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.*

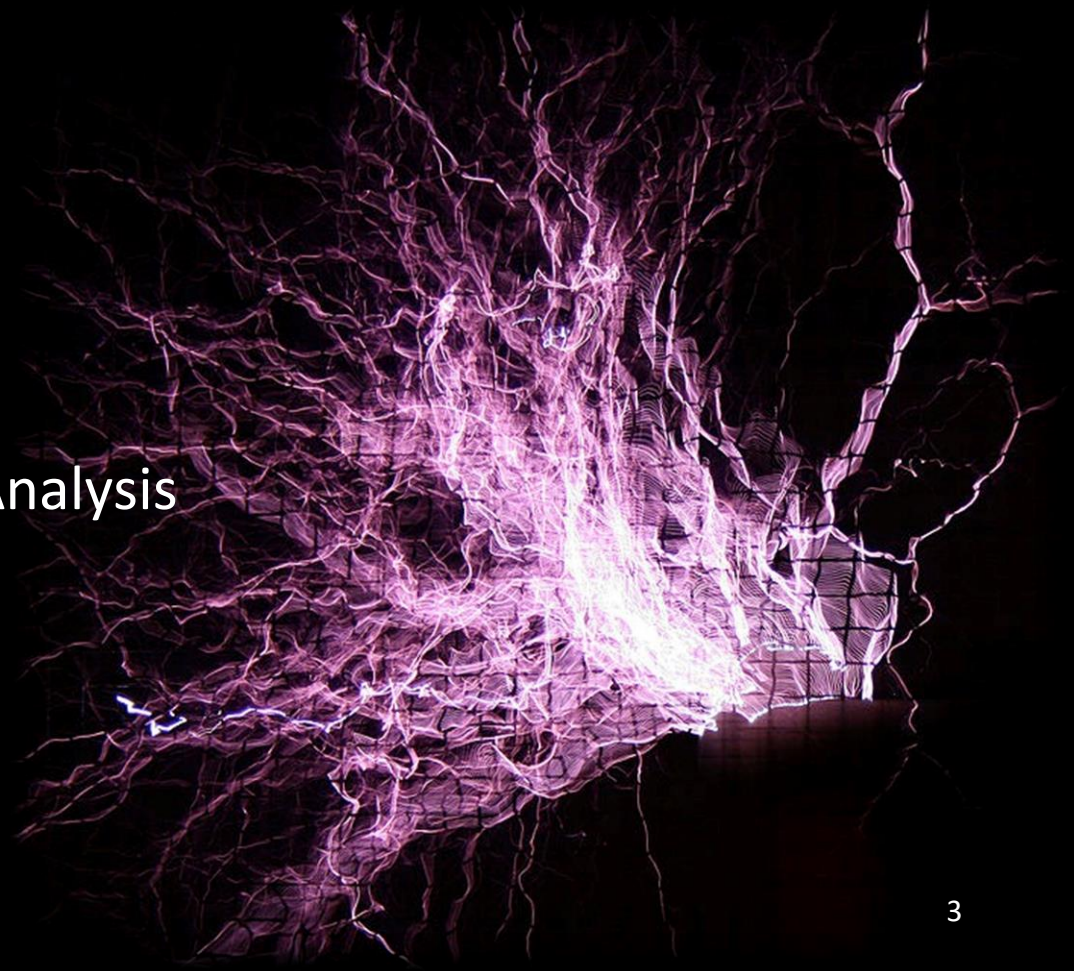
CASA0007: Quantitative Methods

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

Lecture 2 Assignment

- Open questions & no standard answers.
- Four points (and maybe more)
  - Exponential growth (Albert Bartlett)
  - Population forecast (why it fails and what implications?)
  - Oil never running out (reasons and implications)
  - Quantitative predictions (is it really useful and why)
- Breakout room discussion: 5 minutes
- Share your thoughts on the Padlet page

Lecture 2 Assignment

Look at the relationships data sets on moodle.
What do they reveal about the underlying relationships?

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)

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

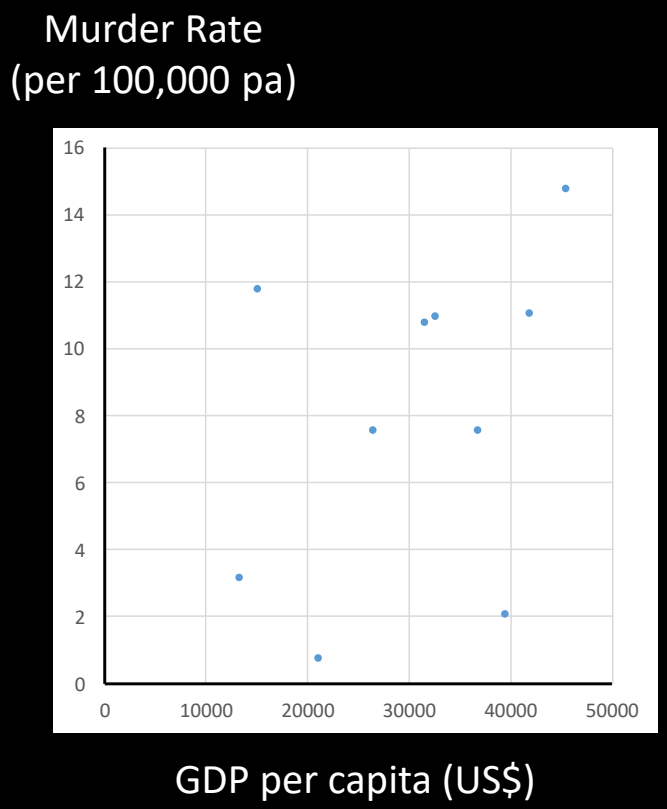
Week 7: Optimising Limited Resources

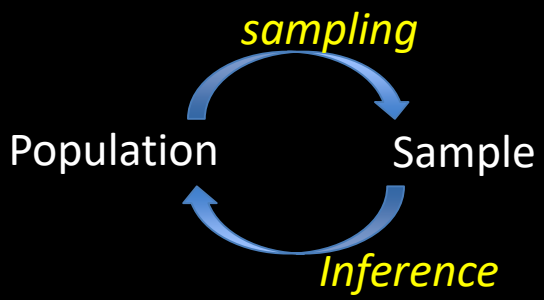
Week 8: Modelling the World

Week 9: Statistical Traps & Advanced Topics

PRESENTATION WEEK

Questions: random variables





- Two random variables here:
  - murder rate
  - GDP per capita
- These points are a sample from a larger population. We don't have access to the population – we can only observe the sample.
- We can estimate some attributes of the population by computing on the sample (this process is called inference)
  - Covariance
  - Average of murder rate
  - The strength of linear relationship between murder rate and GDP per capita (R^2)

It’s (mini) Tea break Time

LECTURE 4

Advanced Regression Techniques

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

OBJECTIVES

1. Learn how to deal with multicollinearity and select variables for multiple regression;
2. Understanding and applying residual analysis;
3. Learn how to interpret a linear regression model;
4. Understanding adjusted R Squared and when to use it;
5. Understanding R Squared for generic regression tasks;
6. Learn how to build and interpret a logistic regression model.

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

Names for y and X

y

Response variable
Outcome variable
Dependent variable
Left-hand variable

relationship

- Linear
- Non-linear

X

Predictor variable
Explanatory variable
Independent variable
Right-hand variable
Feature

Applying linear regression

1. Identify a regression problem
2. Explore the data: on a univariate and bivariate basis, check for outliers, data errors, missing values
3. Variable selection
4. Fit a linear model
5. Diagnose the model using residual analysis
6. Refine the model (by excluding outliers, adding/excluding variables)
7. Interpret the model

Variable selection

- Multicollinearity
  - It exists when two or more predictors in a regression model are highly correlated
- Examples
  - [beach occupation] Farenheit = Celsius * 1.8 + 32.0
  - [office env] (assuming three colours R/G/B) R + G + B = 1
- Consequences
  - Variance of coefficients is large and the certainty is low
  - The model is unstable and unreliable
- Need to detect and eliminate multicollinearity before applying linear regression

Why multicollinearity is bad?

- In the office_env, Fahreheit (F) and Celsius (C) are multicollinear. Assuming that if we fit the environment_score using F and C, and get this result

  $Y = 100 * F + 0.5 * C$ …. (1)
- We know that this holds true for all samples: $F=1.8*C + 32$
- Then, we can formulate many models equivalent to (1): $Y =$ $Y = 100 * F + 0.5 * C + a * (1.8*C + 32-F)$, where a is any number
- If we accidentally get this model: $Y = 100 * F + 0.5 * C + 100000 * (1.8*C + 32-F)$. This is equivalent to (1), but it is very sensitive to a small change in C (e.g. measurement error). So the model is unreliable and unstable. So multicollinearity is bad.

Variance Inflation Factors (VIF)

- Given y and $x_1, x_2, \ldots, x_p$, the VIF for the x_k variable is

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 value obtain by regressing the x_k on the remaining x variables:

$$x_k = \sum_{i=1}^{k-1} b_i x_i + \sum_{i=k+1}^{p} b_i x_i$$

- The larger VIF_k, the higher multicollinearity, as x_k can be largely represented by a linear combination of the other variables and thus x_k is redundant.

Using VIF for variable selection

1. Initialise L as the list of predictor variables. (*HINT*: the response variable is not needed for VIF)
2. Calculate the VIF for each variable in L. (HINT: the order of computing VIF is irrelevant).
3. If the highest VIF is larger than the threshold, remove the corresponding variable from the list L. A threshold of 5 is often used.
4. Repeat Step 2-3, until no VIF is larger than the threshold.
5. Output L.

Optional: if you want to know more about VIF, read this page: https://online.stat.psu.edu/stat501/lesson/12/12.4

Alternatives to VIF

- In statistics, there are always multiple ways to do one thing.
- Alternatives to VIF for dealing with multicollinearity include:
  - Stepwise regression
  - LASSO (would be covered in CASA0006)

Discussion: Python code

Summary: VIF

- Question: multicollinearity between predictors
- A high VIF indicates high multicollinearity. A threshold of 5 is often used.
- Output
  - A subset of predictors, which can be used as input of linear regression.

It’s Tea break Time

Interpreting linear regression

- Linear regression is not causality analysis, it is based on correlation. Don't use declarations like '*X* is the reason for the change of *Y*'.

- The magnitude of coefficients does not imply importance. You can't say a variable with weight=100 is more important than another variable with weight=1.

- It's not meaningful to interpret a model with very low R-squared (like 0.2). It does not explain much of the variance. Try to improve R-squared before you interpret.

Examples – bike rental

Example: predicting the daily number of rented bikes, given weather and calendar information

	Weight
(Intercept)	2399.4
seasonSUMMER	899.3
seasonFALL	138.2
seasonWINTER	425.6
holidayHOLIDAY	-686.1
workingdayWORKING DAY	124.9
weathersitMISTY	-379.4
weathersitRAIN/SNOW/STORM	-1901.5
temp	110.7
hum	-17.4
windspeed	-42.5
days_since_2011	4.9

- Interpretation of temp (numerical)

An increase of the temperature by 1 degree Celsius increases the predicted number of bicycles by 110.7, when all other features remain fixed

- Interpretation of 'workingday' (here 'Weekend' as reference category)

When it is working day, the predicted number of bicycles is 124.9 higher compared to weekend, given all other features remain fixed

https://christophm.github.io/interpretable-ml-book/limo.html#interpretation
https://christophm.github.io/interpretable-ml-book/bike-data.html#bike-data

Example – Airbnb rental

$$\mathrm{LnRent}_{it+1c} = \alpha + \beta_1 \mathrm{Airbnb}_{tc} + \beta_2 \mathrm{Bed}_{it+1c} + \beta_3 \mathrm{Bath}_{i+1tc} + \beta_4 \mathrm{Sqft}_{it+1c} + \beta_5 \mathrm{NC}_{tc} + \delta \mathrm{Month}_t + \eta \mathrm{Tract}_c + \varepsilon_{itc}$$

Table 6
Regression of log of asking rents on Airbnb density.

	(1) All units
Airbnb density	0.627** (0.306)
Bedrooms	0.171*** (0.009)
Bathrooms	0.112*** (0.010)
Sqft. (per 1000)	0.132*** (0.019)
New construction (per 100,000 units)	−0.742 (2.318)
Crime (per resident)	
Building permits (per housing unit)	
Restaurant licenses (per resident)	
Constant	7.373*** (0.015)
N	113,409
Month fixed effects	X
Census tract fixed effects	X

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

- 'To describe the magnitude of the 0.627 coefficient on Airbnb density, a **one standard deviation increase** in Airbnb density in a given census tract raises asking rents by 0.4%.'
- Note that y is the natural log of Rent.
- How does 0.4% come from? (See next page)

http://dx.doi.org/10.1016/j.jhe.2017.08.002

Example – Airbnb rental

$$\mathrm{LnRent}_{it+1c} = \alpha + \beta_1 \mathrm{Airbnb}_{tc} + \beta_2 \mathrm{Bed}_{it+1c} + \beta_3 \mathrm{Bath}_{i+1tc} + \beta_4 \mathrm{Sqft}_{it+1c} + \beta_5 \mathrm{NC}_{tc} + \delta \mathrm{Month}_t + \eta \mathrm{Tract}_c + \varepsilon_{itc}$$

Table 2
Descriptive statistics on Airbnb and rental units by census tract.

	Mean	Standard deviation	Count
Total housing units	1,638	618	832
# of Airbnb listings	11.7	13.5	832
Newly constructed units	1.4	16.4	832
# of rental units listed for rent (weekly)	75.8	100.5	832
Airbnb density	0.007	0.007	832

Table 6
Regression of log of asking rents on Airbnb density.

	(1) All units
Airbnb density	0.627** (0.306)
Bedrooms	0.171*** (0.009)
Bathrooms	0.112*** (0.010)
Sqft. (per 1000)	0.132*** (0.019)
New construction (per 100,000 units)	−0.742 (2.318)
Crime (per resident)	
Building permits (per housing unit)	
Restaurant licenses (per resident)	
Constant	7.373*** (0.015)
N	113,409
Month fixed effects	X
Census tract fixed effects	X

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

- [Table 2] st.dev. of Airbnb density is 0.007
- Let x denote Airbnb density. If x increases by 0.007, then ln(Rent) increases by 0.007*0.627=0.004
  - ln(new_Rent) = ln(Rent) + 0.004
  - new_Rent = Rent * (1+0.004)
  - [exponential calculation]
- So, the relative change of Rent is 0.4%.
- Why using one st dev increase of x?
  - The st dev is a ‘typical’ or ‘reasonable’ deviation from the current value.
  - You can’t meaningfully increase the density by 1.

http://dx.doi.org/10.1016/j.jhe.2017.08.002

Residual Analysis

$$y_i = \beta_1 x_i + \beta_0 + e_i$$

e_i is the residual term. It is very important for checking the linear model.

Necessary Conditions

Linear relationship exists

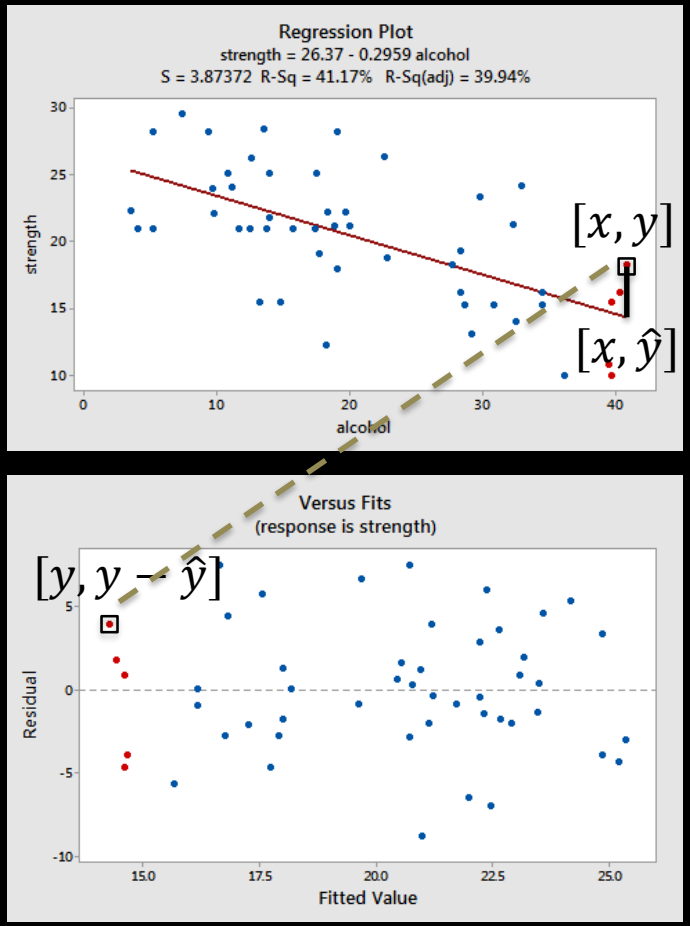
Independent errors

Normally distributed errors

Equal variance for all x values

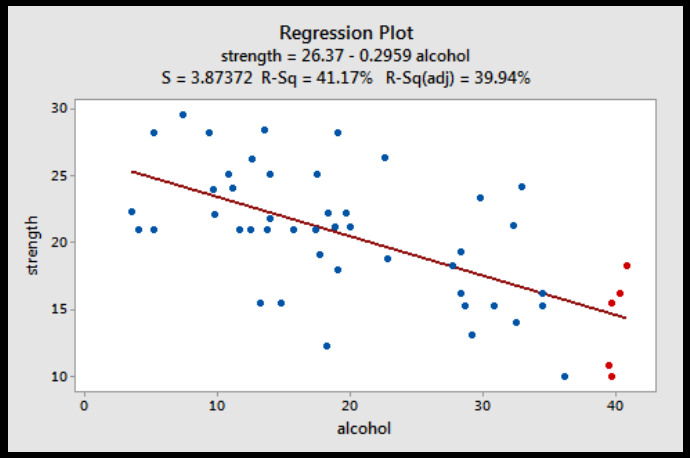
Residual analysis

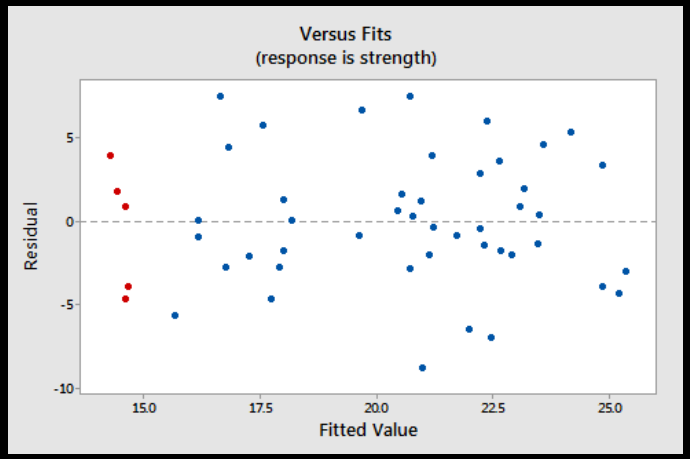
- 'Residual versus fits plot'
- This plot is used to detect
  - Non-linearity
  - Unequal error variances
  - Outliers
- It is important to present the residual plot in the report and explain the plot.
- Examples



Examples of residual analysis

Is alcohol consumption linearly related to muscle strength?





Observation	Condition?
Residuals bounce randomly around 0 line.	Linear relationship: YES
Roughly form a horizontal band around residual = 0	Equal variance of residuals for all x values: YES
No one residual 'stands out' from the overall pattern of residuals	No outliers: YES

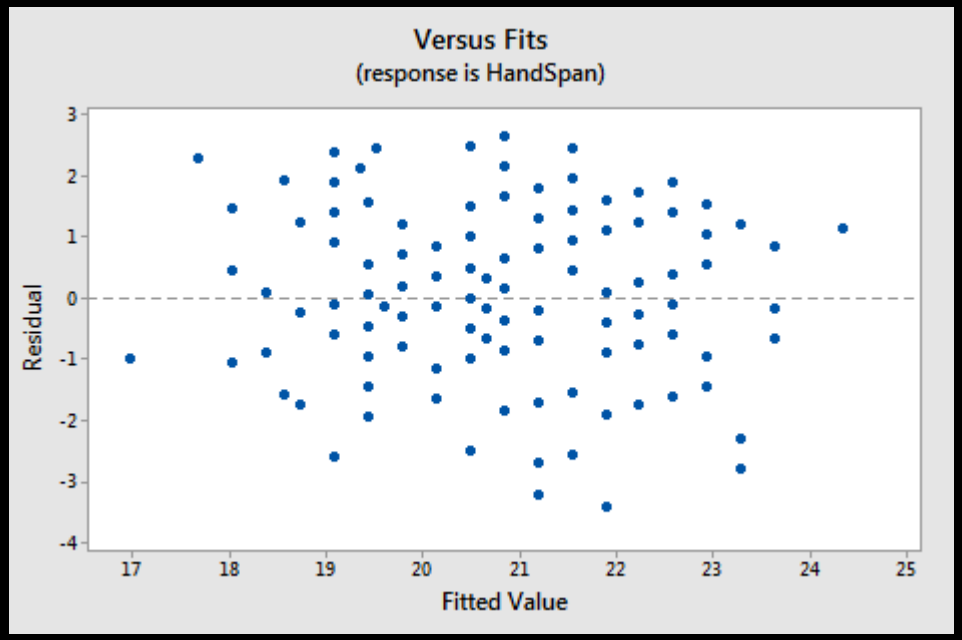
"Residuals versus fits plot"

https://online.stat.psu.edu/stat501/lesson/4/4.2

Another residual plot

Y: handspan

X: height



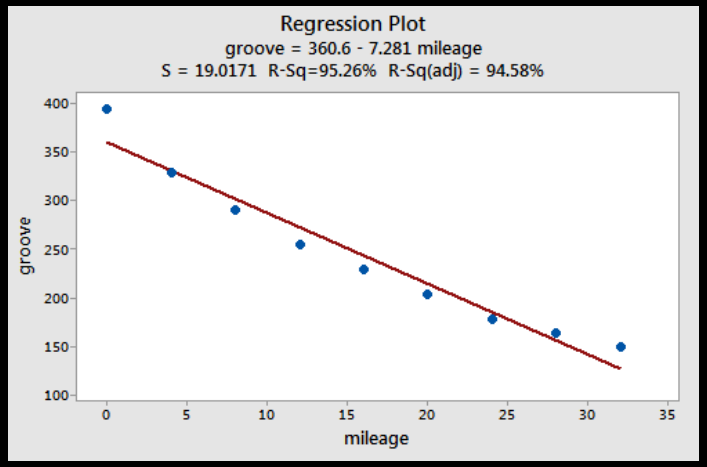
- The variance is roughly the same across.
- There is no systematic patterns of residuals.

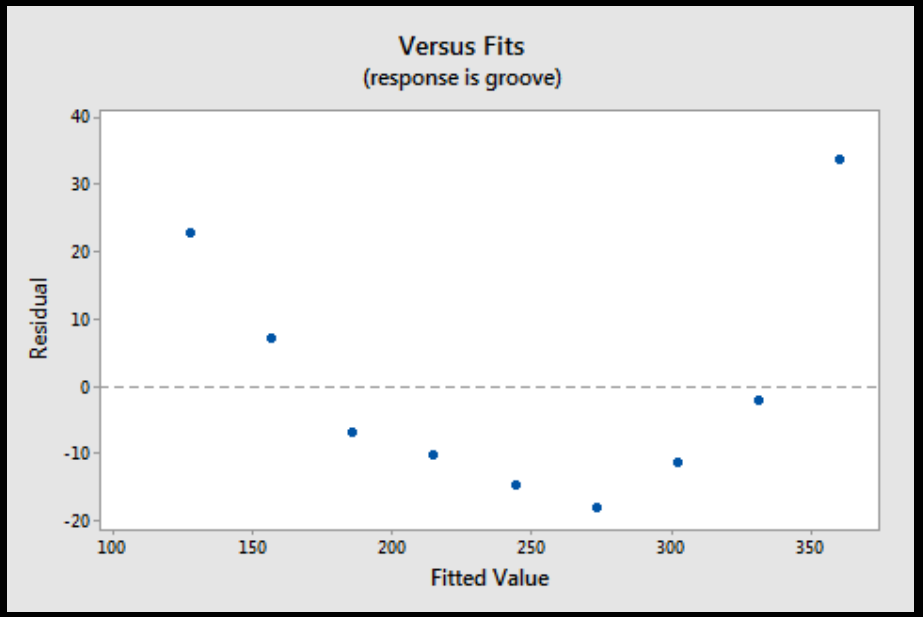
https://newonlinecourses.science.psu.edu/onlinecourses/sites/stat501/files/examples/handheight.txt

Identifying problems using residual plots

A non-linear relationship

Q: Is tire tread wear linearly related to mileage?



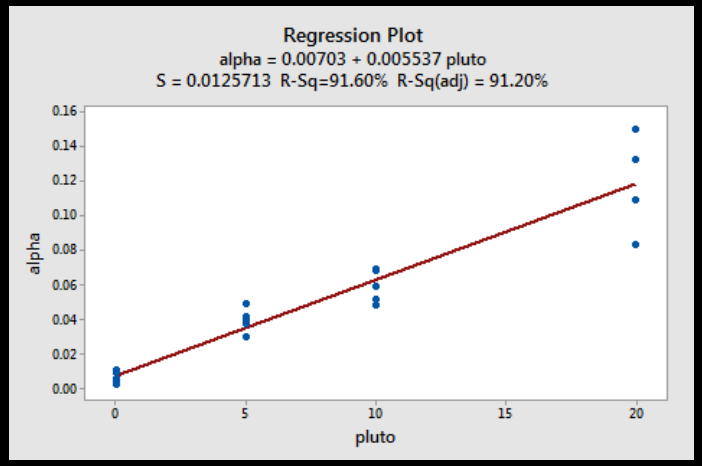


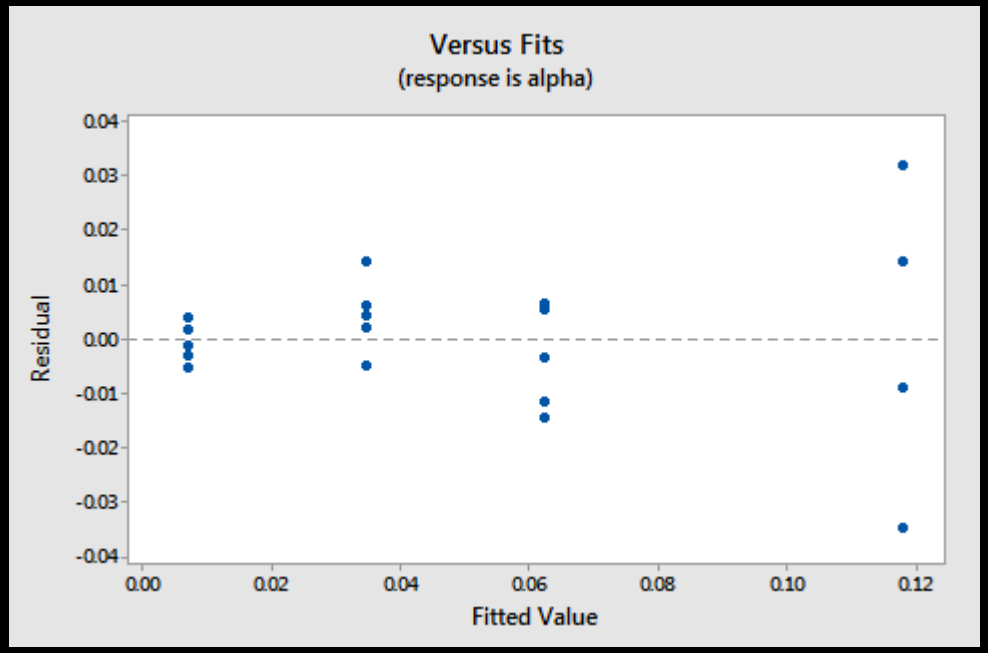
- The residuals depart from 0 in a *systematic manner*.
- Clearly, a non-linear model would better describe the relationship between the two variables.

Identifying problems using residual plots

Non-constant error variance

Q: How is plutonium activity related to the alpha particle counts?



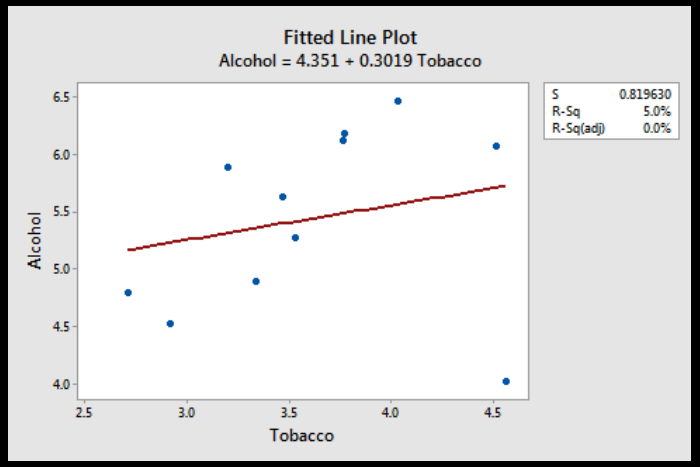


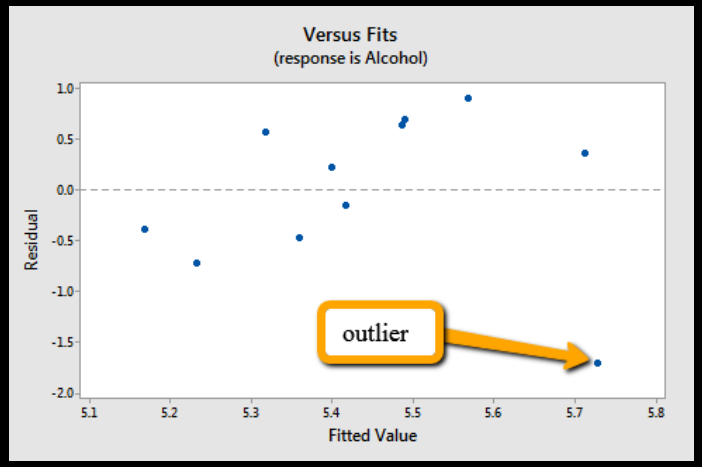
- As the plutonium level increases, not only does the mean alpha count rate increase, but also the variance increases.
- The fitted line plot suggests that the assumption of equal variances is violated.

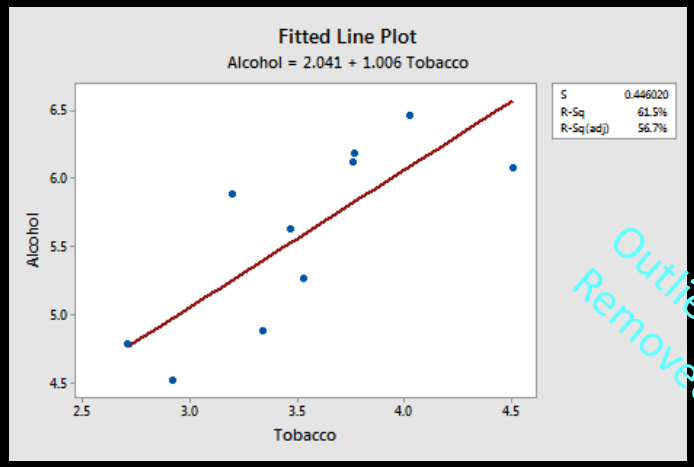
Identifying problems using residual plots

Outliers

Q: Is there a relationship between tobacco use and alcohol use?







Outlier Removed

- r^2 has jumped from 5% to 61.5%.
- One data point can greatly affect the value of r^2
- How large a residual has to be before a data point should be flagged as being an outlier? There are some solutions, but no one "rule of thumb".
- You must explain reasons of removing an outlier

Adjusted R^2

- R^2 is an estimate of the true relationship (called ρ^2) between y and x variables.
- But R^2 would overestimate ρ^2, especially when the sample size is close to variable size.
- If there is no relationship between x and y ($\rho^2=0$), the expected value of R^2 is p/(n-1). (p: number of independent variables; n: sample size.)
- Two points to note
  - The importance of having a large sample size, relative to #variables
  - Additional input variables will make the R-squared stay the same or increase, even if the variable show no relationship with the response variable.

Adjusted R^2

- $R^2 = 1 - \frac{SS_{res}}{SS_{total}}$
- Adjusted R^2: $R^2_{adj} = 1 - \frac{SS_{res}/(n-p-1)}{SS_{total}/(n-1)}$
  - SS_{res} and SS_{total} denote the residual and total sums of squares
- Adj R^2 is useful for comparing a model with and without a specific variable to know whether this variable improves the model.
- When reporting a linear model, it is good to report both R^2 and Adj R^2.

R^2 for generic regression tasks

- $\boldsymbol{R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$
- The range of R^2
  - [0,1] for linear regression model with an intercept term
  - [-Inf,1] for a generic regression task, e.g. decision tree. (As the prediction could be arbitrarily bad)
- *Note: R^2 is not necessarily the square of a number!*

- To understand this, play with this function
  - sklearn.metrics.r2_score

It’s Tea break Time

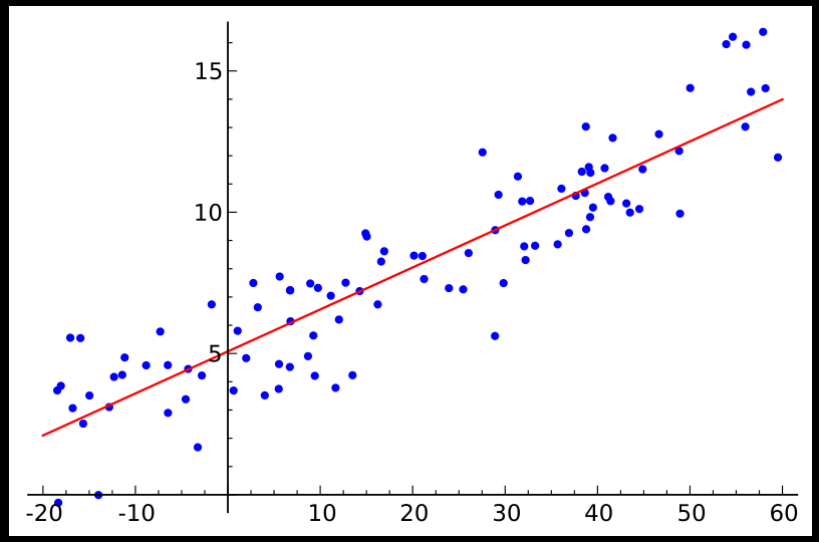
Logistic regression

Regression: use predictors to model/predict a response variable

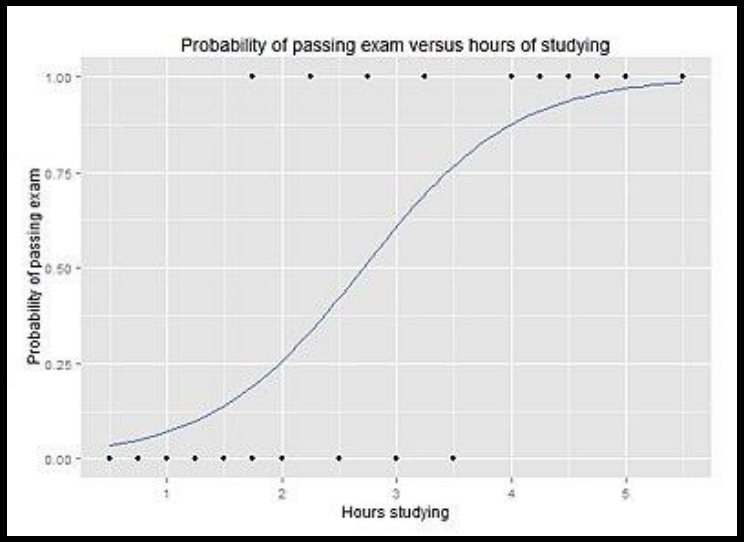
Predictor (X)	Transformation
Continuous	Not required
Categorical	Dummy variables

Response (y)	Regression Model
Continuous, unbounded	Linear regression
Continuous [0,1]	Logistic regression
Binary (0 or 1)	Logistic regression

Linear regression



Logistic linear regression



https://en.wikipedia.org/wiki/Logistic_regression#/media/File:Exam_pass_logistic_curve.jpeg

Logistic regression

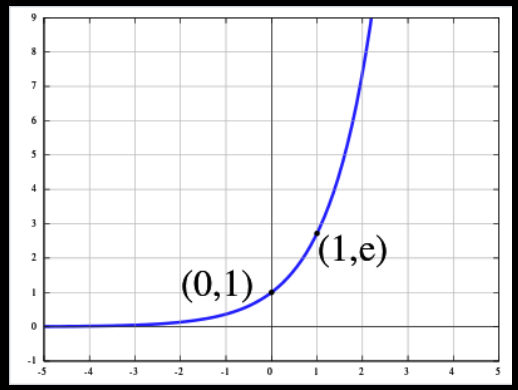
- Linear regression

$$y = \sum_{i=0}^{n} \beta_i x_i + \beta_0$$

- Logistic regression

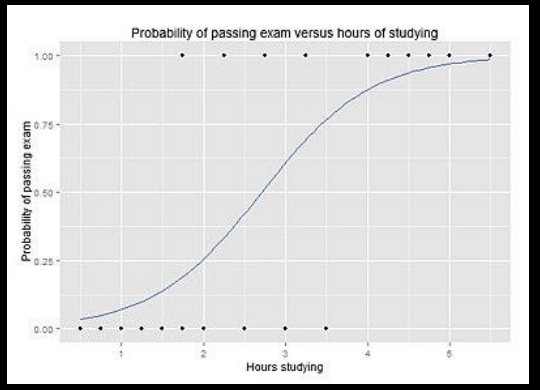
$$y = \frac{exp(\sum_{i=0}^{n} \beta_i x_i + \beta_0)}{1 + exp(\sum_{i=0}^{n} \beta_i x_i + \beta_0)}$$

$$y = \mathbf{exp}(x)$$



$\sum_{i=0}^{n} \beta_i x_i + \beta_0$	$\mathbf{exp}(\sum_{i=0}^{n} \beta_i x_i + \beta_0)$	y
-∞	0	0
+∞	+∞	1

$$y = \frac{\mathbf{exp}(x)}{\mathbf{1} + \mathbf{exp}(x)}$$



Y is a continuous value between [0,1]

Logistic regression

Determine if a student passes or fails? (binary)

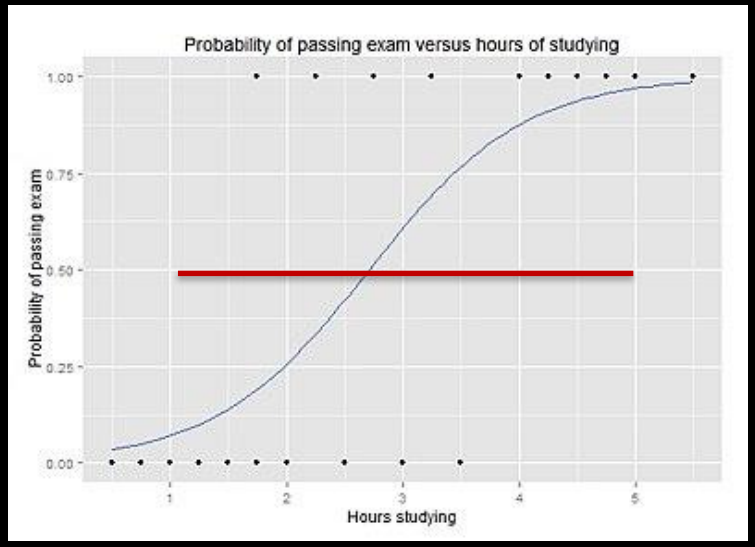
x: study hours, efficiency, …
y: a probability value [0,1]
Adding a cutoff probability value (α)

If y >= α, then predict this student would pass
If y < α, predict fail

- Selection of α: normally 0.5, but not always
- This task is called classification, and is one application of logistic regression
- The mapping is as follows:

x → y → Binary class

(-∞,+∞) [0,1] 0 or 1

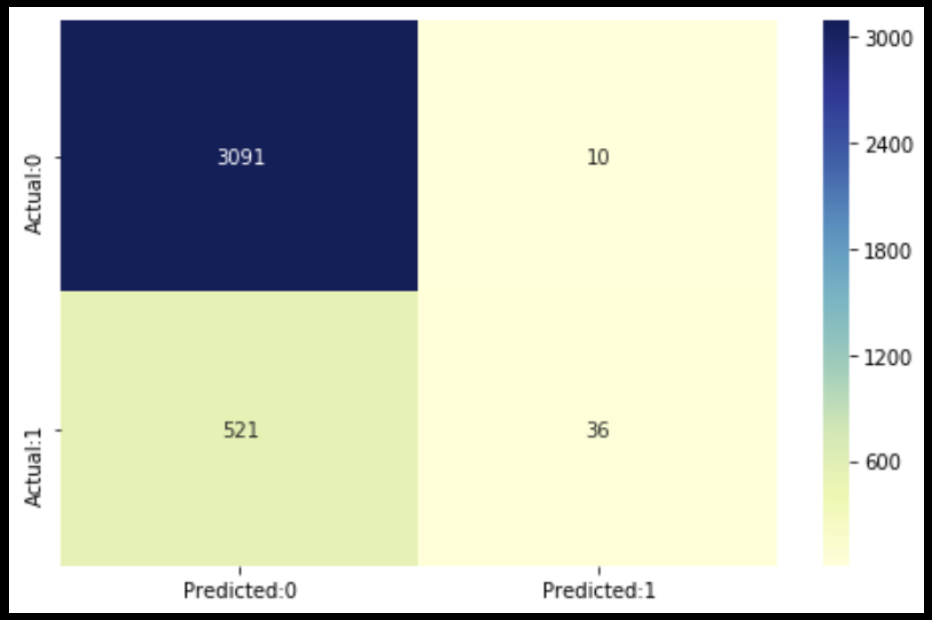


Evaluation

Evaluation of classification

$$\text{Accuracy} = \frac{\#Correct\ prediction}{\#Total\ case}$$

Confusion matrix



Discussion: python code

Logistic regression

- Extensions
  - Nominal logistic regression: when there are three or more categories without natural ordering, e.g. allocating new employees to departments at a business (R&D, marketing, HR)
  - Ordinal logistic regression: three or more categories with a natural ordering to the levels. For example, grade level (distinction, merit, pass, failure)

OBJECTIVES

1. Learn how to deal with multicollinearity and select variables for multiple regression;
2. Understanding and applying residual analysis;
3. Learn how to interpret a linear regression model;
4. Understanding adjusted R Squared and when to use it;
5. Understanding R Squared for generic regression tasks.
6. Learn how to build and interpret a logistic regression model;

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

Lecture 4 Assignment

- Learning one piece of reference management software.
- If you haven't used any, you can start with Mendeley
  - https://www.mendeley.com/guides/, or Google '*Mendeley*'
  - Believe it or not - this will make citing and managing references much easier
  - It is also super useful for building your own knowledge database