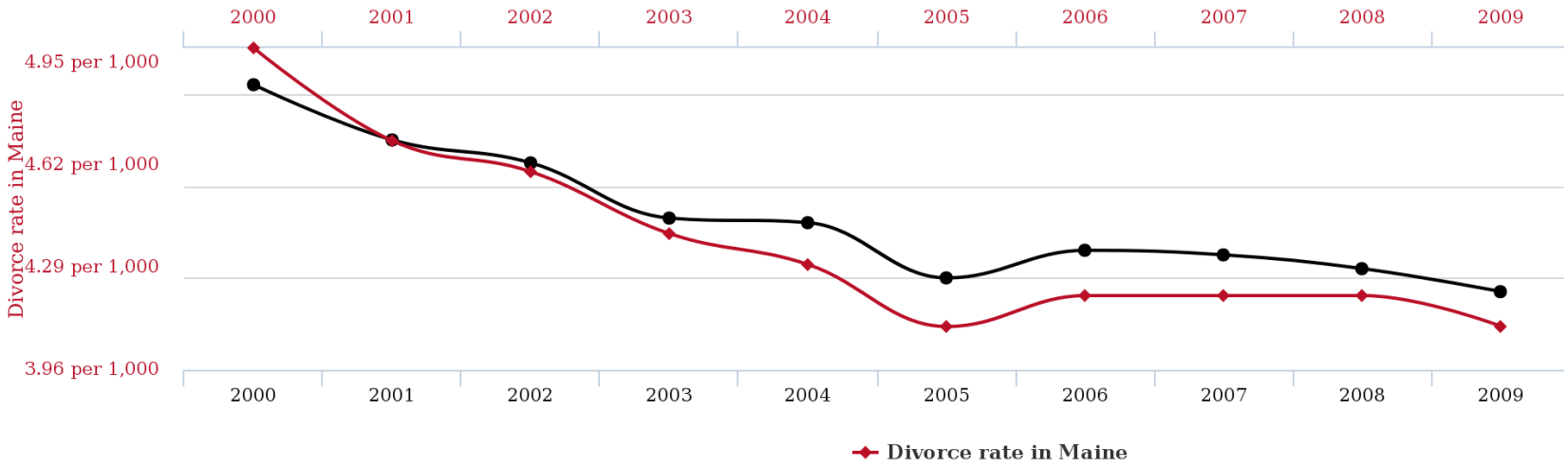
The red line indicates the divorce rate in Maine from 2000 to 2009.

What do you think the black line might represent?



Attendance Recording

- Your attendance is automatically recorded for each lecture on Zoom.
- 70% attendance for each module is required by Home Office, if you are holding a Tier-4 visa.
- You should try to attend all lectures. Please contact me if you are unable to attend any lectures due to time differences.

Consent to record

- We will record each live lecture on Zoom, and videos will be shared on the Moodle page.
- If you have any concerns on the recording, please let me know.
- *When a recording is started by the host, you will be notified and can choose to accept or leave the session.*

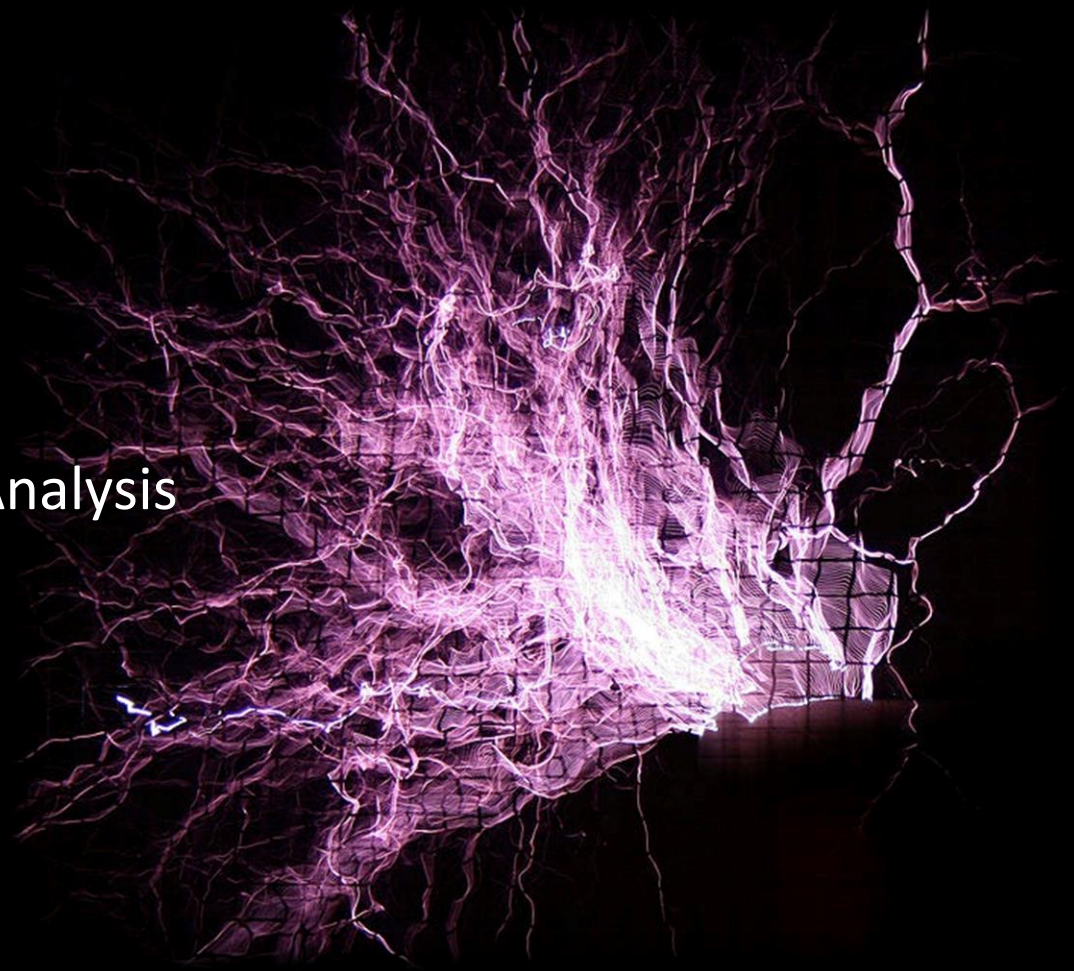
CASA0007: Quantitative Methods

Dr Huanfa CHEN
huanfa.chen@ucl.ac.uk

Dr Hannah Fry
hannah.fry@ucl.ac.uk

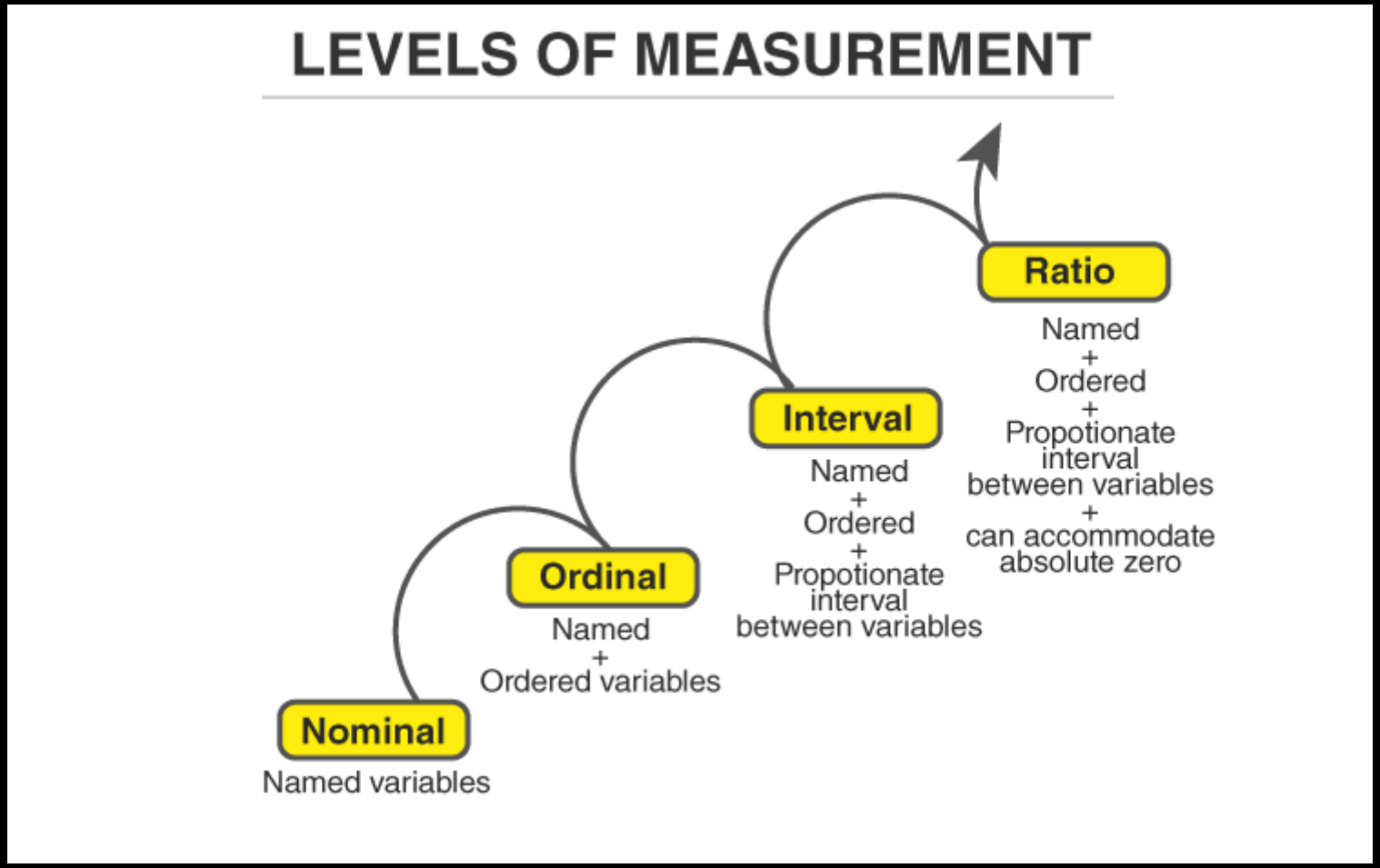
Moodle password: QM2020

Centre for Advanced Spatial Analysis



https://commons.wikimedia.org/wiki/File:Tesla_coil_Vienna_Technical_Museum_2007_a.jpg

Types of Data : Level of measurement

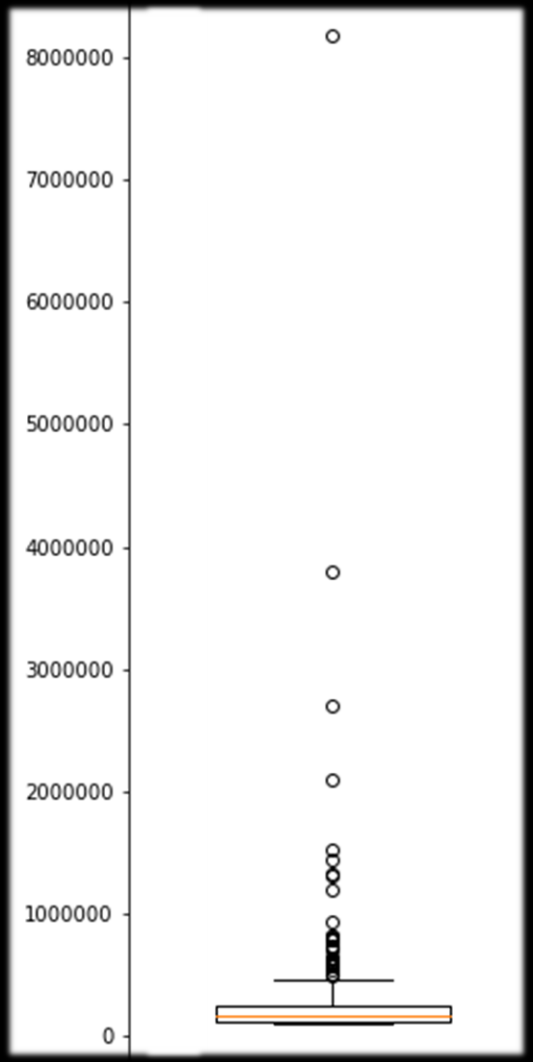


Note: the level of measurement is difference from how they are coded in programming

Developed by psychologist Stanley Smith Stevens
https://en.wikipedia.org/wiki/Level_of_measurement

Data Basics: Summary Statistics

Pop.



Quantity

282 Values

Location

Mean: 303,000

Median: 168,000

Mode: 106,000

Spread

Variance: 3.47×10^{11}

St. Dev.: 589,000

Range: 8,075,000

Boxplot Stats

Min: 100,000

Max: 8,175,000

LQ: 122,000

UQ: 259,000

IQR: 137,000

Lo. Outlier Lim: -84,000

Hi. Outlier Lim: 466,000

(All to nearest 1000)

Data Basics: distribution

Continuous data

- Normal distribution
- Power law distribution
- Exponential distribution

Discrete data

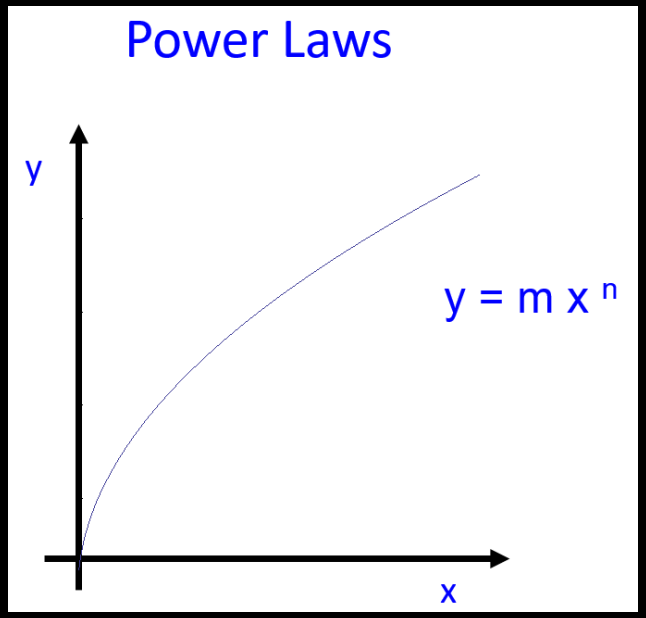
- Poisson distribution (e.g. how many cars pass my office every 10 minutes?)

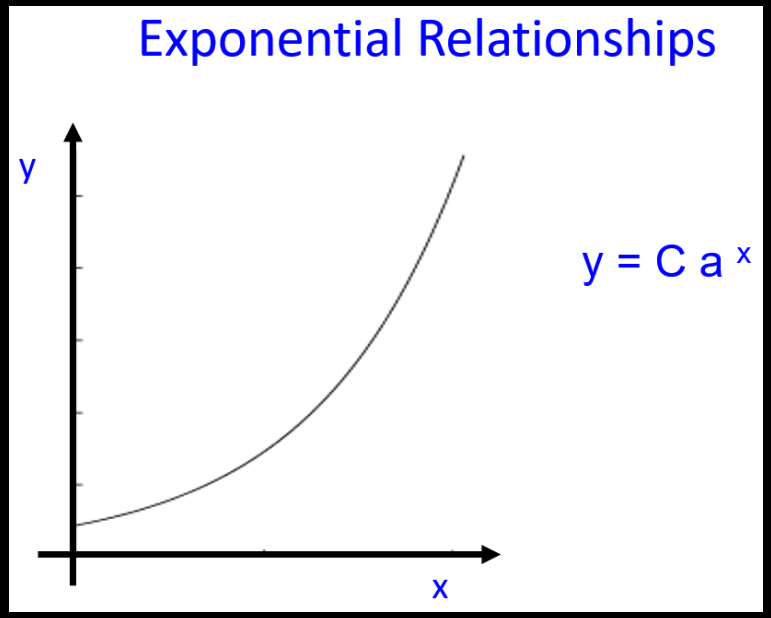
Data Basics: relationship between two vars

- Linear relationship: $y = ax + b$
- Power law relationship
- Exponential relationship

Question:

Is there any metric that evaluates the strength of relationship?





Week 2 – Comments

Very useful comments. Thanks!

Motivations for data transformation

1. To formulate a different question

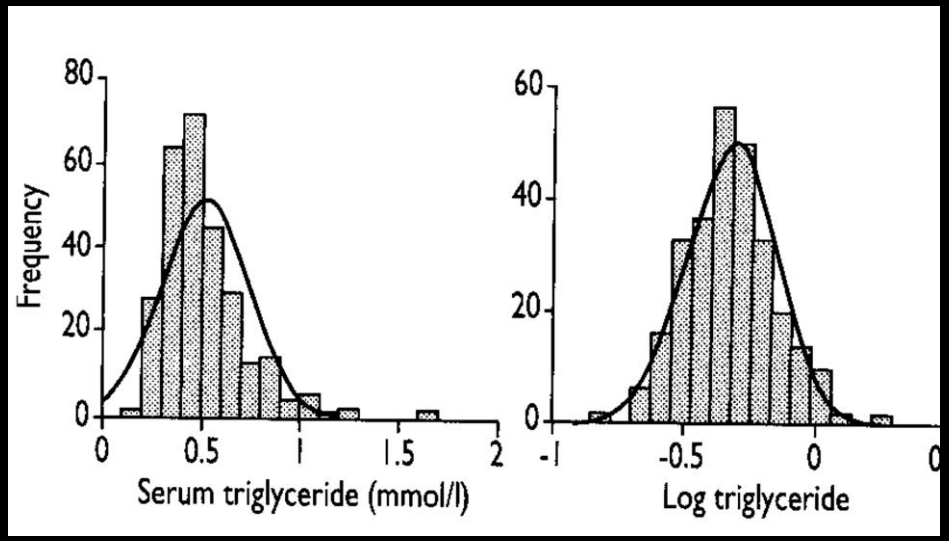
House price vs. number of rooms

1) Original price (absolute change): $Y = aX + b$
2) Log-transformed price (relative change): $\log Y = aX + b$

Motivations for data transformation

2. To meet the assumptions of a statistical inference procedure (today's topic)

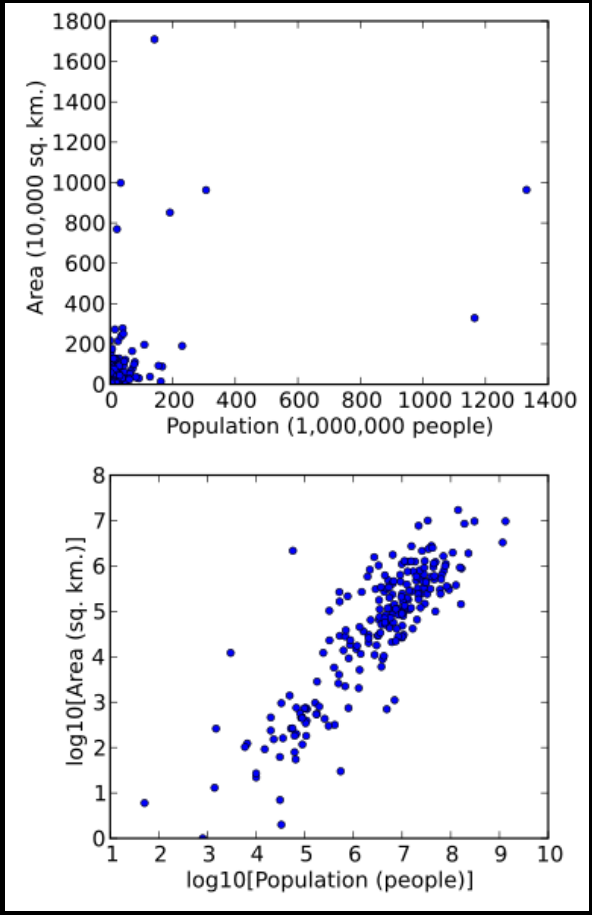
Skewed data into a approximately normal distribution



https://www.bmj.com/content/312/7038/1079

Motivations for data transformation

3. To make data easier to visualise



https://en.wikipedia.org/wiki/File:Population_vs_area.svg

Lecture 2 – Assignment

1) Watch Albert Bartlett's lecture on the exponential function:

http://www.youtube.com/watch?v=F-QA2rkpBSY

(You only need to watch the first 20 minutes or so.)

Read the following articles:

http://news.bbc.co.uk/1/hi/magazine/8000402.stm
http://www.telegraph.co.uk/news/earth/energy/oil/9867659/Why-the-world-isnt-running-out-of-oil.html

Consider what we should conclude about quantitative predictions?

Lecture 2 – Assignment

2) Have a play with the excel data investigation

It will guide you through how to plot and explore data in Excel.

Lecture 2 – Assignment

3) Look at the relationships data sets on moodle.

What do they reveal about the underlying relationships?

Three types of outliers

1. Error
2. Points not following the pattern
   - If you are studying the overall pattern, they need to be removed
   - e.g. City of London (in London)
   - e.g. Vantican City (1000 pop, gender ratio 7:1)

   "Statistical oddities"
3. Points that are essential to the overall pattern
   - Don't remove them
   - e.g. New York in US pop data

https://en.wikipedia.org/wiki/Vatican_City

Dealing with outliers

- There are many approaches to identify outliers (e.g. Tukey fences), but determining an outlier is a subjective exercise.
- Dealing with outliers (not sure type 2 or 3)
  - Retention, but the application should use methods that are robust to outlier points
  - Exclusion. If some outliers are excluded, the reasons should be clearly stated on the report.

Course Objectives

You should...

1. Encounter and develop an understanding of a broad range of quantitative techniques and concepts.

2. Be equipped to expand on and develop these skills through individual research.

3. Become confident in formulating a coherent quantitative argument, built around the results of the techniques encountered in the course.

Week 1: Introduction to Quantitative Problems

Week 2: Approaching & Communicating Data

Week 3: Measuring Relationships

Week 4: Advanced Regression

Week 5: Hypothesis Testing

READING WEEK

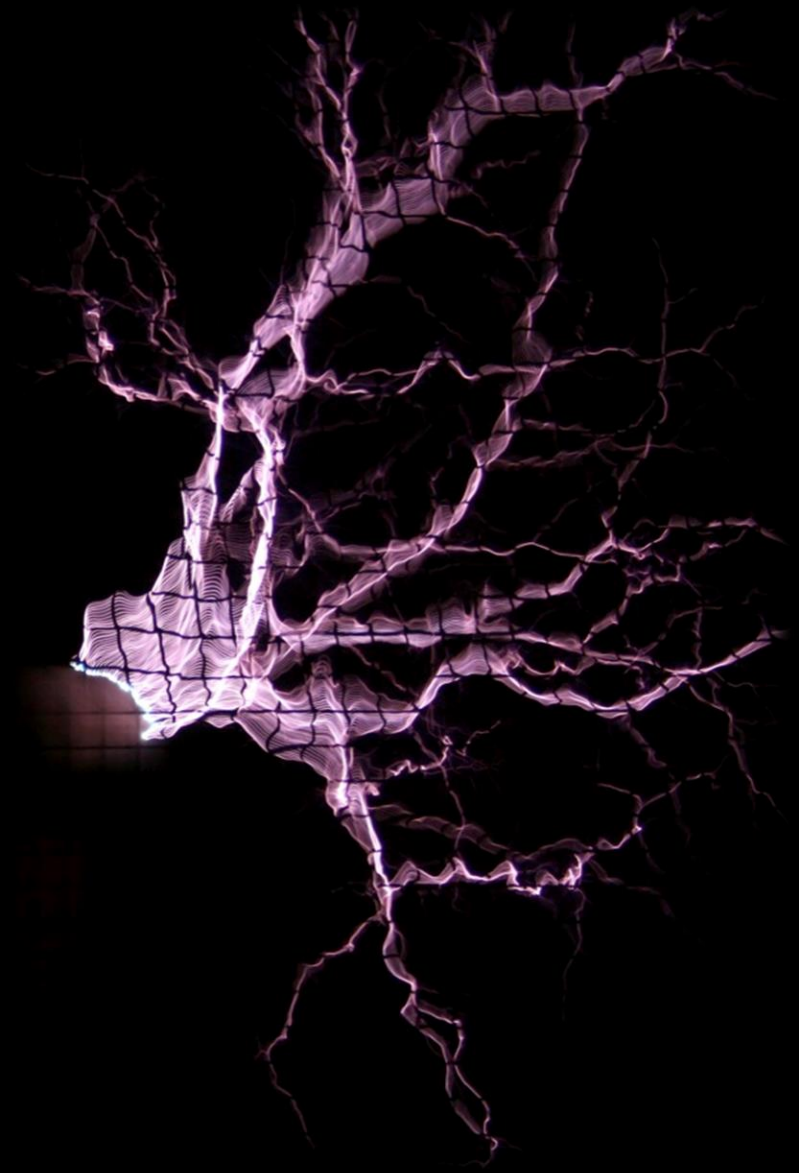
Week 6: Cluster Analysis

Week 7: Optimising Limited Resources

Week 8: Modelling the World

Week 9: Statistical Traps & Advanced Topics

PRESENTATION WEEK



LECTURE 3

Measuring Relationships

https://commons.wikimedia.org/wiki/Category:Tesla_coils#/media/File:Tesla_coil_Vienna_Technical_Museum_2007_b.jpg

OBJECTIVES

1. Understand the concept of **covariance**, **correlation** and what it says about data relationships.

2. Be able to measure a data relationship using simple **linear regression**.

3. Learn how regression can be extended for data with more than two dimensions.

4. Understand how to identify whether additional data provides additional information.

https://commons.wikimedia.org/wiki/File:Tesla_coil_Vienna_Technical_Museum_2007_c.jpg

Part 1: Covariance, Correlation, and Association

Variability of a single variable

Data type	Metric	Visualisation
ratio/interval (continuous/discrete)	Variance, standard deviation, etc.	Boxplot Histogram
nominal	None	Bar chart
ordinal	None	Bar chart

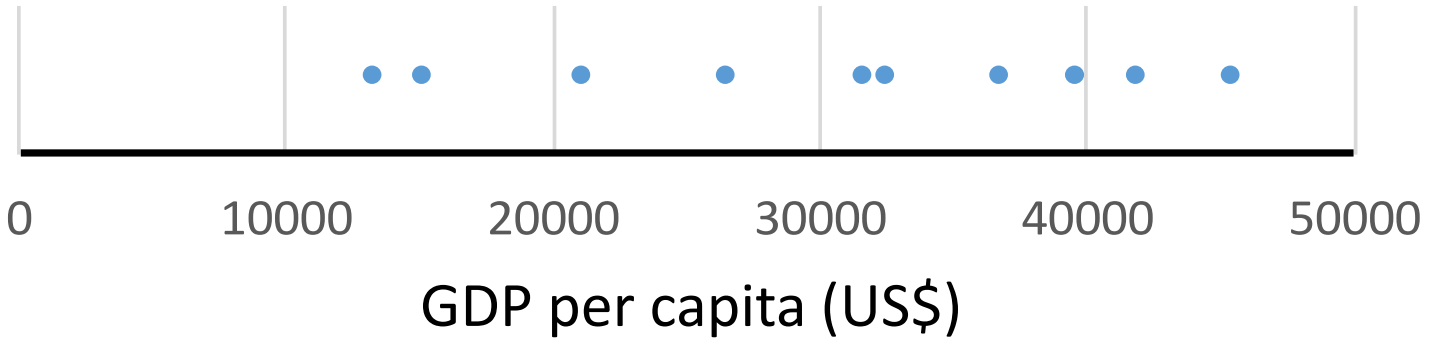
Co-variability of two variables

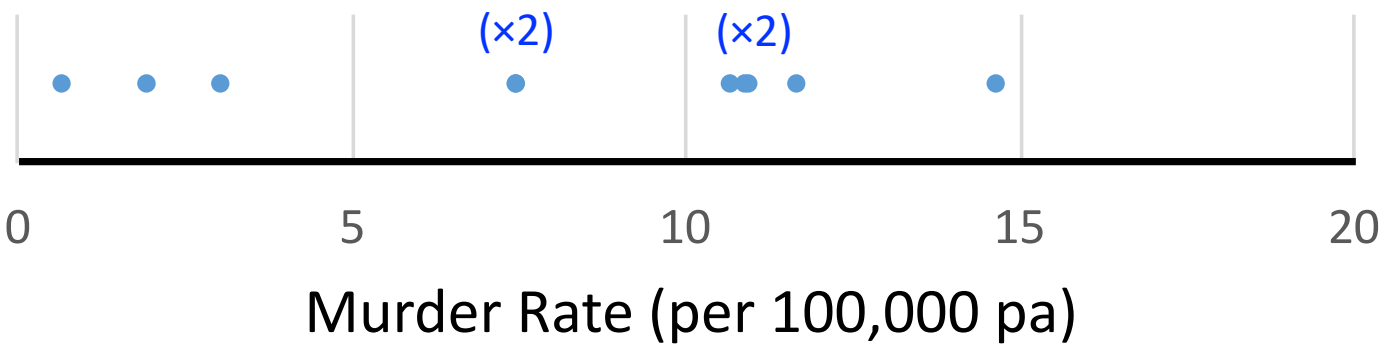
	Ratio	Nominal/ordinal
Ratio	Covariance, correlation	---
Nominal/ordinal	Analysis of Variance (ANOVA)	Confusion matrix

ANOVA (not covered here)

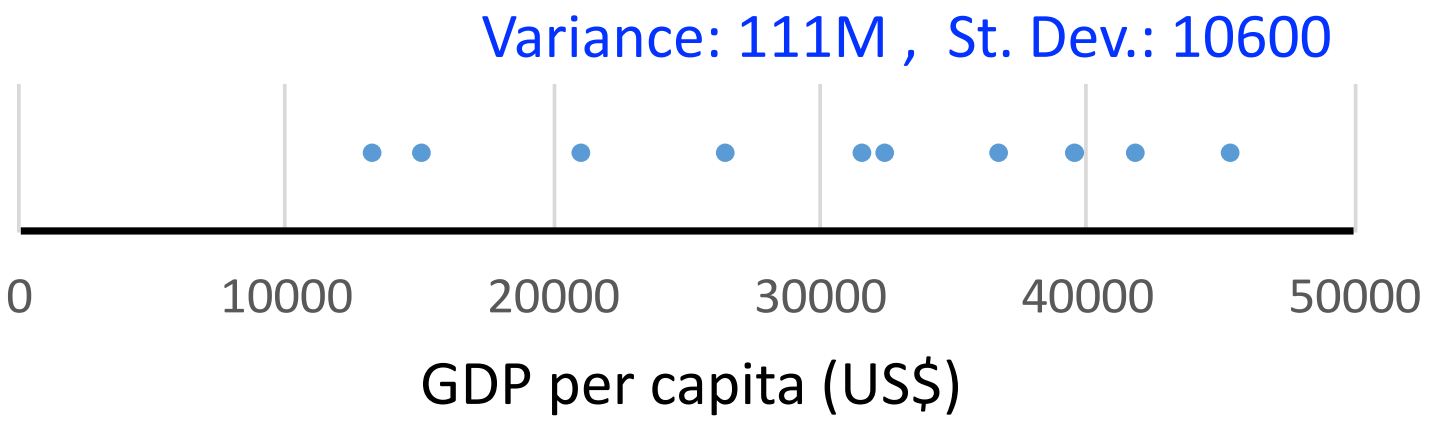
- Whether two or more population means are equal
- When there are only two populations, it is called two-sample t-test (covered in Lecture 5)
- E.g. is the mortality rate of Covid-19 the same across different cities?

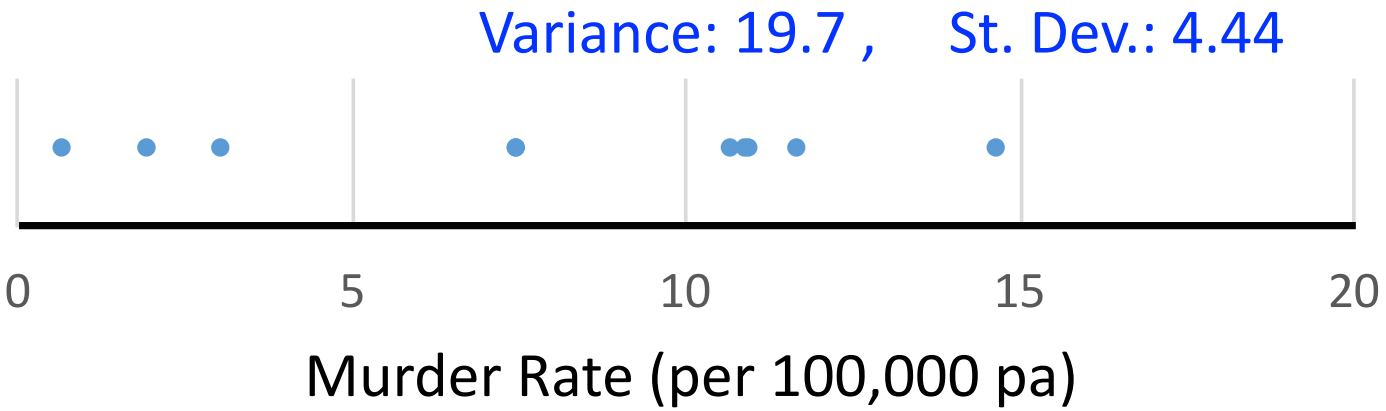
MOTIVATION

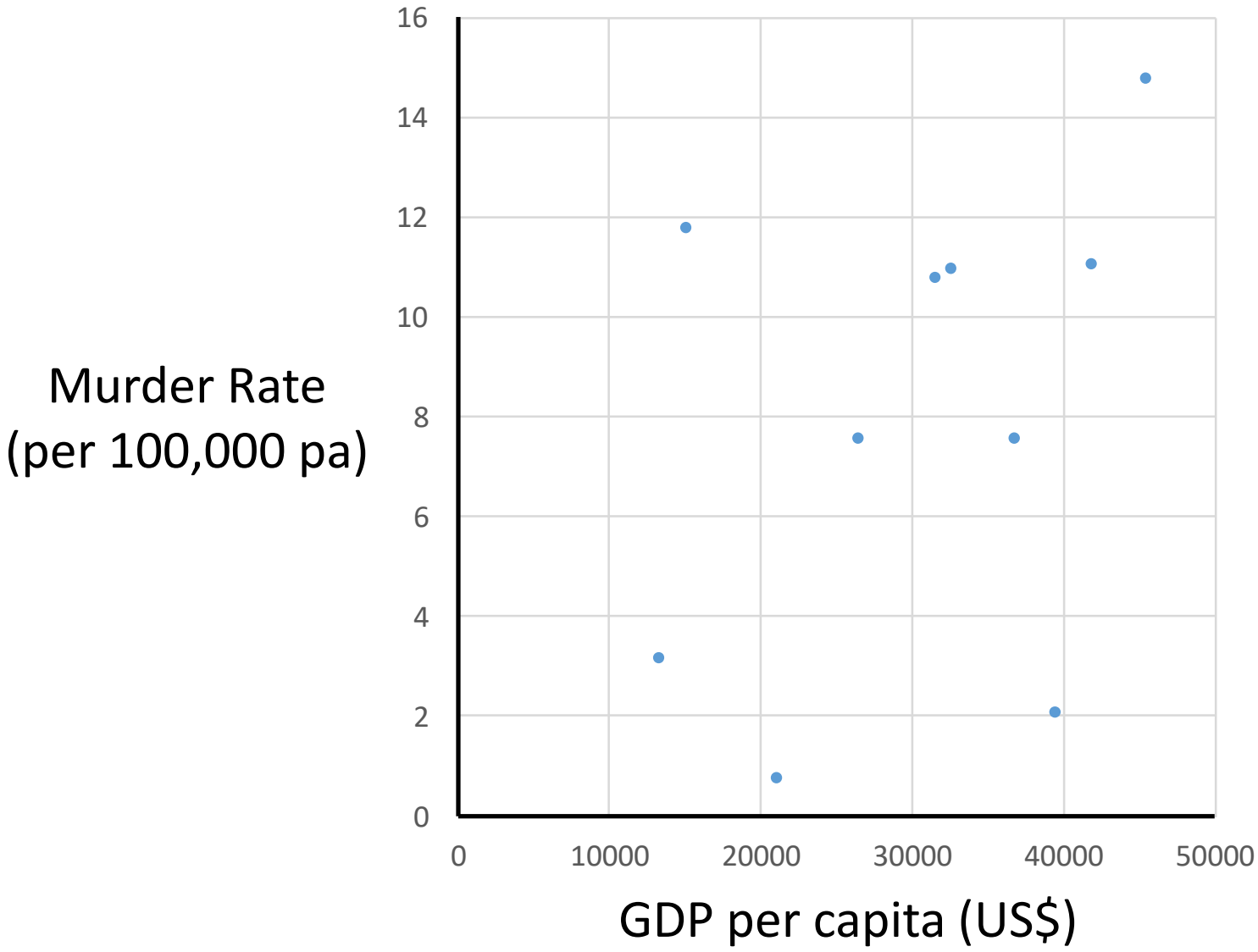

0
10000
20000
30000
40000
50000
GDP per capita (US$)

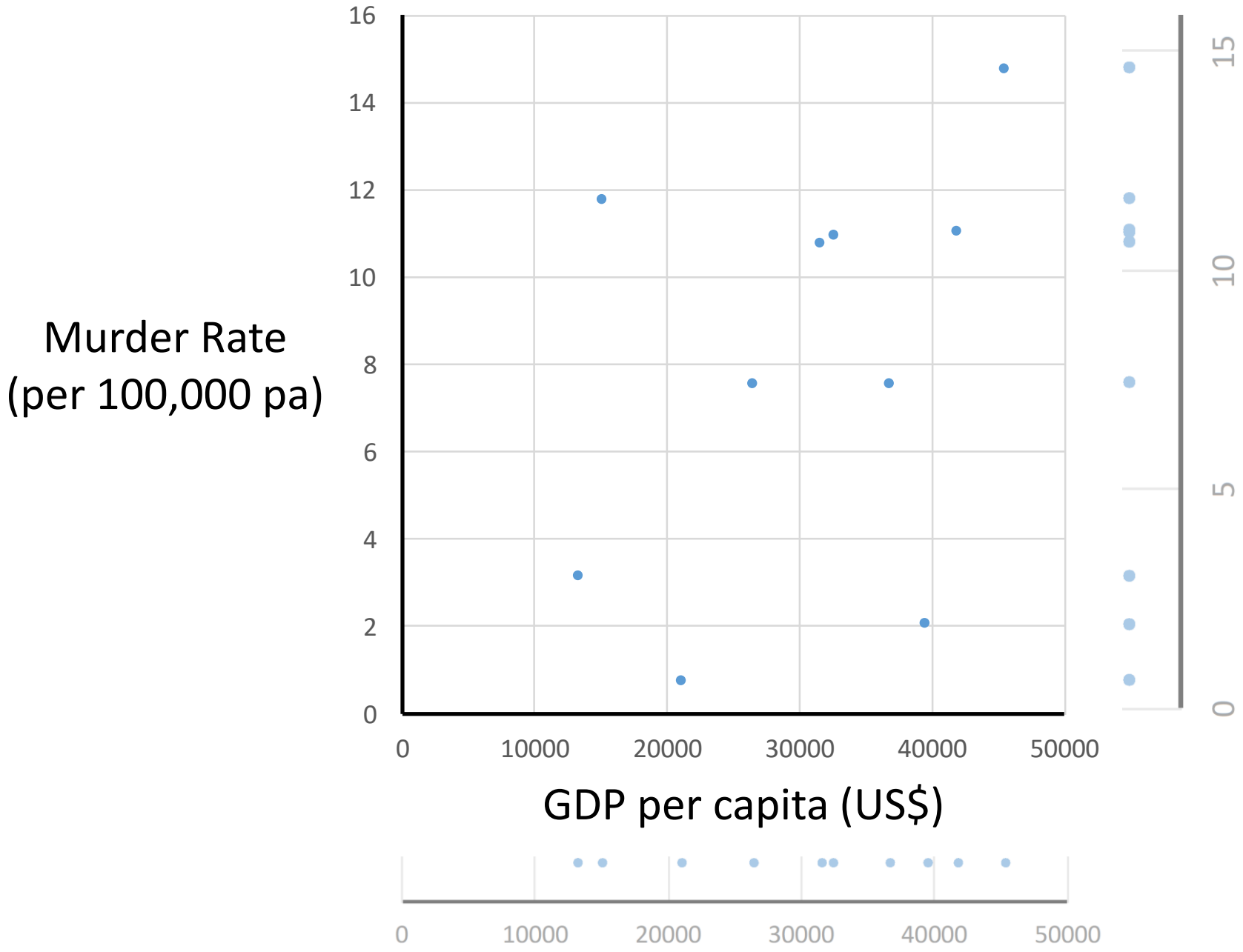

(×2)
(×2)
0
5
10
15
20
Murder Rate (per 100,000 pa)

MOTIVATION

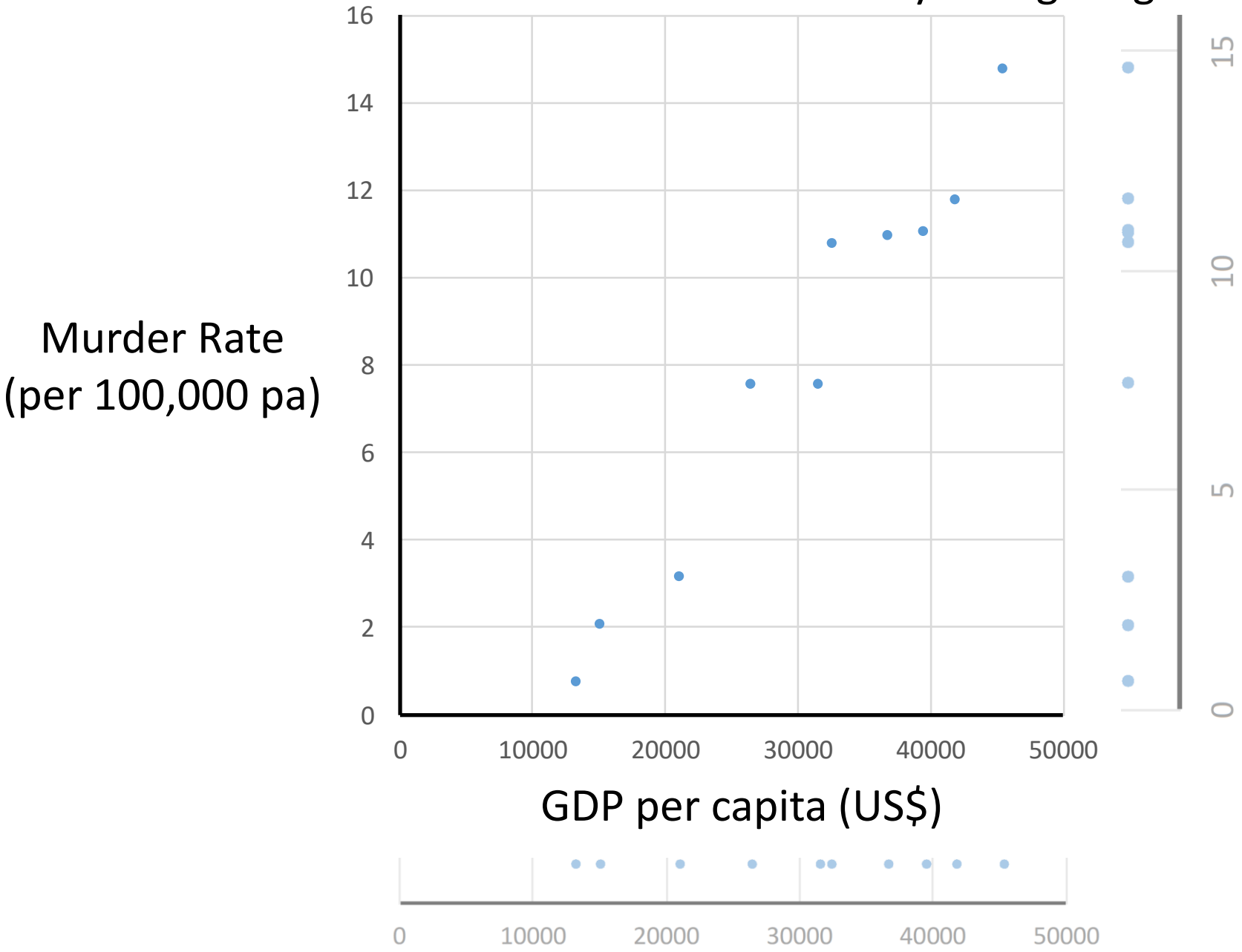

Variance: 111M , St. Dev.: 10600
0
10000
20000
30000
40000
50000
GDP per capita (US$)


Variance: 19.7 , St. Dev.: 4.44
0
5
10
15
20
Murder Rate (per 100,000 pa)

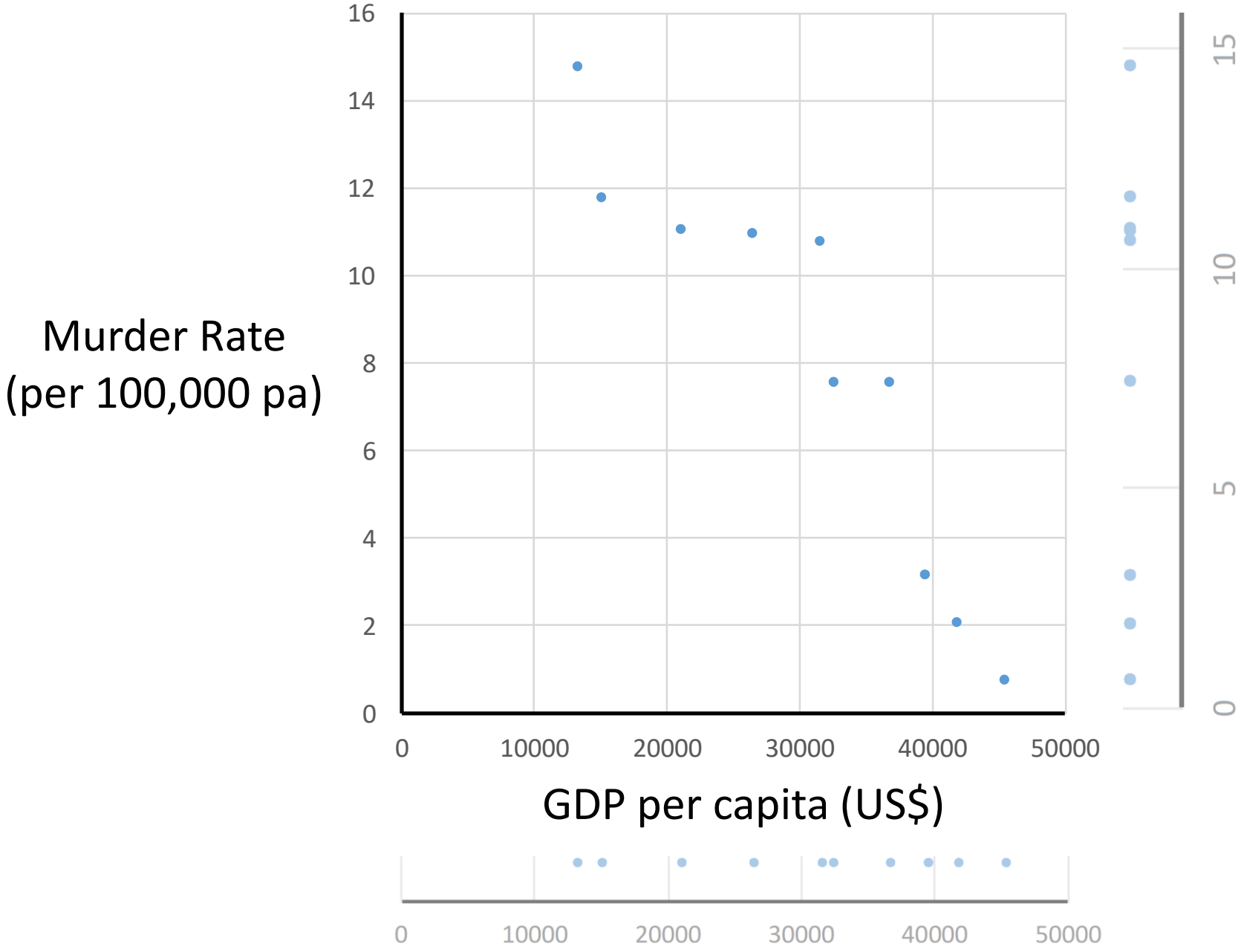

Murder Rate
(per 100,000 pa)
16
14
12
10
8
6
4
2
0
0
10000
20000
30000
40000
50000
GDP per capita (US$)


Murder Rate
(per 100,000 pa)
16
14
12
10
8
6
4
2
0
0
10000
20000
30000
40000
50000
GDP per capita (US$)
15
10
5
0

Permutations: we need to look into how x and y change together

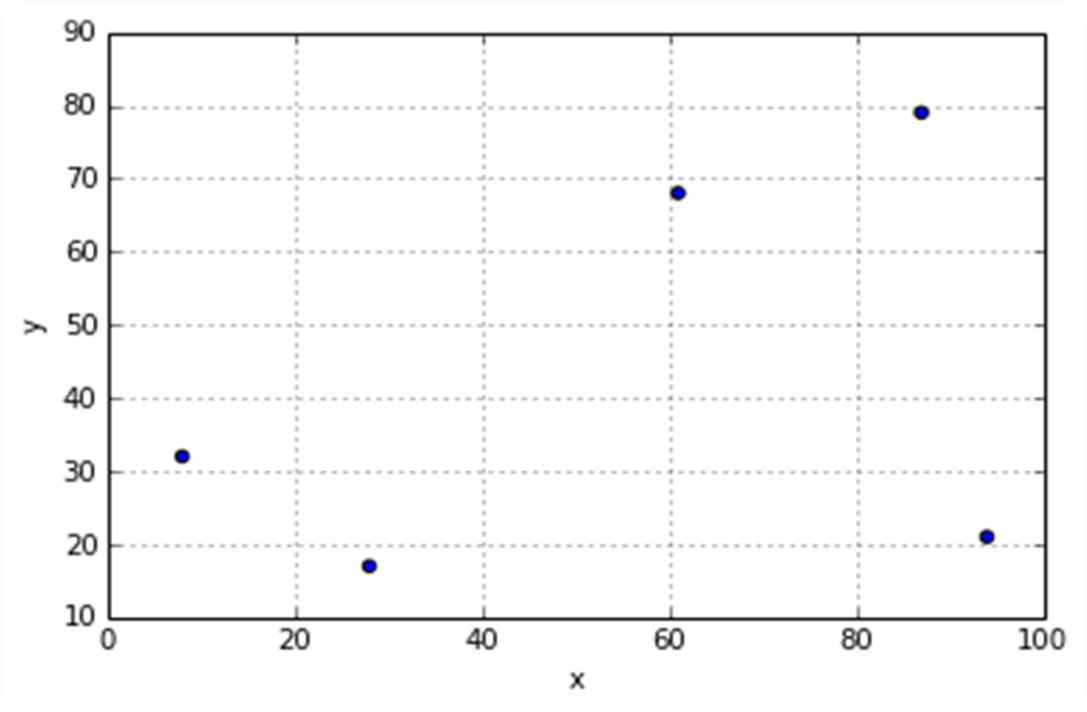


Permutations: we need to look into how x and y change together

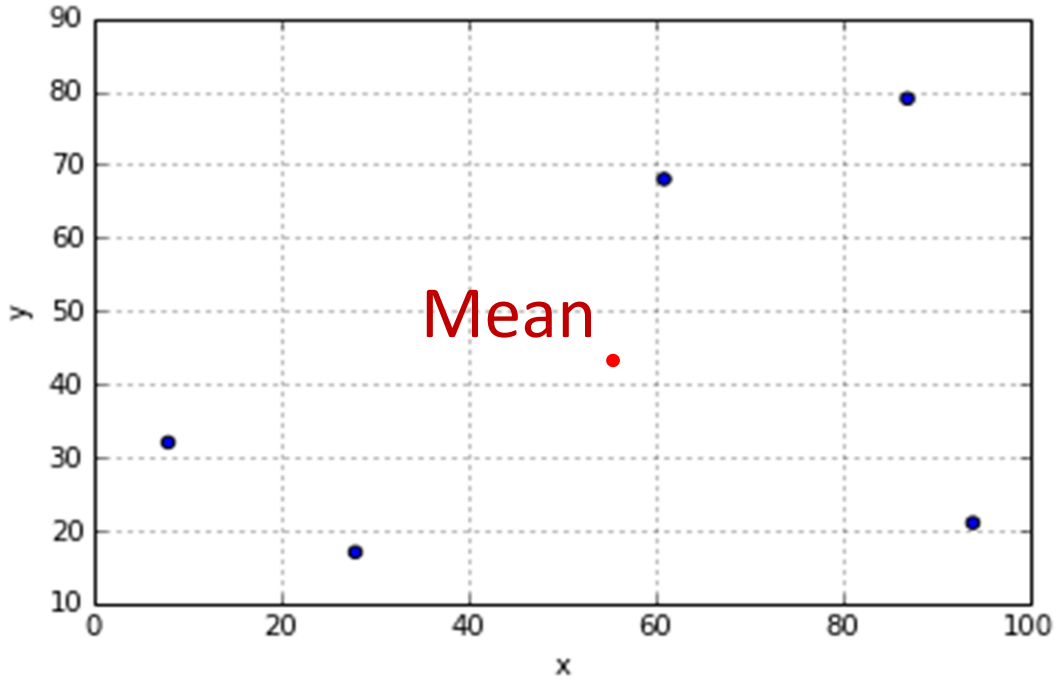


Covariance

“Do the two variables change in the same direction?”
“To what extent do they change together?”



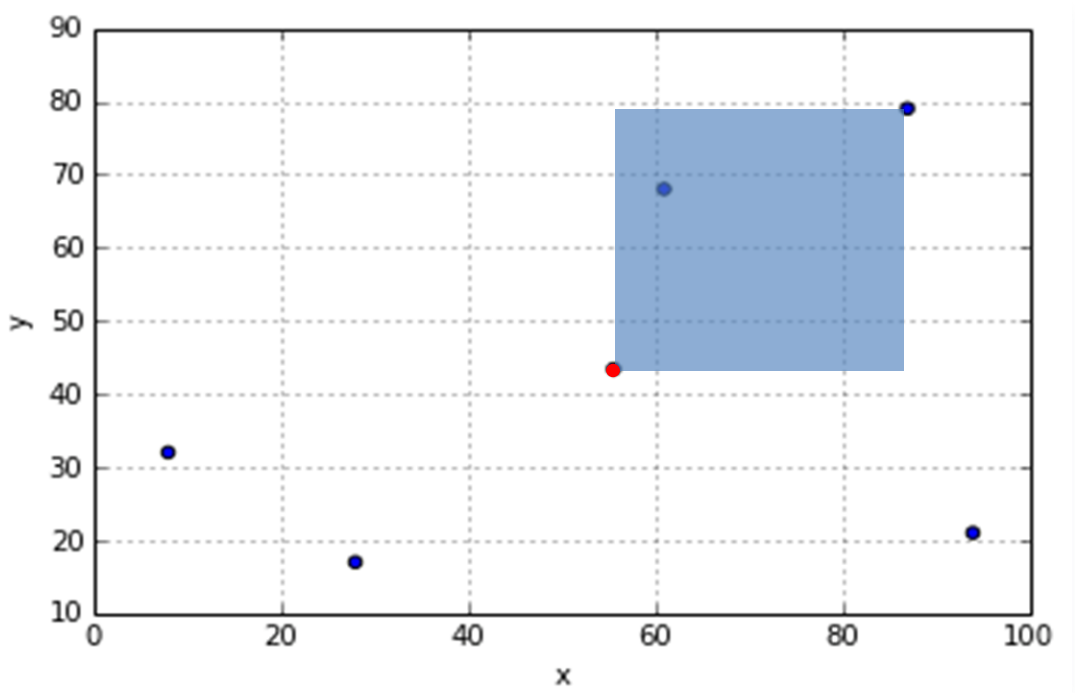
Covariance



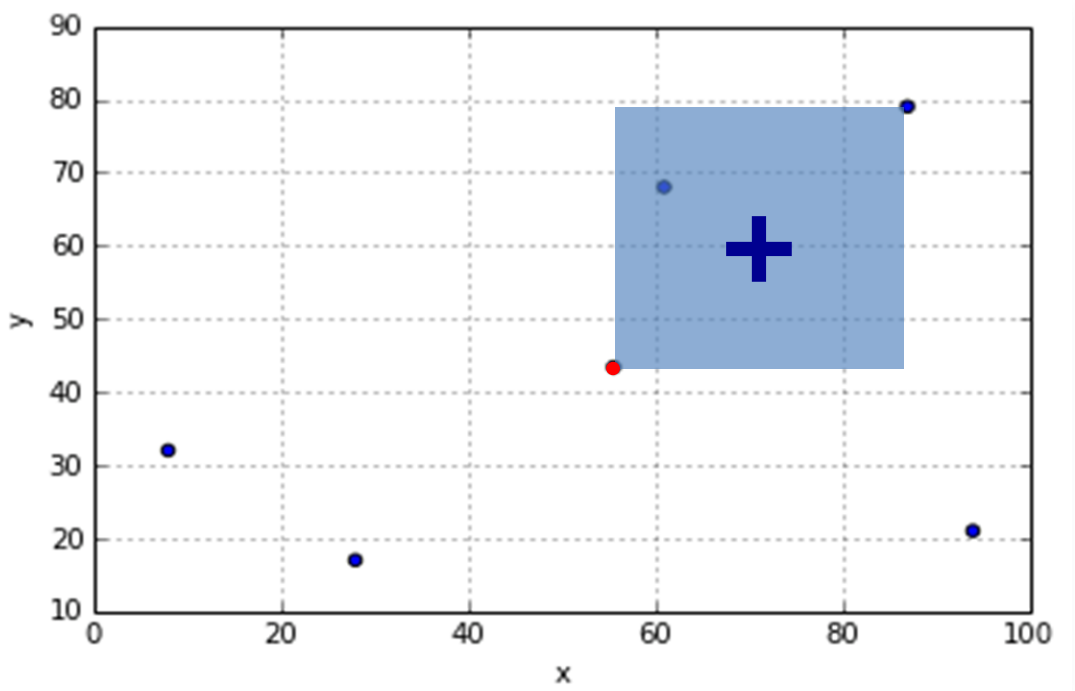
The direction of deviation on two axes

x	y	product
+	+	+
+	-	-
-	+	-
-	-	+

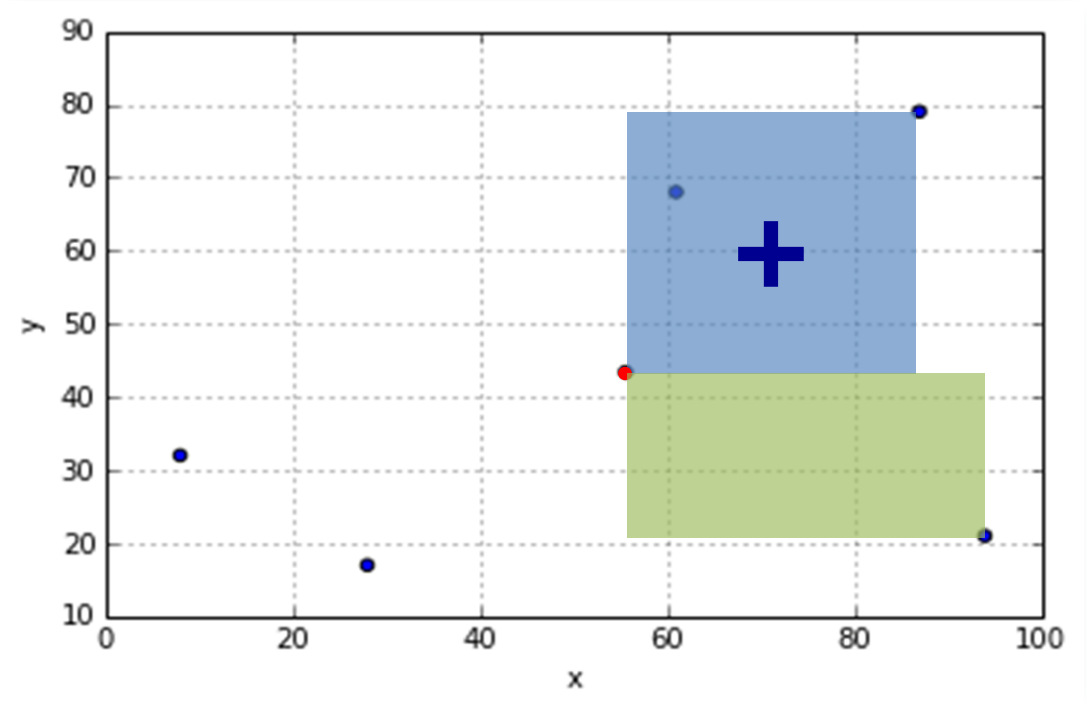
Covariance



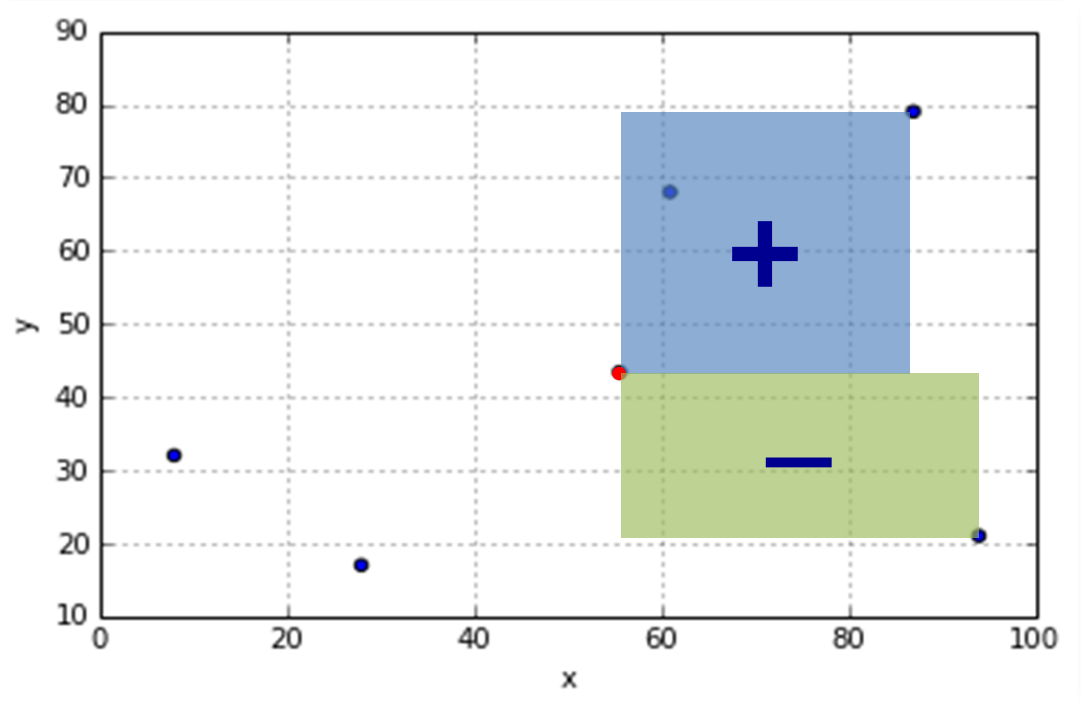
Covariance



Covariance

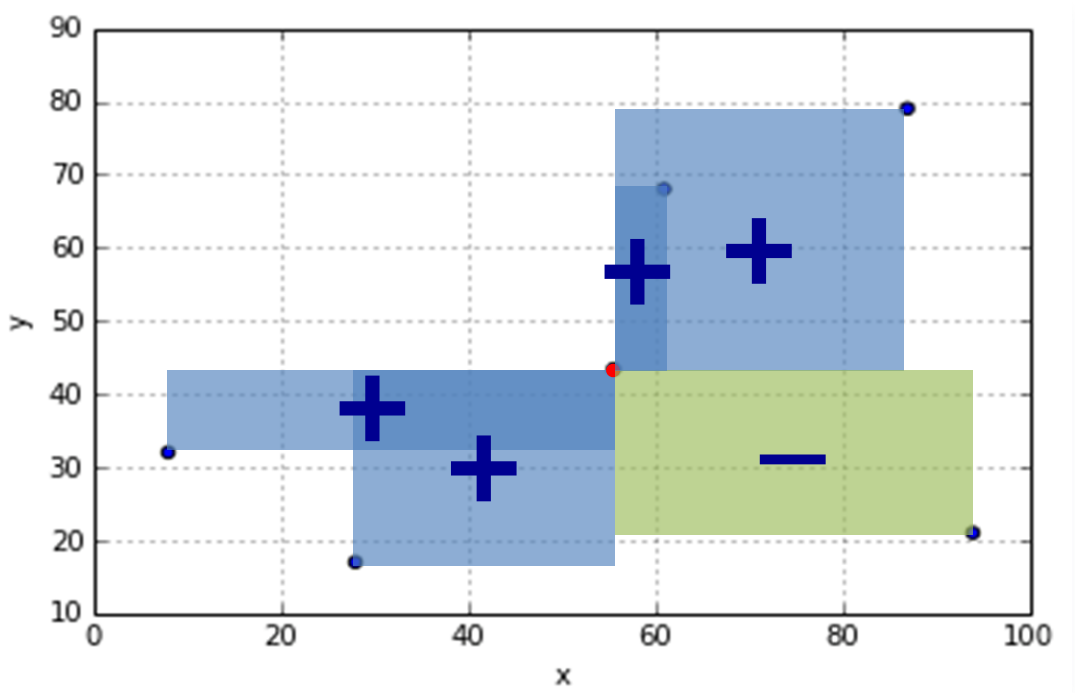


Covariance



Covariance

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$



But this can be negative (which turns out to be really useful)

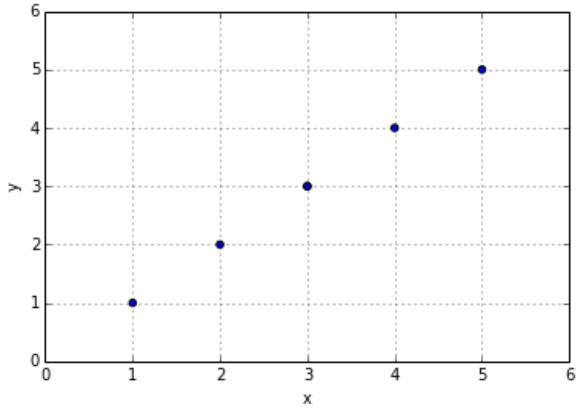
Covariance

$$\mathrm{cov}(x,y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

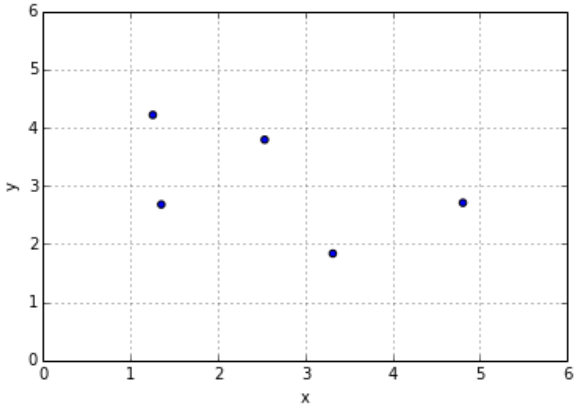
Questions to ask, when you learn a new metric…

- Meaning of the metric?
- Meaning of sign (negative/positive)?
- Range?
- Is it normalised? [-1,1] or [0,1]
- [For a metric with multiple inputs] Is it symmetric? cov(x,y)=cov(x,y)?

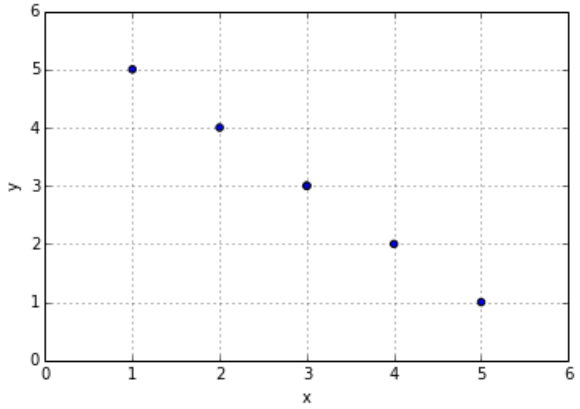
Covariance



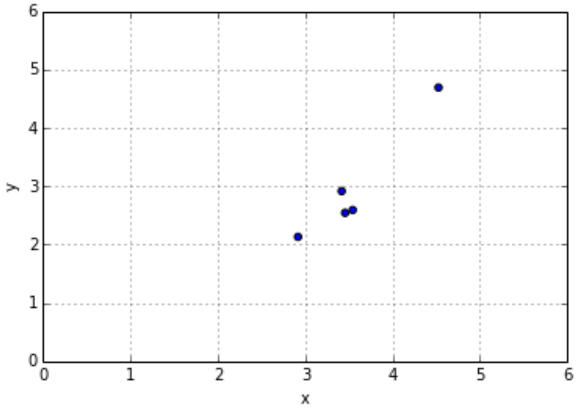
COV(X,Y) = 2.5



COV(X,Y) = -0.69

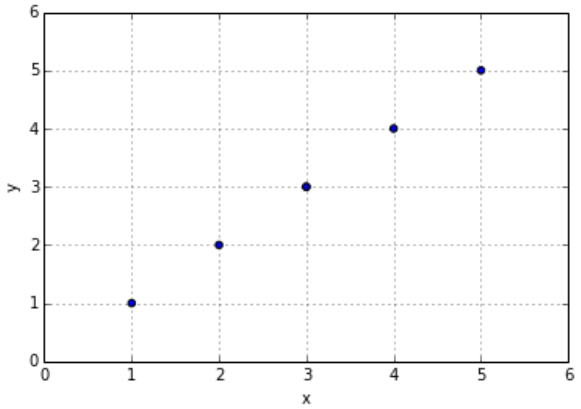


COV(X,Y) = -2.5

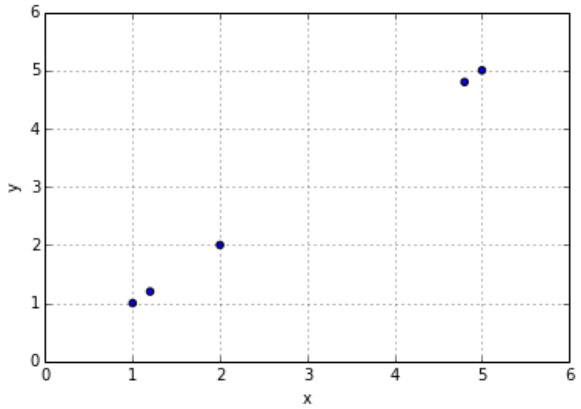


COV(X,Y) = 0.56

From Covariance to Correlation



COV(X,Y) = 2.5



COV(X,Y) = 3.81

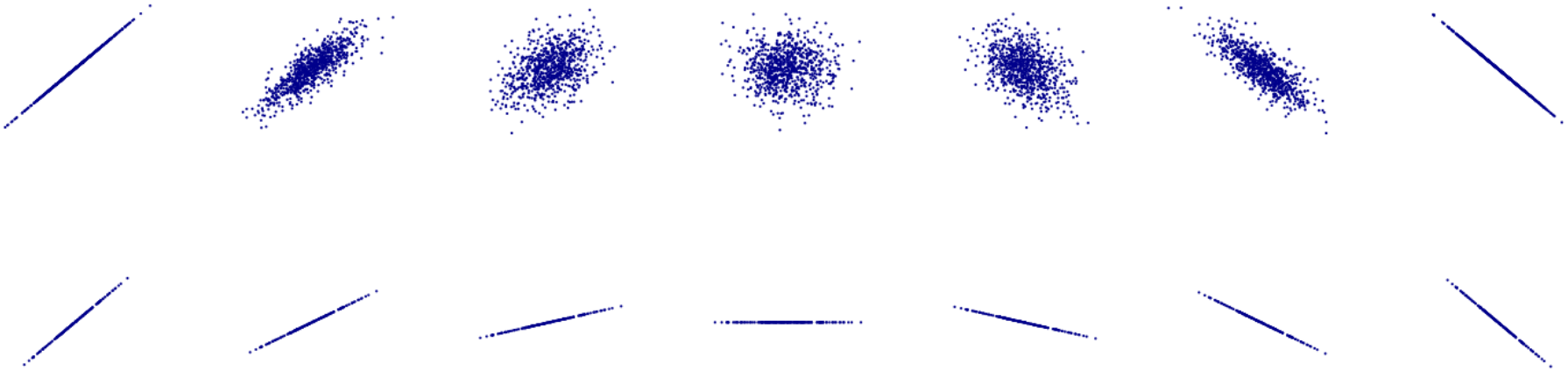
“Looks like they are following the same line.”

“Why is the covariance so different?”

“Note the variance of individual x and y.”

“Can we measure only the linear trend while excluding other factors?”

From Covariance to Correlation



Pearson Correlation Coefficient

Maximum size of the covariance is:

standard deviation of *x*-coordinates

×

standard deviation of *y*-coordinates

Pearson Correlation Coefficient

Maximum size of the covariance is:

standard deviation of *x*-coordinates

×

standard deviation of *y*-coordinates

So if we divide by st.dev.(X) × st.dev.(Y)...

... we get a number between -1 and 1.

Pearson Correlation Coefficient

Maximum size of the covariance is:

standard deviation of *x*-coordinates

×

standard deviation of *y*-coordinates

So if we divide by st.dev.(X) × st.dev.(Y)...
... we get a number between -1 and 1.

$$\mathrm{cor}(\mathrm{X},\mathrm{Y}) = r(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{cov}(X,Y)}{\mathrm{stdev}(X) * \mathrm{stdev}(Y)}$$

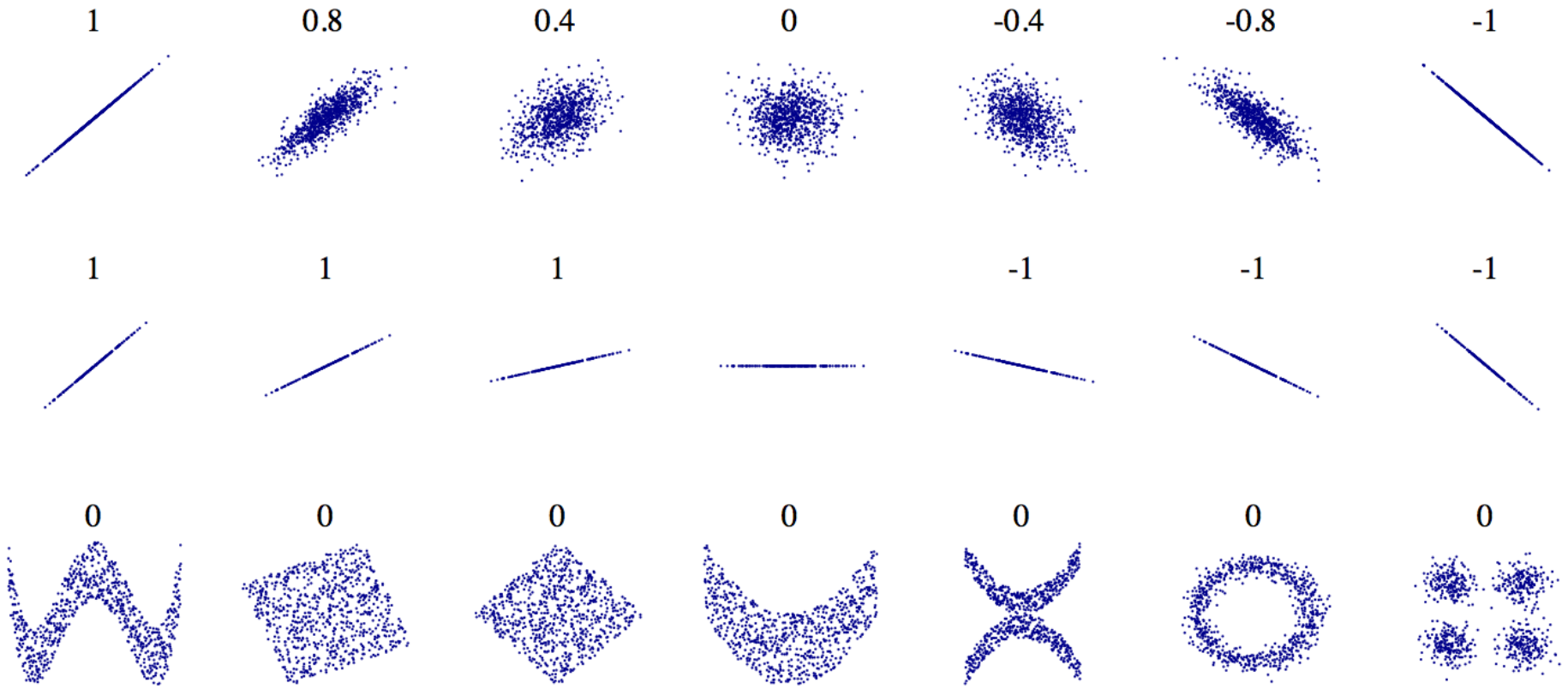
Pearson Correlation Coefficient

$$r = \frac{\operatorname{cov}(X, Y)}{sd(X)´\ sd(Y)}$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Limitations of Pearson Correlation

Question: does a Pearson Correlation of 0 mean no relationship between X and Y?



Correlation vs. association

- The Pearson correlation coefficient measures only linear association: how nearly the data fall on a straight line.
- It is not a good summary of association if the scatterplot has a nonlinear (curved) pattern.
- When you present the correlation, remember to present a scatterplot.

Limitations of Pearson Correlation

- The Pearson correlation coefficient is appropriate only for **interval or ratio** variables, not **nominal or ordinal** variables – even if their values are numerical.
- Note the difference between the data type of the variable and the value.
  - Gender: male: ‘1’, female: ‘0’.
  - *What is the data type of ‘Gender’?*

Spearman’s Rank Correlation Coefficient

Ordered Data

Telegraph
Edinburgh
London
York
Bath
Cambridge
Chester
Oxford
Liverpool
Cardiff
Newcastle

Most Visitors
London
Edinburgh
Liverpool
Oxford
Cambridge
Cardiff
Newcastle
York
Bath
Chester

	City Status
London	Time Immemorial
York	Time Immemorial
Bath	1519
Chester	1541
Oxford	1542
Edinburgh	1700-99
Liverpool	1880
Newcastle	1882
Cardiff	1905
Cambridge	1951

http://www.telegraph.co.uk/travel/destinations/europe/united-kingdom/galleries/Britains-20-best-cities/

http://gouk.about.com/od/getawaysandshorthops/ss/top20.htm

Ordered Data

City	Telegraph Rank	Visitors Rank	City Status Rank
Bath	4	9	3
Cambridge	5	5	10
Cardiff	9	6	9
Chester	6	10	4
Edinburgh	1	2	6
Liverpool	8	3	7
London	2	1	1.5
Newcastle	10	7	8
Oxford	7	4	5
York	3	8	1.5

Spearman's Rank Correlation Coefficient

	Telegraph	No. of Visitors	City Status Date
Telegraph	1.000	0.273	0.596
No. of Visitors	0.273	1.000	-0.109
City Status Date	0.596	-0.109	1.000

Spearman's Rank Correlation Coefficient

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$$

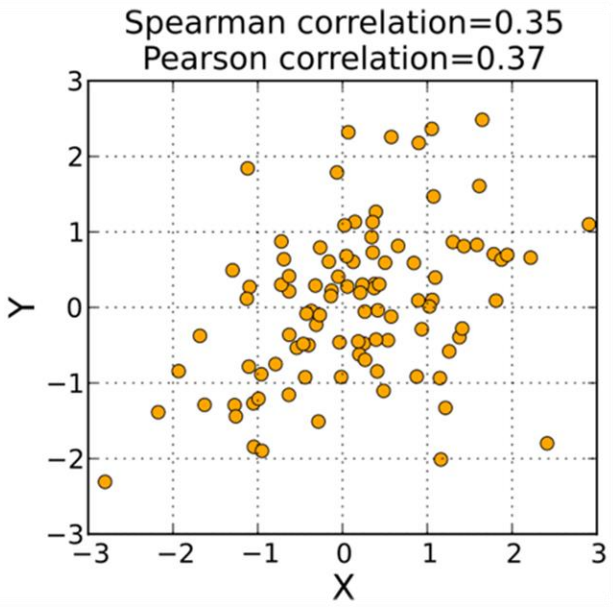
Monotonic relation = perfect Spearman correlation (+1 or -1)

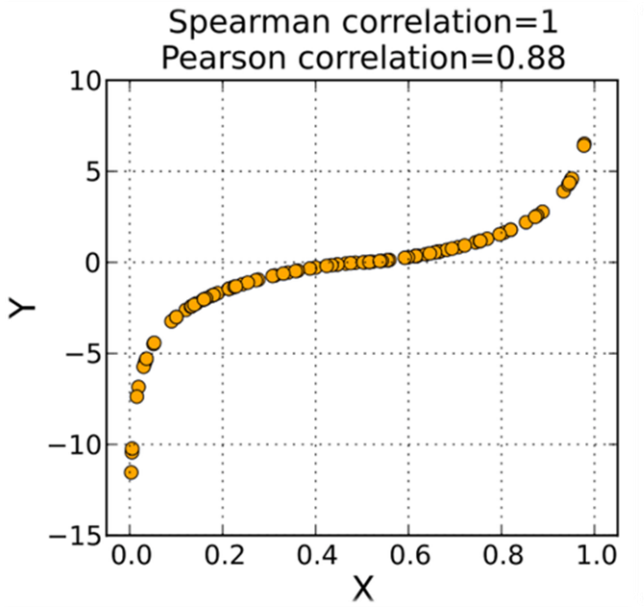
Range: [-1, +1]

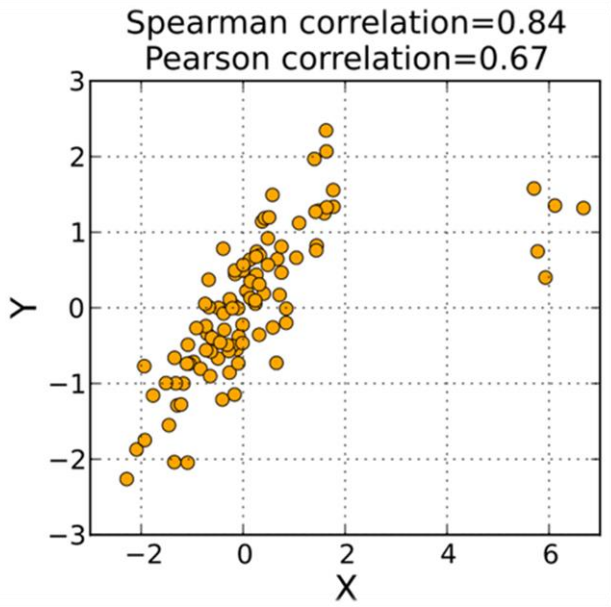
Q: Can we use Spearman's Rank correlation for the interval data or ratio data?

Spearman's Rank Correlation Coefficient

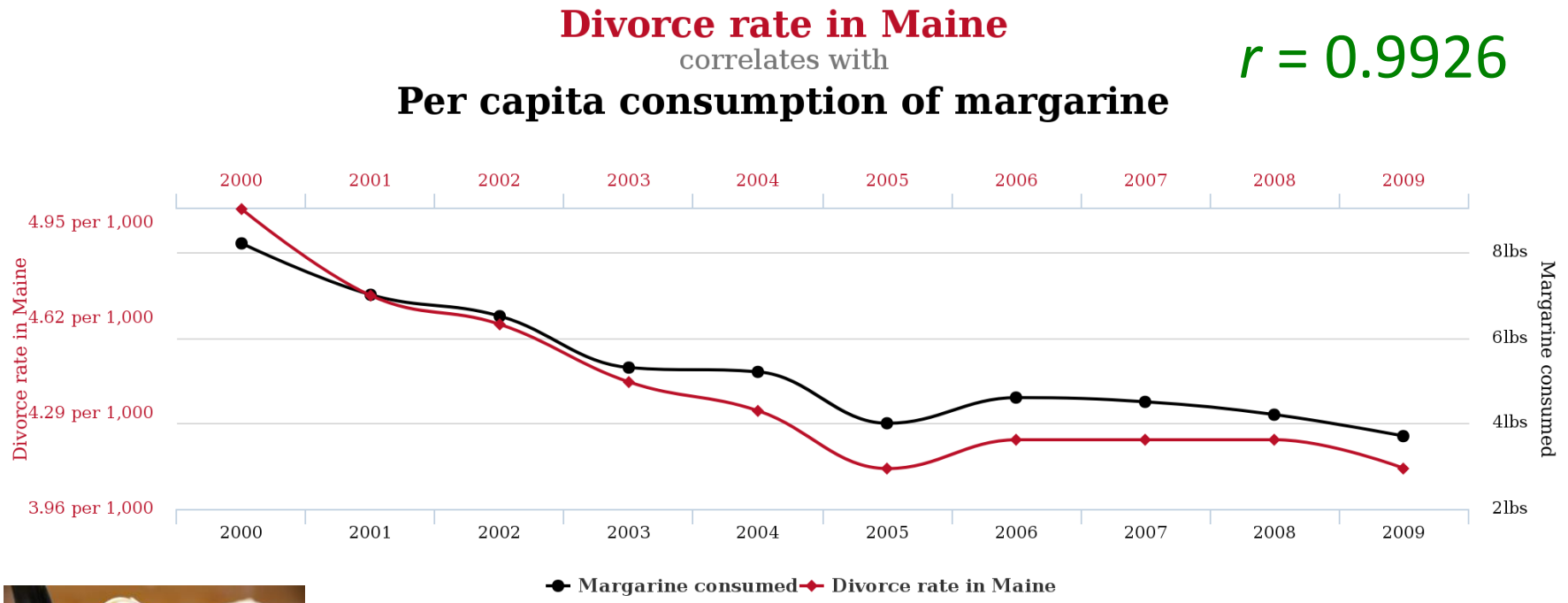
- It is applicable for interval/ratio data
  1. Transform the value into ranks;
  2. Calculate the Rank correlation.
- But it is different from Pearson's correlation, and they are incomparable.
- Less sensitive to outliers than Pearson correlation

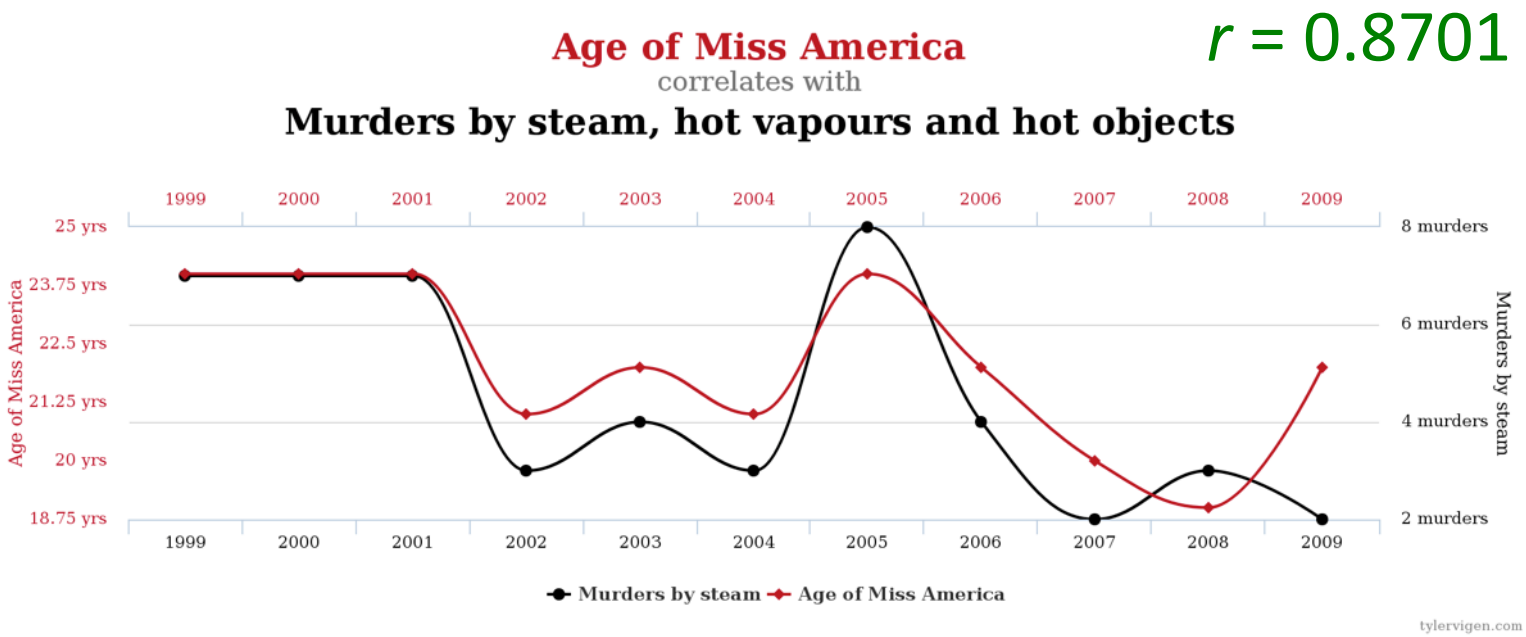


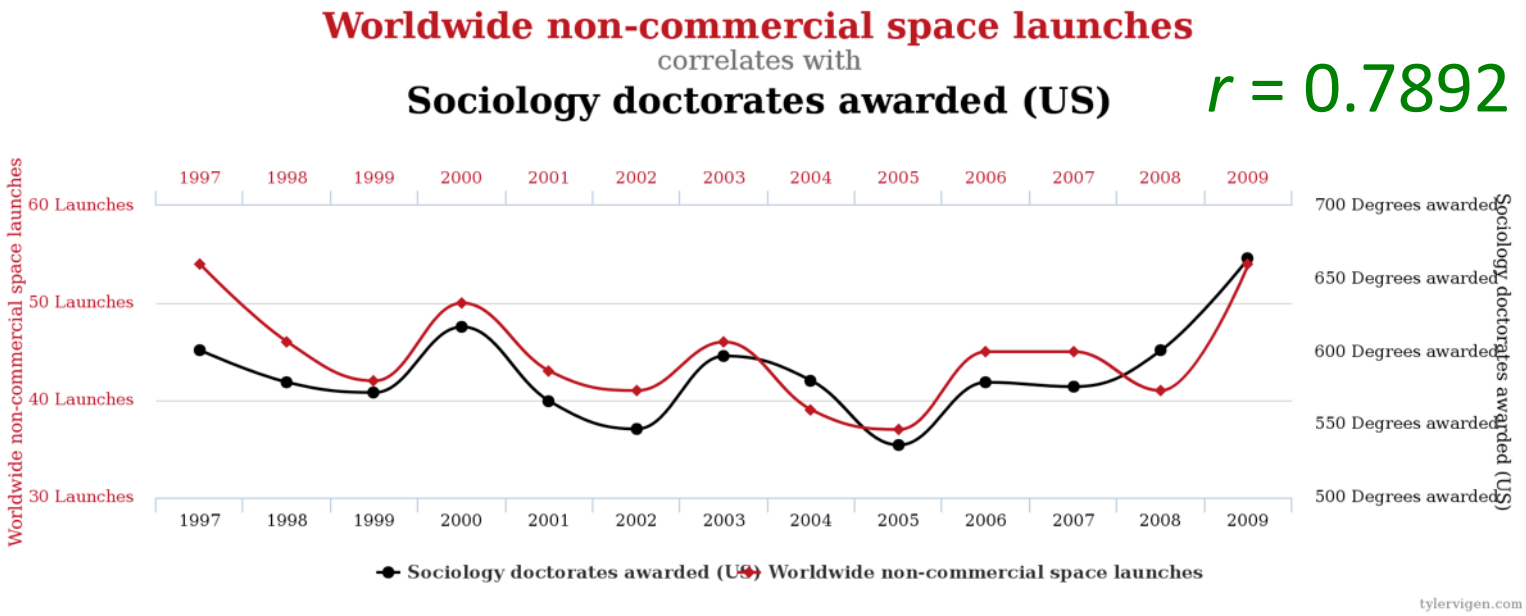




Correlation does not imply causality







Correlation does not imply causality.

Summary of Limitations

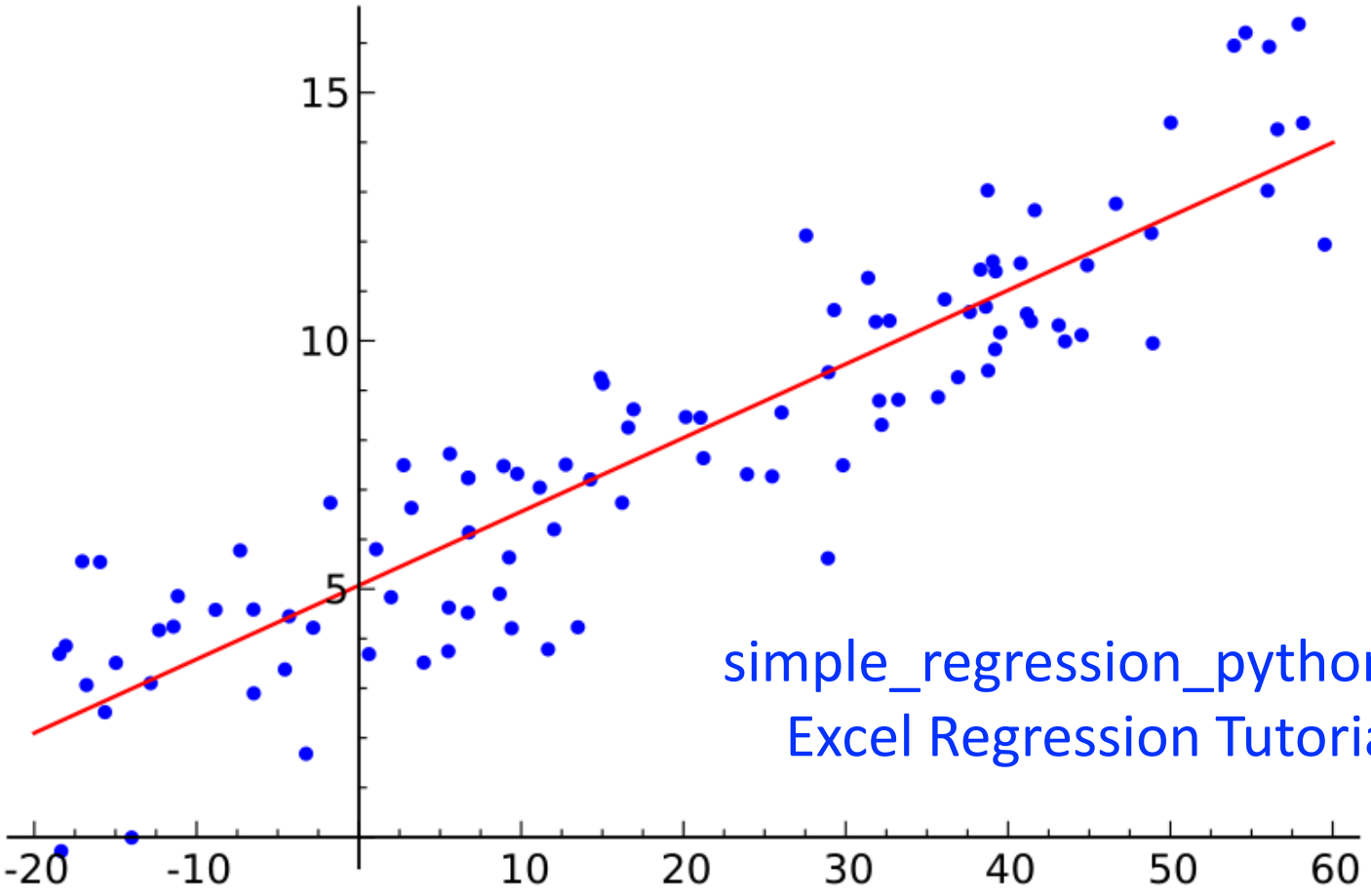
1. Pearson correlation is not applicable for nominal/ordinal data.
2. It measures only linear association, and is not a good summary for non-linear data.
3. It is a measure of association, **not causation**.
4. It is not robust to outliers, less robust than Spearman's rank correlation.

Question

- Can correlation be used to make predictions? Predictions is defined as providing a Y value for a given X value.

Part 2: Linear regression

Linear regression



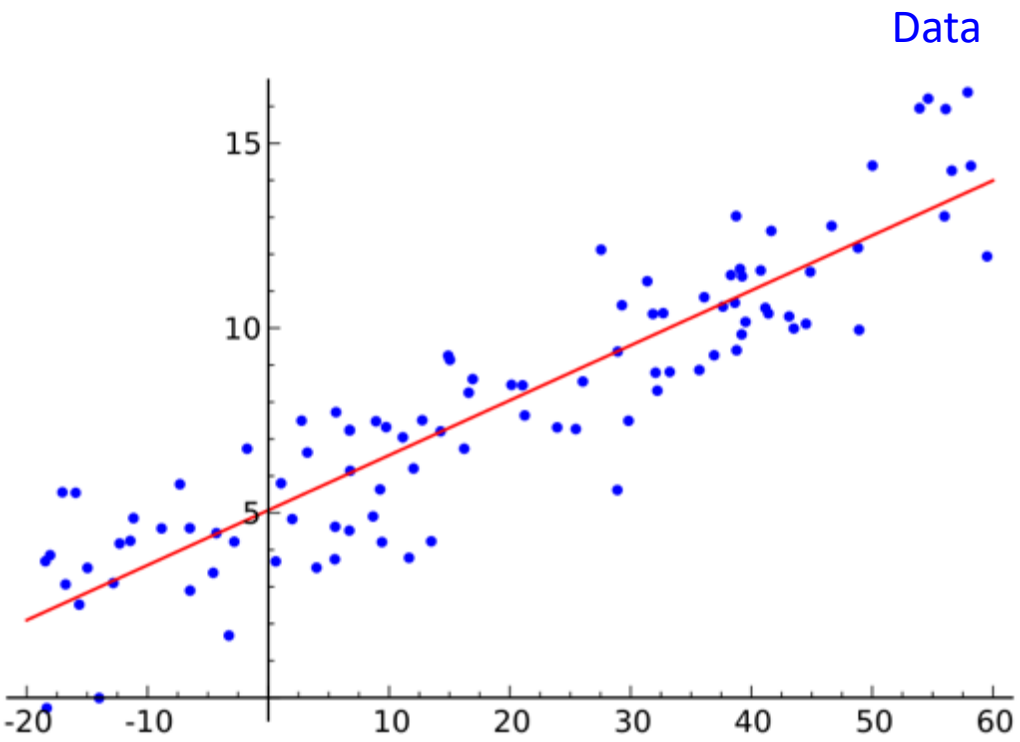
simple_regression_python.py
Excel Regression Tutorial

Linear regression

Key questions:

1. Where to place the line?
2. How well does the line represent the data?

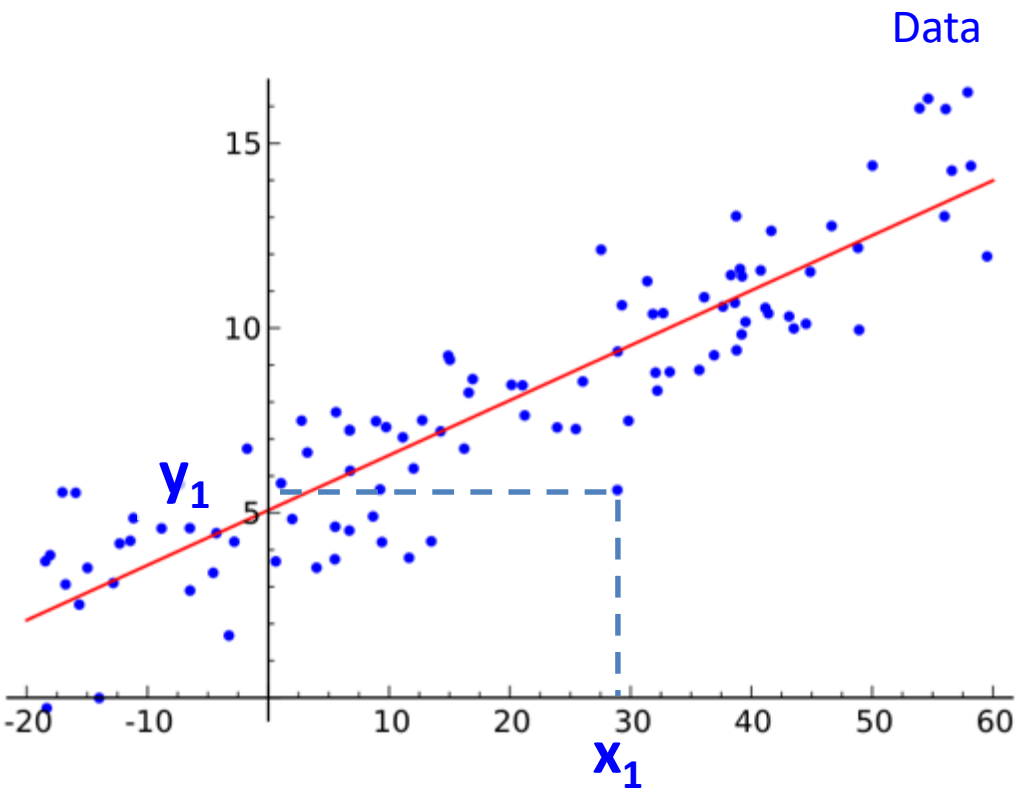
1. Where to place the line?



The modelled relationship:

$\hat{y} = mx + c$

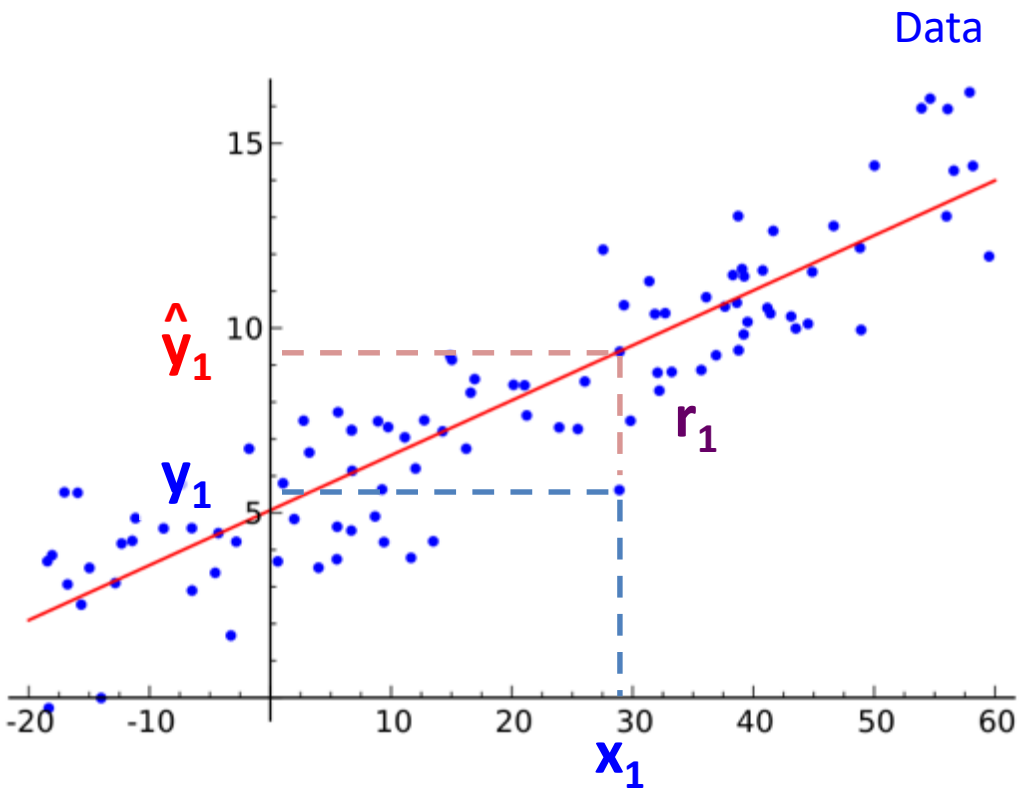
1. Where to place the line?



The modelled relationship:

$\hat{y} = mx + c$

1. Where to place the line?

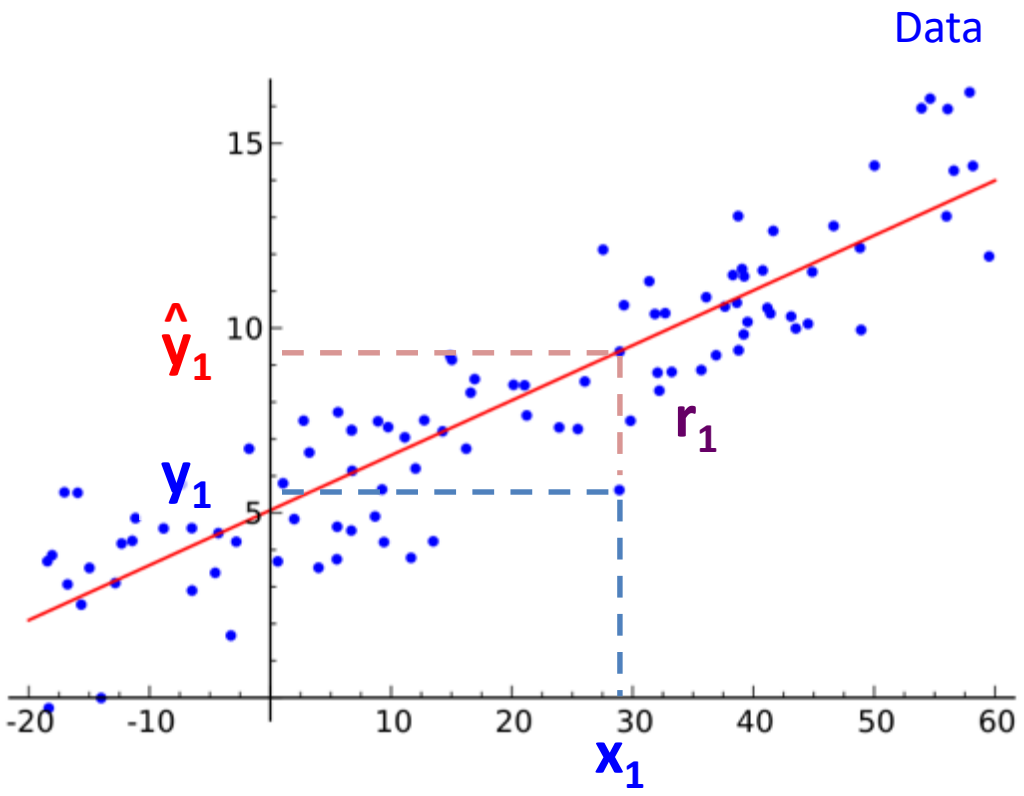


The modelled relationship:

$\hat{y} = mx + c$

$r_1 = \hat{y}_1 - y_1$

1. Where to place the line?



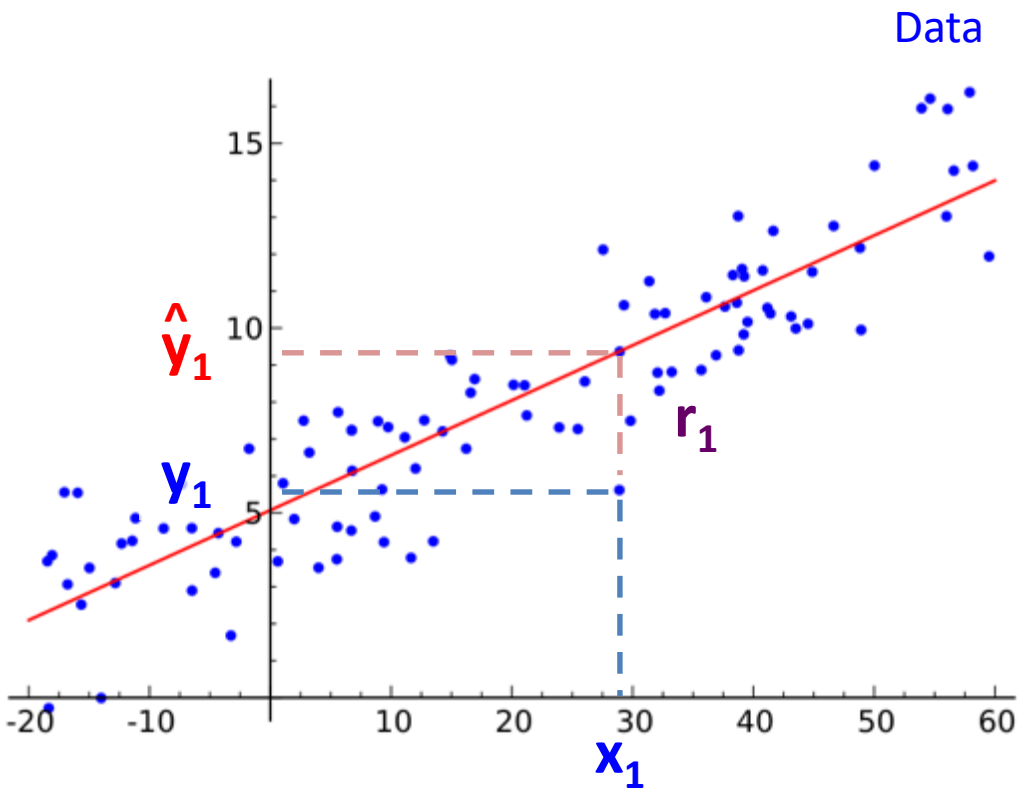
The modelled relationship:

$\hat{y} = mx + c$

$r_1 = \hat{y}_1 - y_1$

Regression minimises: $r_1^2 + \ldots + r_n^2$ OR $\sum_i r_i^2$

1. Where to place the line?



The modelled relationship:

$\hat{y} = mx + c$

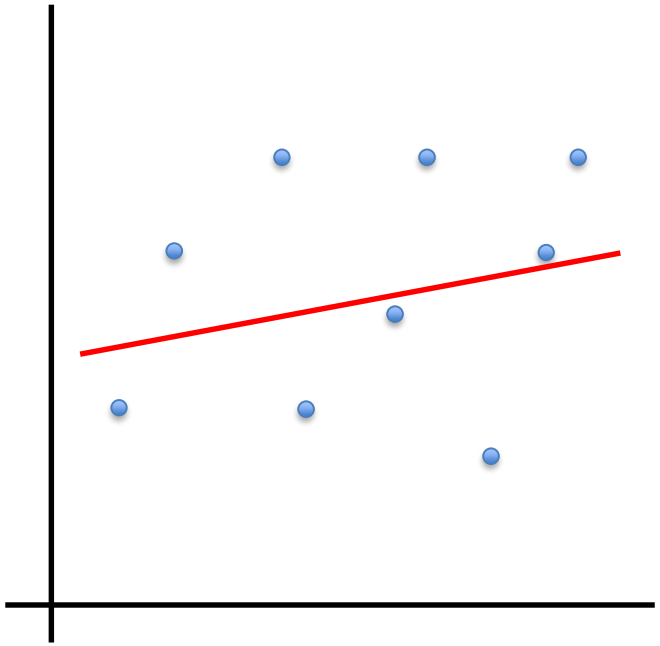
$r_1 = \hat{y}_1 - y_1$

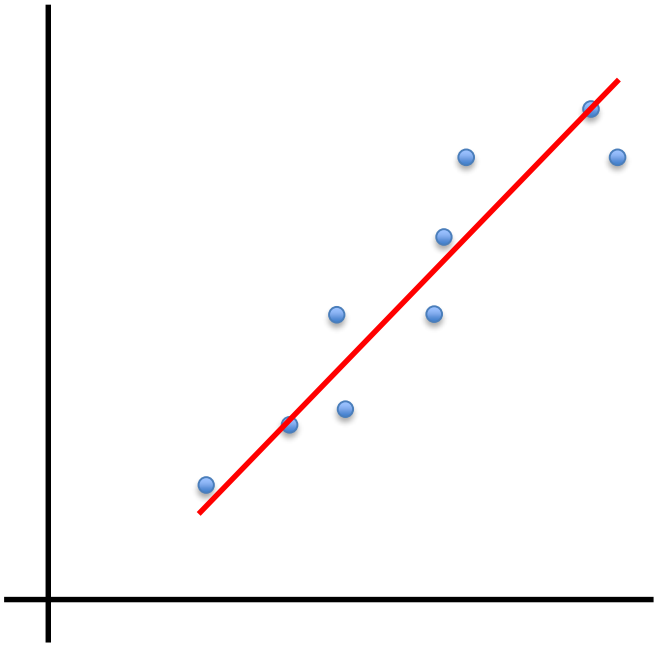
Regression minimises: $r_1^2 + \ldots + r_n^2$ OR $\sum_i r_i^2$

There are formulae to calculate m and c

2. How well does the line represent the data?

Which one better represents the data?





2. How well does the line represent the data?

Coefficient of Determination: R^2 Value

Variation around the line: $\sum_i r_i^2$

Total variation: $(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2$

mean of y

$$R^2 = 1 - \frac{\text{Variation around the line}}{\text{Total variation}}$$

2. How well does the line represent the data?

Coefficient of Determination: R^2 Value

Variation around the line $\sum_i r_i^2$

Total variation: $(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2$

$\bar{y}$: mean of y

$$R^2 = 1 - \frac{\text{Variation around the line}}{\text{Total variation}}$$

2. How well does the line represent the data?

Coefficient of Determination: R^2 Value

Variation around the line $\sum_i r_i^2$

Total variation: $(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2$

mean of y

Divide by $n - 2$ to get the MSE

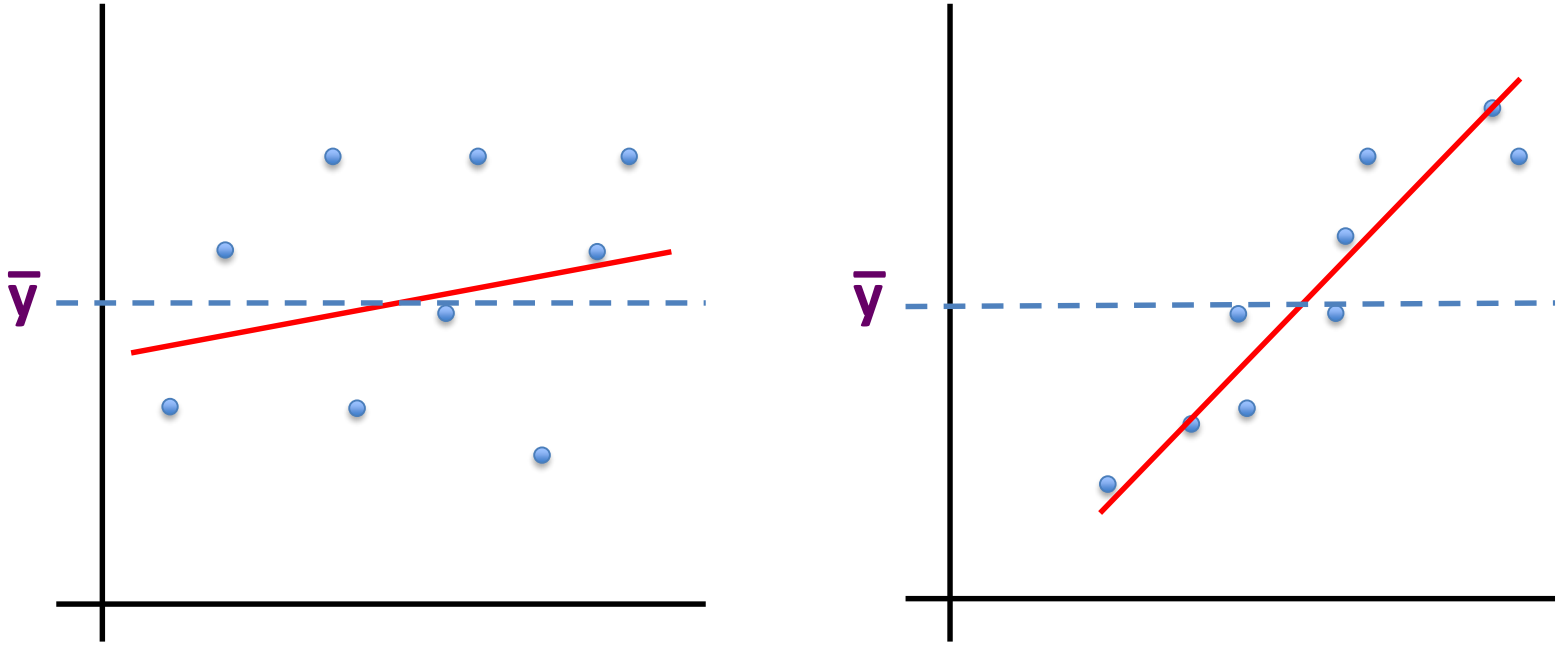
$R^2 = 1 -$ Variation around the line / Total variation

(Mean Square Error)

An estimate of variance around the line

2. How well does the line represent the data?

Total variation of y = variation explained by the line and x + variation of residuals

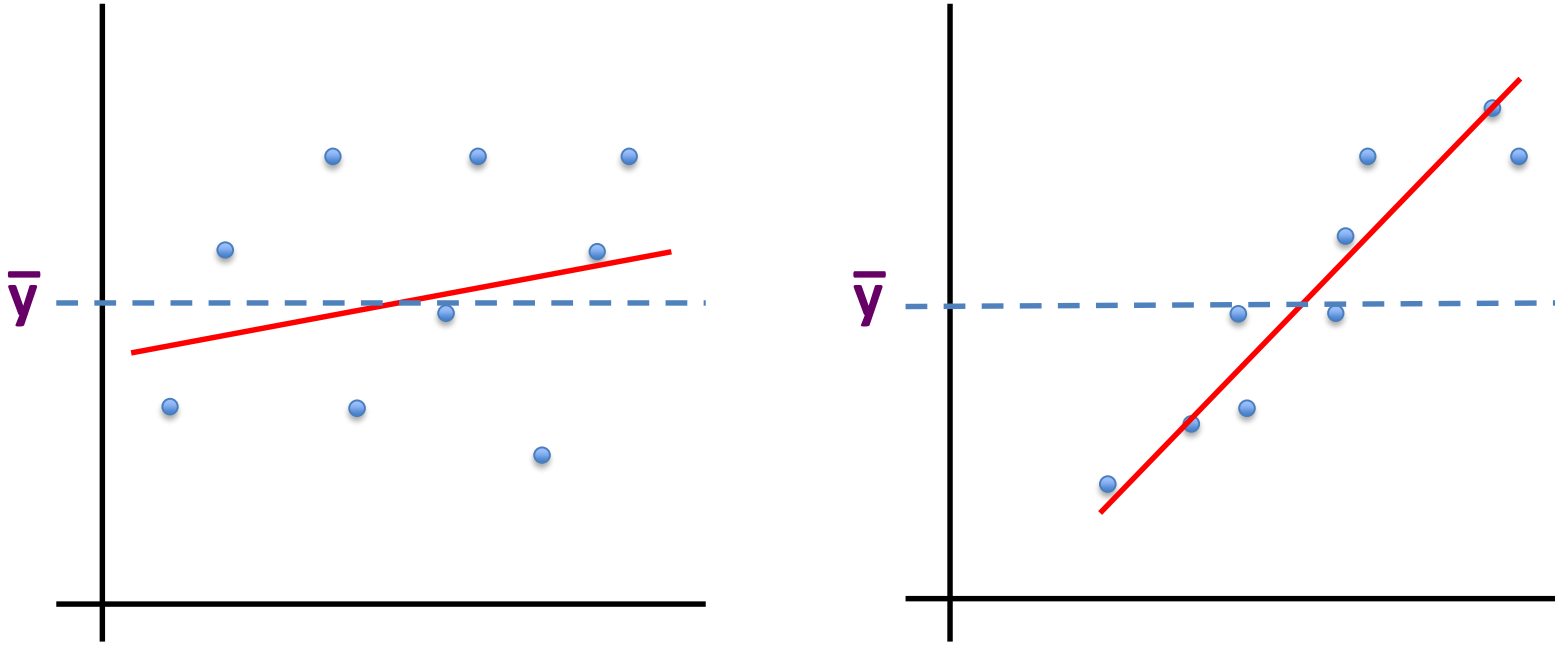


$$R^2 = 1 - \frac{\text{Variation around the line}}{\text{Total variation}}$$

Proportion of variation in *y* that is explained by *x*

2. How well does the line represent the data?

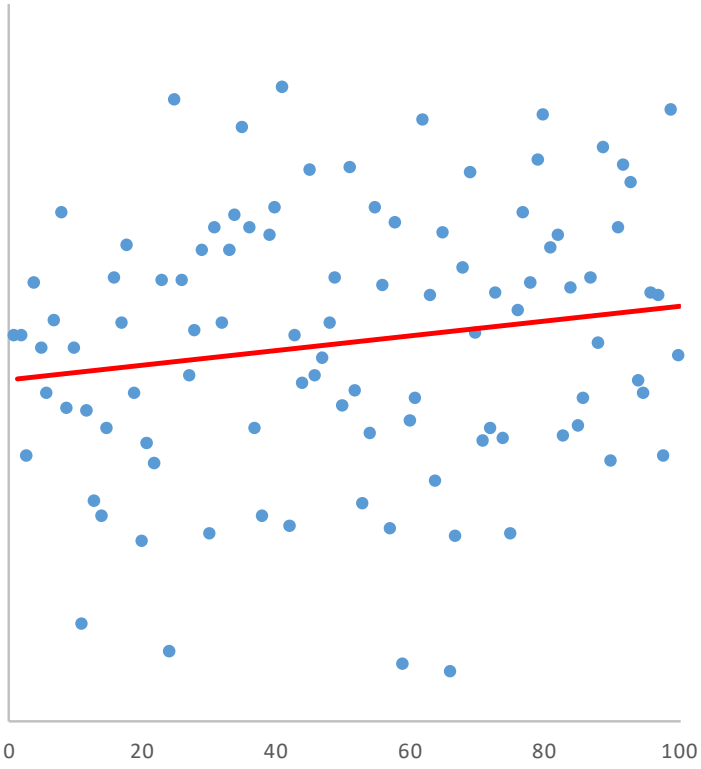
Total variation of y = variation explained by the line + variation of residuals

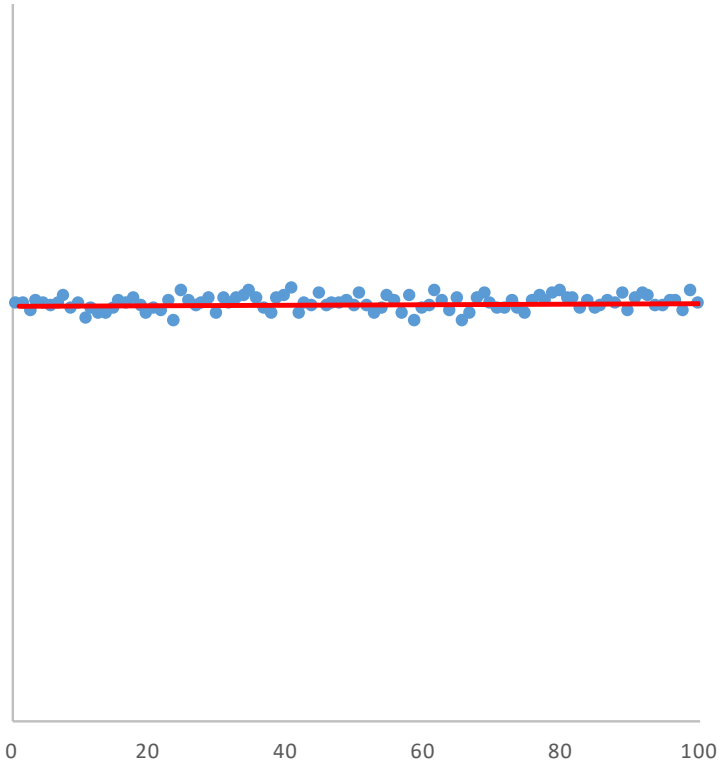


$$\text{MSE} = \frac{\text{Sum of squared errors}}{n - 2}$$

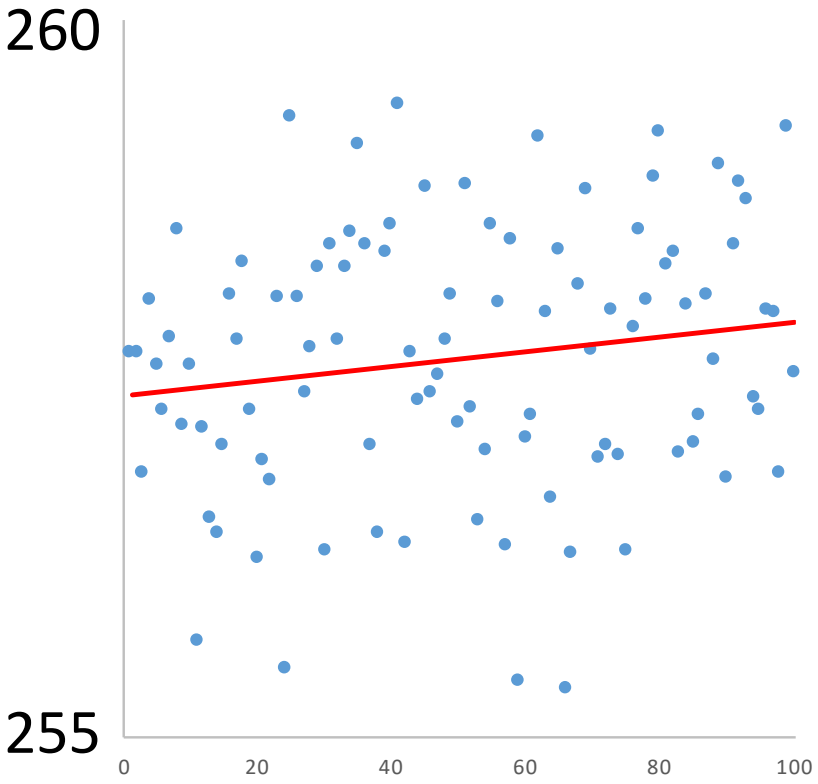
Estimated variance around the line

Which plot depicts a more meaningful relationship?

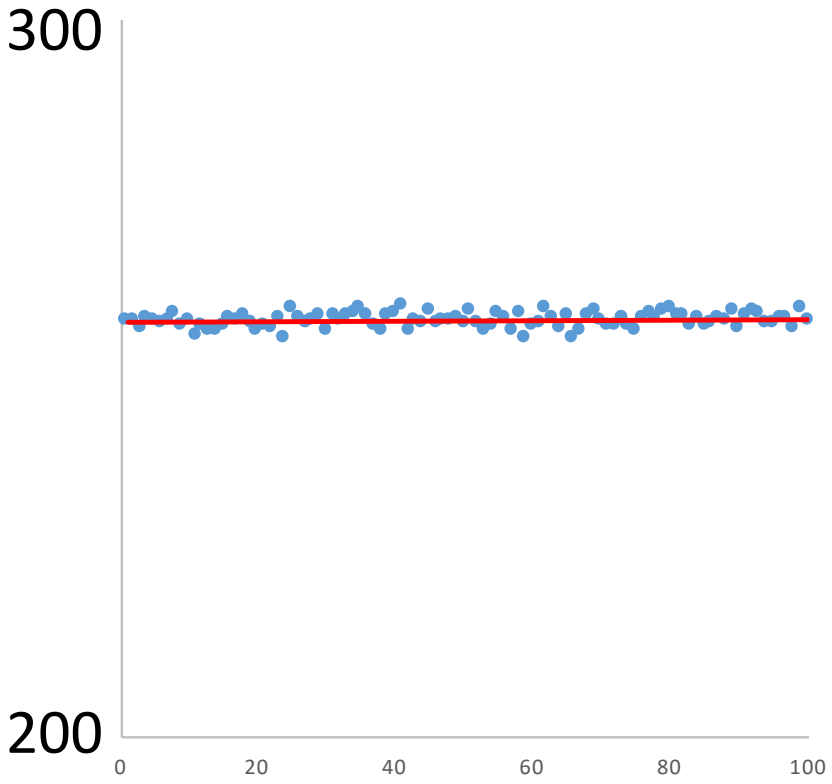




Which plot depicts a more meaningful relationship?



Gradient = 0.006
R^2 = 0.026
MSE = 0.997



Gradient = 0.006
R^2 = 0.026
MSE = 0.997

Useful Excel Functions

Gradient (*m*): = SLOPE(*y_values*, *x_values*)

y-intercept (*c*): = INTERCEPT(*y_values*, *x_values*)

R^2 value: = RSQ(*y_values*, *x_values*)

LINEAR REGRESSION

Necessary Conditions

Linear relationship exists

Independent errors

Normally distributed errors

Equal error variance for all x values

LINEAR REGRESSION

NOT a Necessary Condition!

Normally Distributed Data

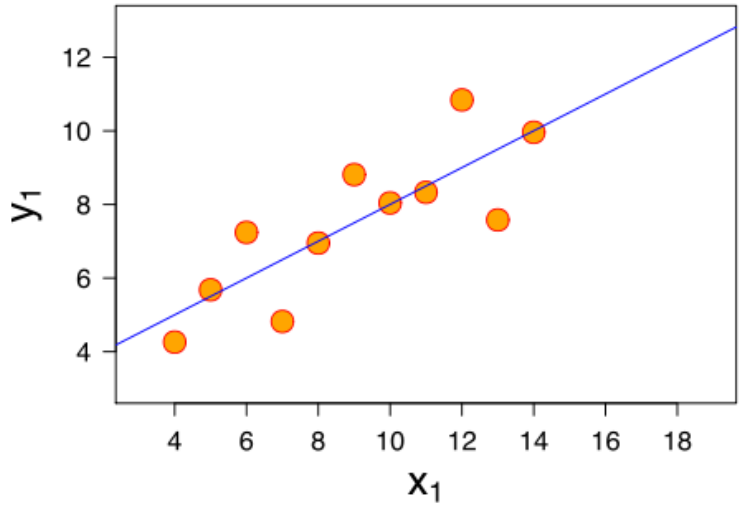
(x and y)

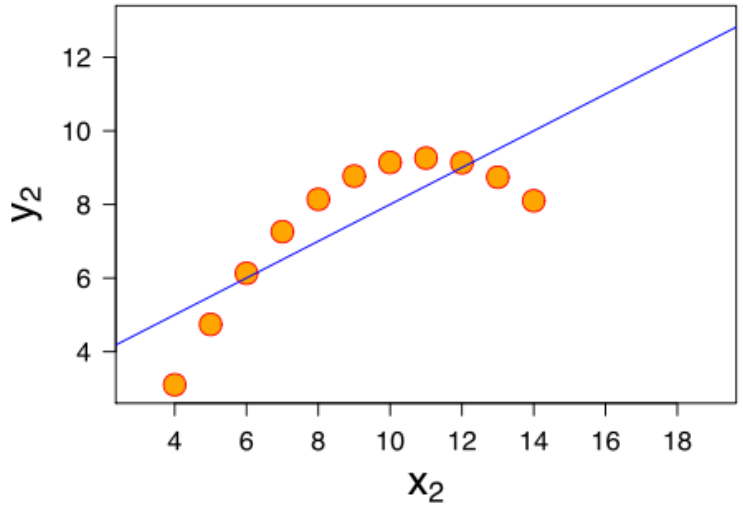
Linear regression exists

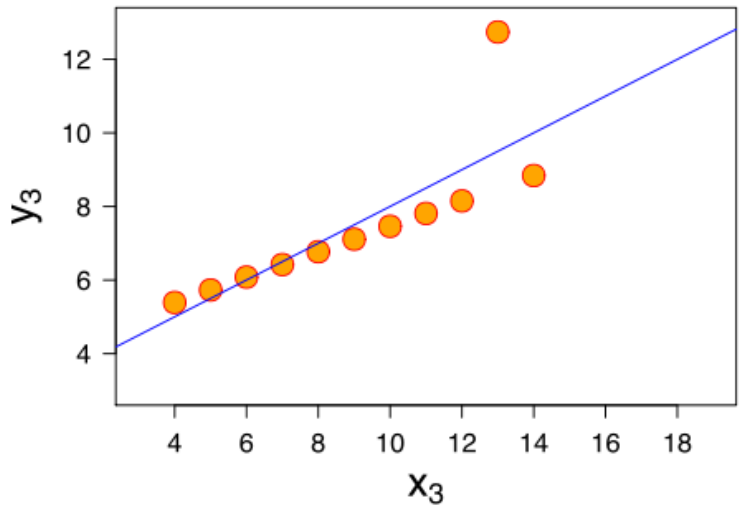
- The relationship between y and x is linear; that is, there is an equation, $y=mx+c+\varepsilon$ that constitutes the population model.
- But… how can we know that?

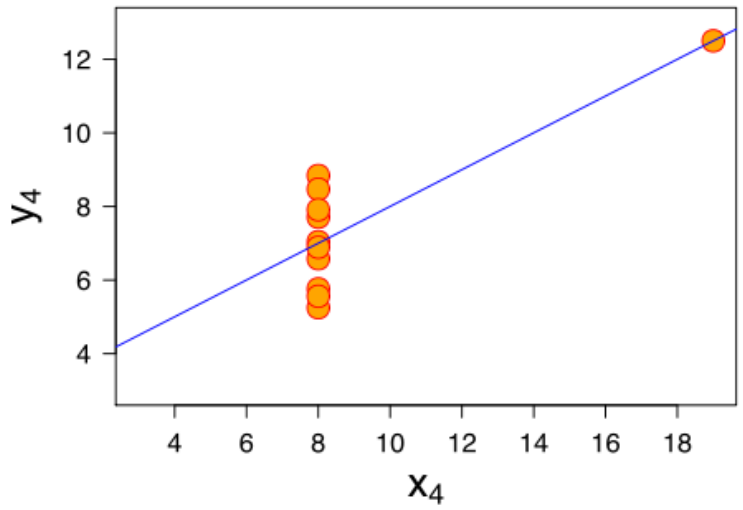
Does a Linear Relationship Exist?

Importance of Visualisation









Independent errors

- The residuals are independent; the value of one error is not affected by the value of another error.
- In probability, two events A and B are independent if knowing that B happens does not alter the probability that A happens. E.g., flipping two coins.
- A and B are indepdent if $Pr[B]=0$ or $Pr[A|B] = Pr[A]$
- But… how can we know that?

Normally distributed errors

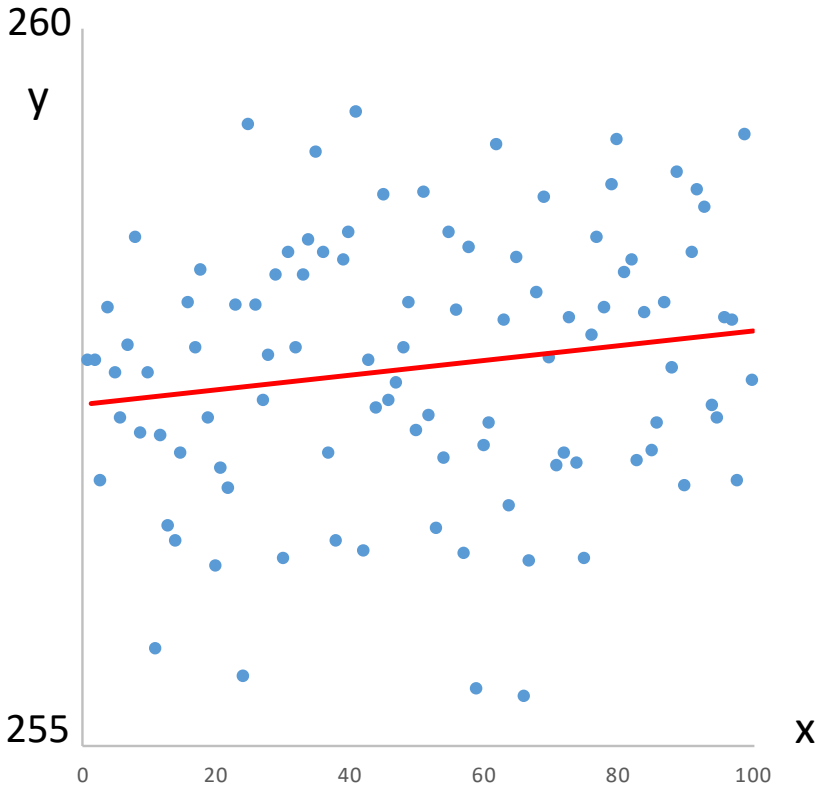
- For each value of x, the errors have a normal distribution about the regression line. This normal distribution is centred on the regression line.
- It may be written as $\varepsilon \sim N(0, \sigma^2)$

Equal error variance for all x values

- The errors about the regression line do not vary with x; that is, $V[\varepsilon|x] = \sigma^2$

Verifying the Conditions (Ex 1)

"Residuals vs Fits Plot"



residual (or error) is defined as

$$e_i = y_i - \hat{y}_i$$

Where y_i is a data point, and $\hat{y}_i$ is its fitted value

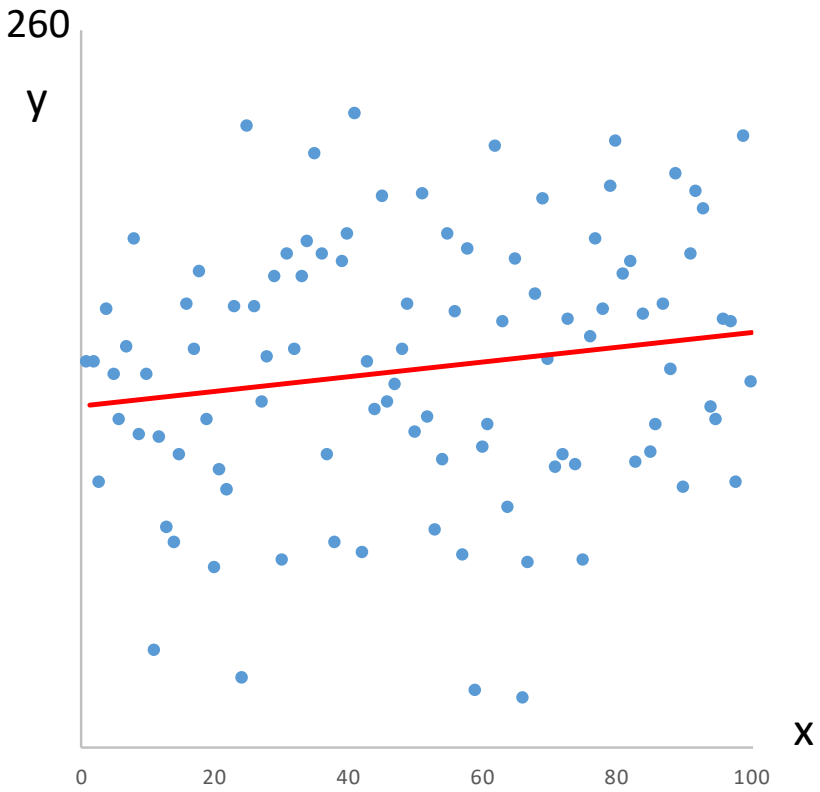
Gradient = 0.006

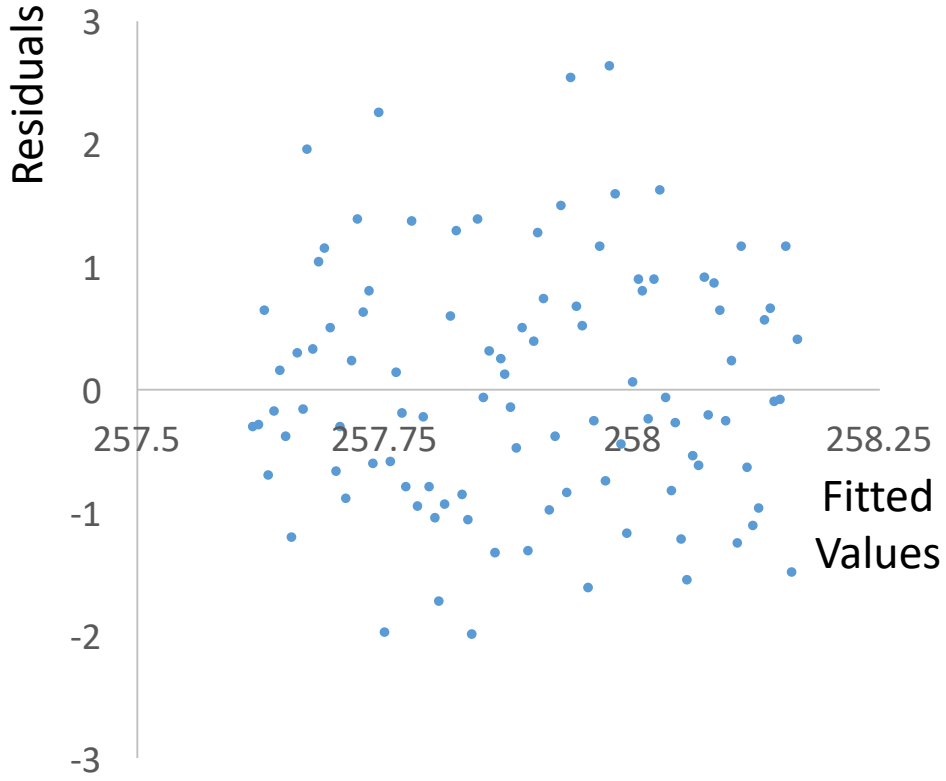
R^2 = 0.026

MSE = 0.997

Verifying the Conditions (Ex 1)

"Residuals vs Fits Plot"





Gradient = 0.006

R^2 = 0.026

MSE = 0.997

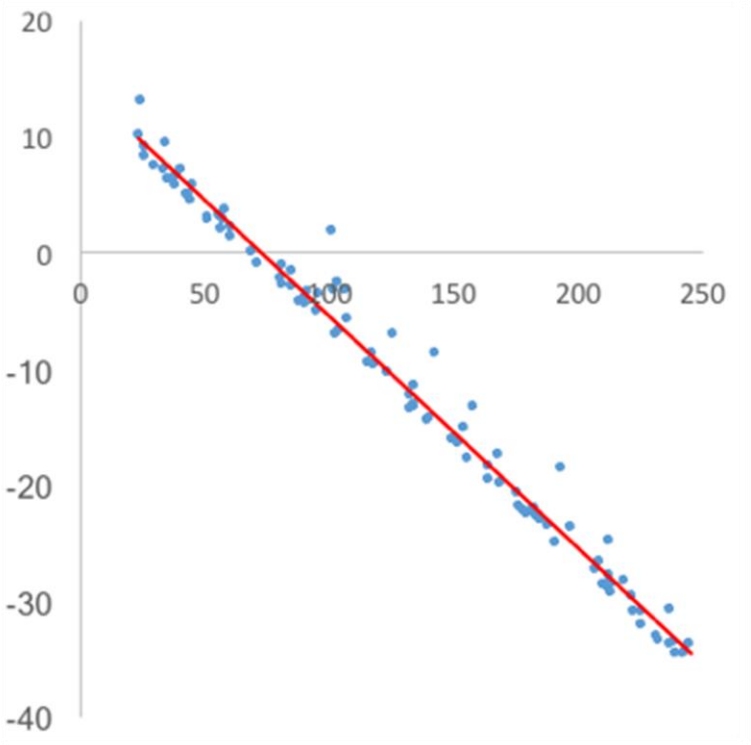
Are residuals independent?

Are they normally distributed?

Do they have equal variance?

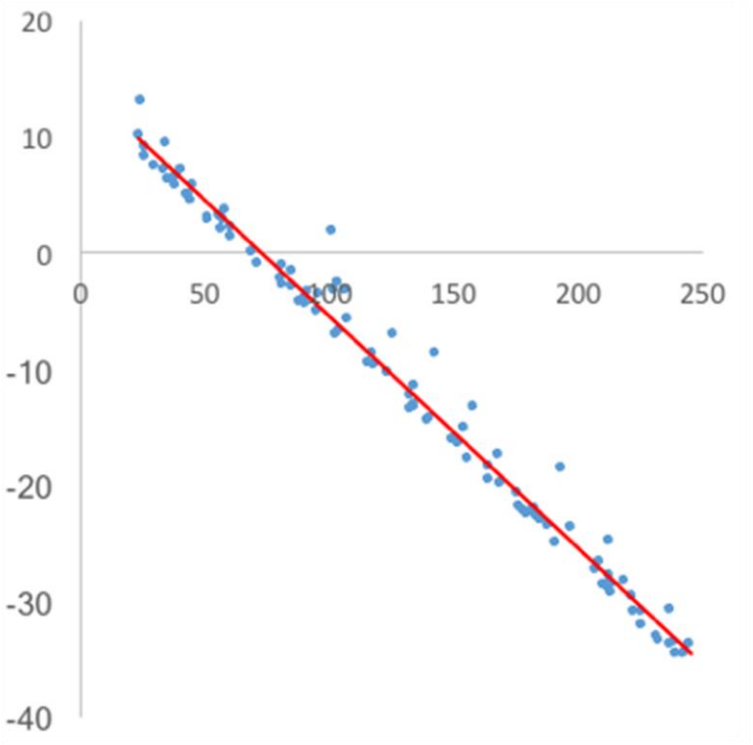
Verifying the Conditions (Ex 2)

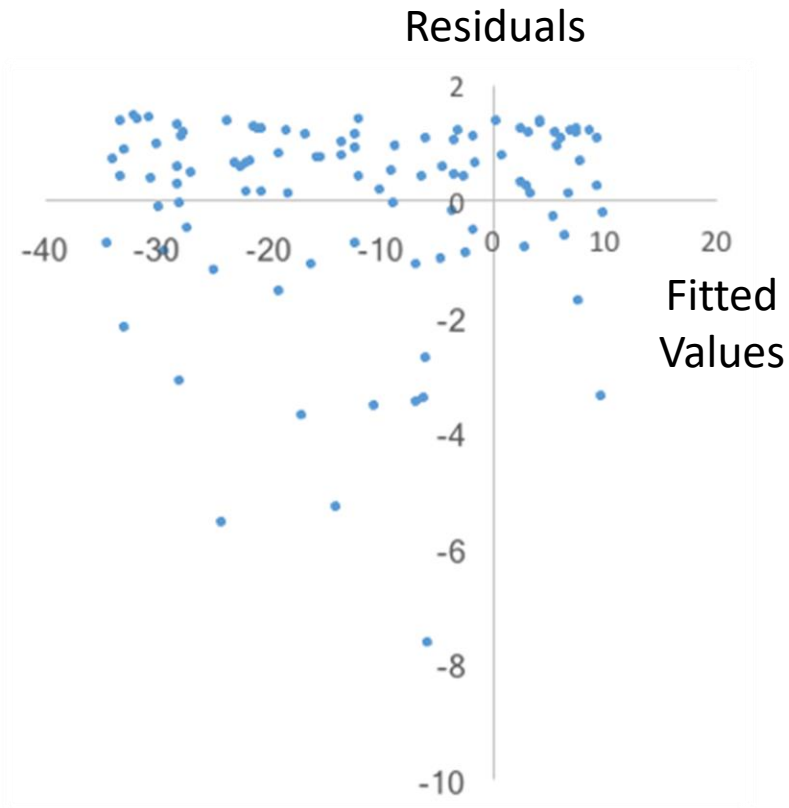
"Residuals vs Fits Plot"



Verifying the Conditions (Ex 2)

"Residuals vs Fits Plot"





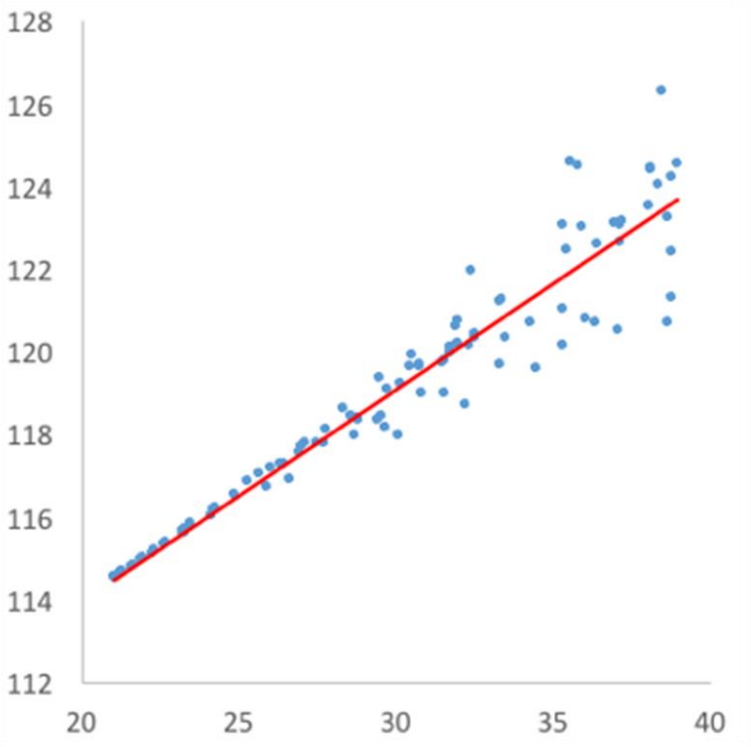
Are residuals independent?

Are they normally distributed?

Do they have equal variance?

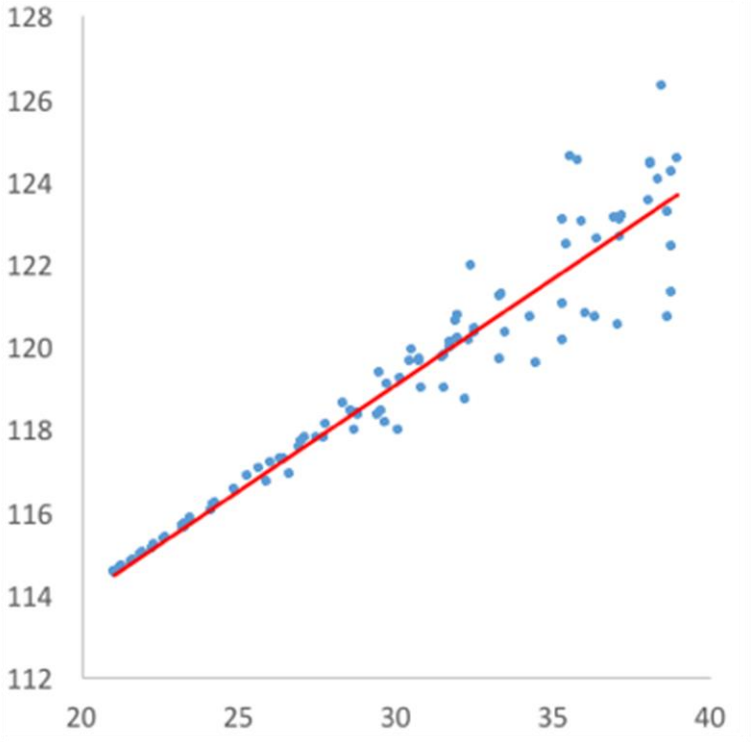
Verifying the Conditions (Ex 3)

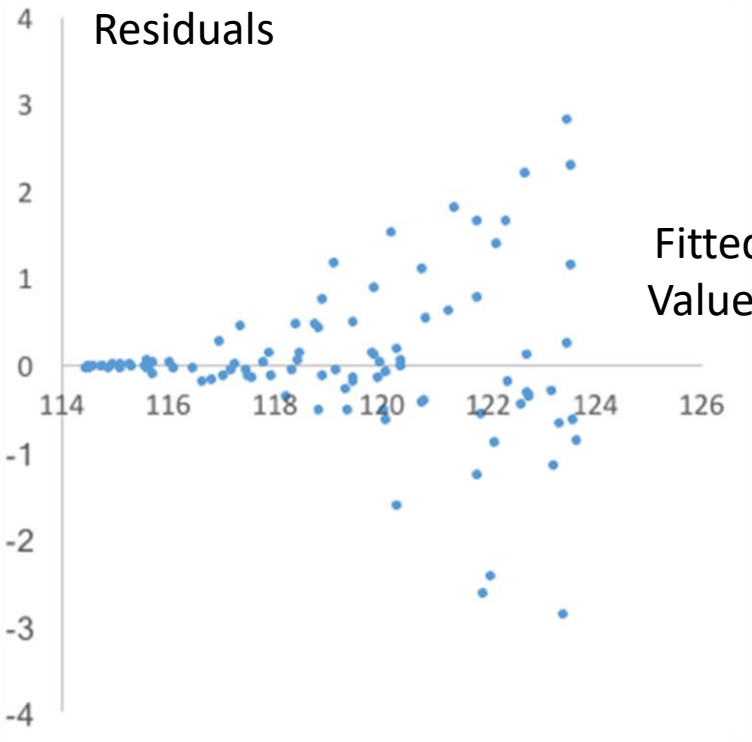
"Residuals vs Fits Plot"



Verifying the Conditions (Ex 3)

"Residuals vs Fits Plot"





Are residuals independent?

Are they normally distributed?

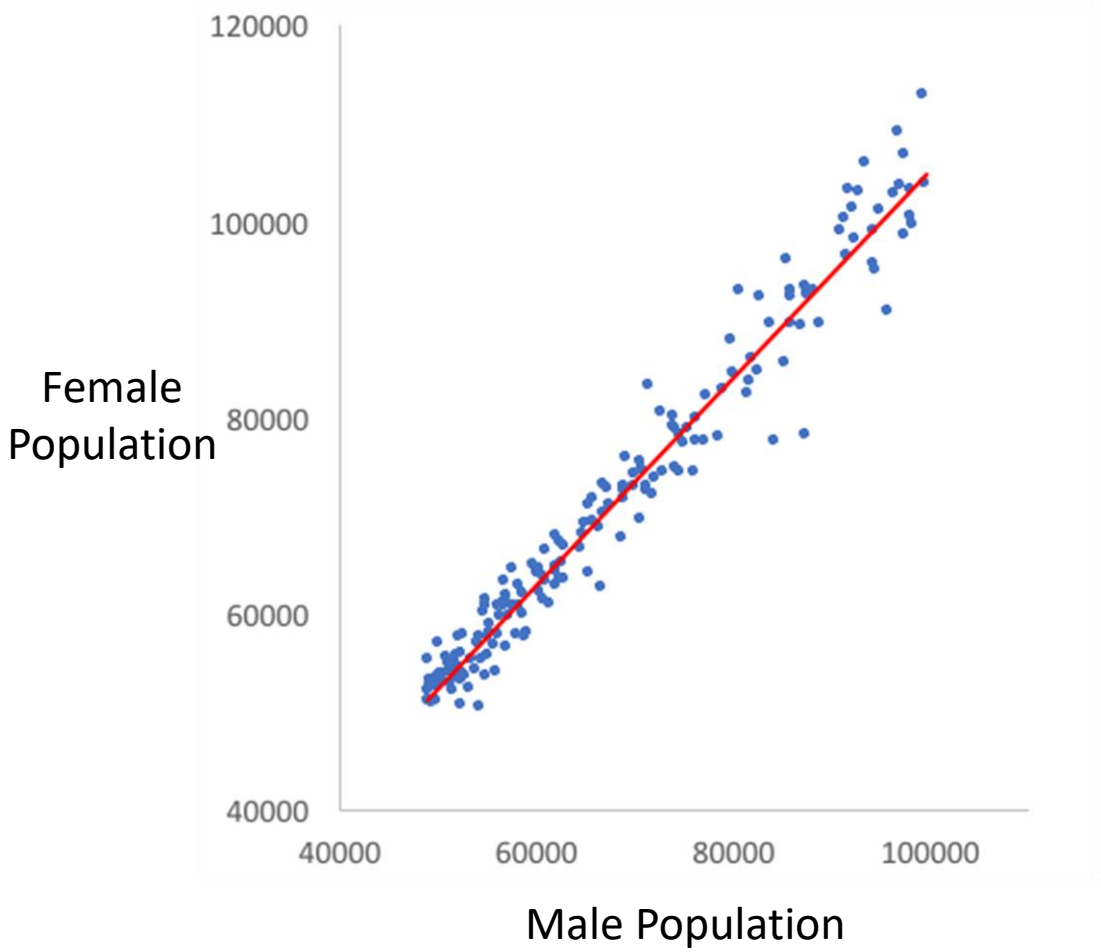
Do they have equal variance?

Download the Excel tutorial:

Simple_Regression_Tutorial_Basics_2018

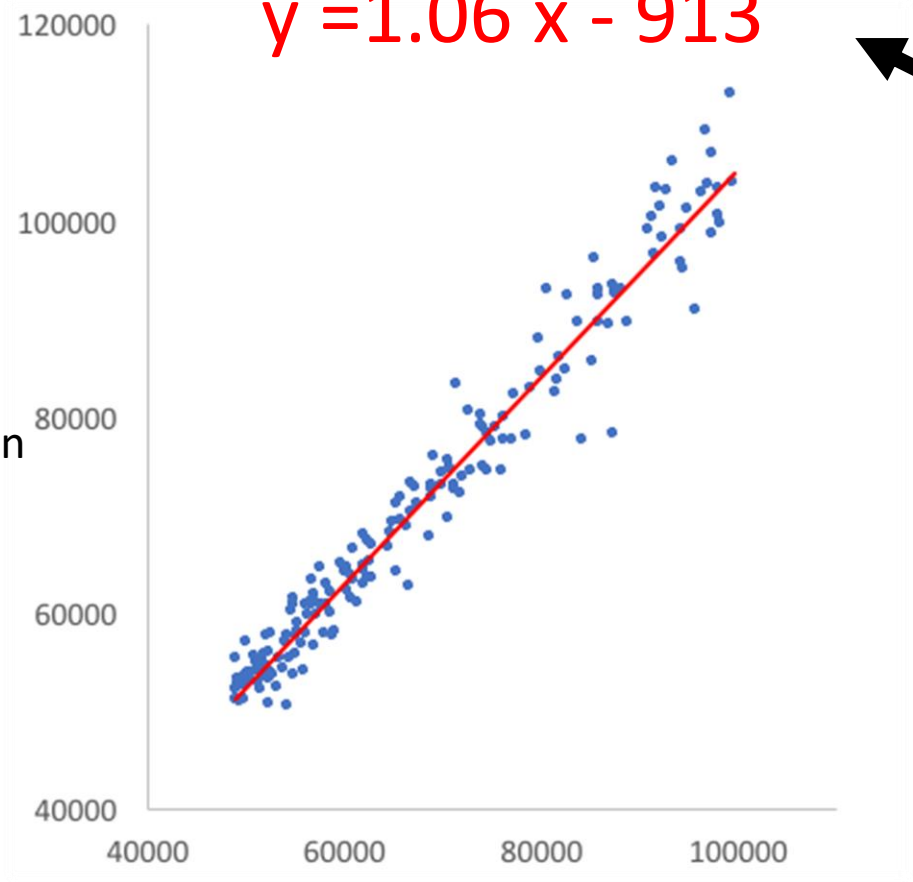
Why Regression? Making Predictions

Why Regression? Making Predictions



Why Regression? Making Predictions

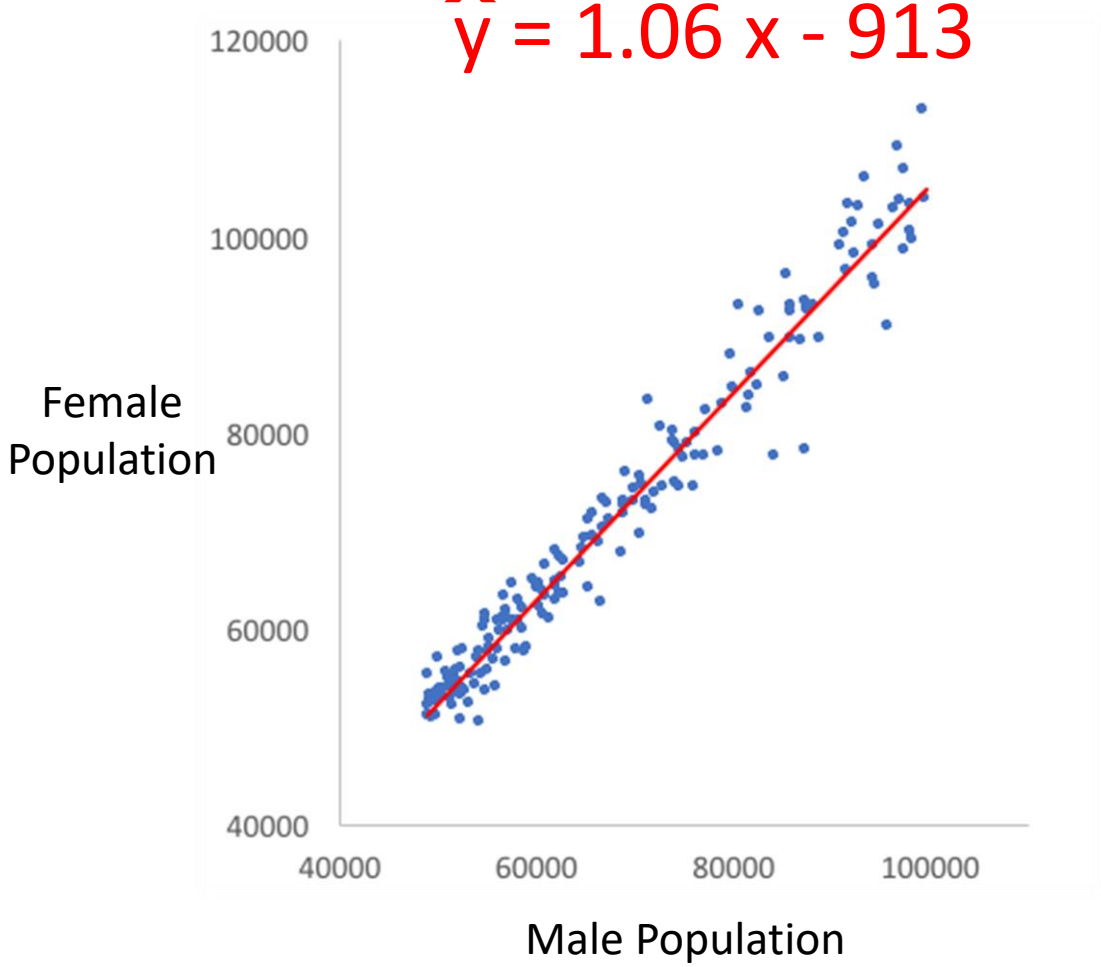
$\hat{y} = 1.06x - 913$



Based only on cities with <100k males

Why Regression? Making Predictions

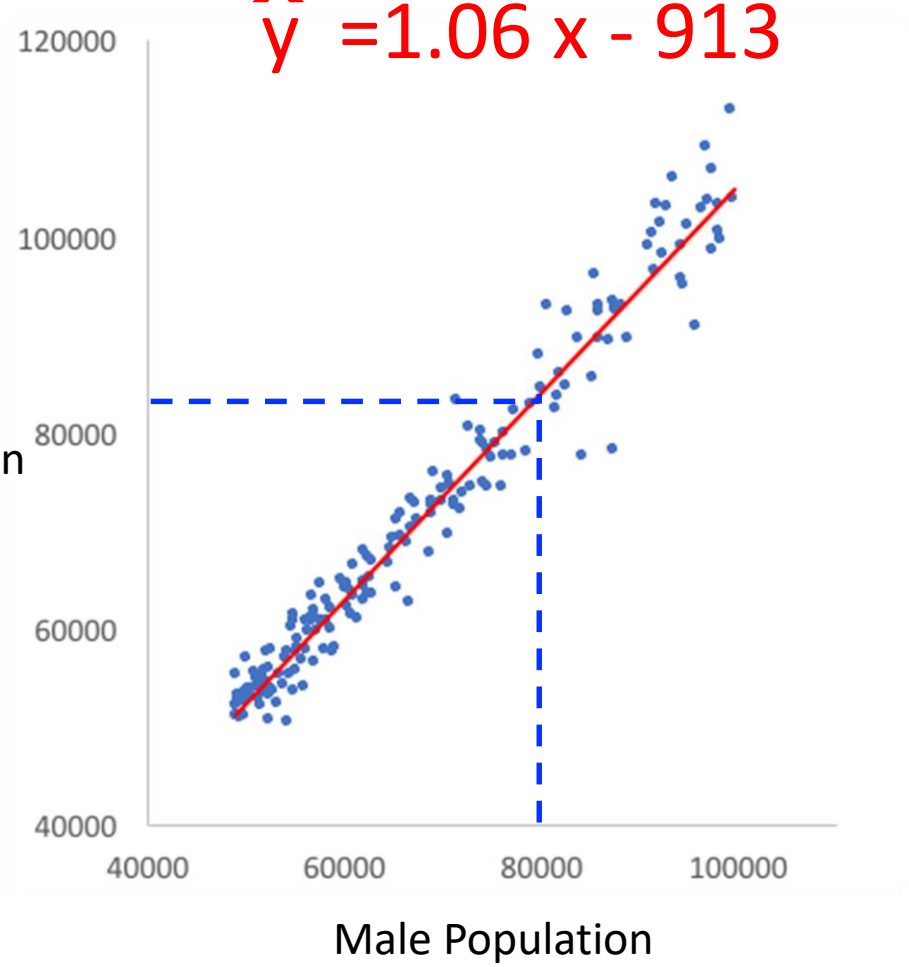
$\hat{y} = 1.06 x - 913$



What is the predicted female population of a city with 80000 males?

Why Regression? Making Predictions

$\hat{y} = 1.06x - 913$



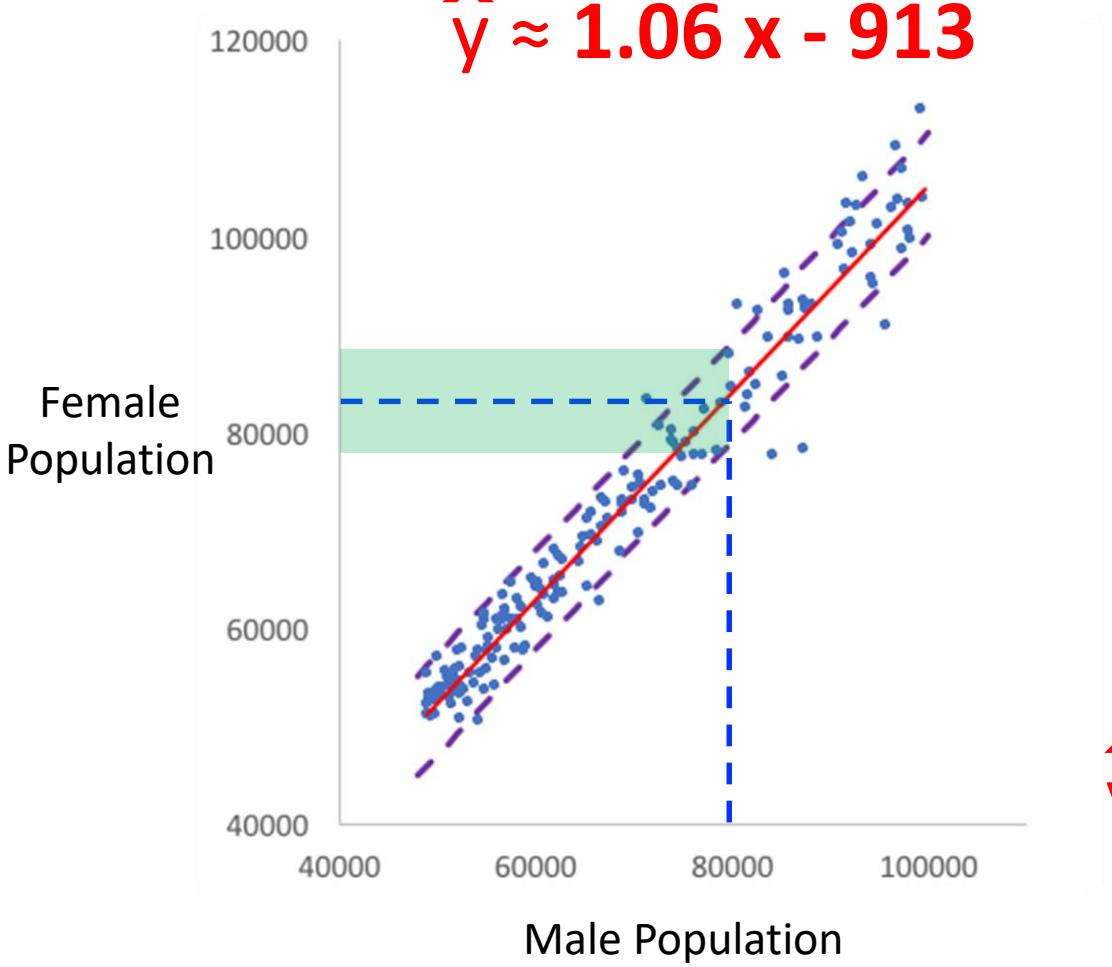
What is the predicted female population of a city with 80000 males?

$\hat{y} = 1.06(80000) - 913$

$\hat{y} = 83\,900$

Why Regression? Making Predictions

$\hat{y} \approx 1.06\,x - 913$



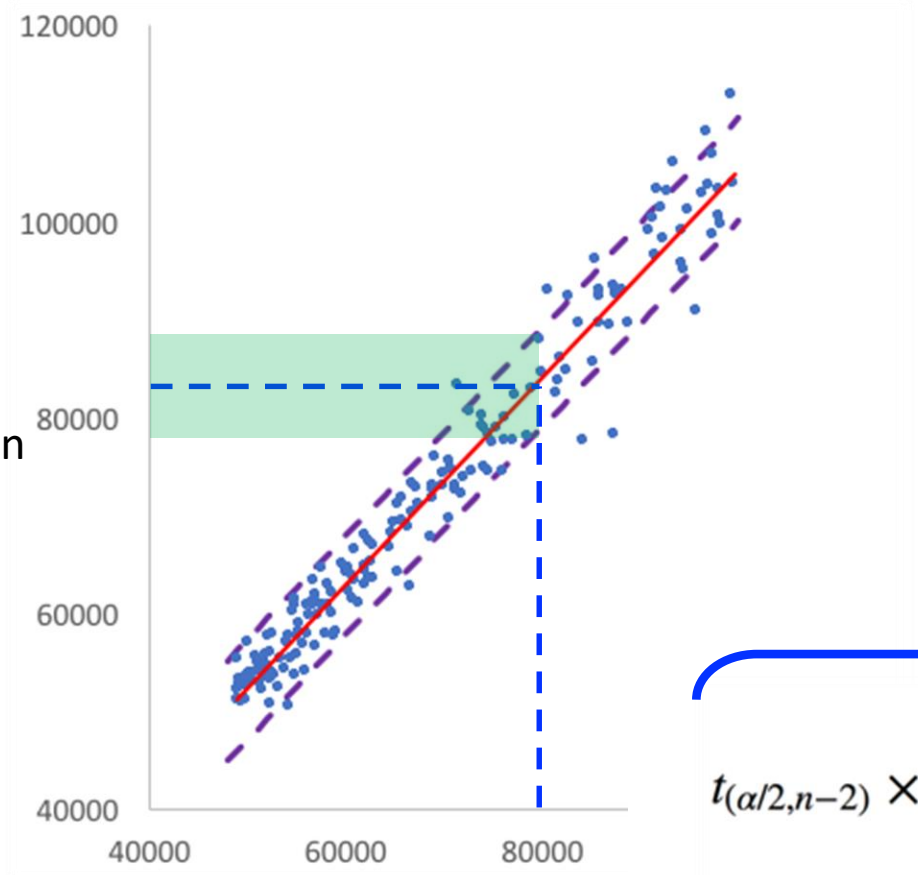
What is the predicted female population of a city with 80000 males?

(95% confidence)

$\hat{y} = 1.06(80000) - 913$

$\hat{y} = 83\,900$

Why Regression? Making Predictions

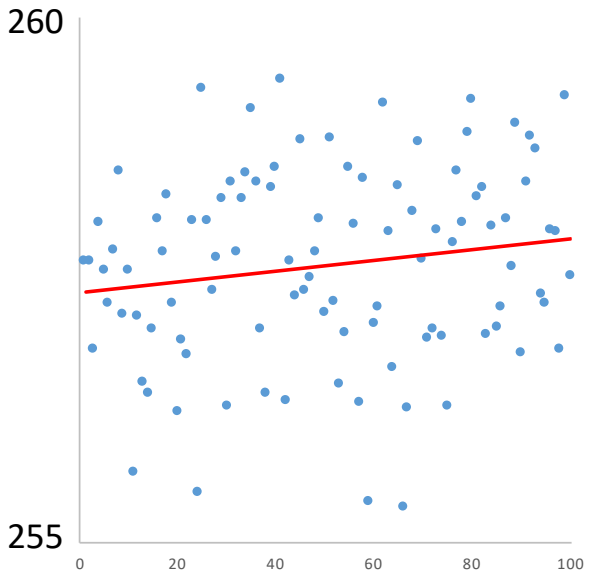


What is the predicted female population of a city with 80000 males?

$$t_{(\alpha/2,n-2)} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

Testing the Model

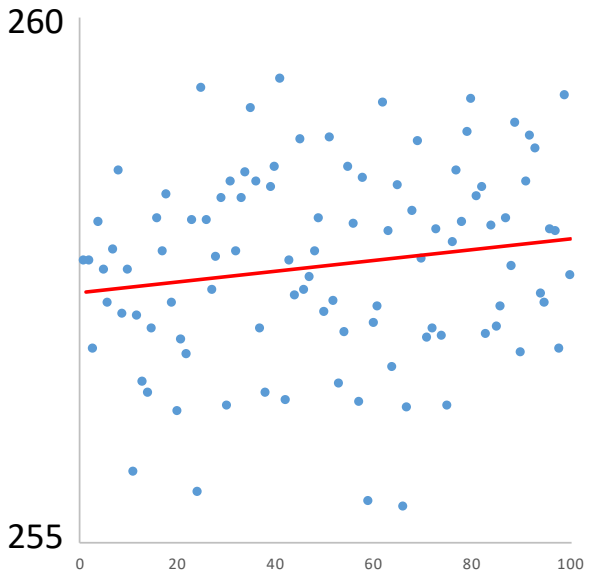
T-Test / F-Test: Is the gradient significantly different from zero?



Testing the Model

T-Test / F-Test: Is the gradient significantly different from zero?

The associated p-value measures the unlikeliness of the gradient arising from random noise.

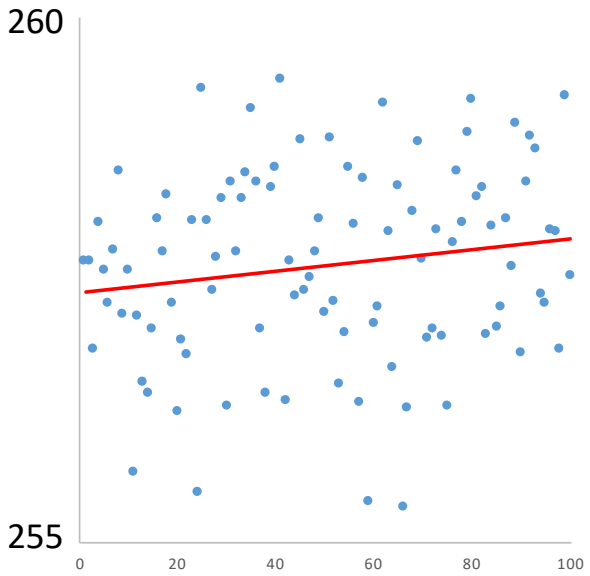


Testing the Model

T-Test / F-Test: Is the gradient significantly different from zero?

The associated p-value measures the unlikeliness of the gradient arising from random noise.

Generally, a p-value below 0.05 is evidence for a genuine relationship.

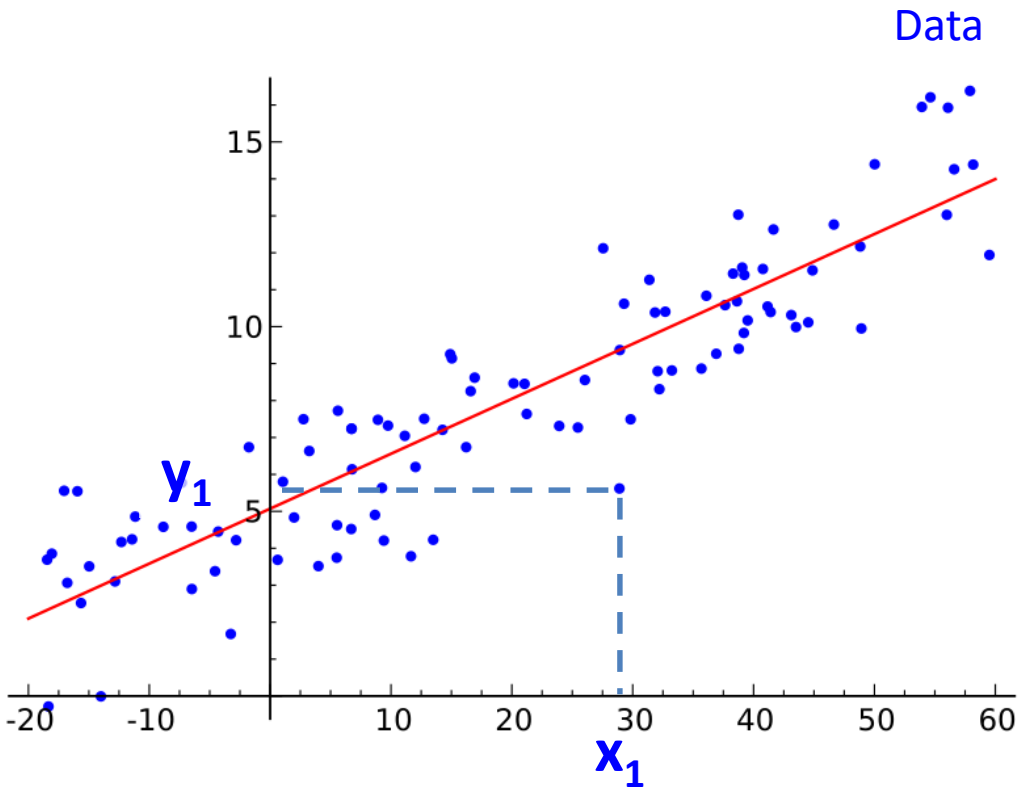


Simple Linear Regression

What to Report and Visualise

Fitted Equation:	$\hat{y} = mx + c$
p-value:	Is the result significant?
R-Squared Value:	Goodness of fit
Data Scatter Plot:	With fitted line
Residuals vs Fits Plot:	Consider LINE conditions

Part 3: Multiple Regression



The modelled relationship:

$$\hat{y} = mx + c$$

Part 3: Multiple Regression

Image: Teepetersen

Image: Public Domain

Motivation

	gender	greenery index	natural light index	reported well-being
0	M	124.7	28.4	6
1	F	67.9	64.0	6
2	M	129.4	79.5	7
3	F	111.1	130.7	8
4	F	168.2	79.1	8
5	F	130.5	100.5	8
6	M	104.1	78.0	7
7	F	149.5	85.8	7
8	F	116.2	140.3	8
9	M	112.7	103.0	7
10	M	111.0	124.5	7
11	F	65.5	115.8	7
12	M	141.1	134.6	8
13	M	47.7	116.6	6

office_environ_data_1.csv

Survey responses on quality of office environment, with measures of respondent gender, greenery (e.g. pot plants) & natural light.

office_environ_investigation.ipynb

Key questions

	gender	greenery index	natural light index	reported well-being
0	M	124.7	28.4	6
1	F	67.9	64.0	6
2	M	129.4	79.5	7
3	F	111.1	130.7	8
4	F	168.2	79.1	8
5	F	130.5	100.5	8
6	M	104.1	78.0	7
7	F	149.5	85.8	7
8	F	116.2	140.3	8
9	M	112.7	103.0	7
10	M	111.0	124.5	7
11	F	65.5	115.8	7
12	M	141.1	134.6	8
13	M	47.7	116.6	6

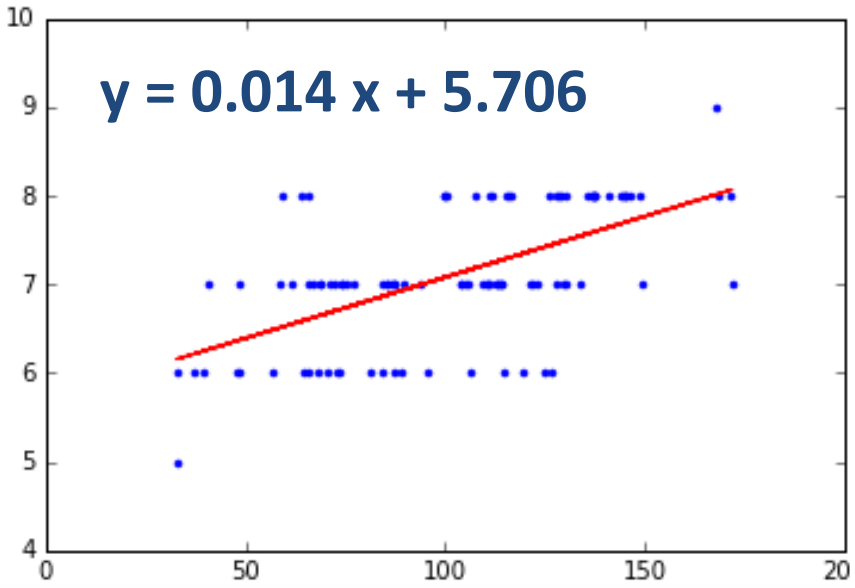
1. Can we combine greenery and natural light to create a model for predicting well-being?
2. How does the combined model compare with individual models?
3. Can we put categorical variables into the model?

```
slope, intercept, r_value, p_value, std_err = sps.linregress(X, Y)

b = plt.plot(X, X*slope + intercept, 'r') # Plot the regression line.

print "Rsq = ", r_value**2
print "St Err. = ", std_err
print "p-value = ", p_value
```

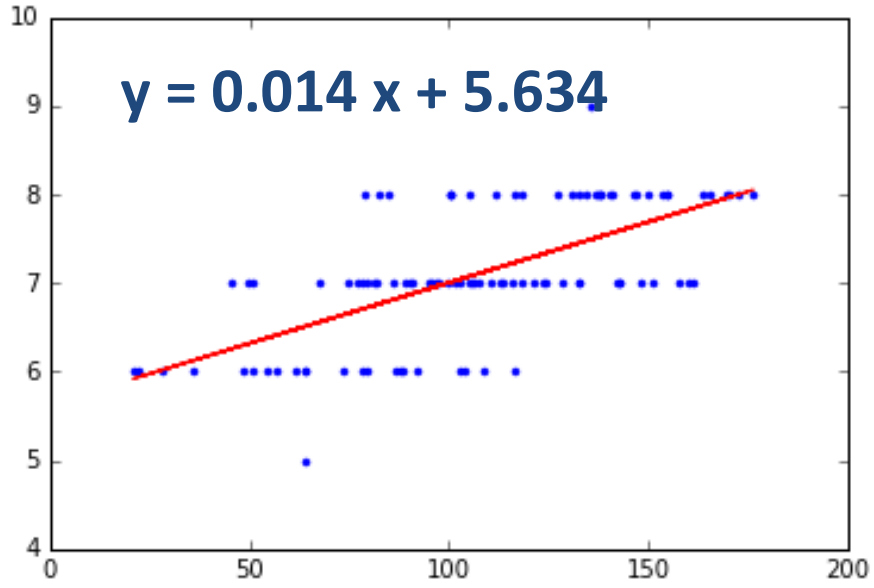
```
Rsq =  0.345235165782
St Err. =  0.00191058863189
p-value =  1.30754858938e-10
```

```
Rsq =  0.410481587673
St Err. =  0.00165780706628
p-value =  7.02929172235e-13
```

reported well-being



greenery index

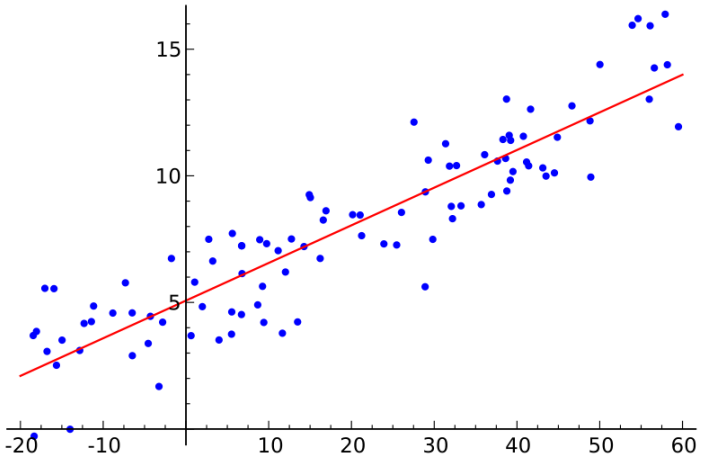


natural light index

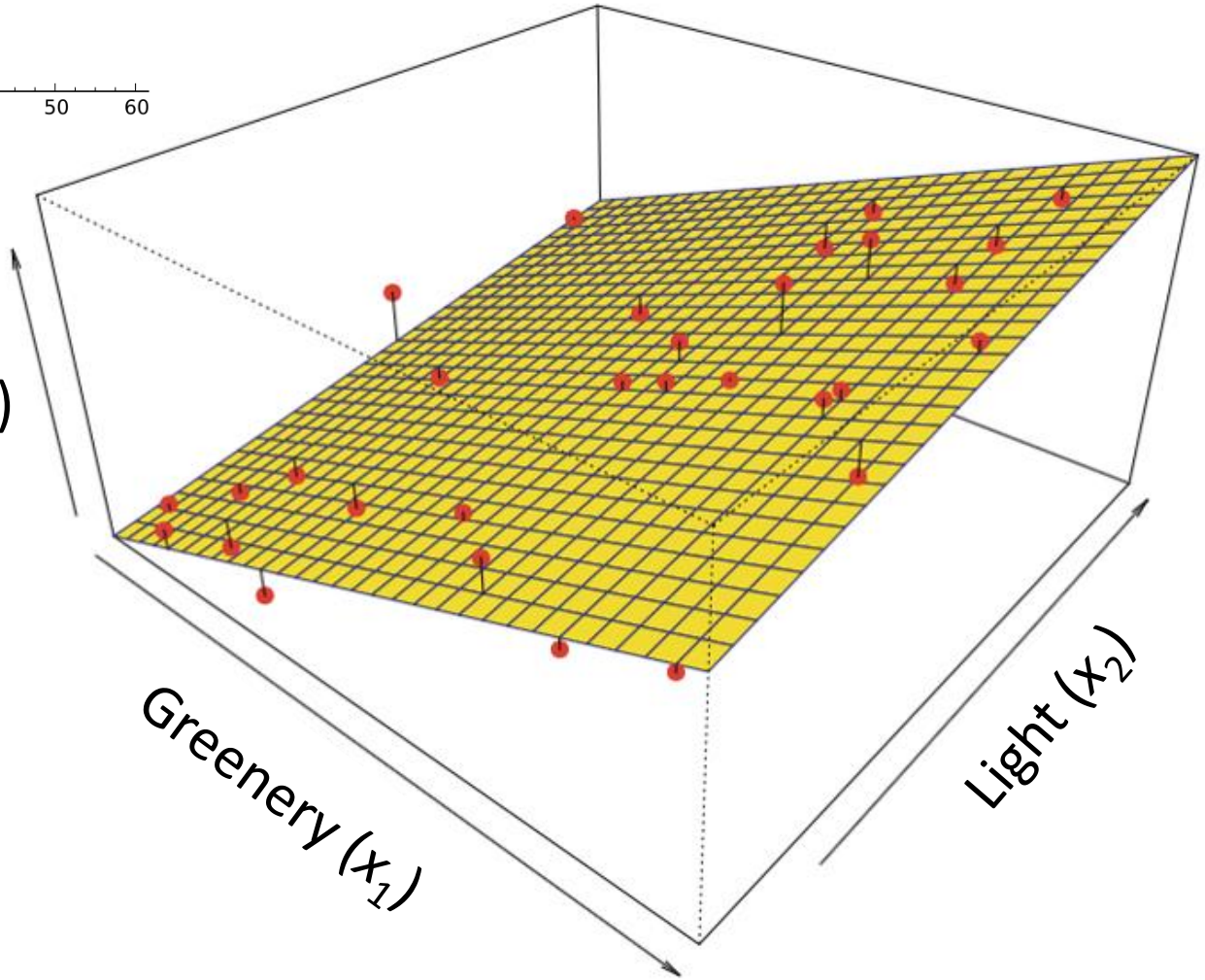
Key Question 1

With R^2 values of 30-40%, both models leave a lot of variation unaccounted for.

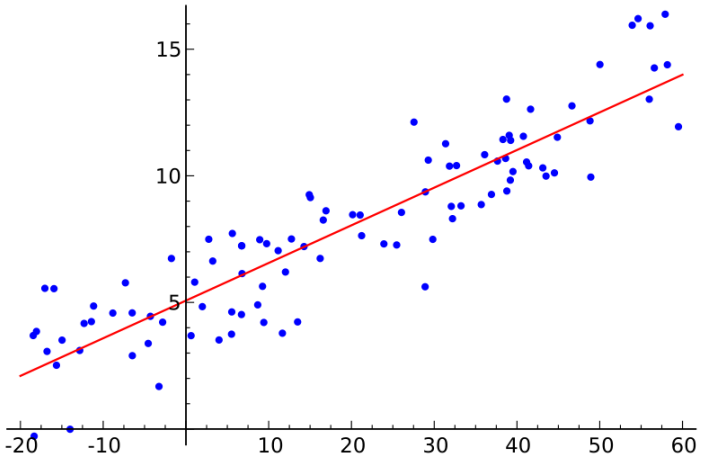
Could we do better, by using BOTH data series to explain the variation in the reported well-being?



Well-being (*y*)



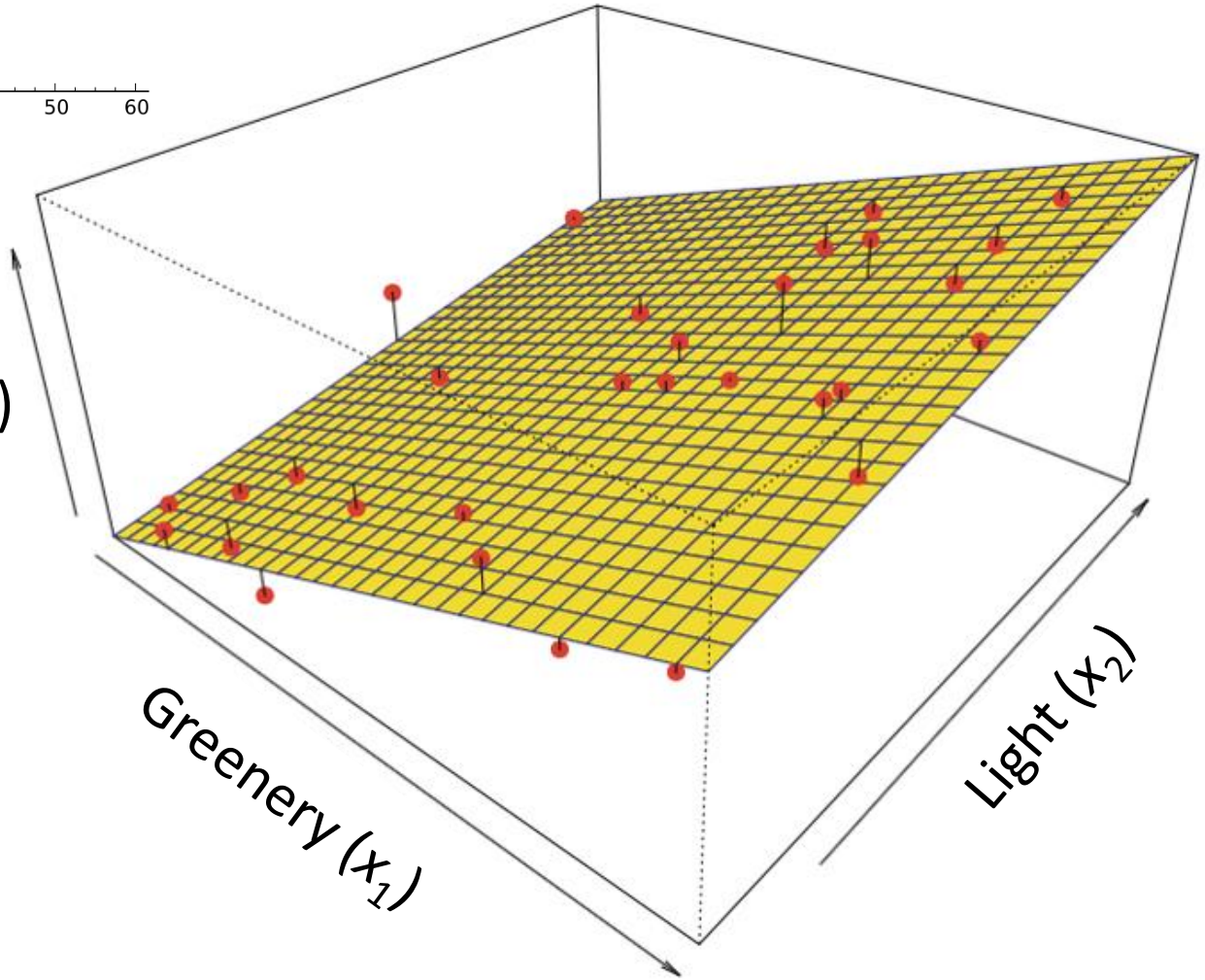
http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/MultipleLinearRegression-Plane.png



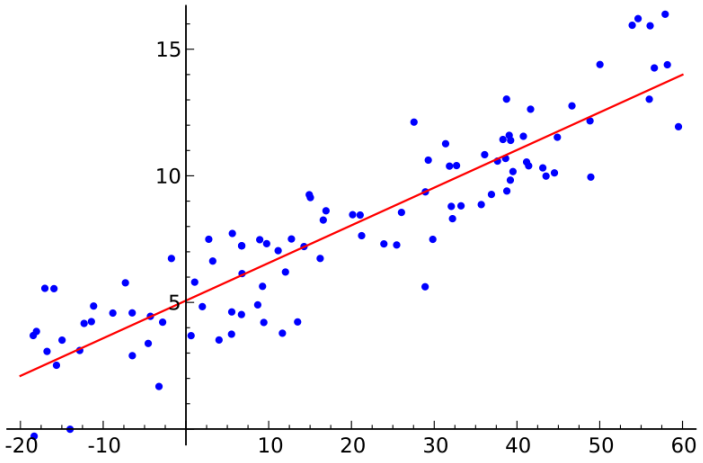
The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Well-being (y)



http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/MultipleLinearRegression-Plane.png

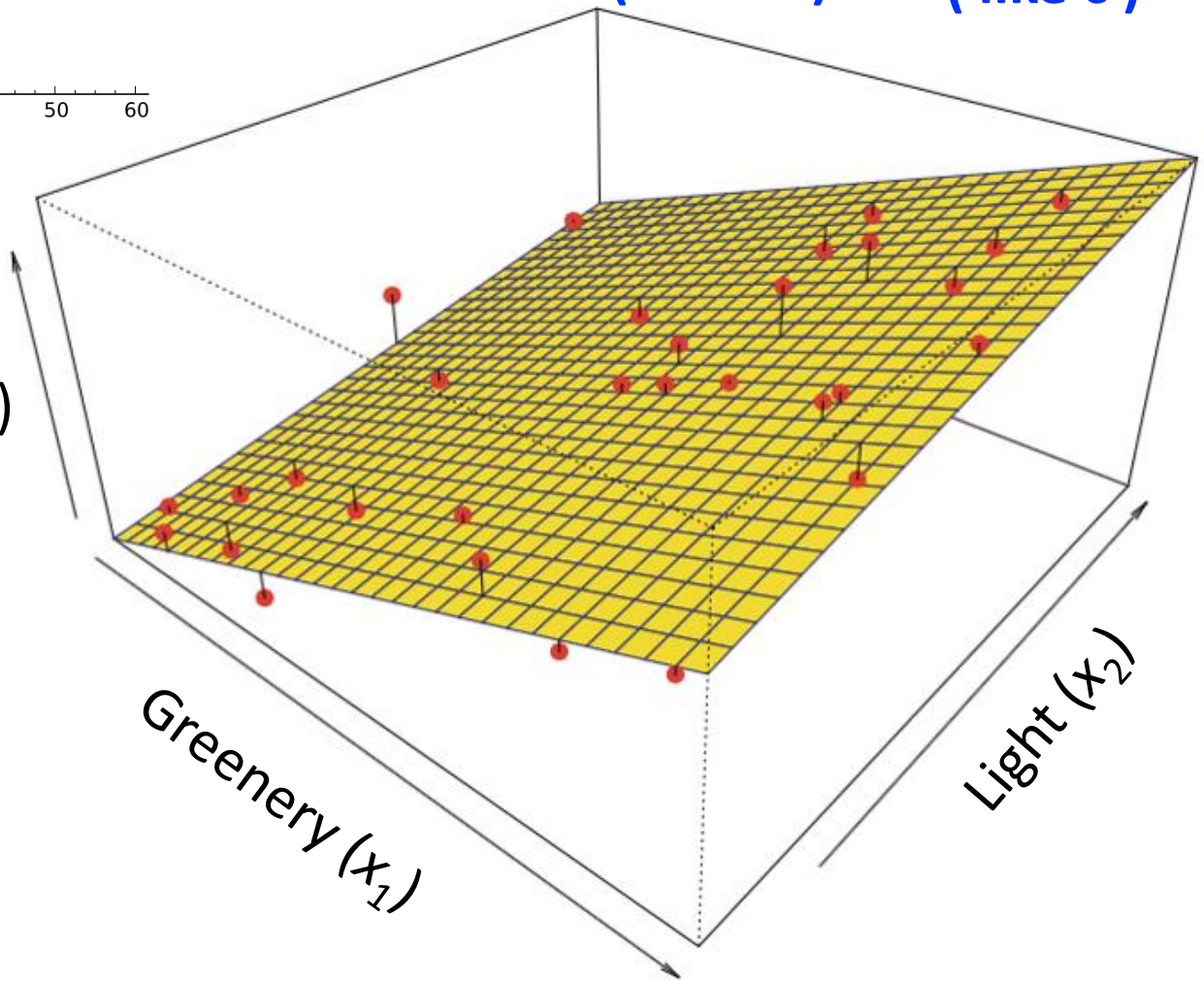


The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

(like m) (like m) (like c)

Well-being (y)



http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/MultipleLinearRegression-Plane.png

The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

well-being greenery light

	gender	greenery index	natural light index	reported well-being
0	M	124.7	28.4	6
1	F	67.9	64.0	6
2	M	129.4	79.5	7
3	F	111.1	130.7	8
4	F	168.2	79.1	8
5	F	130.5	100.5	8
6	M	104.1	78.0	7

Answer to Key Question 1

```
import statsmodels.formula.api as smf
```

```
y = 0.014 x1 + 0.014 x2 + 4.267
R^2       =  0.748588335477
p-value_1 =  0.0
p-value_2 =  0.0
```

Combined model explains about 75% of variance.
Close to the sum (35% + 41%) of the individual R^2 values.

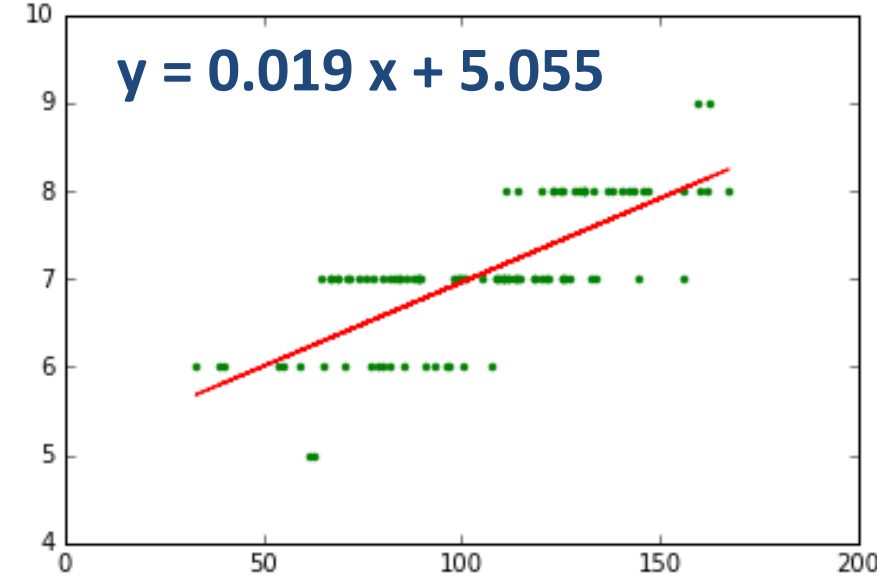
Alternative Data Set

	gender	greenery index	natural light index	reported well-being
0	M	86.1	93.8	7
1	M	110.5	70.0	7
2	F	77.6	93.0	7
3	M	130.7	99.6	8
4	M	68.7	110.3	7
5	M	55.0	32.1	6
6	M	123.0	162.7	8
7	M	121.3	111.8	7
8	F	122.9	91.3	8
9	F	62.7	43.4	5
10	F	114.9	99.1	7
11	F	67.1	86.3	7
12	F	118.6	111.1	7
13	F	125.2	108.3	8

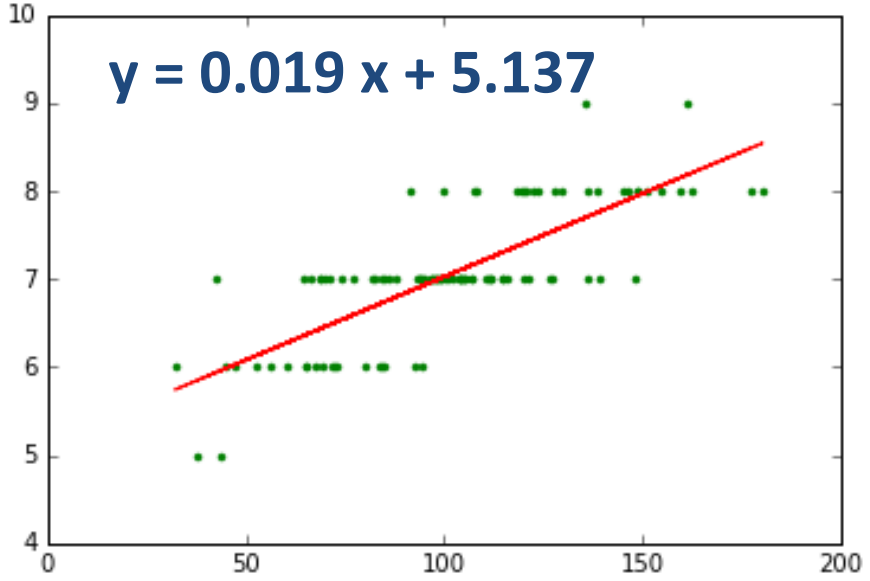
Q: a multiple regression model is always as good as a sum of the individual models?

Try this adjusted dataset:
office_environ_data_2.csv

```
y = 0.019 x + 5.055
Rsq =  0.561033874644
p-value =  3.21002614461e-19
```




```
y = 0.019 x + 5.137
Rsq =  0.599178170721
p-value =  3.61575389149e-21
```



Key Question 2

How does the combined model compare with individual models?

Answer to Key Question 2

office_environ_data_2.csv

```
y = 0.012 x1 + 0.013 x2 + 4.56
R^2       =   0.741004206603
p-value_1 =   0.0
p-value_2 =   0.0
```

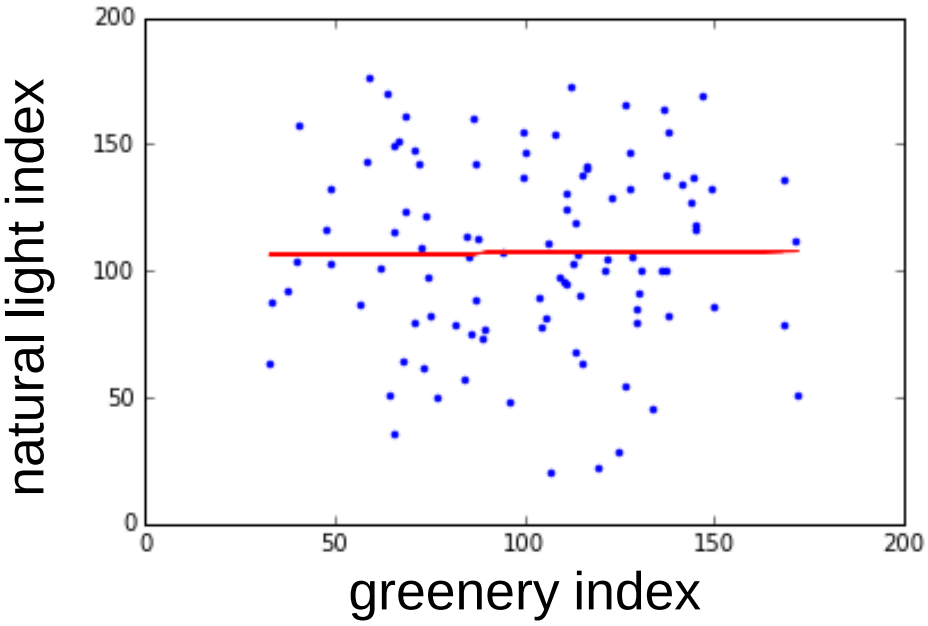
Combined model explains about 74% of variance.
Less than the combined weaker models in data set 1.
Much less than sum (56% + 60%) of individual R^2 values.

What is going on?

Resolution

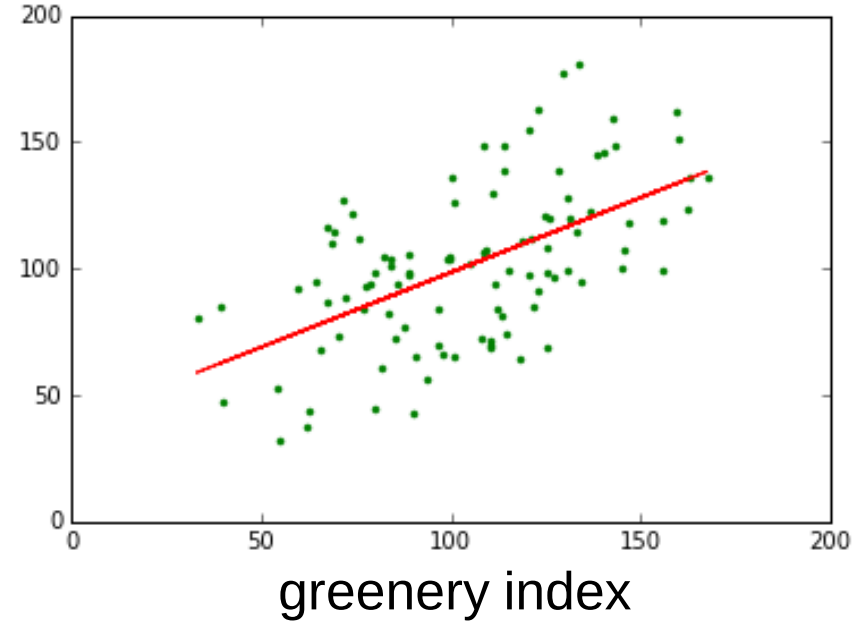
office_environ_data_1.csv

```
y = 0.01 x + 105.976
Rsq =  9.13718302054e-05
p-value =  0.924800176063
```



office_environ_data_2.csv

```
y = 0.591 x + 39.401
Rsq =  0.321304180146
p-value =  7.8542077053e-10
```



Resolution

office_environ_data_1.csv

	greenery	light	wellbeing
greenery	1.00	0.01	0.59
light	0.01	1.00	0.64
wellbeing	0.59	0.64	1.00

correlations

office_environ_data_2.csv

	greenery	light	wellbeing
greenery	1.00	0.57	0.75
light	0.57	1.00	0.77
wellbeing	0.75	0.77	1.00

correlations

(See correlation_matrix_python.py)

Answer to Key Question 2

The combined model is not always as good as a sum of the individual models.
One factor is the correlation between the independent variables.

Key Question 3

	gender	greenery index	natural light index	reported well-being
0	M	124.7	28.4	6
1	F	67.9	64.0	6
2	M	129.4	79.5	7
3	F	111.1	130.7	8
4	F	168.2	79.1	8
5	F	130.5	100.5	8
6	M	104.1	78.0	7
7	F	149.5	85.8	7
8	F	116.2	140.3	8
9	M	112.7	103.0	7
10	M	111.0	124.5	7
11	F	65.5	115.8	7
12	M	141.1	134.6	8
13	M	47.7	116.6	6

office_environ_data_1.csv

Could the gender information explain even more variation?

But how can we include it?

Categorical Data

	gender	greenery index	natural light index	reported well-being	gender_cat
0	M	124.7	28.4	6	1
1	F	67.9	64.0	6	0
2	M	129.4	79.5	7	1
3	F	111.1	130.7	8	0
4	F	168.2	79.1	8	0
5	F	130.5	100.5	8	0
6	M	104.1	78.0	7	1
7	F	149.5	85.8	7	0
8	F	116.2	140.3	8	0

Categorical Data

	gender	greenery index	natural light index	reported well-being	gender_cat
0	M	124.7	28.4	6	1
1	F	67.9	64.0	6	0
2	M	129.4	79.5	7	1
3	F	111.1	91.7	7	0
4	F	168.2	79.1	8	0
5	F	130.5	100.5	8	0
6	M	104.1	78.0	7	1
7	F	149.5	85.8	7	0
8	F	116.2	140.3	8	0

The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_0$$

well-being greenery light gender

Categorical Data

colour
blue
green
green
blue
blue
red
blue

What if there were 3 categories?

Categorical Data

colour	is_blue	is_green
blue	1	0
green	0	1
green	0	1
blue	1	0
blue	1	0
red	0	0
blue	1	0

If there were
3 categories,
you would need
2 new variables

"dummy variable"

Categorical Data

The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0$$

well-being greenery light is_blue is_green

The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_0$$

well-being greenery light gender

Answer to Key Question 3

```
y = 0.014 x1 + 0.014 x2 + 0.021 x3 + 4.256
R^2       =  0.748763383852
p-value_1 =  0.0
p-value_2 =  0.0
p-value_3 =  0.796
```

Previously…

R^2 = 0.748588335477

Combined model explains a little more of the variance.
But… p-value suggests gender adds no information.

The best equation to explain well-being,
would involve greenery and natural light only.

Which variables to include

- We will talk about variable selection next week

- Visualise data and check correlations between variables.
  Highly correlated variables provide less information.

- Variables with high p-values *probably* do not add information.
  Try the model without them (but check the new R^2 value).

- A large improvement in R^2 suggests that a new variable adds information; a small change in R^2 may not mean anything.
  - R^2 always increases when new variables are added

- For more on how to choose variables, see:
  https://onlinecourses.science.psu.edu/stat501/node/283

MULTIPLE LINEAR REGRESSION

FIVE Necessary Conditions

Linear relationship exists

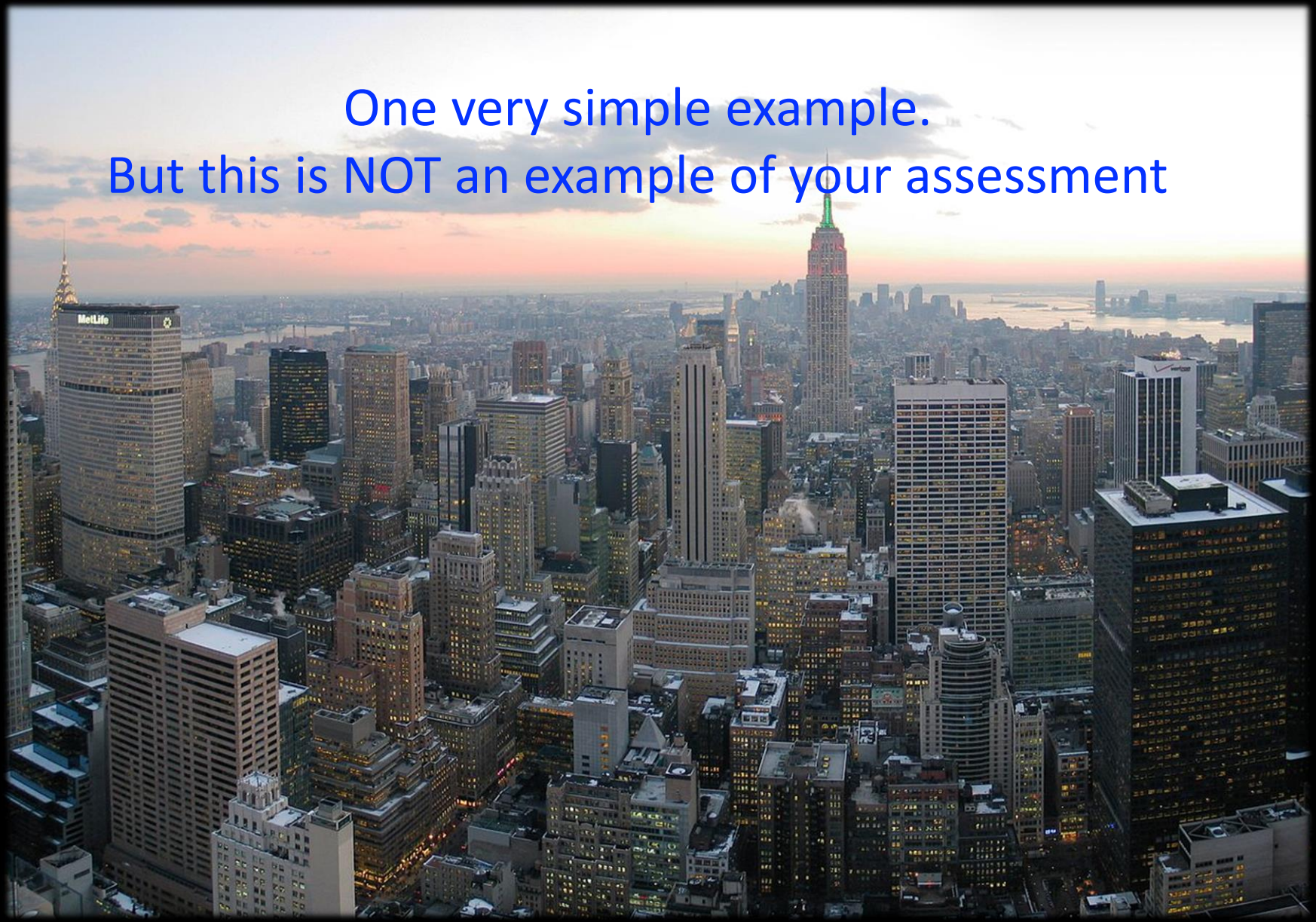
Independent errors

Normally distributed errors

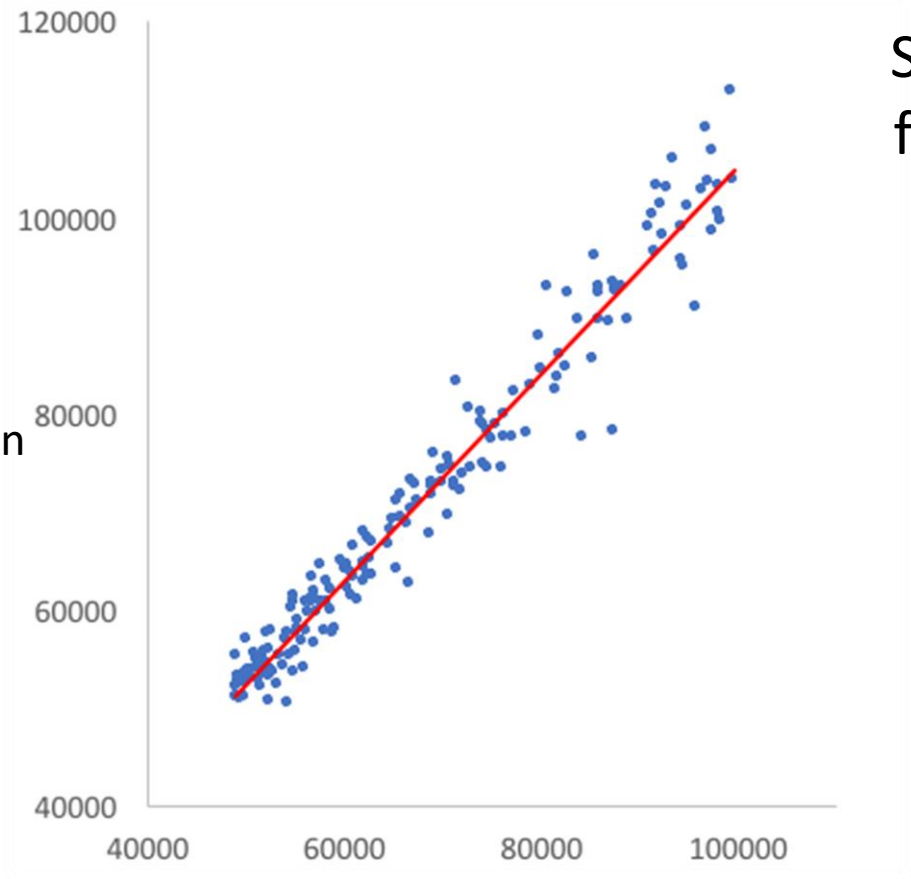
Equal error variance for all x values

No multicollinearity between the X variables

(Again, check with a Residuals vs. Fits Plot)

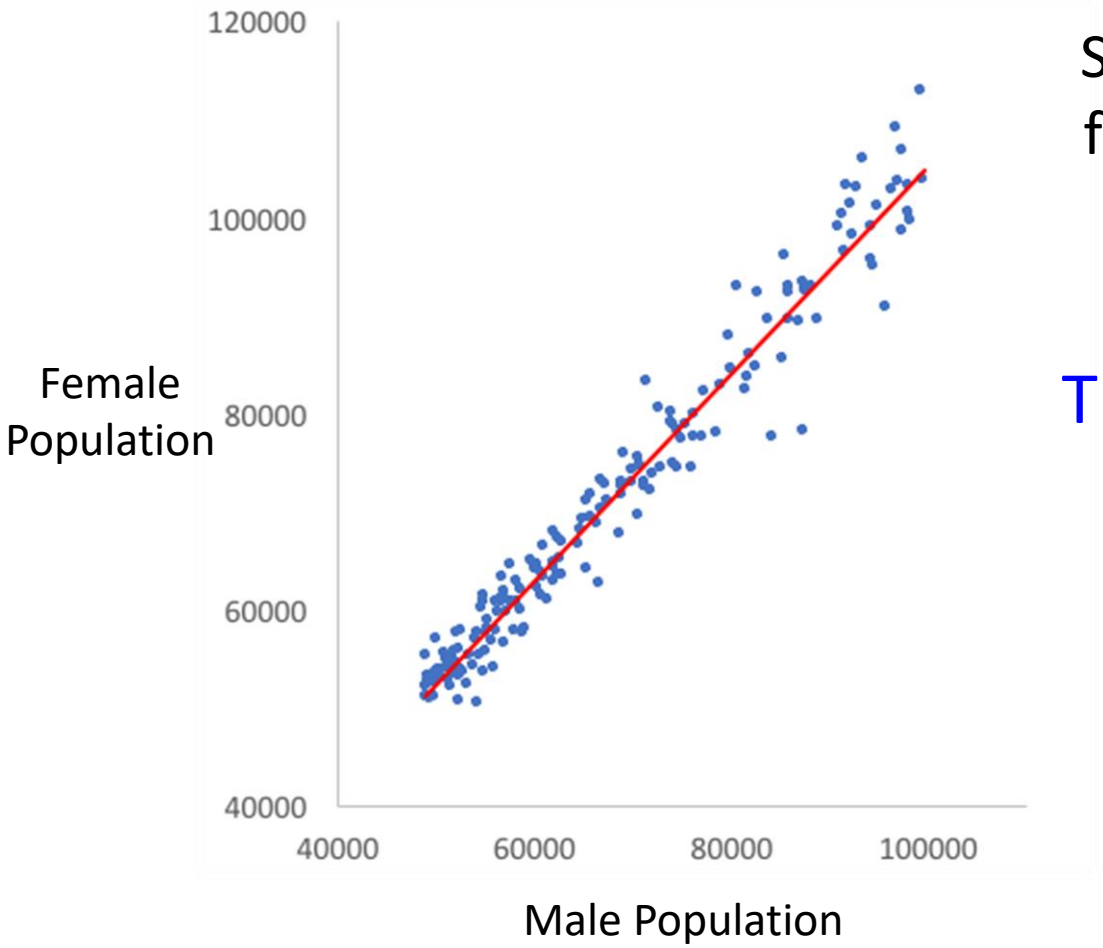

One very simple example.
But this is NOT an example of your assessment
MetLife

US Cities Data



Suppose we use the male and female populations to predict the overall populations

US Cities Data

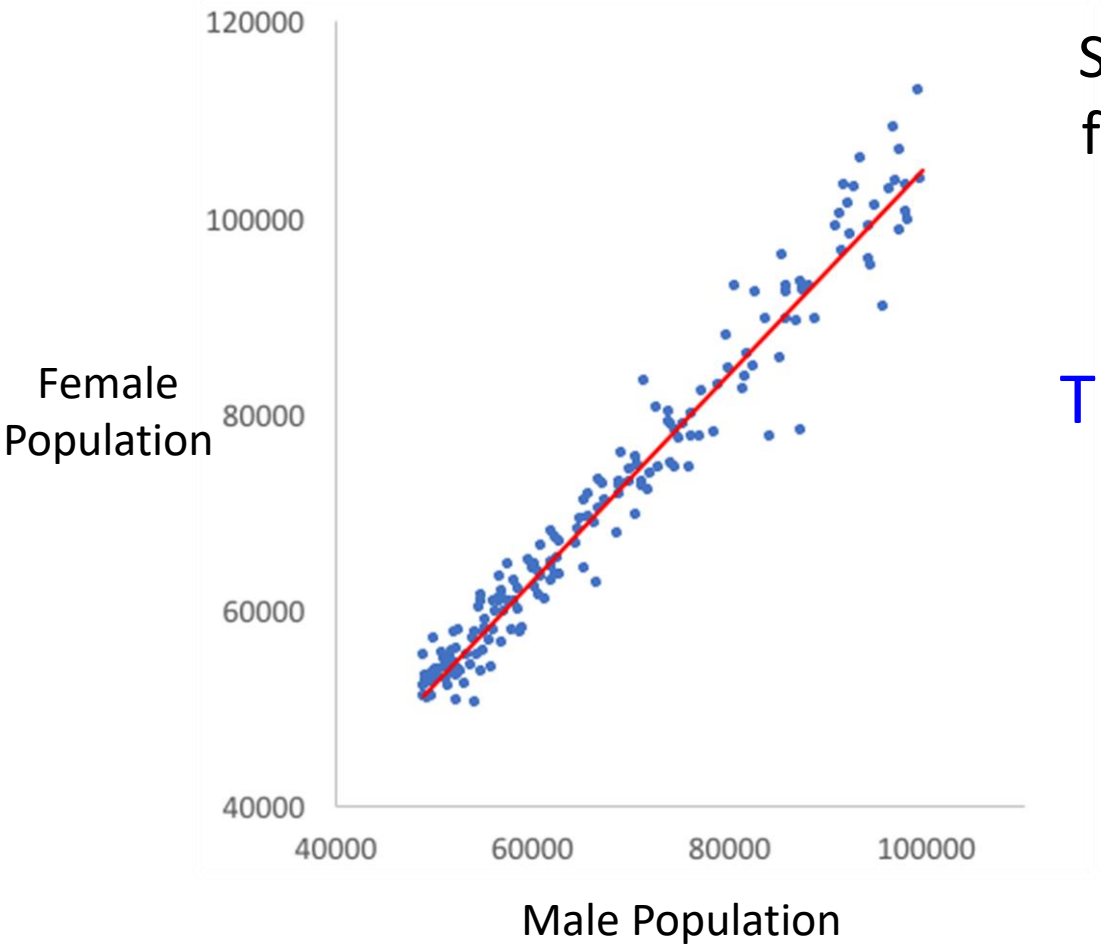


Suppose we use the male and female populations to predict the overall populations

The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

US Cities Data



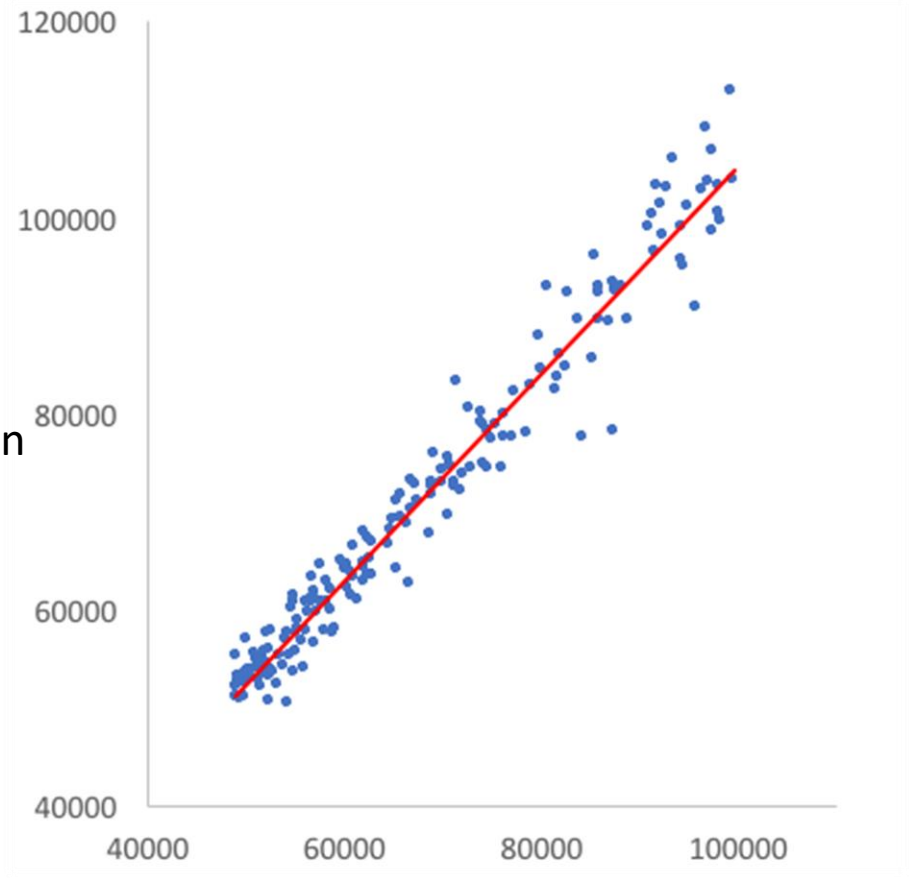
Suppose we use the male and female populations to predict the overall populations

The modelled relationship:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Total Pop. | Male Pop. | Fem. Pop.

US Cities Data



$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Total Pop. | Male Pop. | Fem. Pop.

β_1 = ???

β_2 = ???

β_0 = ???

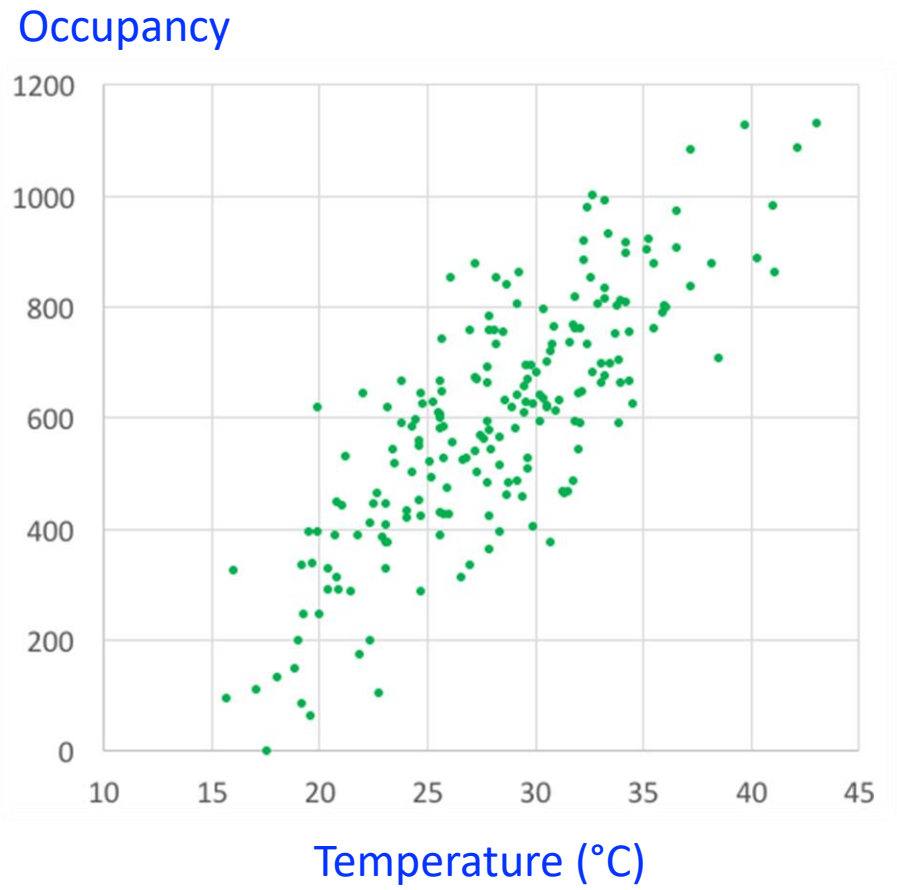
R^2 = ???

MSE = ???

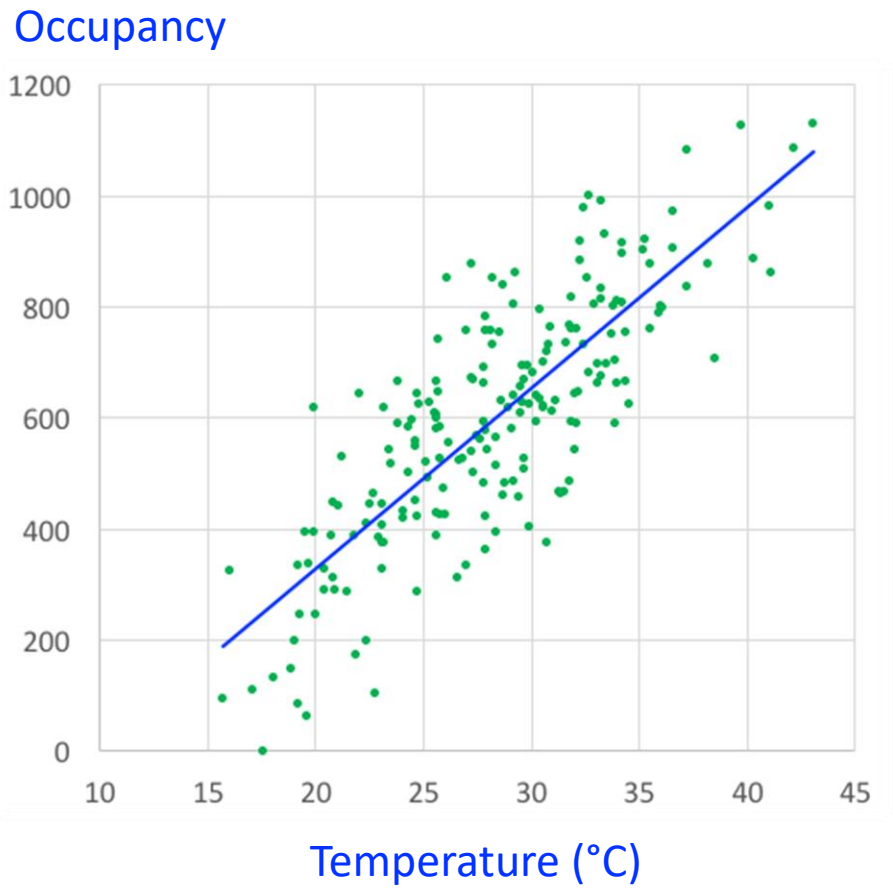
p-value = ???

Another example:
Consequence of multicollinearity between variables

Beach Occupancy Data



Beach Occupancy Data



$$\hat{y} = m\,x + c$$

Occupancy Temp. (°C)

m = 32.6

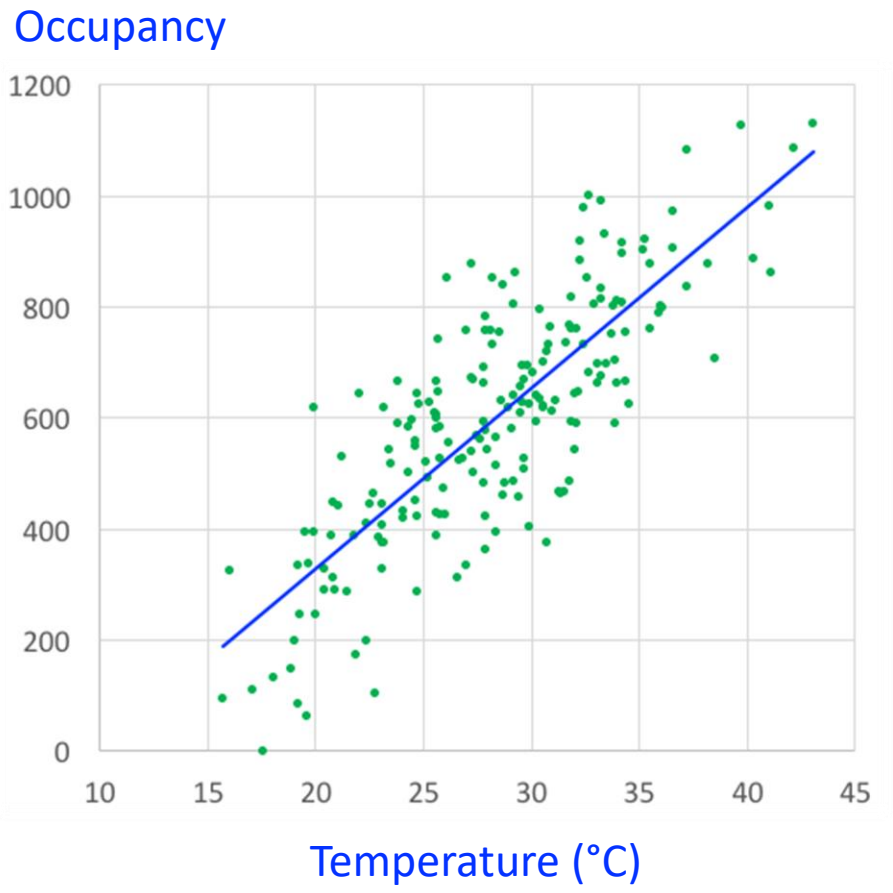
c = -323

R^2 = ???

MSE = ???

p-value = ???

Beach Occupancy Data



$$\hat{y} = m\,x + c$$

Occupancy Temp. (°C)

m = 32.6

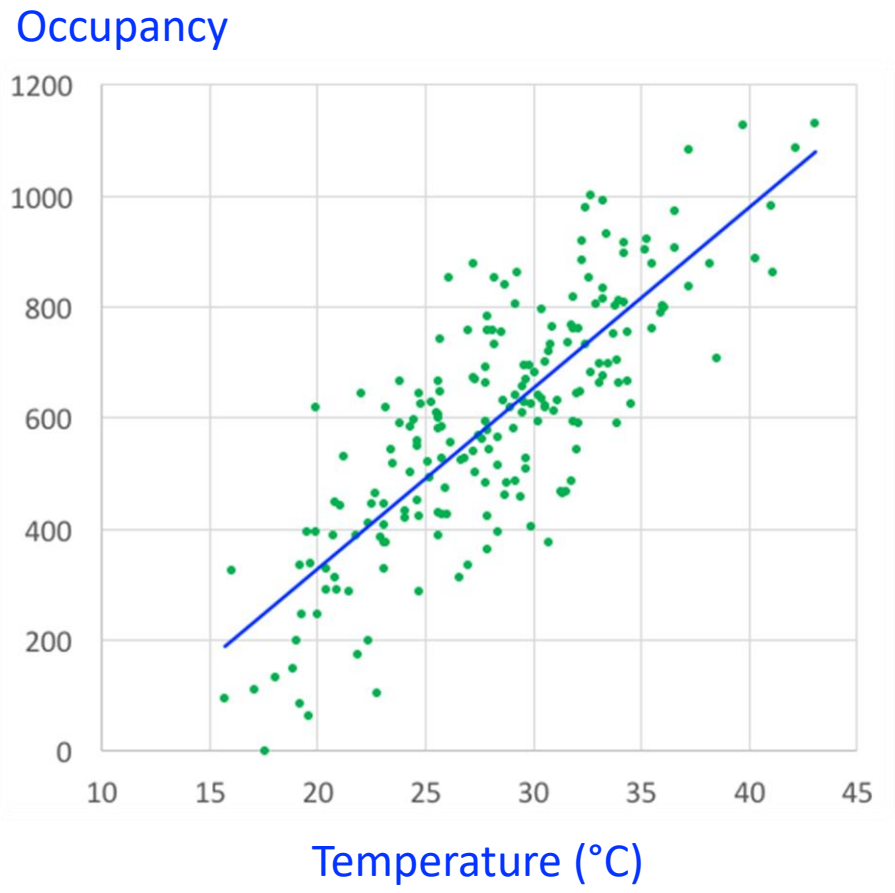
c = -323

R^2 = ???

MSE = 16 800

p-value = ???

Beach Occupancy Data



$$\hat{y} = m\,x + c$$

Occupancy Temp. (°C)

m = 32.6

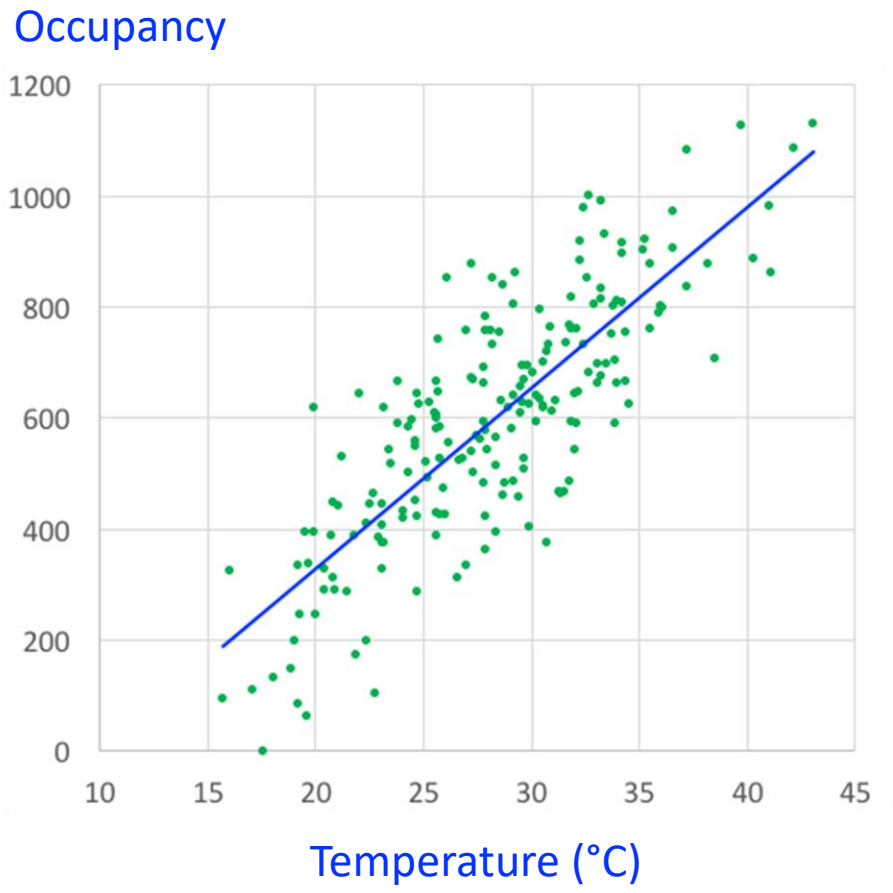
c = -323

R^2 = 0.650

MSE = 16 800

p-value = ???

Beach Occupancy Data



$$\hat{y} = m\,x + c$$

Occupancy Temp. (°C)

m = 32.6

c = -323

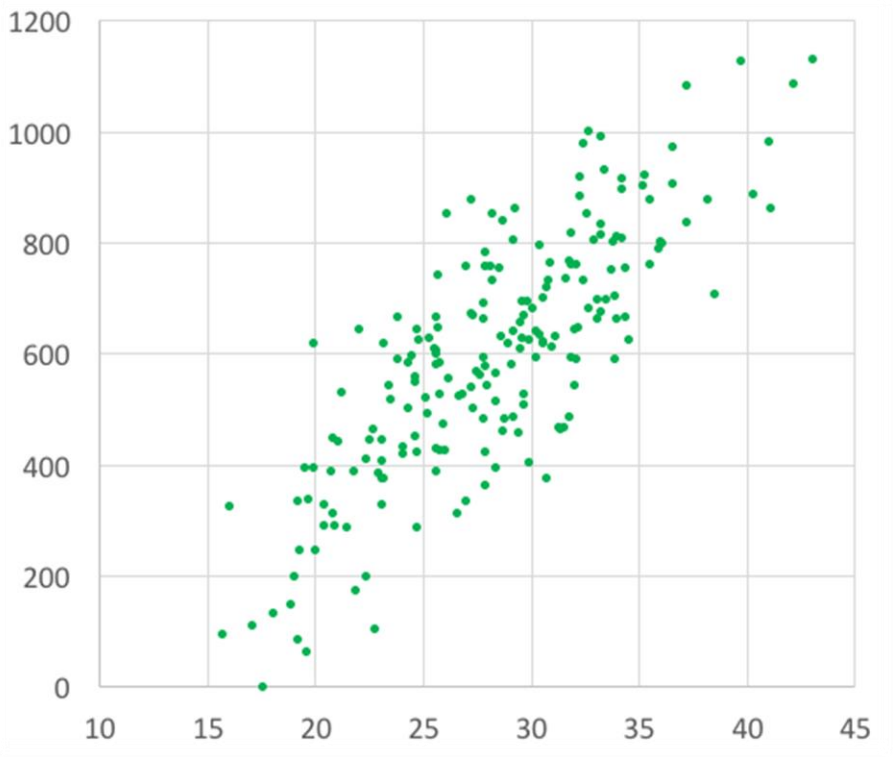
R^2 = 0.650

MSE = 16 800

p-value ≈ 0

Beach Occupancy Data

Occupancy

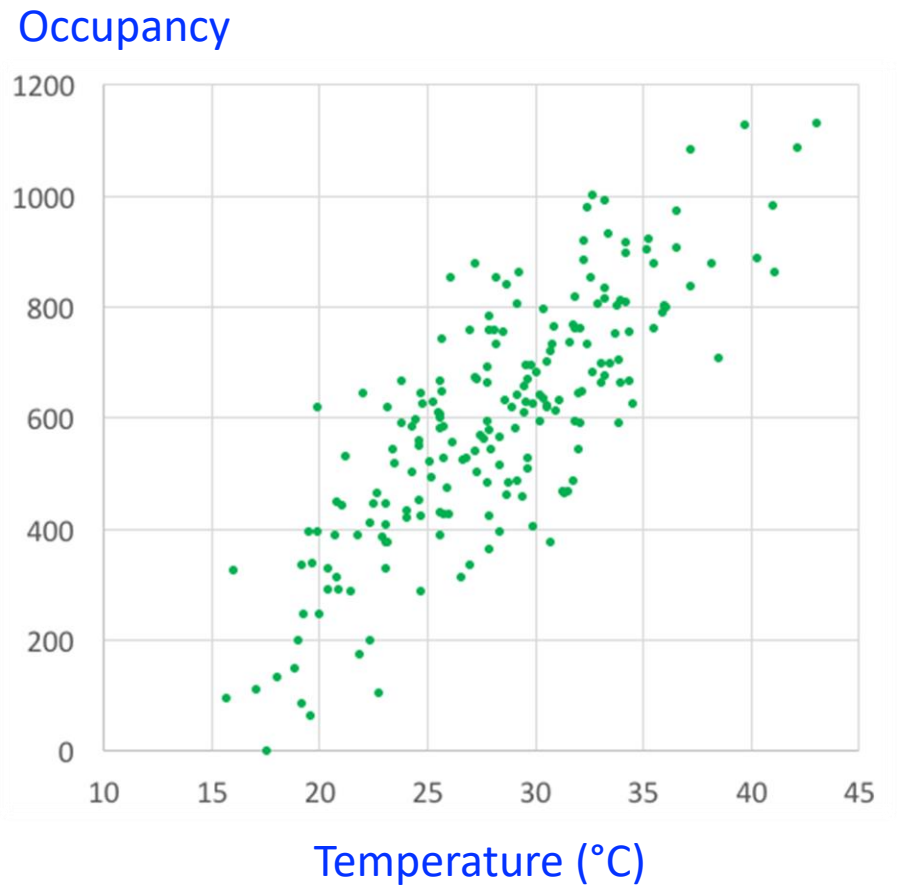


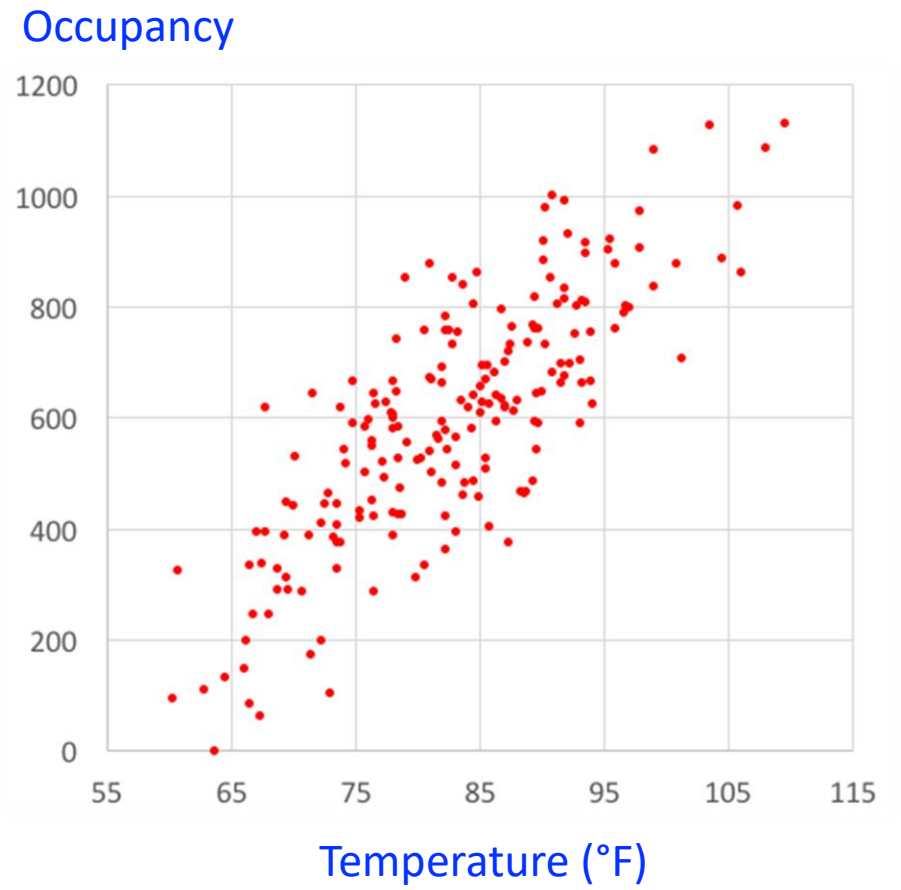
Temperature (°C)

Occupancy

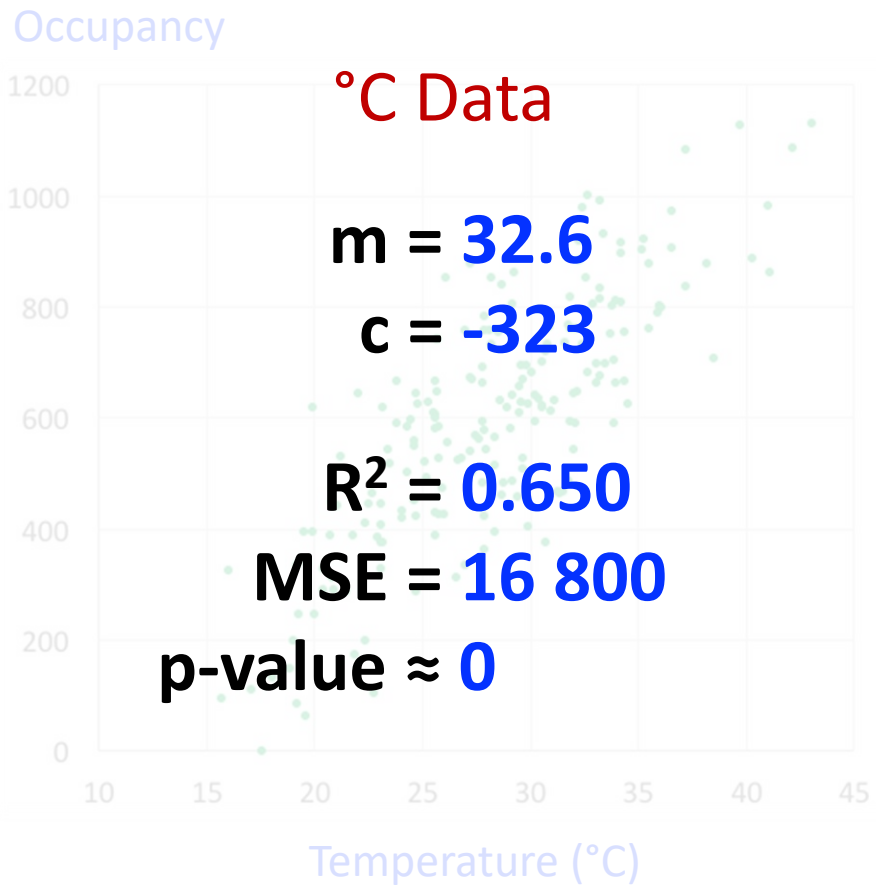
Temperature (°F)

Beach Occupancy Data





Beach Occupancy Data



°C Data

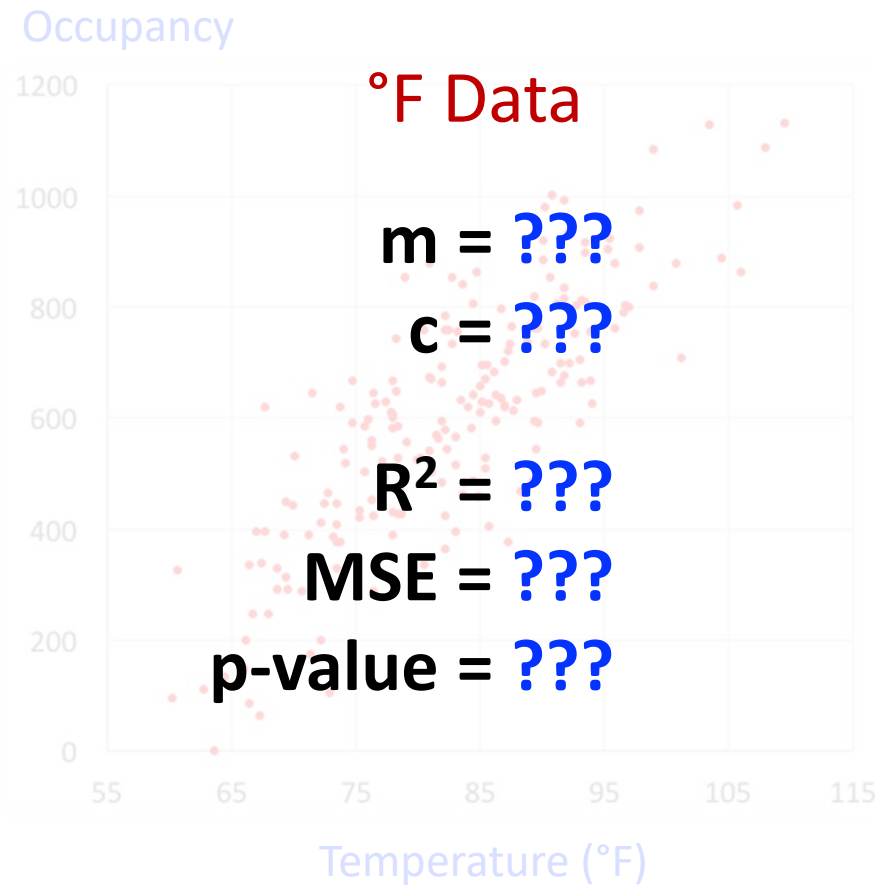
m = 32.6

c = -323

R^2 = 0.650

MSE = 16 800

p-value ≈ 0



°F Data

m = ???

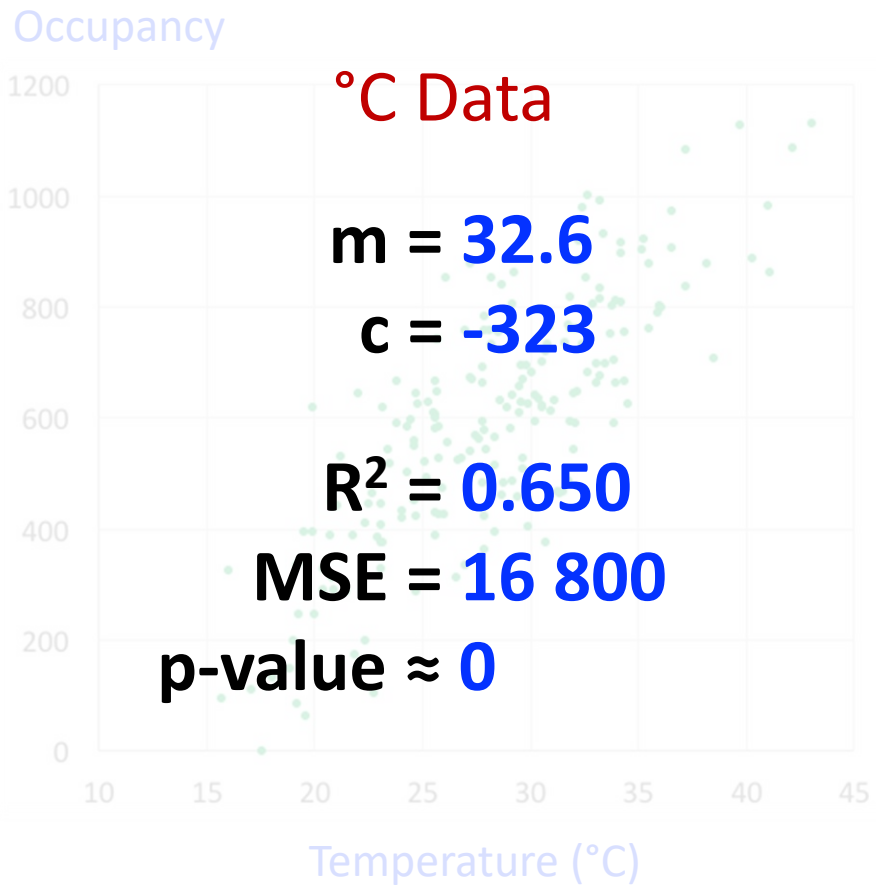
c = ???

R^2 = ???

MSE = ???

p-value = ???

Beach Occupancy Data



°C Data

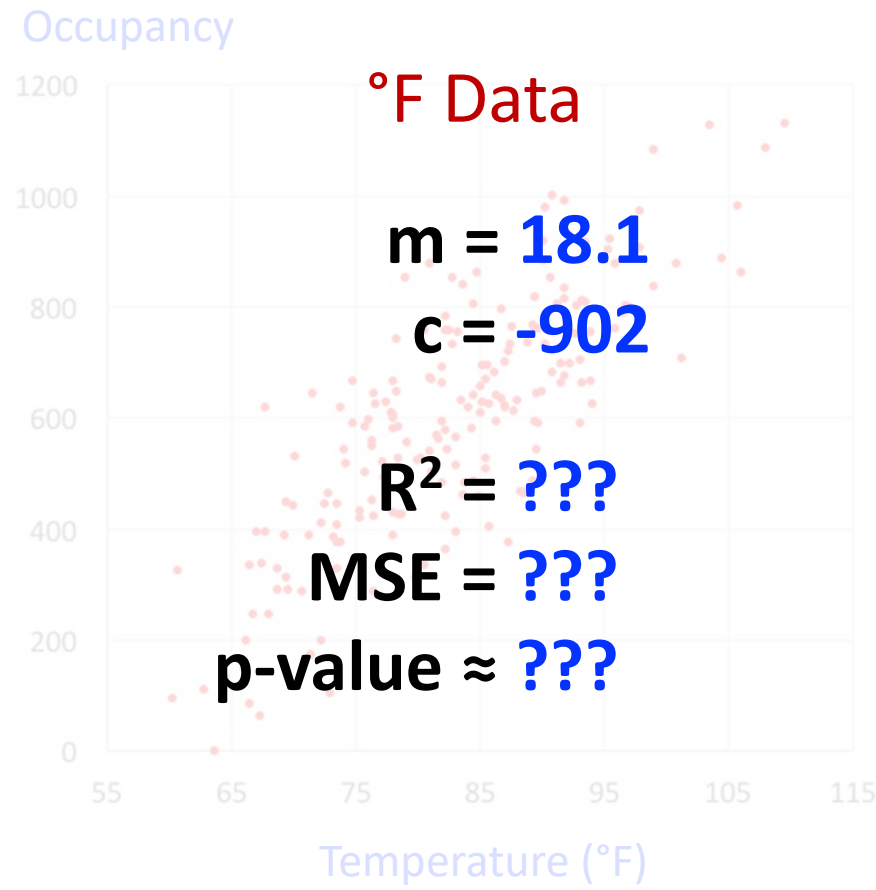
m = 32.6

c = -323

R^2 = 0.650

MSE = 16 800

p-value ≈ 0



°F Data

m = 18.1

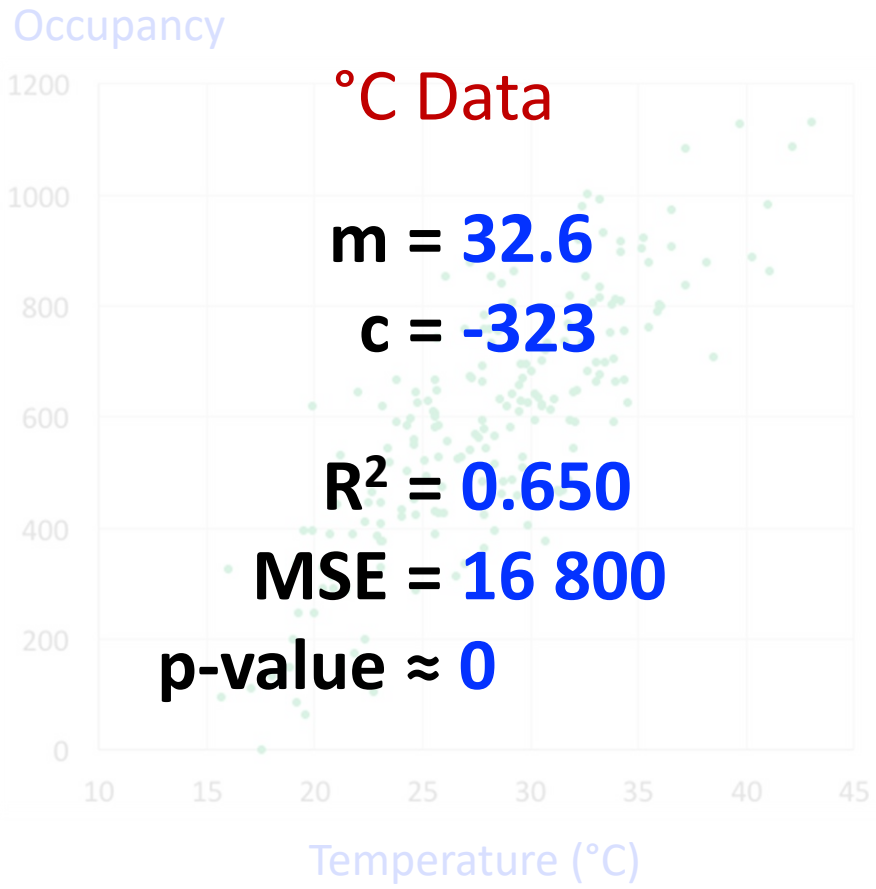
c = -902

R^2 = ???

MSE = ???

p-value ≈ ???

Beach Occupancy Data



°C Data

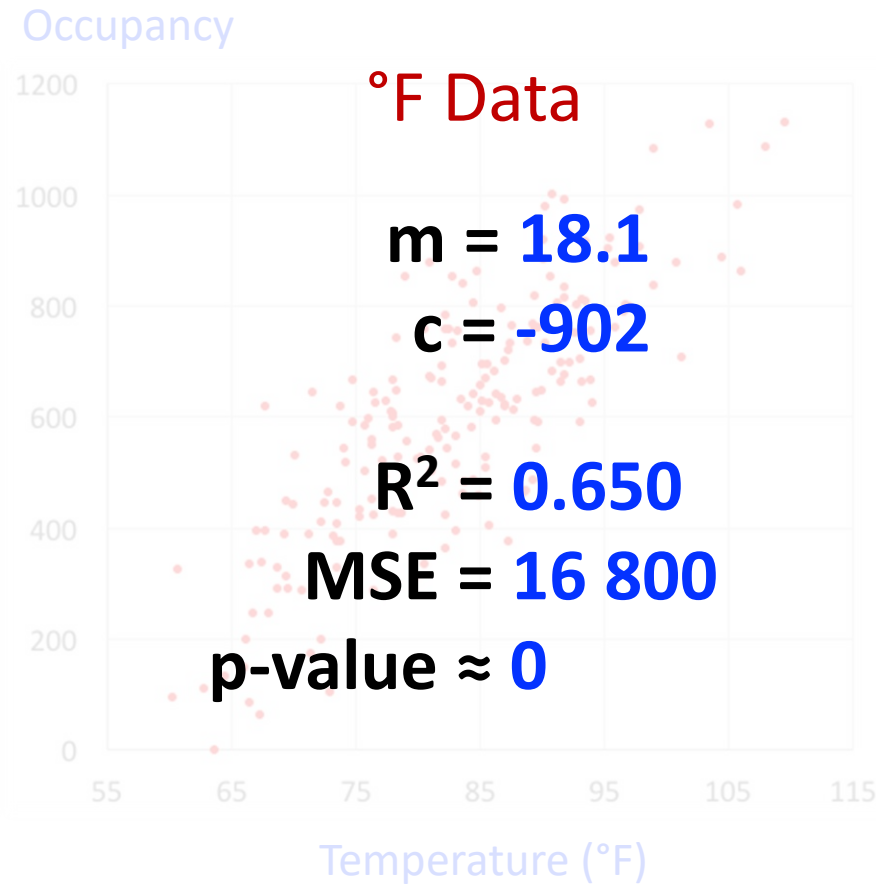
m = 32.6

c = -323

R^2 = 0.650

MSE = 16 800

p-value ≈ 0



°F Data

m = 18.1

c = -902

R^2 = 0.650

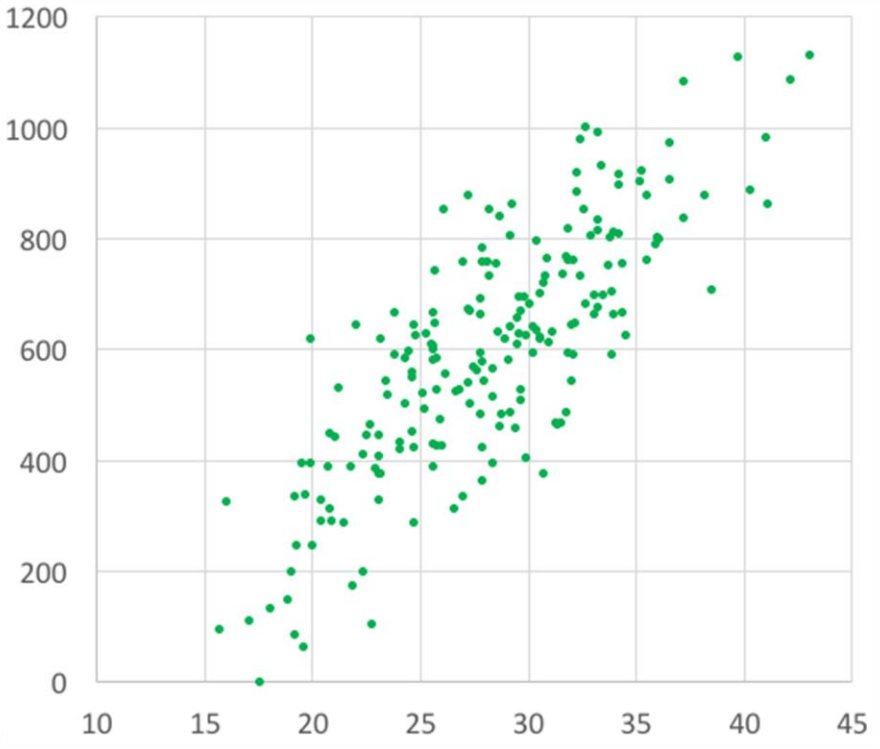
MSE = 16 800

p-value ≈ 0

Key Question:

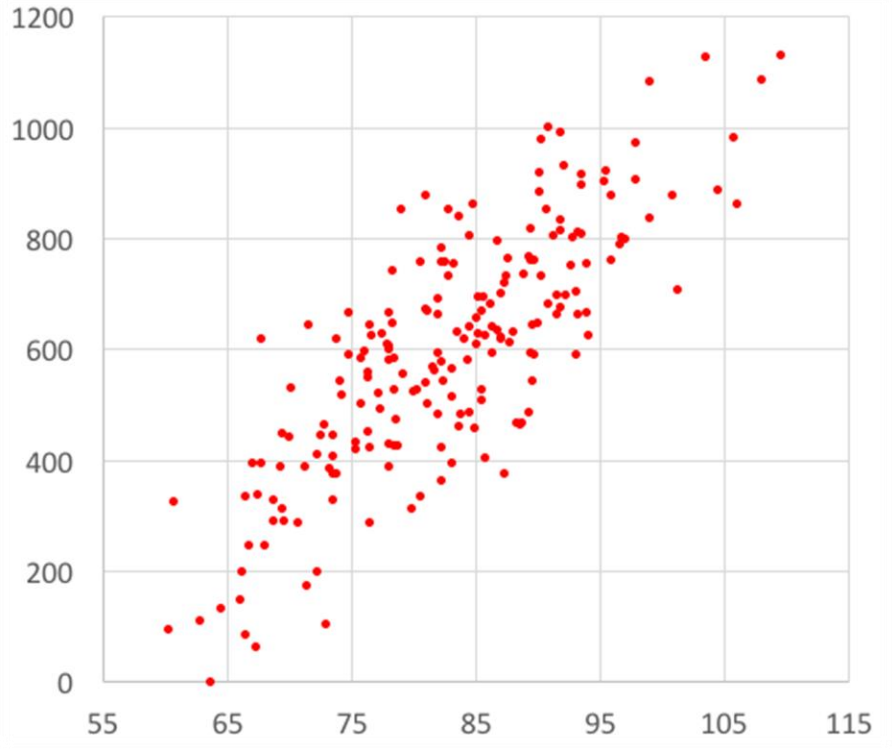
What if we use BOTH sets of temperature data to predict occupancy?

Occupancy



Temperature (°C)

Occupancy



Temperature (°F)

Key Question:

What if we use BOTH sets of temperature data to predict occupancy?

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Occ. Temp (°C) Temp (°F)

Key Question:

What if we use BOTH sets of temperature data to predict occupancy?

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Occ. Temp (°C) Temp (°F)

```
predictor coefficients = [ 50.55816851  -9.99972651]
constant               = -3.15638843237
Rsquared               = 0.649562424832
MSE                    = 16777.4197842
T-test pvalues         = [  1.24311659e-23   4.00502754e-10]
F-test pvalue          = 5.78150445886e-47
```

Key Question:

What if we use BOTH sets of temperature data to predict occupancy?

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Occ. Temp (°C) Temp (°F)

```
predictor coefficients = [ 50.55816851  -9.99972651]
constant               =  3.15638843237
Rsquared               = 0.649562424832
RSE                    = 10777.4197612
T-test pvalues         = [  1.24311659e-23   4.00502754e-10]
F-test pvalue          = 5.78150445886e-47
```

Key Question:

What if we use BOTH sets of temperature data to predict occupancy?

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Occ. Temp (°C) Temp (°F)

Correlation between x_1 and x_2 = **???**

Key Question:

What if we use BOTH sets of temperature data to predict occupancy?

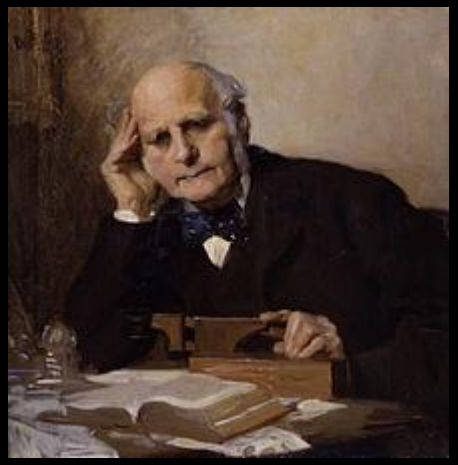
$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Occ. Temp (°C) Temp (°F)

Conclusions

1. With multicollinearity, the model is instable and unreliable, as there are many equivalent models.
2. **Be aware of multicollinearity. We will cover it next week.**

Some stories



Sir Francis Galton

- Develop the concept and coefficient of correlation
- Invent the use of regression line
- ‘Regression towards the mean’
  - If Liverpool won the Premier League last year, what does that mean for their chances for winning this season?
  - *"Sophomore slump“*

Some stories

Karl Pearson

- Founder of the statistics department at UCL
- Pearson correlation coefficient
- Chi-squared test
- Standard deviation

Eugenicist

UCL denames buildings named after eugenicists

19 June 2020

UCL has today announced it will dename spaces and buildings named after two prominent eugenicists Francis Galton and Karl Pearson.

Follow

"…… the names have been changed to Lecture Theatre 115 (formerly the Galton Lecture Theatre), Lecture Theatre G22 (formerly the Pearson Lecture Theatre) and the North-West Wing (formerly the Pearson Building)."

Galton

EUGENICS

"IS THE STUDY OF THE AGENCIES UNDER SOCIAL CONTROL, THAT IMPROVE OR IMPAIR THE RACIAL QUALITIES OF FUTURE GENERATIONS EITHER PHYSICALLY OR MENTALLY."

SIR FRANCIS GALTON.

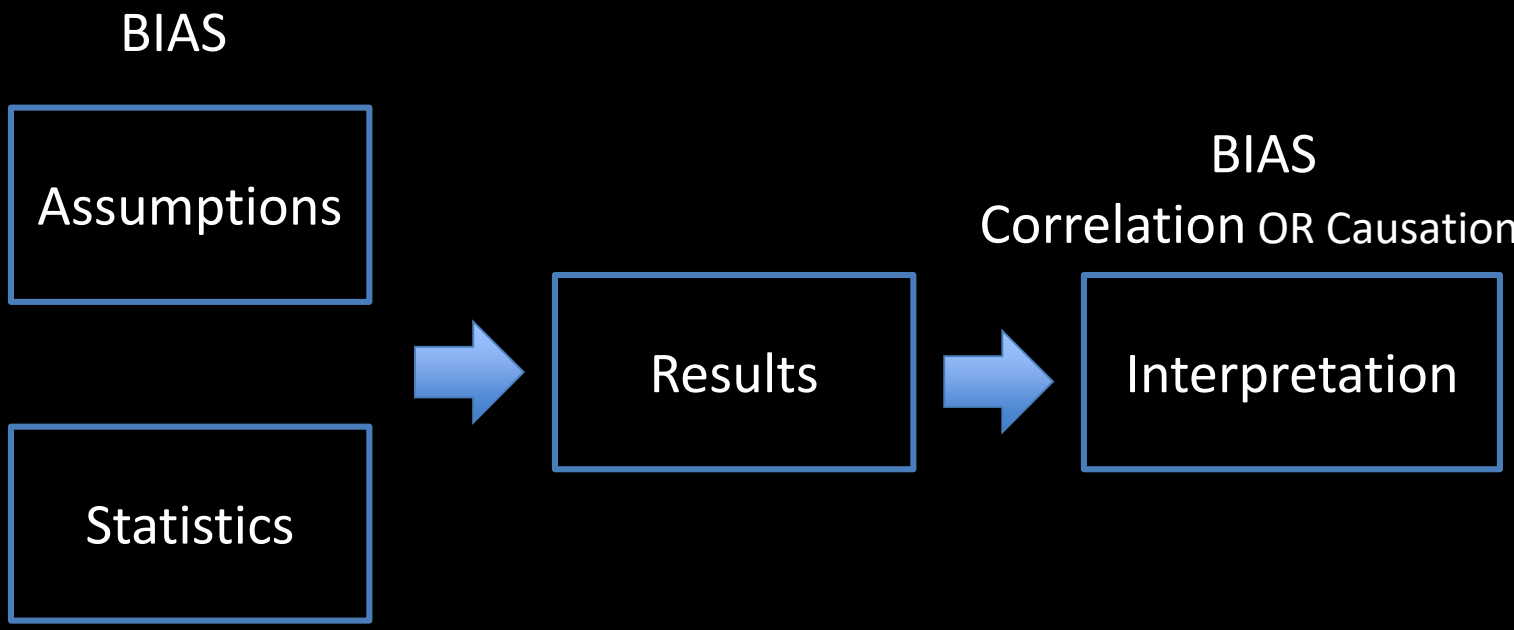
Pearson

- ‘the standard of the Jewish aliens in the matter of personal cleanliness is substantially below that of even the poor Gentile children’
- ‘the cold light of statistical inquiry’
- ‘we firmly believe that we have no political, no religious and no social prejudices … we rejoice in numbers and figures for their own sake … to find out the truth that is in them’

- **So what is going wrong?**

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1925.tb02037.x

Correlation doesn’t imply causation



- The issue is not with the statistics. It is in the assumptions and the interpretation where people can embed bias.
- Quantitative methods and statistics are powerful tools, but they don't necessarily generate unbiased and correct answers. The key issue is how and why you use them.

Lecture 3 – Assignment – Part A

Download datasets from Moodle.

Perform a regression on at least two of the datasets and consider what your results tell you about the relationship between the data series.

You can use Python code (recommended) or the Excel file.

There is no need to write up your results neatly, but if you have time, you can try this.

Lecture 3 – Assignment – Part B

Find a research paper that uses regression.

Read it and look at how it uses the technique, how it interprets and communicates the results, and how it incorporates regression into a wider quantitative argument.

Come to the next lecture with the paper and some notes on your thoughts.
Be prepared to discuss it.

(Recommend Web of Science to search for research papers)