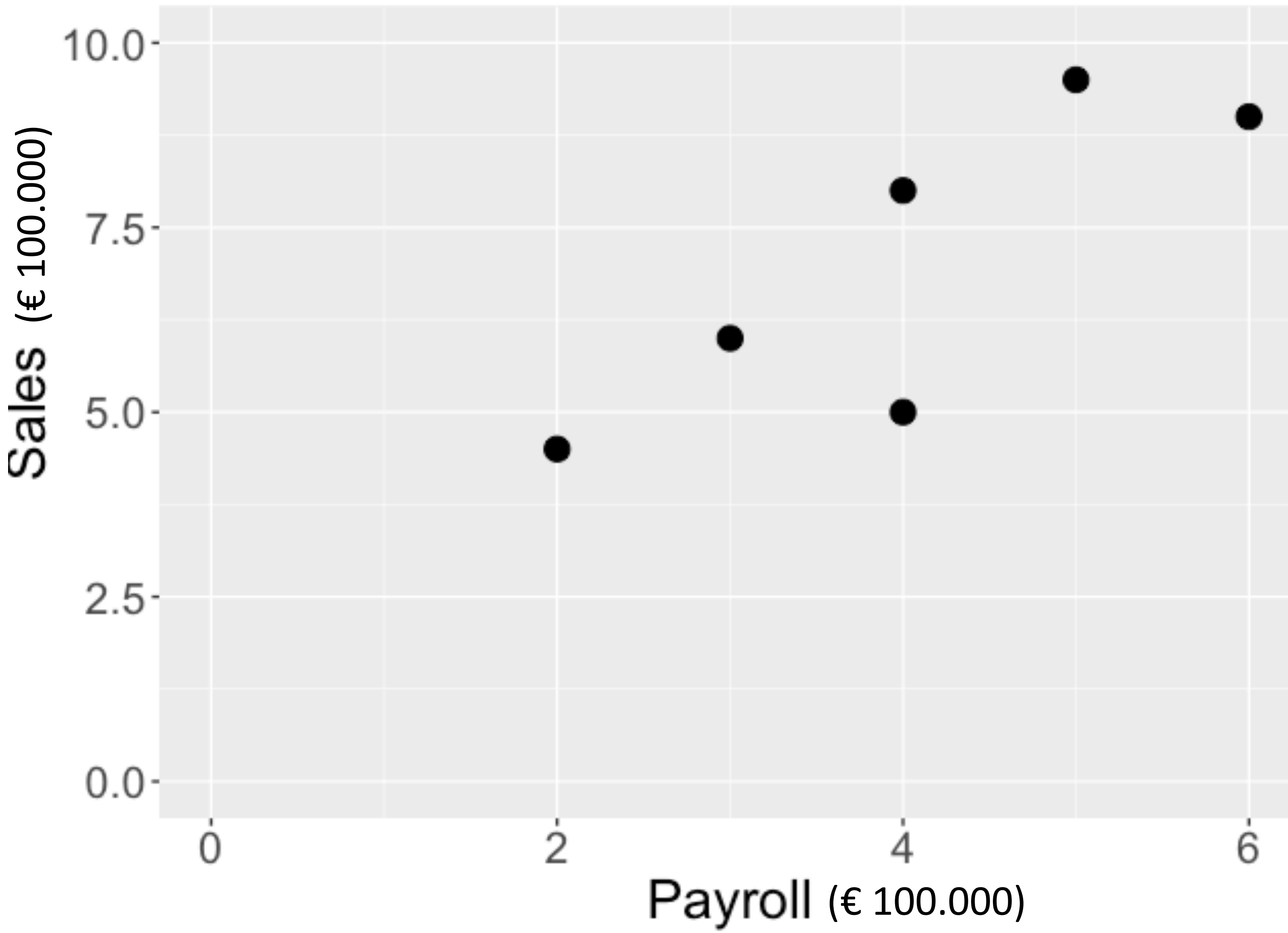
City	Sales (Y)	Payroll (X)
Tilburg	6	3
Eindhoven	8	4
Utrecht	9	6
Nijmegen	5	4
Maastricht	4.5	2
Amsterdam	9.5	5

*€ 100.000

Regression Model

- Inspect your data (Create a Scatter Plot)

Scatter Plot



Regression Model

- Inspect your data (Create a Scatter Plot)
- Perform a Regression Analysis: f(X) = AX + B

$$Y = B + AX$$

Dependent Variable (Target)

Intercept (Value of Y when X = 0)

Slope

Independent Variable / predictor (features)

Practical has notation: $Y = \beta_0 + \beta_1 \cdot X$

Regression Model

Training data are used to determine the values for the intercept and slope (~sample data).

$$Y = \beta_0 + \beta_1 X$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Regression Example

City	Sales (Y)	Payroll (X)	$(X - \bar{X})^2$	$(X - \bar{X}) * (Y - \bar{Y})$
Tilburg	6	3	1	1
Eindhoven	8	4	0	0
Utrecht	9	6	4	4
Nijmegen	5	4	0	0
Maastricht	4.5	2	4	5
Amsterdam	9.5	5	1	2.5
Σ	42	24	10	12.5

$\bar{Y} = 42 / 6 = 7$

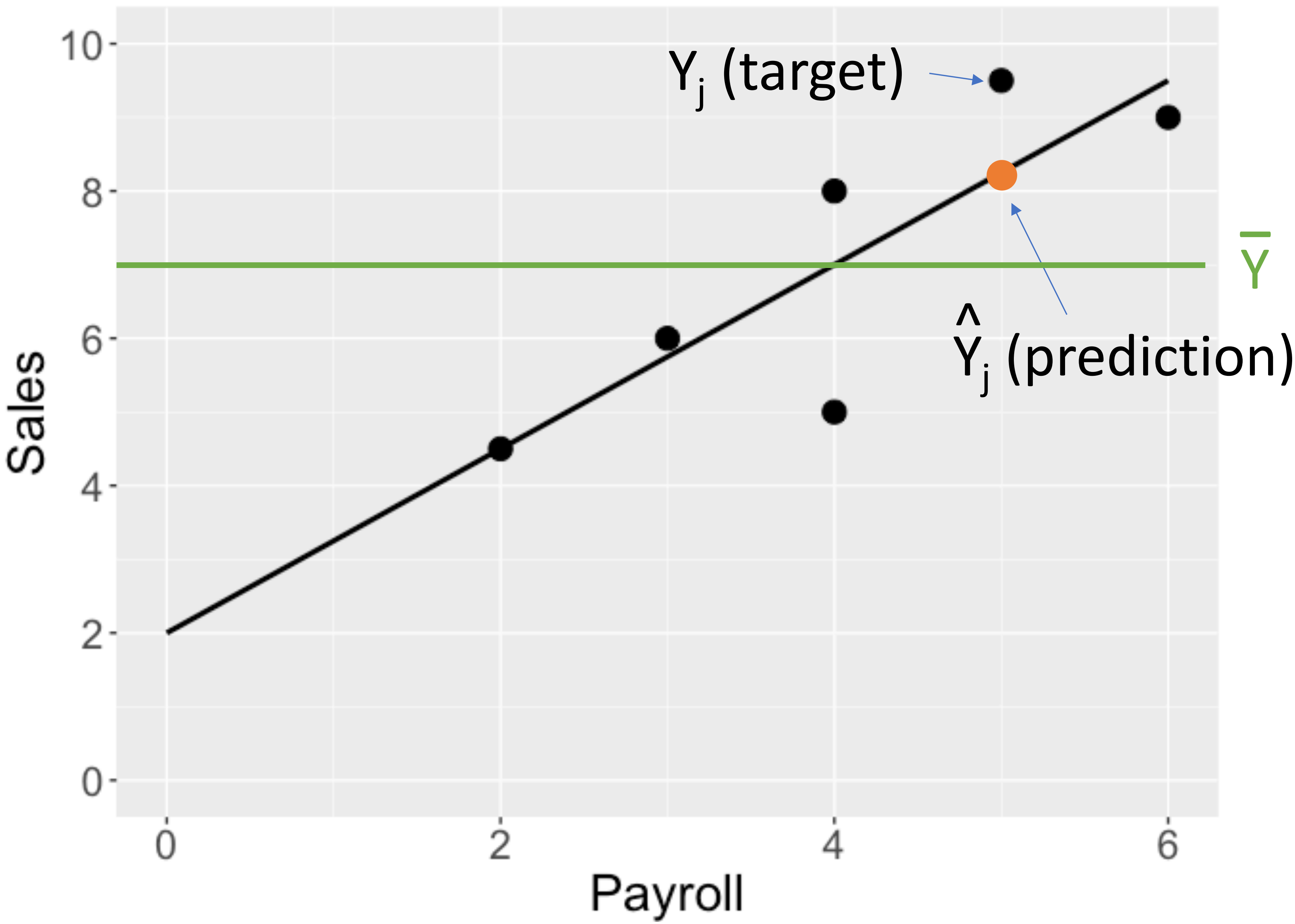
$\bar{X} = 24 / 6 = 4$

$Y = \beta_0 + \beta_1 X$

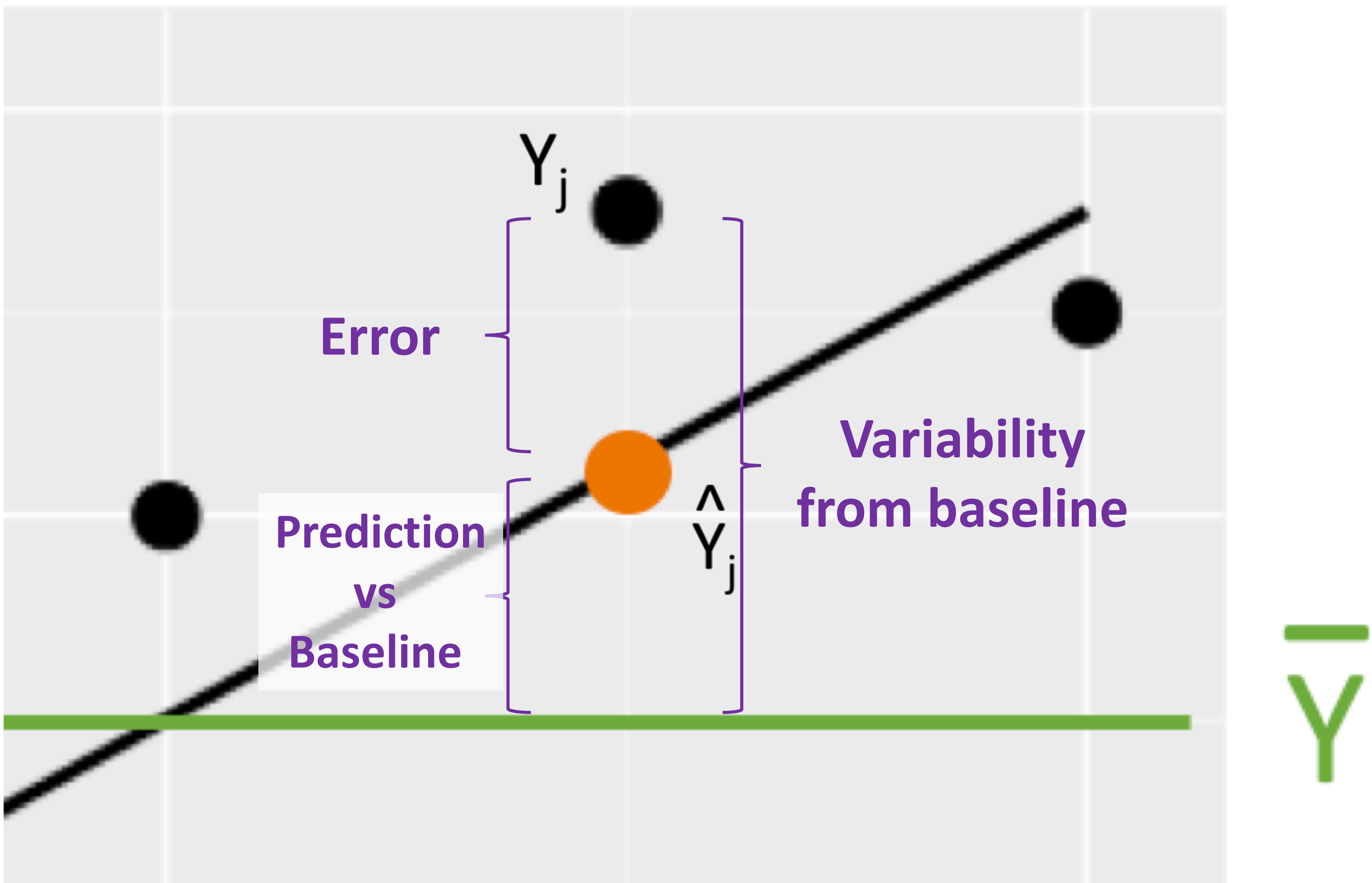
$Y = 2 + 1.25\ X$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{12.5}{10} = 1.25$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 7 - (1.25 * 4) = 2$$



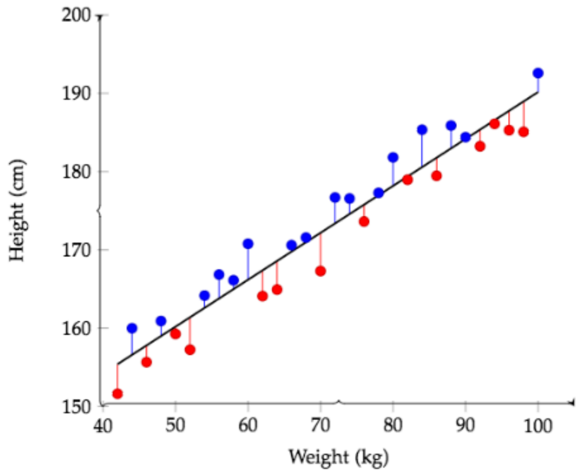
Y_j (target)
$\hat{Y}_j$ (prediction)
$\bar{Y}$
Sales
Payroll
0
2
4
6
8
10



For each observation (j) you can calculate: Error = Target - Prediction $Y_j - \hat{Y}_j = error_j$

Best Fit is Minimal Error

- Residual (Error): difference between true value y and predicted value ($\hat{y}$) = f(x)



- Errors may be positive or negative
  Summing errors can be misleading
  Square terms prior to summing

- Find Line that minimizes the Mean Squared Error (MSE)

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

- MSE: how much values deviate from the regression line. Or how much of the total variance is unexplained

MSE vs RMSE

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

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

MSE and variance "explained"

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

- Total variability in Y = Sum of the Squares Total (SST).

$$SST = \Sigma(Y-\bar{Y})^2$$

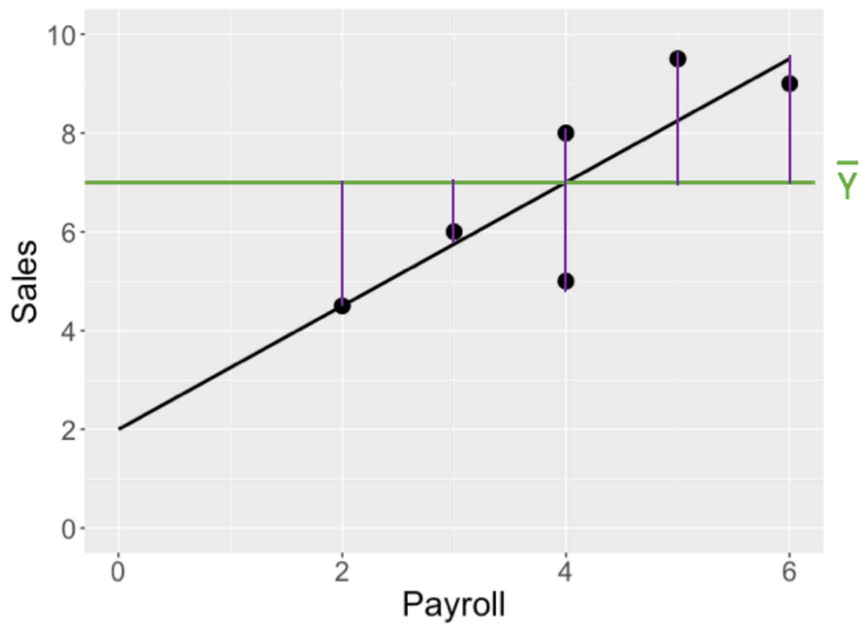
- Sum of Squared Errors (SSE)

$$SSE = \Sigma e^2 = \Sigma(Y-\hat{Y})^2$$

- Sum of Squared due to Regression (SSR)

$$SSR = \Sigma(\hat{Y}-\bar{Y})^2$$

Variability of our example



$$SST = \Sigma(Y-\bar{Y})^2 = 22.5$$

$$SSE = \Sigma e^2 = \Sigma(Y-\hat{Y})^2 = 6.857$$

$$SSR = \Sigma(\hat{Y}-\bar{Y})^2 = 15.625$$

$$SST = SSR + SSE$$

Explained Variance (Prediction) *Unexplained Variance (Error)*

Proportion of "Explained" Variance

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SST} = R^2$$

$$R^2 = \frac{15.62}{22.5} = 0.6944$$

R^2: coefficient of determination

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

$$R^2 = 1 - \frac{MSE(f)}{MSE(\text{mean})} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- How well model f predicts targets relative to mean
- Equivalent to proportion of variance explained by f
- Mean is not always the baseline prediction

Overview: Regression

- Covariance and Correlation
- Correlation vs Causation
- Linear Regression with two variables
- **Mean Squared Error**
- Binary Predictor
- Multiple Predictors
- Simpson’s paradox
- Interpreting regression coefficients

Interim Summary: the fit of a linear regression

To **describe** how well the X predicts Y, you need to evaluate:

- The variability in the target (Y)
- Correlation coefficient (R)
- Proportion of variance explained (R^2)
- Standard Error (standard deviation around the regression line)
- Analyze the residuals (errors)
- Test of Linearity (significance)

To demonstrate how well X **predict** Y, you need to evaluate

- The variability of the variables (X,Y)
- Proportion variance explained (R^2)
- Root Mean Squared Error (RMSE)

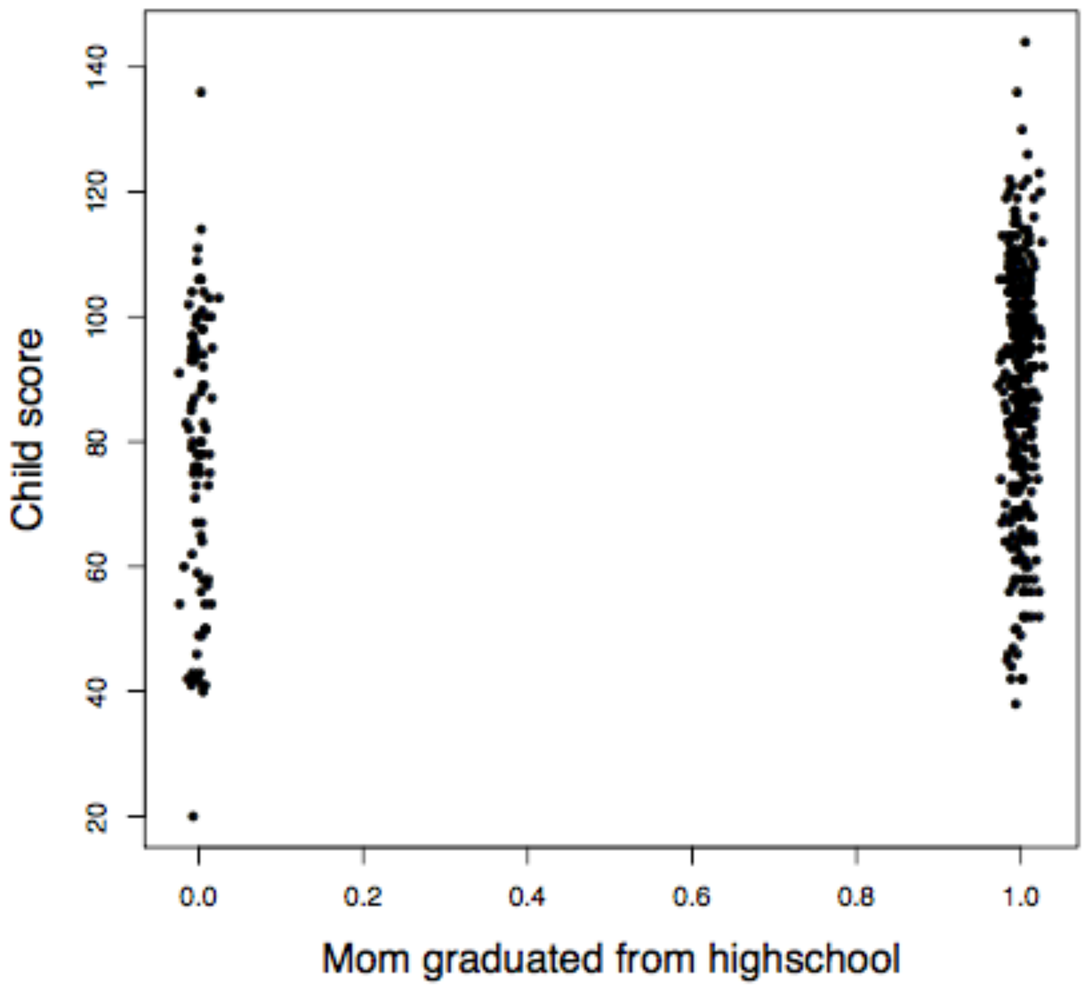
Overview: Regression

- Covariance and Correlation
- Correlation vs Causation
- Linear Regression with two variables
- Mean Squared Error
- **Binary Predictor**
- Multiple Predictors
- Simpson’s paradox
- Interpreting regression coefficients

Binary Prediction

- Simplest categorical variable: binary value
- Code False → 0 and True → 1

Binary Prediction



Binary Prediction

Equation: ($Y = \beta_0 + \beta_1 X$)

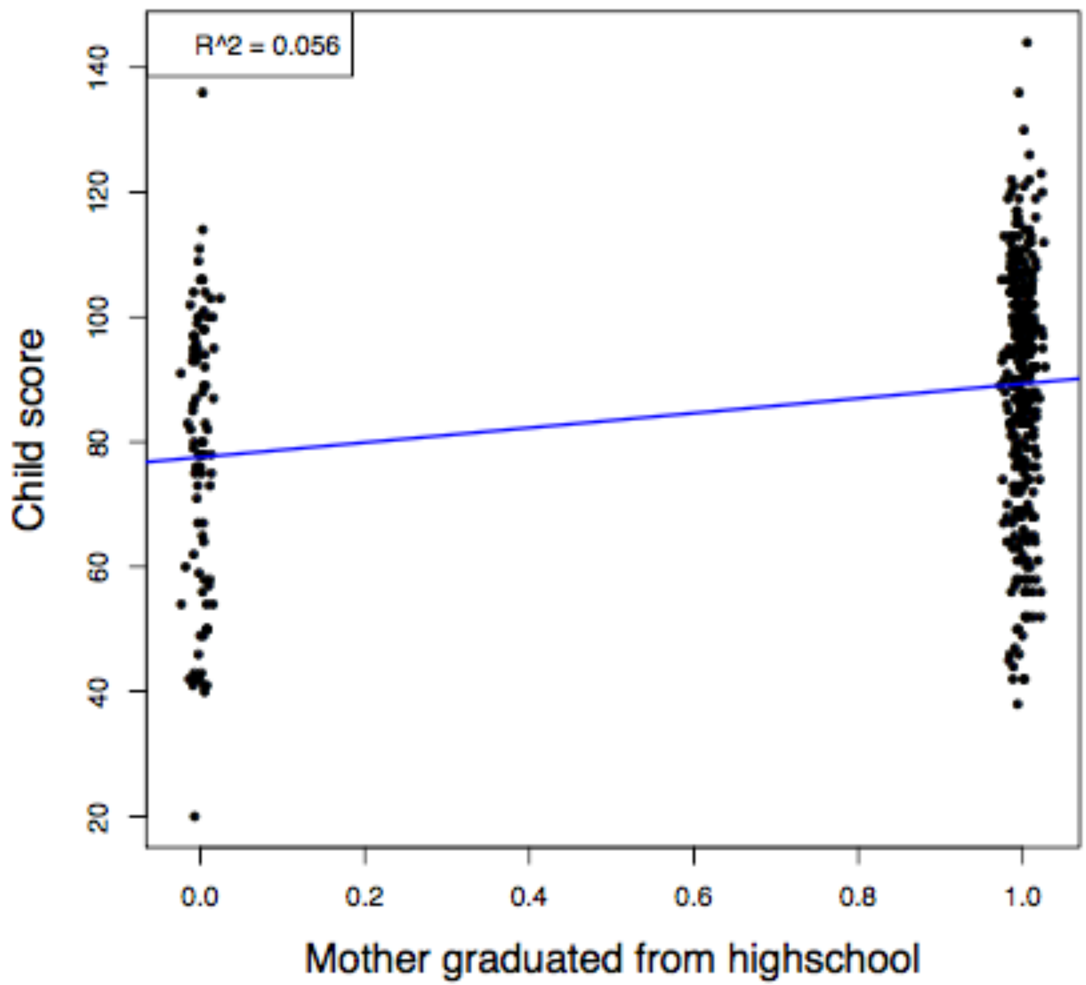
child score = $\beta_0 + \beta_1$* highschool

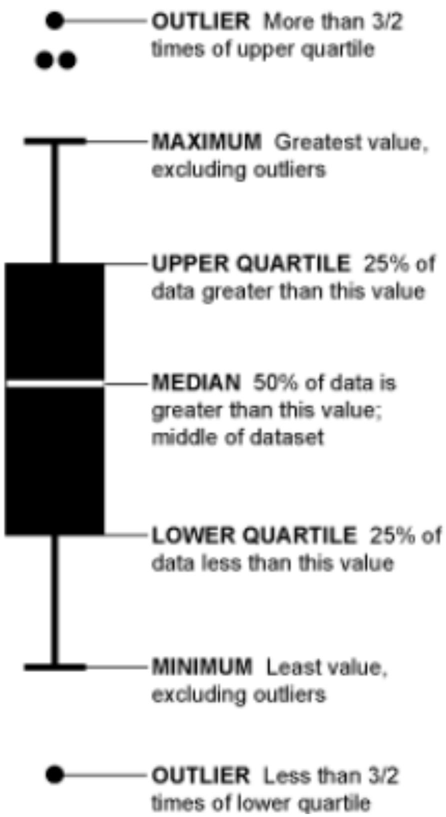
child score = 78 + 12 * highschool

78 + 12 * 0 = mean score of children whose mothers have no high school

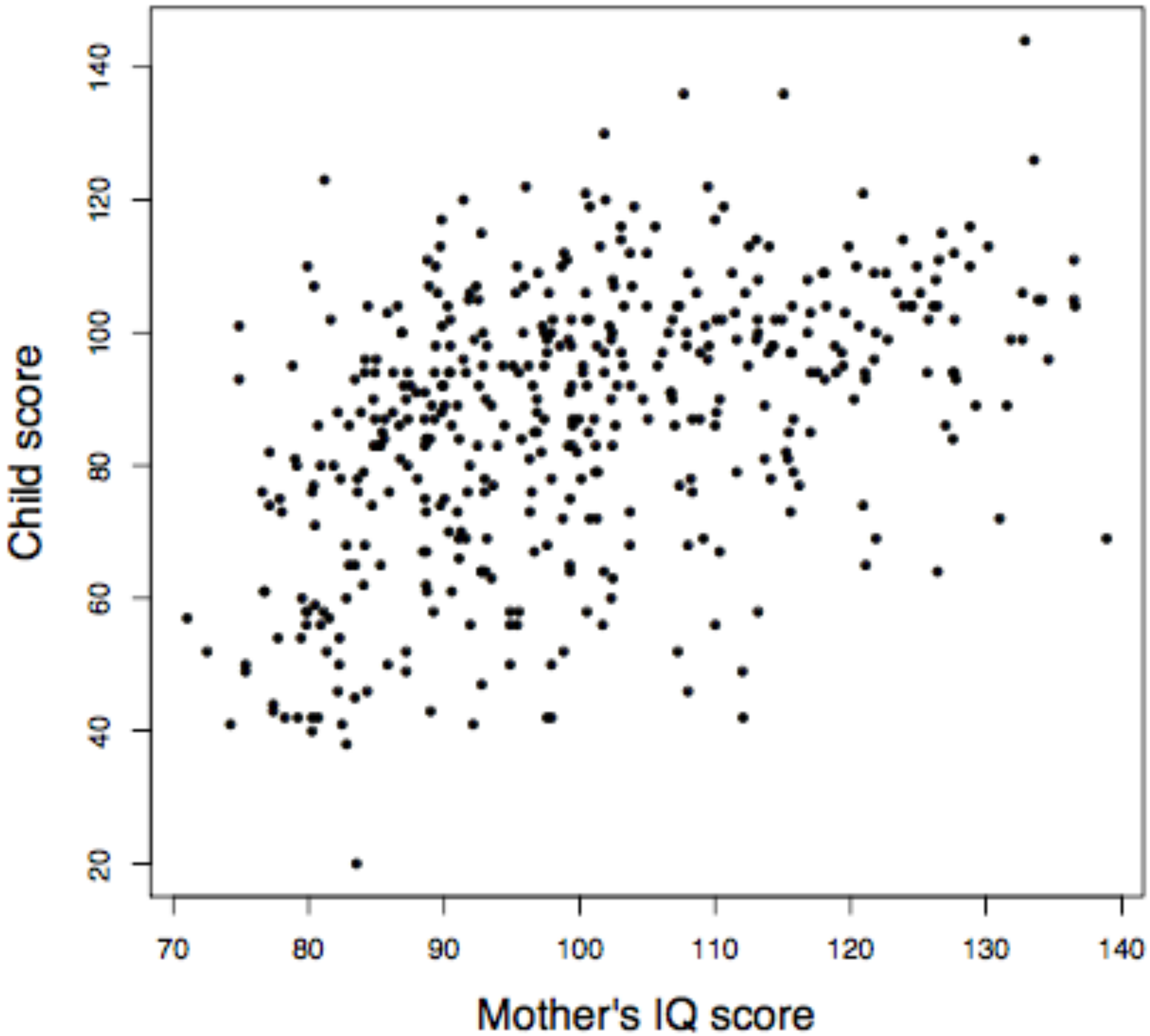
78 + 12 * 1 = mean score of children whose mothers do have highschool

Binary Prediction





Example: Child score as a function of mother's IQ

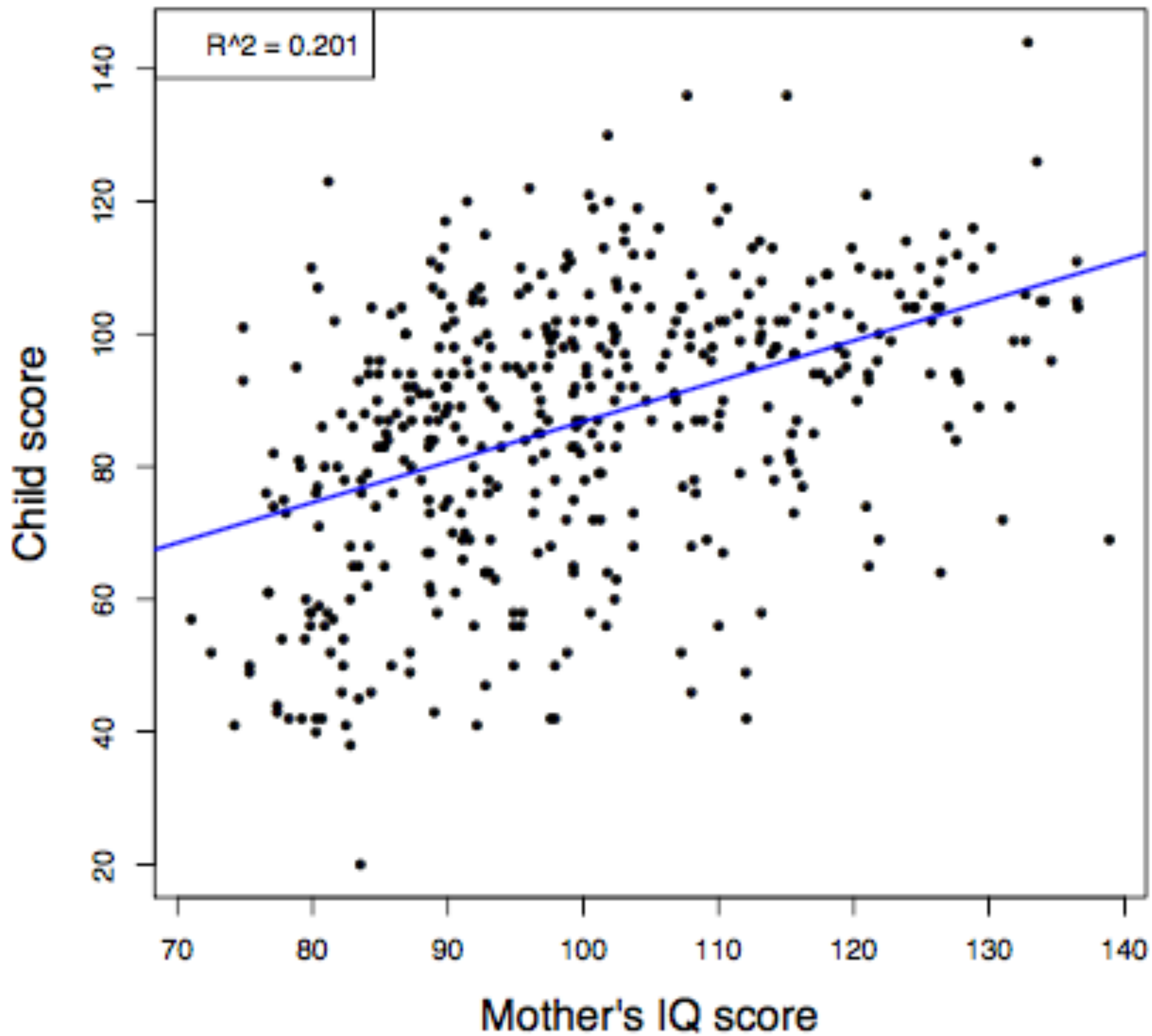


Example: Child score by Mothers IQ

$$\text{child score} = \beta_0 + \beta_1 * \text{mother's IQ}$$

$$\text{child score} = 26 + 0.6 * \text{mother's IQ}$$

- $\beta_1 = 0.6$ “for each additional 10 points of mother’s IQ, child’s score goes up by 6”
- What is the interpretation of $\beta_0 = 26$?


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

Overview: Regression

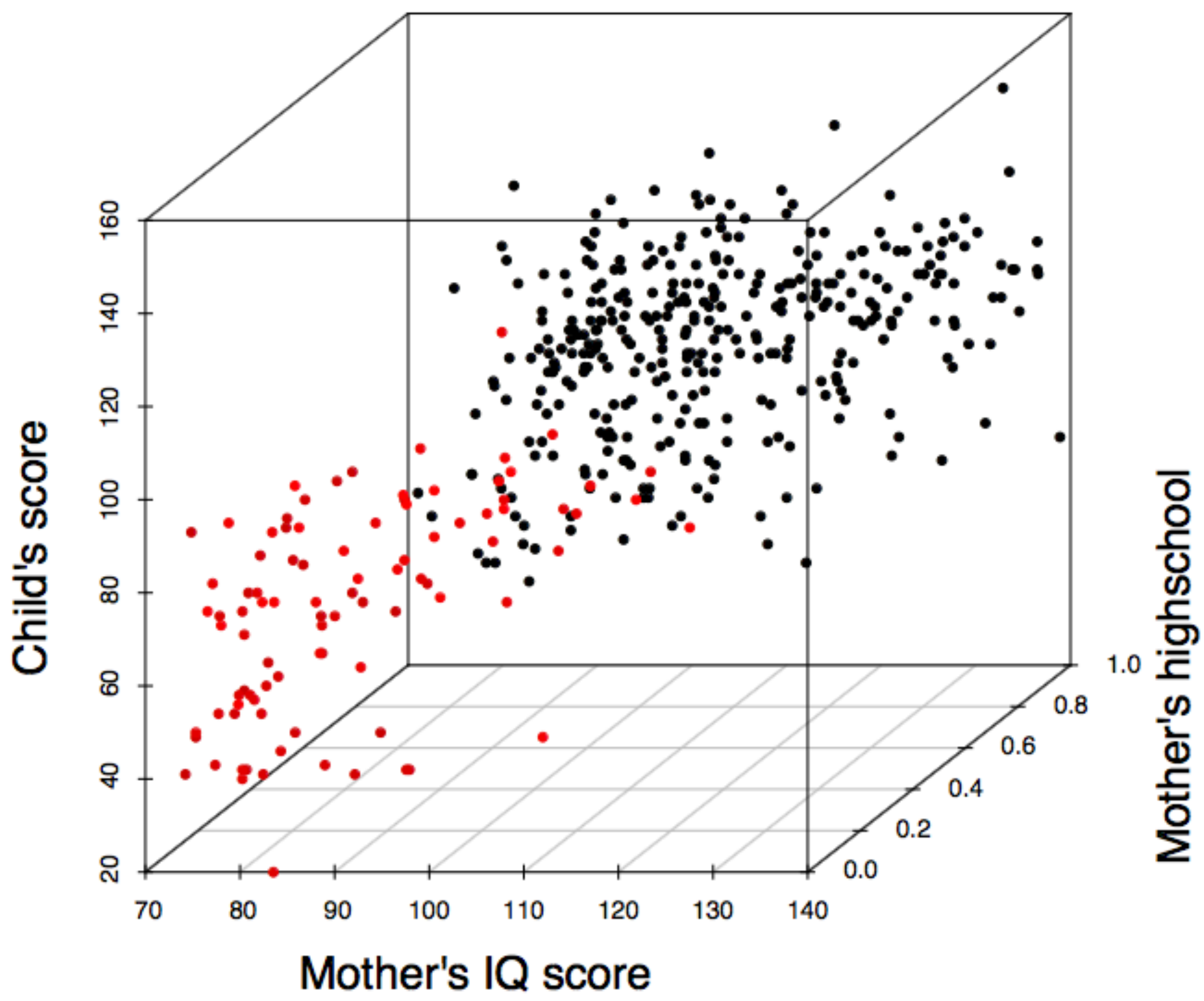
- Covariance and Correlation
- Correlation vs Causation
- Linear Regression with two variables
- Mean Squared Error
- Binary Predictor
- **Multiple Predictors**
- Simpson’s paradox
- Interpreting regression coefficients

Multiple Predictors

Child score = Mother's High school + Mother's IQ

(Binary) *(Continuous)*

Multiple Predictors



Multiple Predictors

Child score = Mother's High school + Mother's IQ

(Binary) *(Continuous)*

$$\hat{Y} = f(X) = \beta_0 + \sum_{i=1}^{j} \beta_i X_i$$

- $\hat{Y}$ = outcome (prediction)
- β_0 = intercept (bias)
- $X_1 \ldots X_j$ = independent variables (predictors)
- $\beta_1 \ldots \beta_j$ = regressions coefficients (~slope)

Multiple Predictors

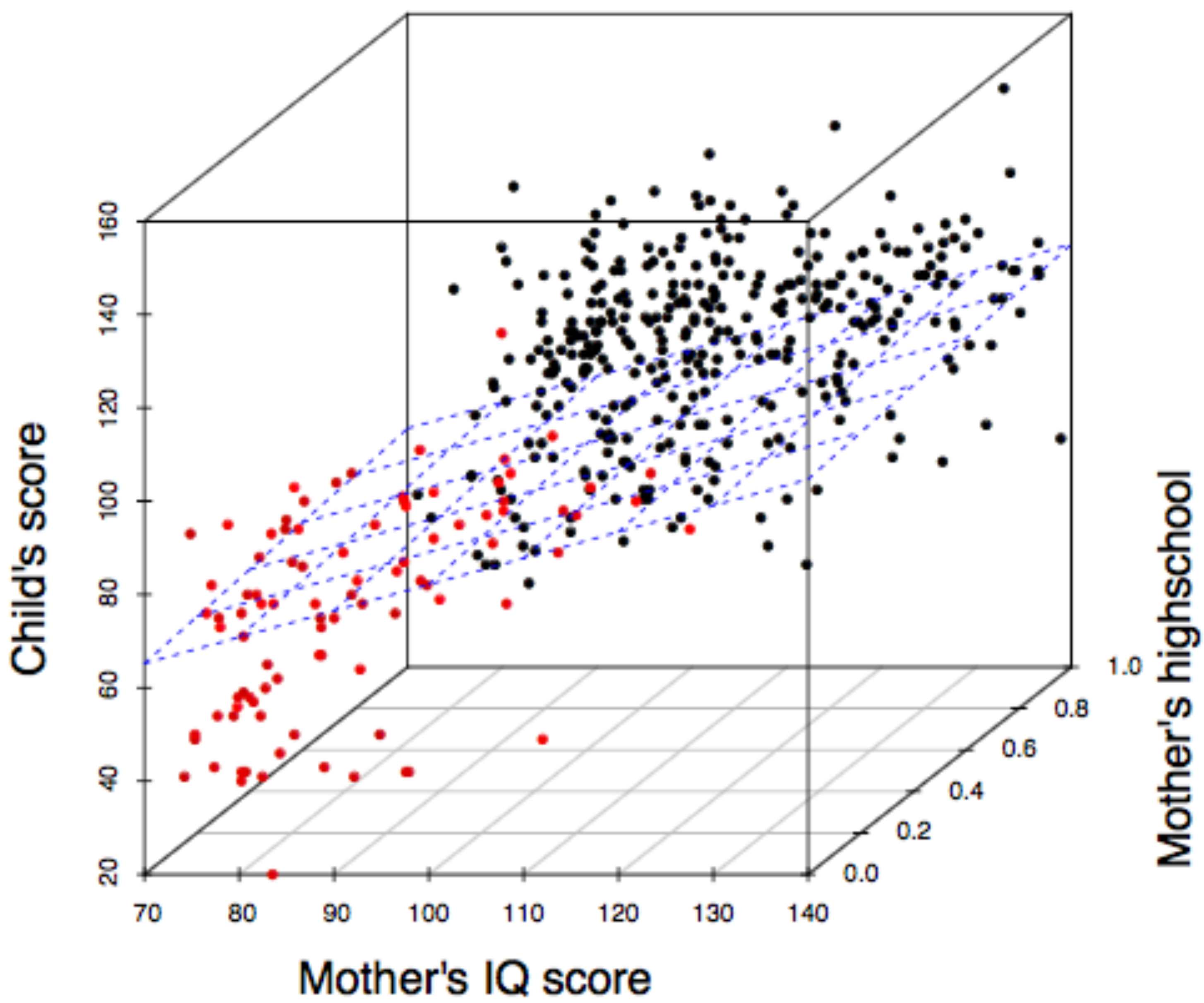
Child score = Mother's Highschool + Mother's IQ

(Binary) *(Continuous)*

$$\hat{Y} = f(X) = \beta_0 + \sum_{i=1}^{j} \beta_i X_i$$

Child score = β_0 + β_1* Highschool + β_2 * mother's IQ

Child score = 26 + 6 * Highschool + 0.6 * mother's IQ



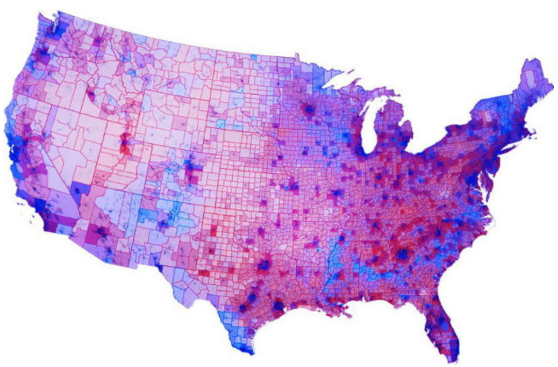
Child's score
20
40
60
80
100
120
140
160
70
80
90
100
110
120
130
140
Mother's IQ score
0.0
0.2
0.4
0.6
0.8
1.0
Mother's highschool

Overview: Regression

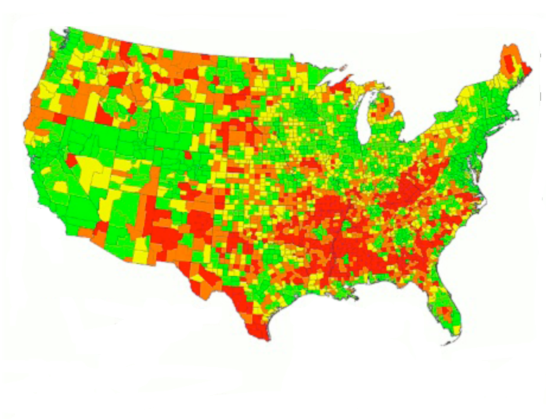
- Covariance and Correlation
- Correlation vs Causation
- Linear Regression with two variables
- Mean Squared Error
- Binary Predictor
- Multiple Predictors
- **Simpson's paradox**
- Interpreting regression coefficients

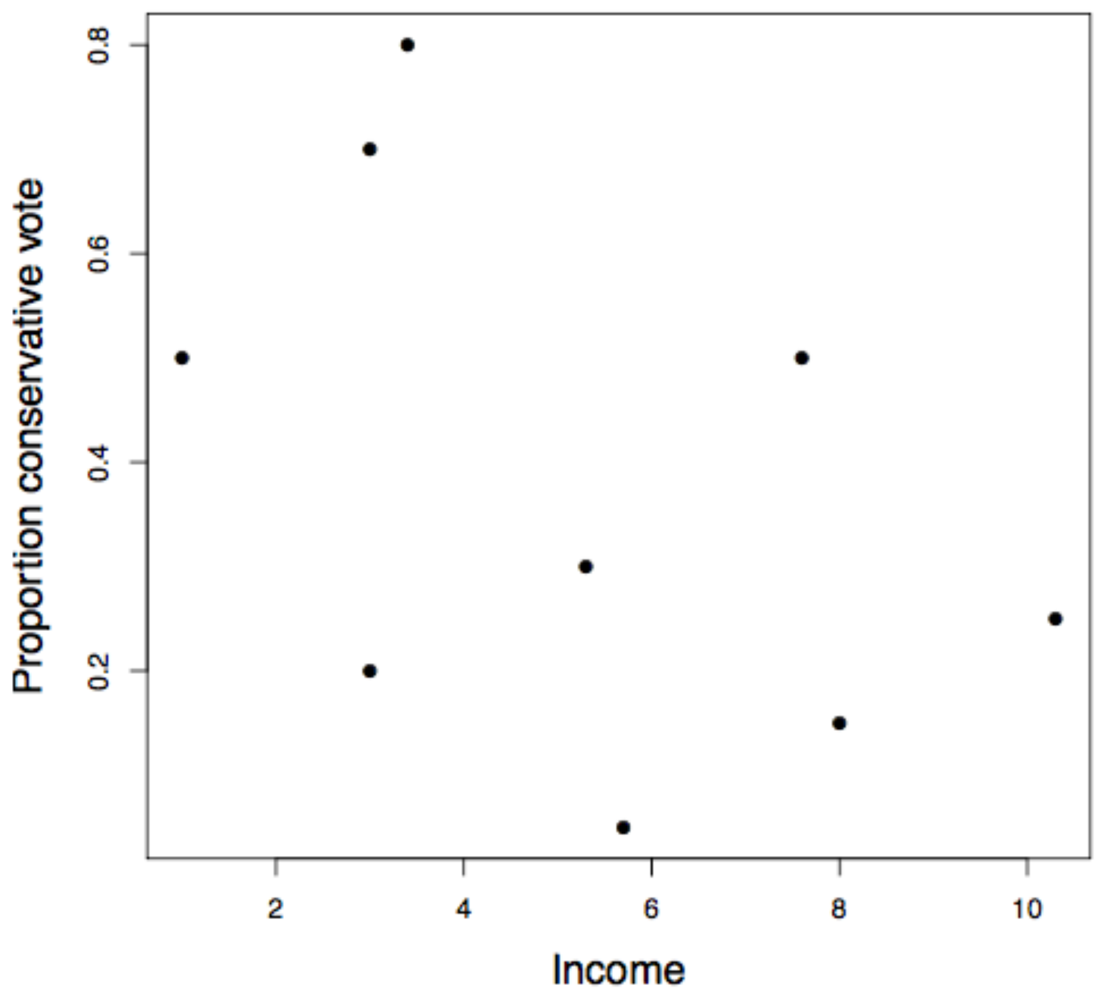
How does vote depend on income?

Proportional conservative vote



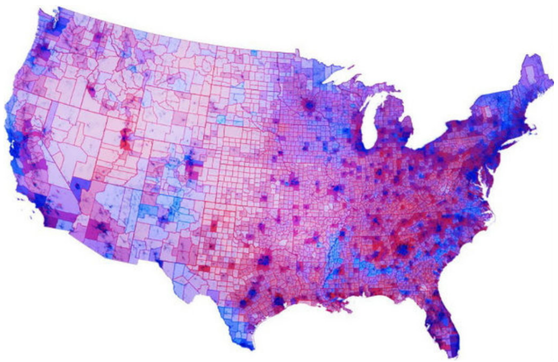
Income



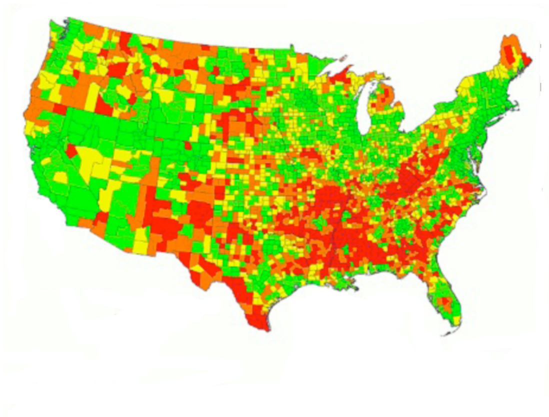


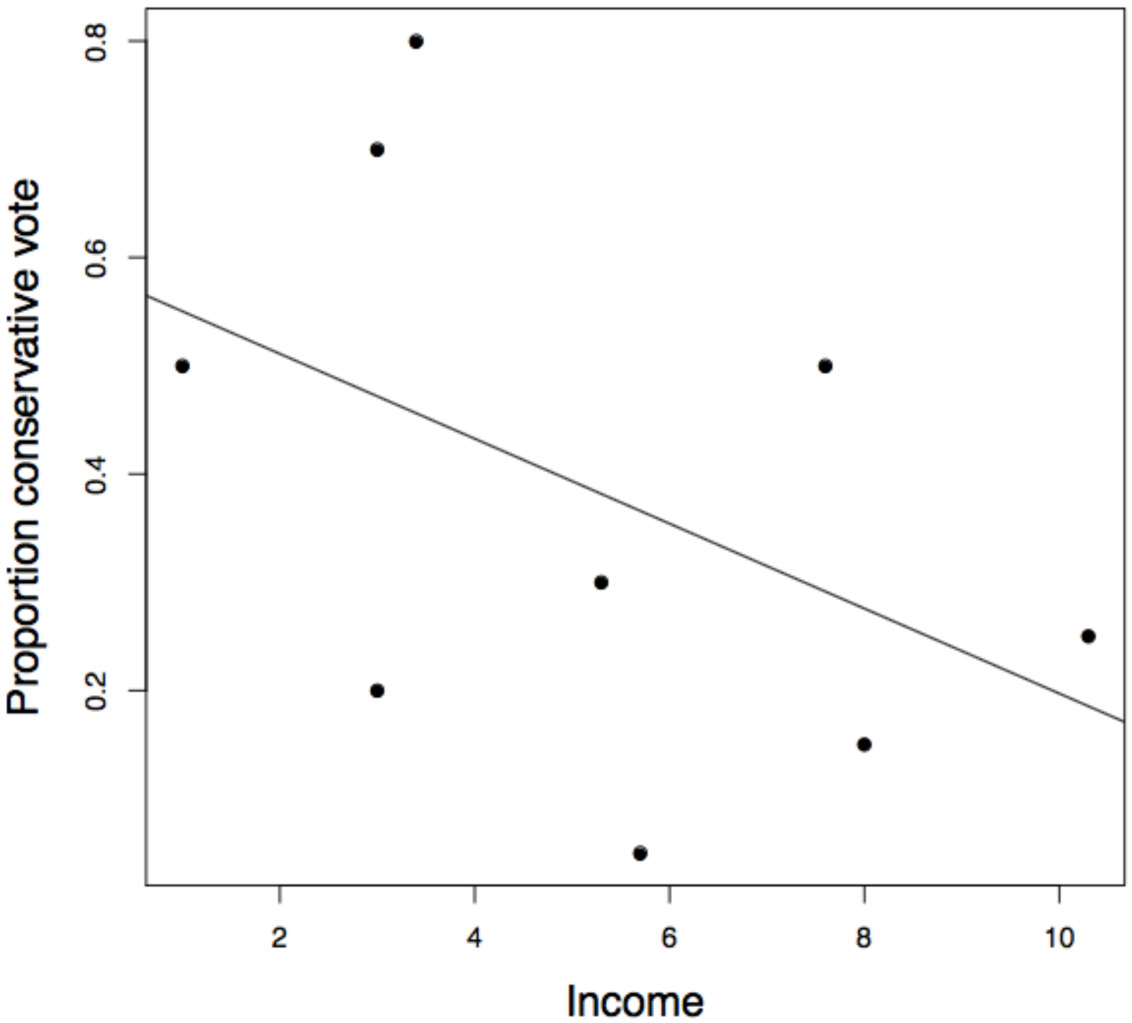
How does vote depend on income?

Proportional conservative vote



Income

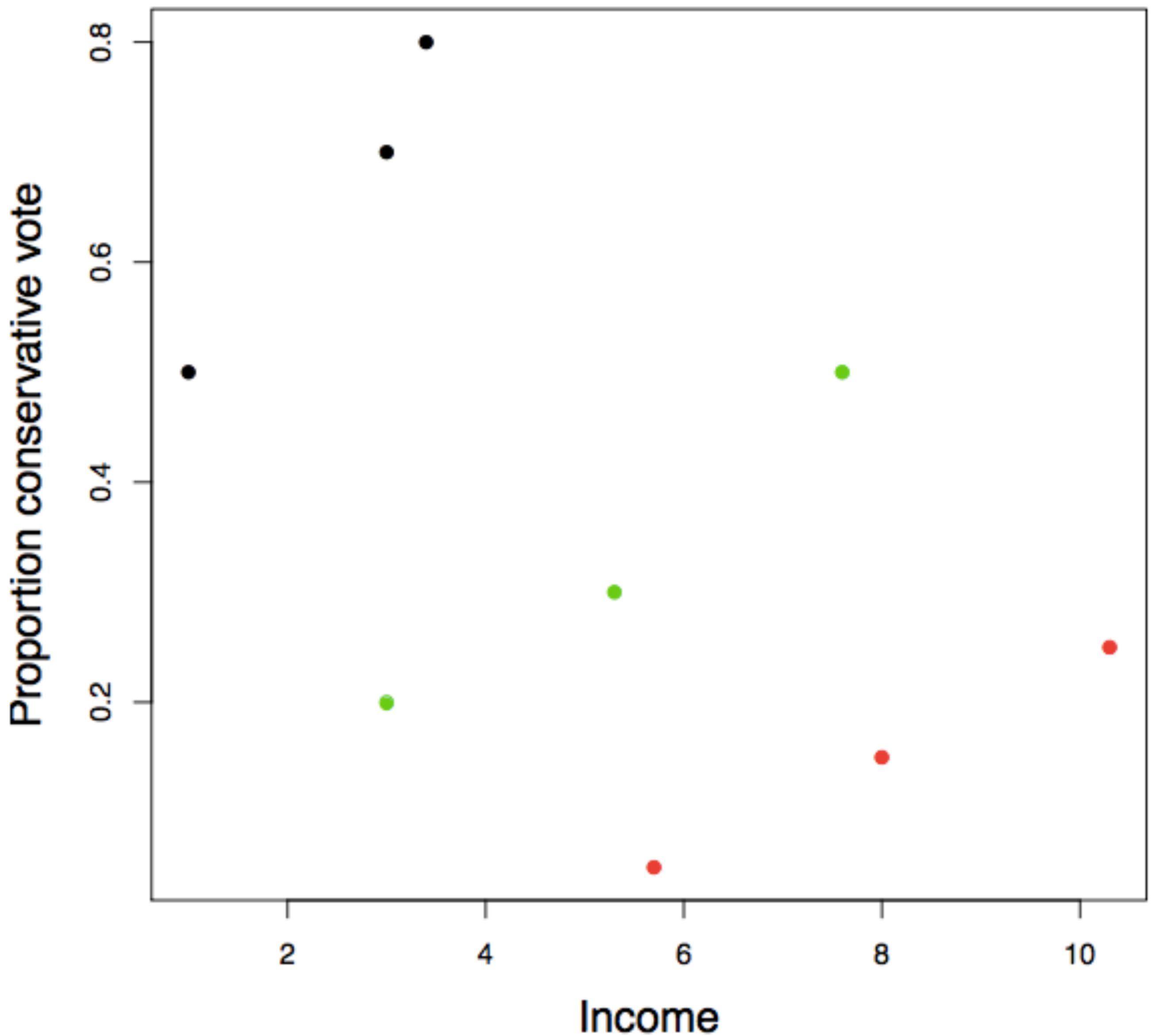




How does vote depend on income?

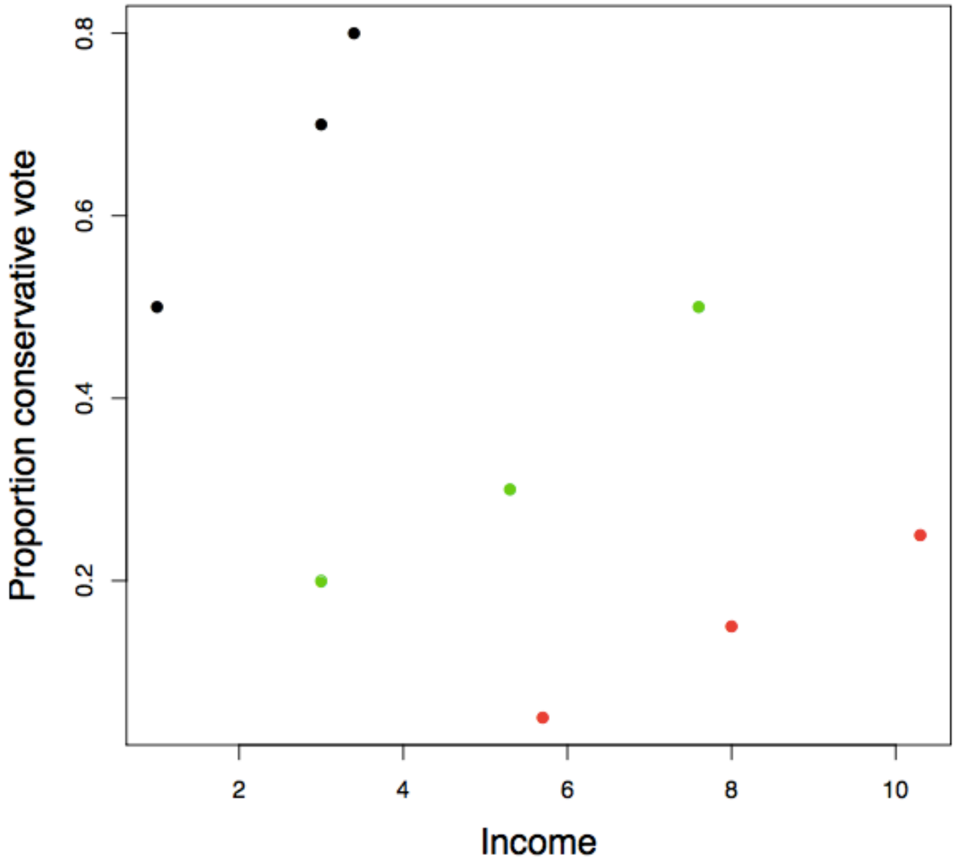
$$\text{cons} = 0.59 \;+\; -0.04 * \text{income}$$

Add another predictor: region



Model with region + income

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

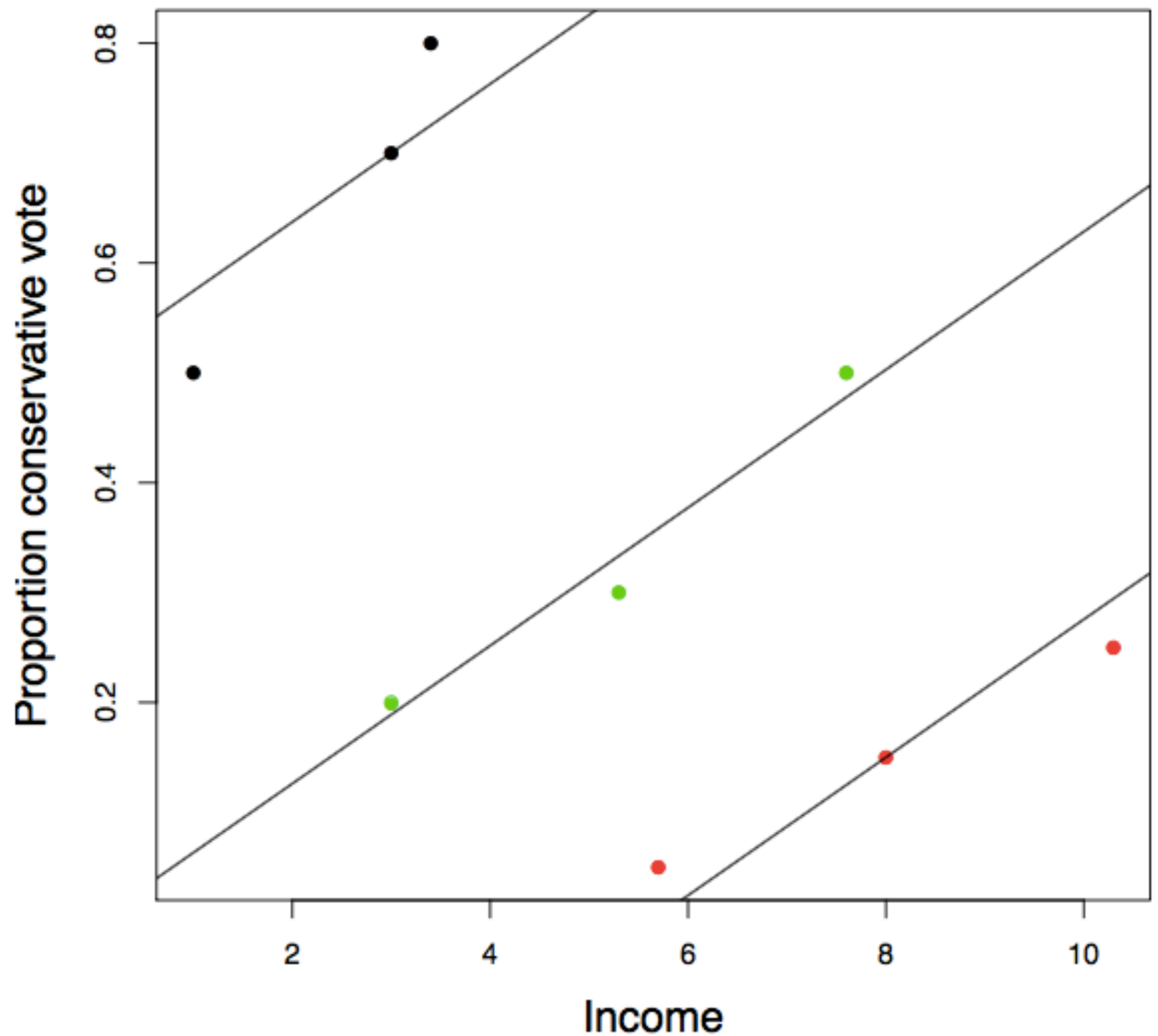


Model with region + income

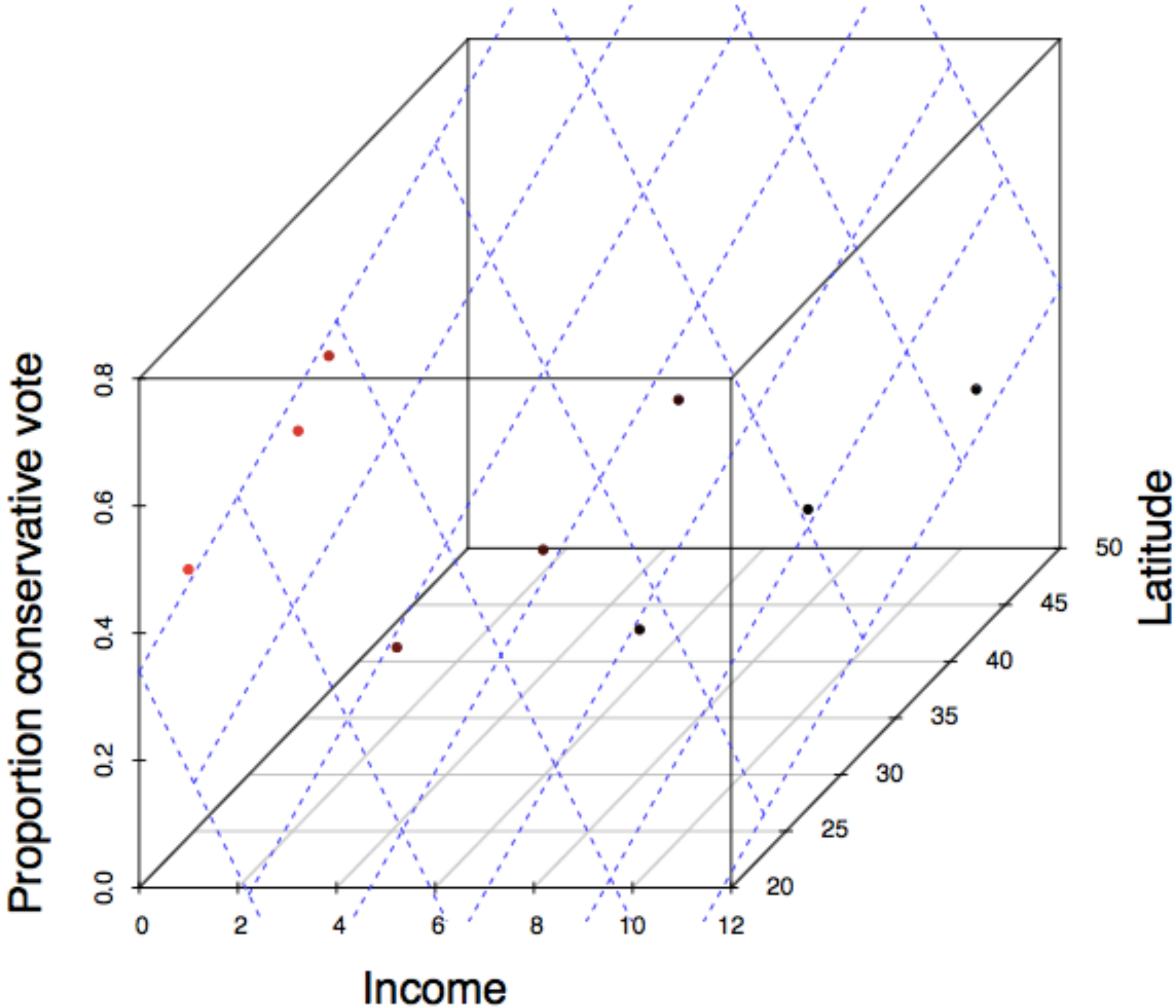
$$\text{cons} = 0.51 - 0.51 * \text{green} - 0.86 * \text{red} + 0.06 * \text{income}$$

- 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



Controlling for Latitude



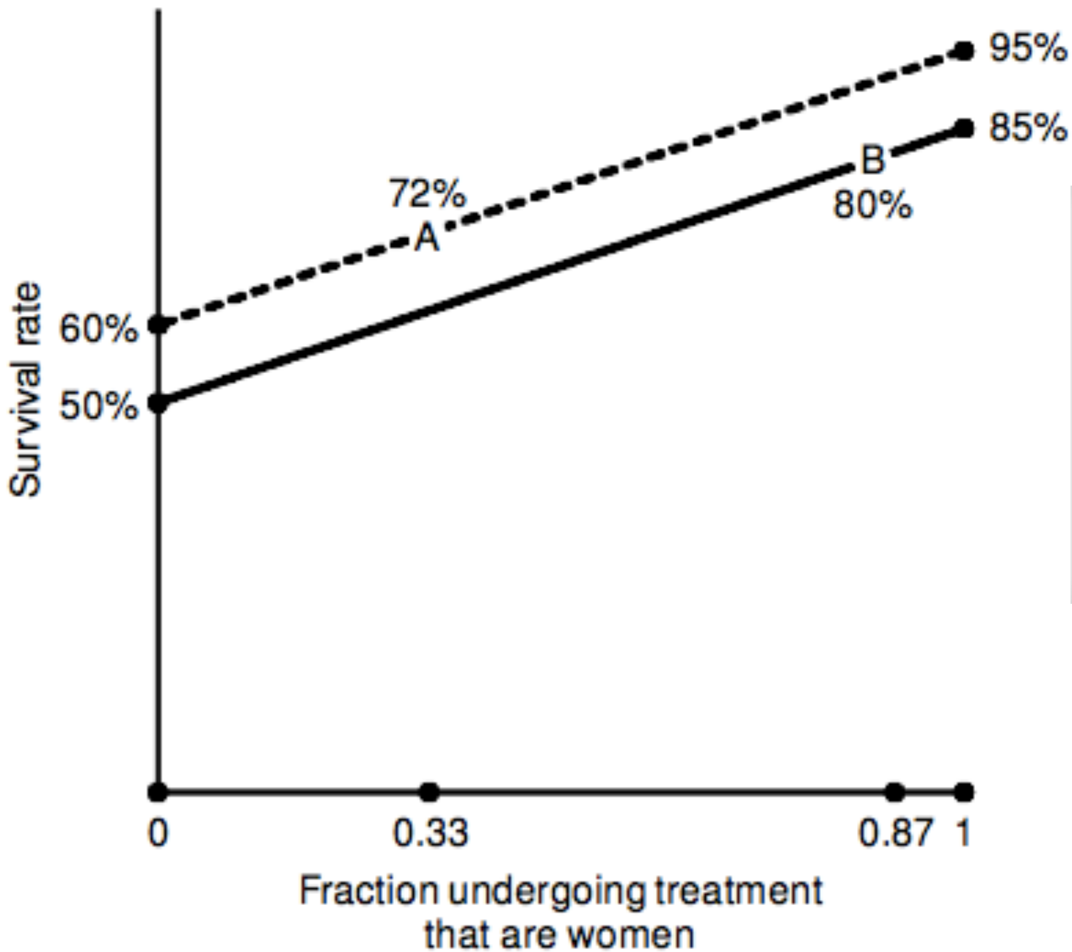
Simpsons Paradox

$$cons = 1.4 + 0.14 * income - 0.05 * latitude$$

- Adding a control variable (or variable of interest) can change the sign of your regression coefficients

Simpsons Paradox is a real problem

“Good for Women, Good for Men, Bad for People”



Drug A > Drug B

Simpsons Paradox Examples

Stanford University

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

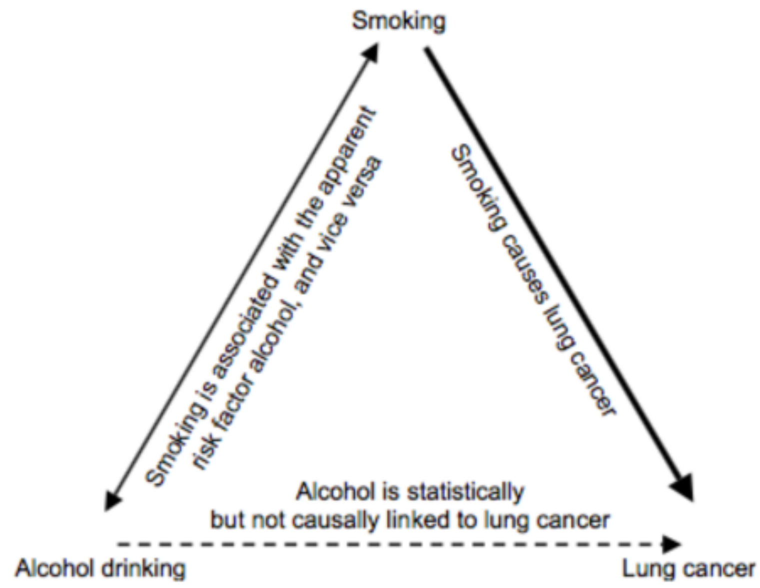
1973

Overview: Regression

- Covariance and Correlation
- Correlation vs Causation
- Linear Regression with two variables
- Mean Squared Error
- Binary Predictor
- Multiple Predictors
- Simpson’s paradox
- **Interpreting regression coefficients**

Interpreting 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

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

Course Schedule

`v05.09.2017` (subject to change – always check the latest version!)

#	Date	Lectures (Theory - Willem)	Date	Video Lectures (Applications - Chris)	Video Practicals & Notebooks
1	29-08	Introduction to Data Mining	31-08	Introduction to Data Science	Introduction to jupyter, pandas, and scikit-learn
2	05-09	Regression	07-09	Representing Data: Vectors, Types, Databases	Handling & Interpreting Data, Plotting
3	12-09	Classification	14-09	Working with Text Data Part 1 (17-09)	DIY Pandas + scikit-learn
4	19-09	Algorithm Fitting & Tuning	21-09	Working with Text Data Part 2	**No practical** -> time to prepare for midterm.
5	26-09	**Midterm**	28-09	Best Practices, Common Pitfalls & Research	Preprocessing + Pipelines, MNIST Challenge
6	03-10	Data Reduction & Decomposition	05-10	Mining Massive Data, Ensemble Methods	Online / Out-of-Core Learning
7	10-10	Time Series Analysis	12-10	Applications of Deep Learning	Social Media and Multi-modal Data
8	17-10	Clustering and Graphs	19-10	Explaining Models, Ethics, Privacy	Unsupervised Learning: Intuitions and Metrics