The Strategy Unit.

Midlands DSU Network
Decision Support Centre

Evaluating Complex Interventions using Statistical Models

Session 1.1 Introduction to course & Regression Modelling

Anastasiia Zharinova, Healthcare Analyst at the Strategy Unit

Session structure

Session 1 (25th and 27th of May): Review of regression modelling

1.1. Introduction to the course and Regression modelling

- Types of models?
- How to build?
- How to assess?

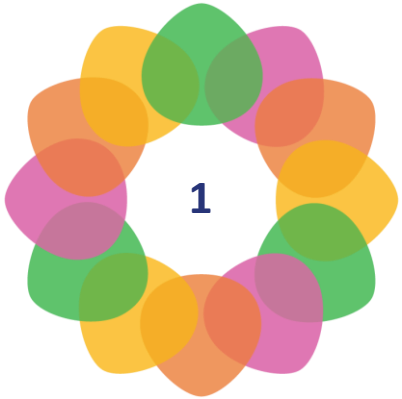
1.2 Multiple regression models – practical exercise

1.3 Time series models – practical exercise

In today's session 1

1. What this course is about?
2. Regression modelling: introduction
3. Regression modelling: multivariable regression
4. Time-Series Analysis

SECTION



What this course is about?

Evaluation exercise

Evaluating (complex) interventions/the effect of an 'event'. Which alternative is better?

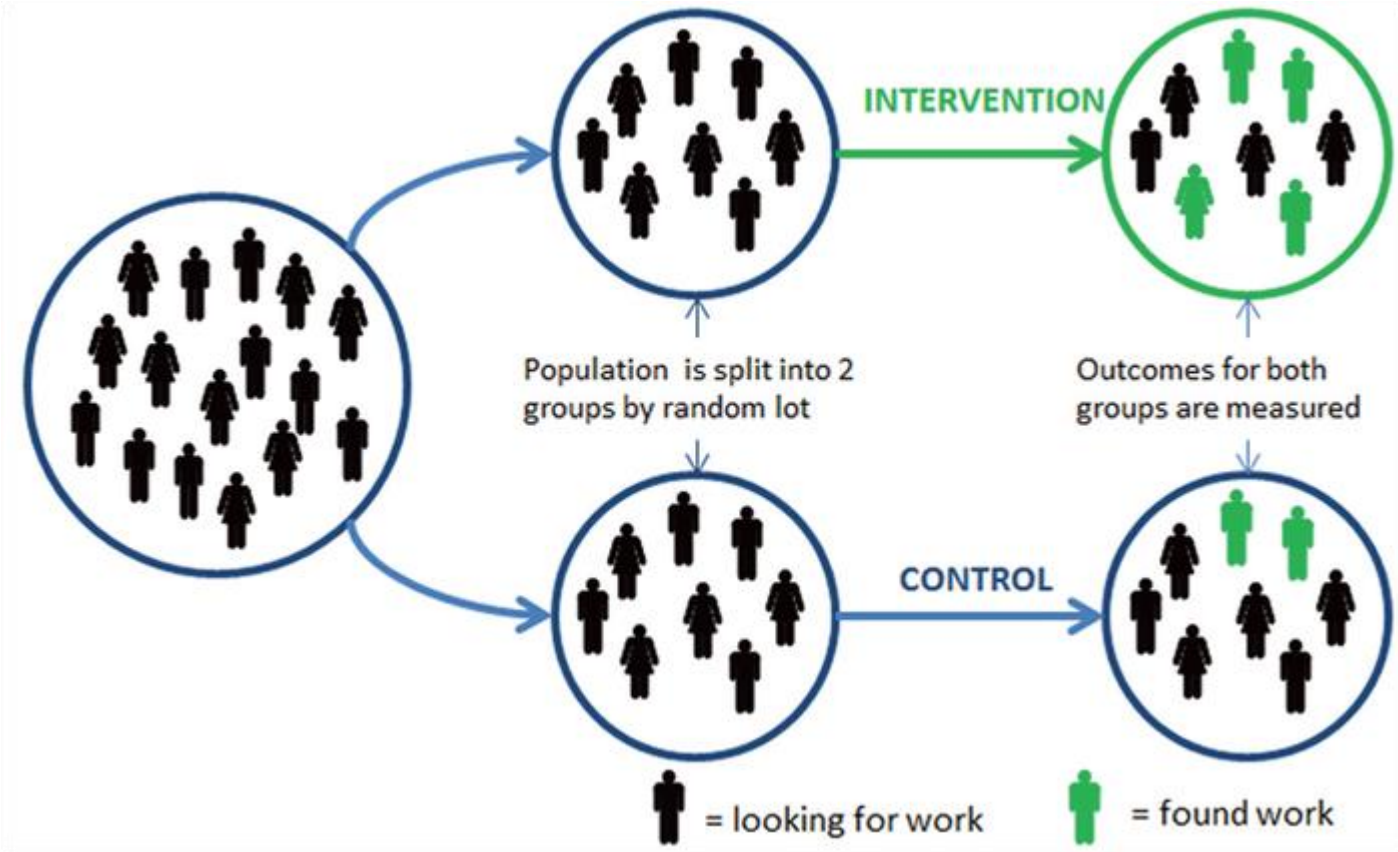
On average,
Drug A – 4.5 QALYs per participant
Drug B – 4.2 QALYs per participant

No Participant	Drug	Gain in Quality-Adjusted life years (QALY)*
1	A	5.0
2	A	4.5
3	A	4.0
4	A	4.0
5	A	5.0
6	B	4.0
7	B	4.0
8	B	4.5
9	B	4.5
10	B	4.0

*standard measure which represents the quantity and quality of life saved. Calculated as Utility (from 0 to 1)*number of years saved

Evaluation exercise

The truth is – we cannot know without knowing the details about the randomised control trial.



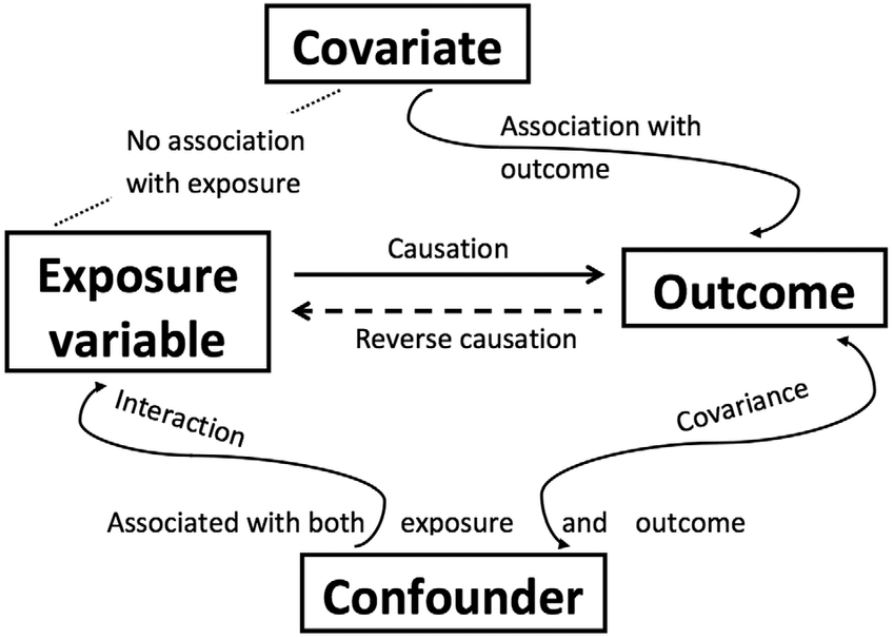
The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme (Haynes et. al, 2012, p.4).

Randomised Control Trials: calculate an effect

$$Outcome = \alpha + \beta * Treatment + error$$

However, different factors may influence outcome

- Covariates - variables that explain a part of the variability in the outcome but have no effect on the intervention
- Confounders - variables that are related to both the intervention and the outcome.



What variables can it be?

What should we do to deal with confounding?

Quasi-Experimental methods

When ‘perfect’ randomisation is not possible, quasi-experimental methods - statistical methods which 'recreate' experiments – are being used.

1. Interrupted Time Series
2. *Panel Data*
3. Propensity Score Matching and Retrospective Cohort Study
4. Synthetic Controls
5. Regression Discontinuity Design
6. *Instrumental variables*

What is this programme about?

Learning theory

1. Review of regression modelling - May
2. Directed Acyclic Graphs (DAGs) - June
3. Introduction to building quasi-experimental models and Difference in Difference analysis - July
4. Interrupted Time Series - September
5. Panel Data - November
6. Propensity Score Matching and Retrospective Cohort Study - January
7. Synthetic Controls - February
8. Regression Discontinuity Design & Instrumental variables - April

Putting knowledge into practice

1. Hands-on workshops in R and optional homework
2. Exercises after each module using health and social care data
3. Group coaching sessions
4. Optional: support to build a model of choice at the end of the course using a real life problem and data from the hosting organisation

Course will be run for 12 months. By the end of the programme, participants will be able to:

1. Discuss the concept of causation and various quasi-experimental approaches
2. Examine effectiveness and suitability of different quasi-experimental methods to different scenarios
3. Design evaluation studies on real-world NHS problems using quasi-experimental methods

Action Learning Set

- Literally, learning from action
- Smaller group coaching sessions after learning material
- Homework/extra exercise discussions
- Setting the problem and finding the ways to approach it

What this course is not about?

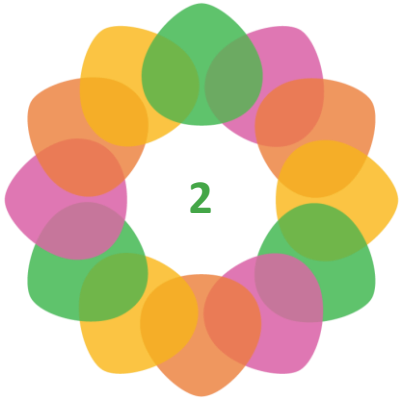
1. All the ‘preparation stages’
- Data Collection
- Data wrangling
- Data quality
- Missing values imputation

2. All possible regression methods – there are too many of them

3. Mixed method evaluation/qualitative research

However, if there are particular questions you are interested in, we can always build them into exercise/homework.

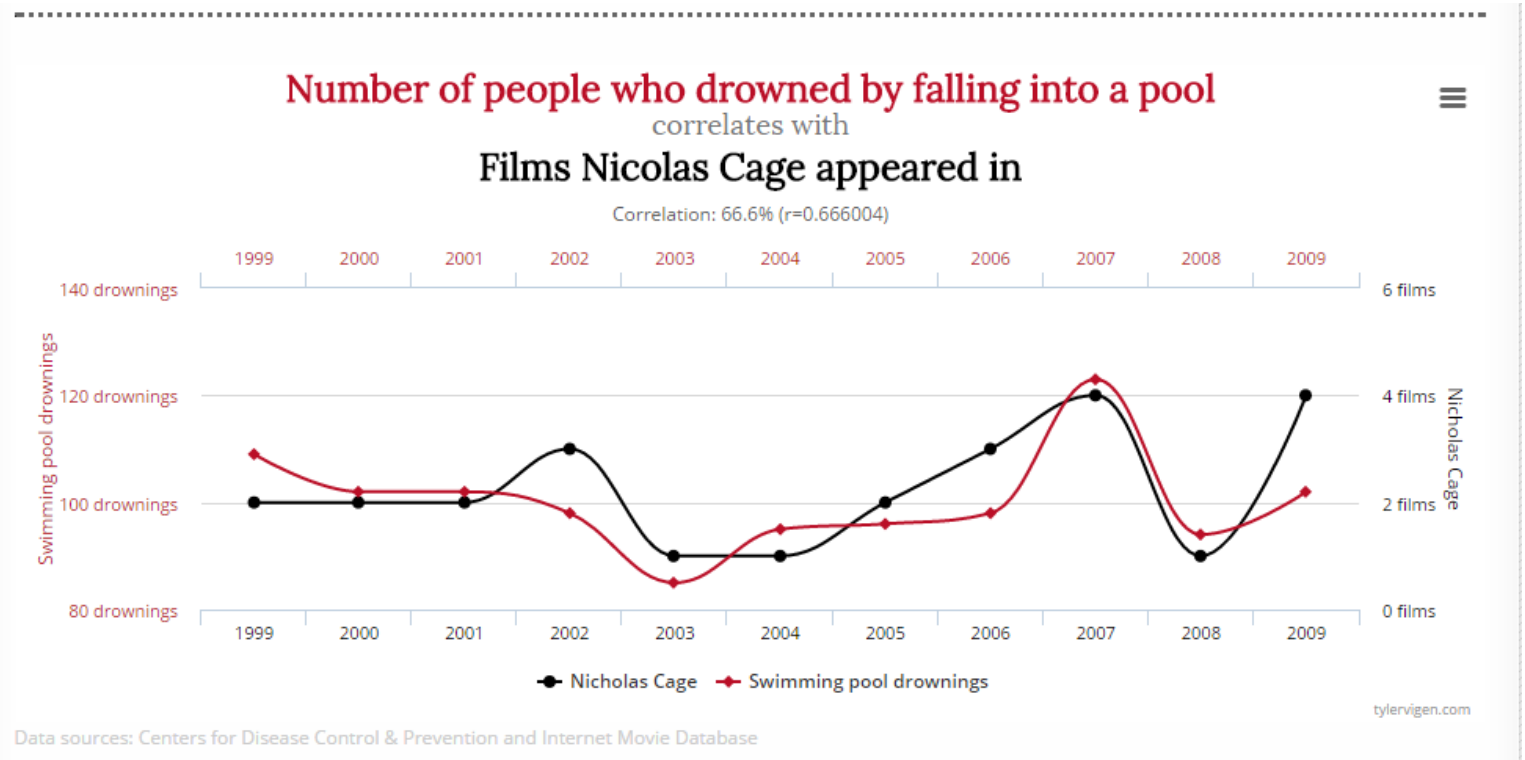
SECTION



Introduction to regression modelling

Regression modelling – why do we need it?

- Identifying which variables have impact on a topic of interest. What drives waiting time? What are the factors of longer length of stay?
- Forecasting future activity. What will A&E Demand look like?
- Explore Causality – not correlation!



Regression modelling

Negative effect of taking a medicine?

	Drug A	No drugs
Average QALYs increase	4.30 years (50 people)	4.85 years (100 people)

If we add more factors…

	Drug A	No drugs
Self-reported health >=3.5 out of 5	5.5 years (10 people)	5 years (95 people)
Self-reported health <3.5 out of 5	4 years (40 people)	2 years (5 people)
Total	4.3 years (50 people)	4.85 years (100 people)

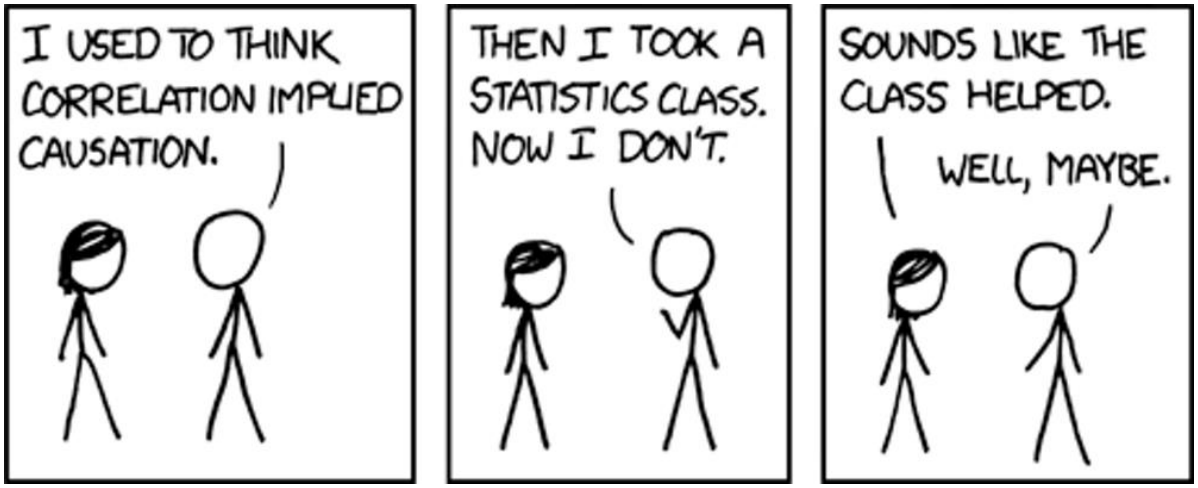
Positive effect for each Group:

1. Healthier group – 5 years VS 5.5 years
2. Less healthy group - 2 years VS 5 years

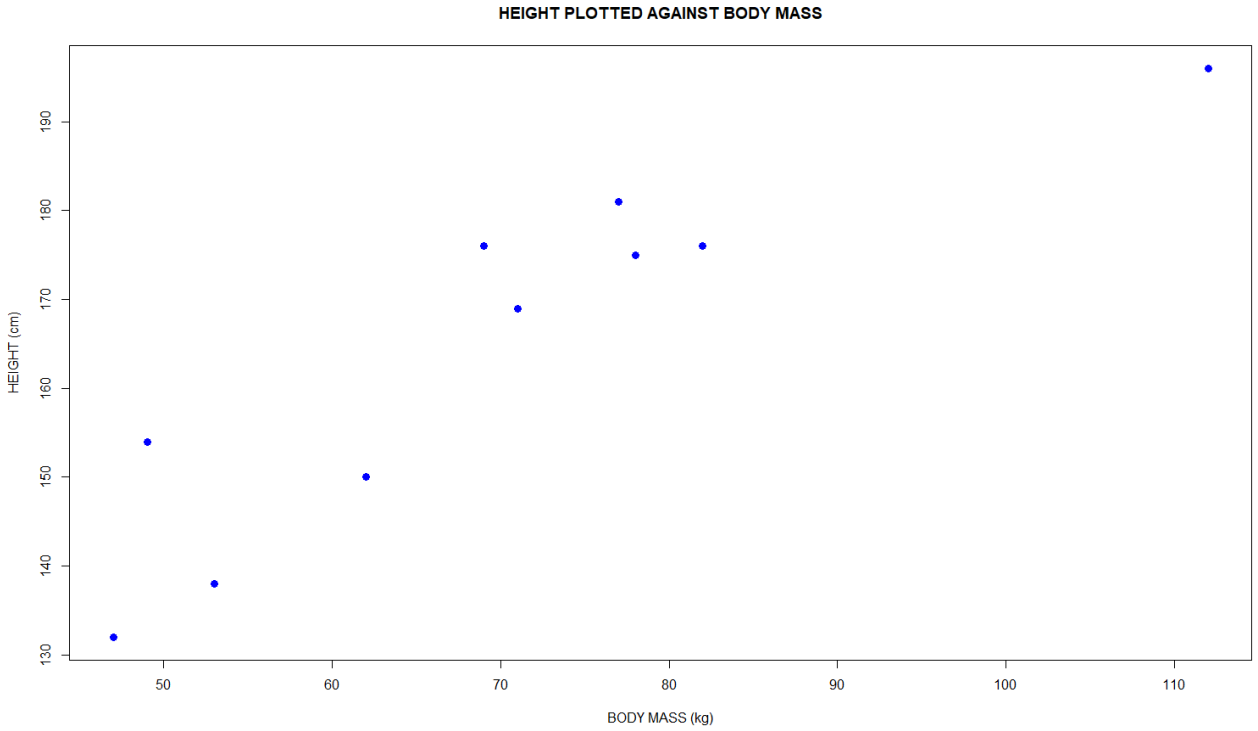
Regression modelling

Lessons so far

1. Correlation is not causation – negative correlation between taking a drug and QALYs does not mean, that the drug is not effective.
2. If we ignore significant variables, we will have omitted-variable bias (self-reported health and future health gain – healthier people are more likely to stay healthy)
3. If we add many variables into regression, it might cause overfitting and make very little sense (e.g. Nicholas Cage variable might well correlate with the NHS waiting times!)
4. There is also a reverse causation which may lead to endogeneity. E.g. Health and wealth



Simple linear regression



height	bodymass
176	82
154	49
138	53
196	112
132	47
176	69
181	77
169	71
150	62
175	78

What is the relationship between Height and Bodymass?

*Linear Models in R: Plotting Regression Lines (the analysis factor)

Ordinary Least Square

$$Height = \alpha + \beta * Bodymass$$

So, we need to find such α and β, that the estimated $Height$ ($\widehat{Height}$) will be as close to actual $Height$ as possible. Or, in other words, for each observation I

$$e_i = Height_i - \widehat{Height}_i = Height_i - \alpha - \beta * Bodymass$$

As e_i can be positive or negative, we need to minimise the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_3^2 = \sum_{i=1}^{n} e_i^2$$

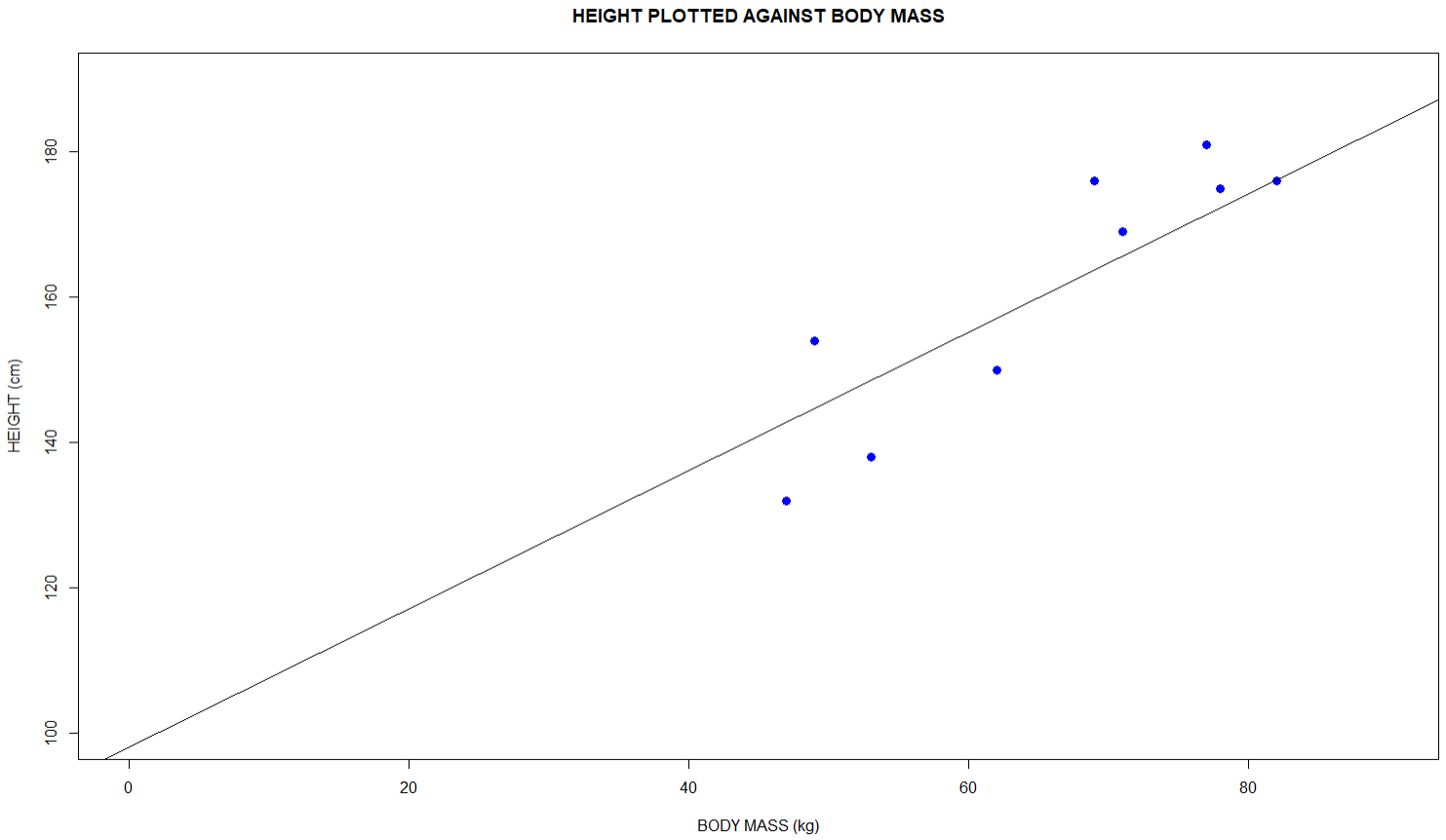
Hence the name!

Luckily, R can do it for us (but I am happy to talk more about the math!)

Ordinary Least Square

In our case,

$$Height = \ 98.00 + 0.95 * Bodymass$$



Ordinary Least Square

However, real-life coefficients α and β will never be the same as the modelled 98.00 and 0.95.

Therefore, statisticians are looking at confidence intervals. Confidence intervals are being calculated using standard error of the coefficient (radical of its variation) and t-value.

$$Height = \underset{(11.70)}{98.00} + \underset{(0.16)}{0.95} * Bodymass$$

Confidence intervals

$$\hat{\beta} - t * se(\hat{\beta}) < \hat{\beta} < \hat{\beta} + t * se(\hat{\beta})$$
$$0.95 - 1.96 * 0.16 < \hat{\beta} < 0.95 + 1.96 * 0.16$$
$$0.95 - 1.96 * 0.16 < \hat{\beta} < 0.95 + 1.96 * 0.16$$
$$0.63 < \hat{\beta} < 1.26$$

Which means that with 95% confidence interval, each additional kg of bodymass increases height by from 0.63 to 1.26 cm.

It is important to know, how different estimated coefficient $\hat{\beta}$ from the real world β

Econometrics software (e.g. R) uses P-value or Probability. It shows the significance level at which Hypothesis β=0 is being accepted. In simpler terms, p-values shows the probability that variable bodymass has no effect on Height. The 'acceptable' probability is usually 5% or 1% therefore, we want our p-value to be <0.05 or < 0.01

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.0054    11.7053   8.373 3.14e-05 ***
bodymass      0.9528     0.1618   5.889 0.000366 ***
```

Ordinary Least Square

If we predict height using our model, we will get

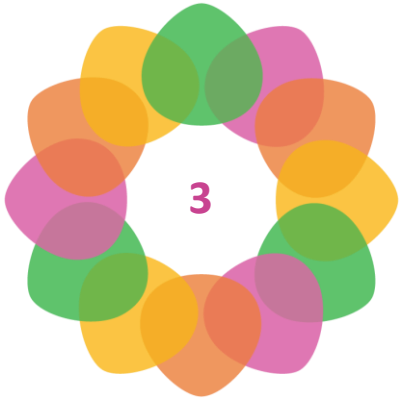
No	height	bodymass	height_modelled	e_i^2
1	176	82	176.13	0.0
2	154	49	144.69	86.6
3	138	53	148.50	110.3
4	196	112	204.72	76.0
5	132	47	142.79	116.3
6	176	69	163.75	150.1
7	181	77	171.37	92.7
8	169	71	165.65	11.2
9	150	62	157.08	50.1
10	175	78	172.32	7.2

Which can help us assess how 'good' model is.

- R^2. R^2 shows which proportion of the variation of Y can be explained by the linear regression of Y.
  In our case, $R^2 = \frac{var(\widehat{Height})}{var(Height)} = 1 - \frac{ESS(error\ sum\ of\ squares)}{TSS(total\ sum\ of\ squares)}$=0.81
- Residual Standard Error. $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Height_i - \widehat{Height_i})^2}{n}}$=9.4
- F-statistic. Similar to p-value, but we test hypothesis that all regression coefficients are equal to 0.

```
F-statistic: 34.68 on 1 and 8 DF,  p-value: 0.0003662
```

SECTION



Regression modelling: multivariable regression

Linear multivariable regression: overview

In real life, most of the regression models will be multivariable

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + error$$

For example,

$$LOS = 0.2 + 0.1 * Age + 0.05 * BMI + 1.5 * NumberOfComorbidities$$

The estimation principles are the same – we are trying to minimise the error. The ways to estimate the quality of the regression are largely the same:

- R^2. In multilinear regression, we should use R^2 adjusted (for number of variables)
- Residual Standard Error (RMSE).
- F-statistic. Similar to p-value, but we test hypothesis that all regression coefficients are equal to 0.
- Restricted VS Unrestricted regression test. Instead of testing Hypothesis that all coefficients are equal to 0, this test allows us to check the significance of specific variables. E.g, we can test which regression is a better fit

$$HealthCareCost = b_o + b_1Age + b_2MLTC + b_3IMD \textbf{ VS}$$
$$HealthCareCost = b_o + b_1Age + b_2MLTC + b_3IMD + b_4BMI + b_5Cigarettes$$

If probability that the b_4 and b_5 equal to 0 > 0.05 then with 95% confidence interval we can say that the restricted regression is 'better'

Multicollinearity

What if we add variables to the model but they correlate with each other?

E.g $HealthCareCost = b_o + b_1 Age + b_2 MLTC + b_3 BMI + b_4 Weight$

If correlation is high enough, it may cause problems, such as biased estimates of coefficients.

How to detect?

1. Build a regression with independent variables of interest. E.g. $BMI = b_o + b_1 Weight$
2. Calculate R^2 for the regression 1.
3. Calculate Variance Inflation Factor (VIF). $VIF = \frac{1}{1-R^2}$

VIF >= 10 shows the problem (but in weaker models, values >=5 might also be a sign of concern)

Causes of multicollinearity:

- Wrong choice of independent variables
- Dummy variables being used incorrectly (extra category has not been excluded)
- Insufficient data

What else can we do? Have more observations, exclude variable of concern, use non-linear models or aggregators.

Dummy variable

In regression modelling, we quite often use categorical (dummy) variables. E.g. coding of gender in SUS or marital status.

$$LifeSatisfaction = \beta_0 + \beta_1 Income + \beta_2 Marital\ Status + error$$

$Marital\ Status$= 0 if single (divorced, separated, etc)

$Marital\ Status$= 1 if not single (married, in relationship)

Let's say Life Satisfaction is a self-reported value from 0 to 10 and all coefficients in the regression below are significant.

$LifeSatisfaction = 1.5 + 0.2 * Income + 1.5 * Marital\ Status$. How would we interpret the regression then?

Single people: $LifeSatisfaction = 1.5 + 0.2 * Income + 1.5 * 0 = 1.5 + 0.2 * Income$

Married people: $LifeSatisfaction = 1.5 + 0.2 * Income + 1.5 * 1 = 3 + 0.2 * Income$

What if there are 'structural' changes and the dummy variable also affect the slope?

$$LifeSatisfaction = 1.5 + 0.2Income + 1.5Marital\ Status + 0.1Marital\ Status * Income + error$$

Dummy variable

$$LifeSatisfaction = 1.5 + 0.2Income + 1.5Marital\ Status + 0.1Marital\ Status * Income + error$$

Single people: $LifeSatisfaction = 1.5 + 0.2 * Income$

Married people: $LifeSatisfaction = 1.5 + 0.2 * Income + 1.5 + 0.1 * Income = 3 + 0.3 * Income$

To check if dummy affects a slope, an intercept or both, we can do restricted VS unrestricted test (Chow Test) and test that coefficients β_2 and β_3 equals 0. Alternatively, we can test that the coefficients are equal for both subsets

Single: $LifeSatisfaction = \beta_0 + \beta_1 Income + error$

Married: $LifeSatisfaction = \alpha_0 + \alpha_1 Income + error$

We are testing that $\alpha_0 = \beta_0$ and $\alpha_1 = \beta_1$. This can be done in R manually (split dataset=> run regression => run F test) or automatically using strucchange package.

Dummy variables

What if there are more than 2 categories of a categorical variable? E.g. more than 2 genders, ethnicity, categories of education (primary school/secondary school/College/University)

If we have 5 wide ethnic groups: White, Asian, Black, Mixed, Other

$$CovidLOS = \beta_0 + \beta_1 Age + \beta_2 White + \beta_3 Asian + \beta_4 Black + \beta_5 Mixed + \beta_6 Other + error$$

$White$=1 if White, 0 if otherwise

$Asian$=1 if Asian, 0 if otherwise

And so on

Any problem?

Dummy variables

$$CovidLOS = \beta_0 + \beta_1 Age + \beta_2 White + \beta_3 Asian + \beta_4 Black + \beta_5 Mixed + error$$

If there are n categories in the variable, we always add (n-1) to the model. Otherwise, it will lead to multicollinearity.

How to interpret results?

$$CovidLOS = 1 + 0.1 * Age - 0.5 * White + 1.5 * Asian + 1.2 * Black + 1.3 * Mixed + error$$

White: $CovidLOS = 1 + 0.1 * Age - 0.5 = 0.5 + 0.1 * Age$

Asian: $CovidLOS = 1 + 0.1 * Age + 1.5 * 1 = 2.5 + 0.1 * Age$

Black: $CovidLOS = 1 + 0.1 * Age + 1.2 * 1 = 2.2 + 0.1 * Age$

Mixed: $CovidLOS = 1 + 0.1 * Age + 1.3 * 1 = 2.3 + 0.1 * Age$

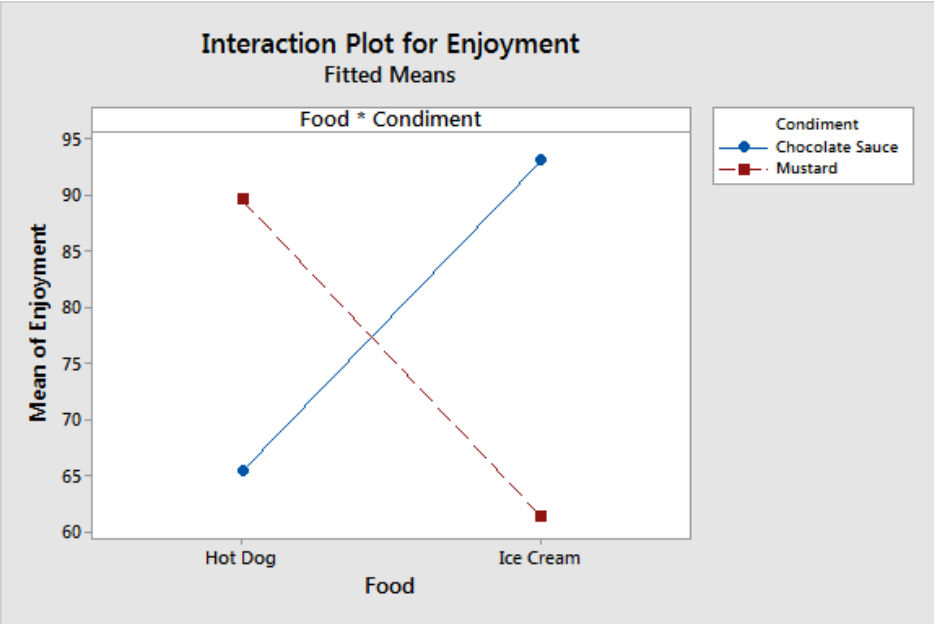
Other: $CovidLOS = 1 + 0.1 * Age$

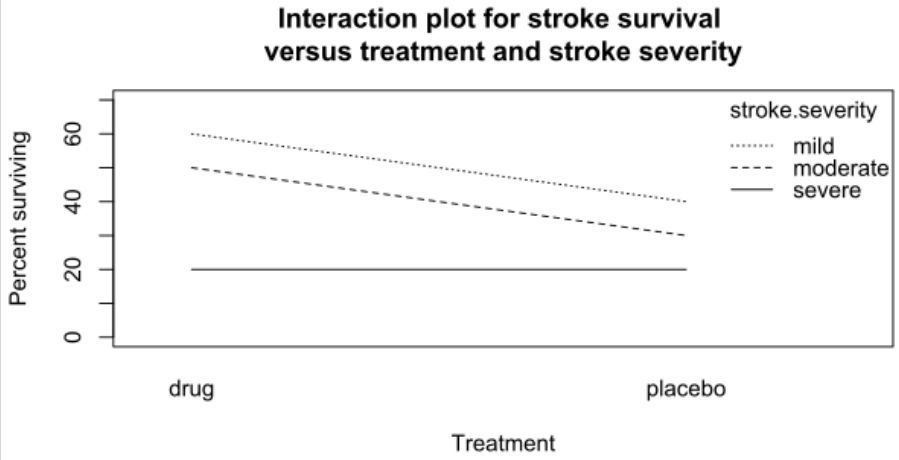
If other factors are the same (Age), White patient will spend in hospital 2 less days than the Asian patient.

Interaction terms

If an independent variable has a different effect on the outcome depending on the values of another independent variable, we need to add interaction terms. This is similar to the 'slope change' in the dummy variable example.

$$LifeSatisfaction = 1.5 + 0.2Income + 1.5Marital\ Status + 0.1Marital\ Status * Income + error$$



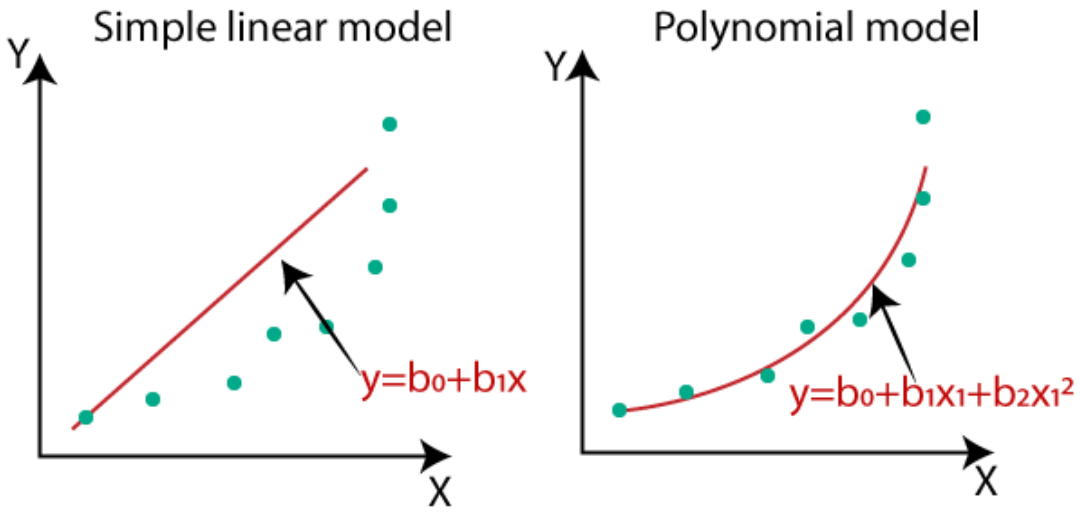


In the exercise, we will look at interaction both between dummy variables and continuous variables.

*https://statisticsbyjim.com/

Non-linear regression

The relationship between variables are not always linear, e.g. they can be logarithmic or polynomial. For example, education and salary or age and number of comorbidities/number of appointments



$App = 0.5 + 0.05 * Age$

At age 30, $App = 0.5 + 1.5 = 2$

At age 60, $App = 0.5 + 3 = 3.5$

At age 70, $App = 0.5 + 3.5 = 4$

$App = 0.5 + 0.01 * Age + 0.002 * Age^2$

At age 30, $App = 0.5 + 0.3 + 1.8 = 2.6$

At age 60, $App = 0.5 + 0.6 + 7.2 = 8.3$

At age 70, $App = 0.5 + 0.7 + 9.8 = 11$

Exercise: interpret regression

What impacts self-reported health?

1 - Excellent

2 - Good

3 – Average

4 – Poor

5 – Very poor

	OR	-95%CI	+95%CI	P
Age	1.05	1.04	1.06	<0.0001
Education:				
primary	1			
skilled	0.95	0.64	1.41	0.8
high school	0.57	0.39	0.83	0.003
college/university	0.41	0.25	0.67	0.0003
Income:*				
<75	1			
75–124	0.67	0.46	0.99	0.04
125–249	0.68	0.47	0.98	0.04
250–374	0.80	0.46	1.40	0.4
>374	0.30	0.14	0.63	0.002
Poor control over life score	1.59	1.46	1.74	<0.0001
Chronic illnesses:				
absence	1			
presence	7.18	5.74	9.00	<0.0001

*The association between income, education, lifestyle and psychosocial stressor and obesity in elderly

Logistic regression

In healthcare analysis, we are often interested in the binomial variable

- got readmitted – yes/no
- discharged alive – yes/no

If we use linear model:

$$ReadmissionRisk = \beta_0 + \beta_1 Age + \beta_2 MLTC + error$$
$$ReadmissionRisk = -0.3 + 0.01 * Age + 0.02 * MLTC$$

For 60 years old with 6 MLTC, the risk is 0.42. For 20 years old with 0 comorbidities, it is -0.1

Using maximum likelihood estimation, we can assume that the probability is following logistic regression. So, we can use Logit-analysis

$$\Pr(Readmission = 1) = F(Age, MLTC) = \frac{1}{1+e^{-(\beta_0+\beta_1 Age+\beta_2 MLTC+error)}} = \frac{1}{1+e^{-(-4+0.5Age+1.1MLTC+error)}}$$

For 20 years old with 0 MLTC, it will mean

$$\Pr(Readmission = 1) = \frac{1}{1+e^{-(-4+0.5*20+1.1*0)}} = \frac{1}{1+e^{-(-6)}} = \frac{1}{1+e^{-(-0.1)}} = 0.02$$

Odds ratios

$$\Pr(Readmission = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 Age+\beta_2 MLTC+error)}}$$

So, what is the effect of extra age or 1 extra multimorbidity?

If we do loads of math, we will find out for the age variable, its effect on probability can be calculated as $\exp(\beta_1)$. This value is called an odds ratio.

The results of fitting a logistic regression model on the cervical cancer dataset.

	Weight	Odds ratio	Std. Error
Intercept	-2.91	0.05	0.32
Hormonal contraceptives y/n	-0.12	0.89	0.30
Smokes y/n	0.26	1.30	0.37
Num. of pregnancies	0.04	1.04	0.10
Num. of diagnosed STDs	0.82	2.27	0.33
Intrauterine device y/n	0.62	1.86	0.40

If all other factors are the same, an increase in the number of diagnosed STDs changes the odds of cancer vs. no cancer by a factor of 2.27

If all other factors are the same, For women using hormonal contraceptives, the odds for cancer vs. no cancer are lower by 0.89, compared to women without hormonal contraceptives

*https://christophm.github.io/interpretable-ml-book/logistic.html

Average marginal effect

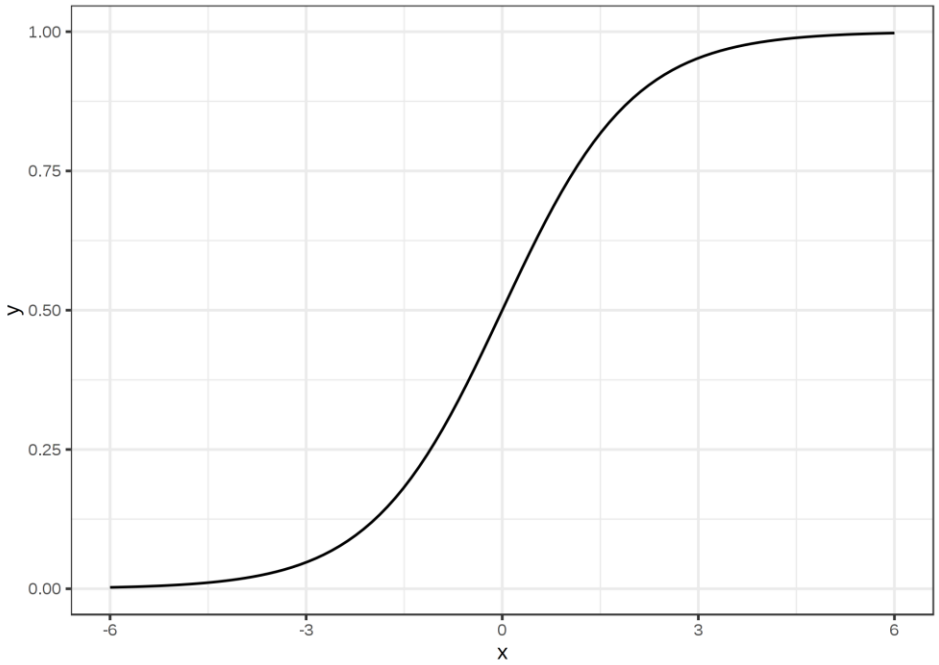
However, we are also keen to know what is the effect of an extra unit of independent variable on the outcome. In econometrics, it is called a marginal effect. From the visual representation of the logistic regression, we can see that this effect (slope) will be different subject to value of x. Mathematically, it can be calculated as a differential.

For example, for someone aged 20, the effect of 1 extra year is

$$\frac{dPr}{dAge} = \frac{e^{-(-4+0.5Age)}}{(1+e^{-(-4+0.5Age)})^2} * 0.01 = \frac{0.0024}{1.004} * 0.5 = 0.0019 \text{ or } 0.19\%$$

For someone aged 70:

$$\frac{dPr}{dAge} = \frac{e^{-(-4+0.5Age)}}{(1+e^{-(-4+0.5Age)})^2} * 0.01 = \frac{0.67}{2.79} * 0.01 = 0.0000000001$$



Average marginal effect

So, to avoid calculating effects for each age, we can calculate **average marginal effect**. Average marginal effect is the marginal of effects for all observation (please do not mix with marginal effect for a mean, although in the good model these values are close!). R can calculate it automatically using the package margins.

For a dummy variable, we don't need to do it, because it can only have 2 values: 0 and 1.

$$\Pr(Death = 1) = F(Stroke) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Stroke)}} = \frac{1}{1 + e^{-(-2 + 2 * Stroke)}}$$

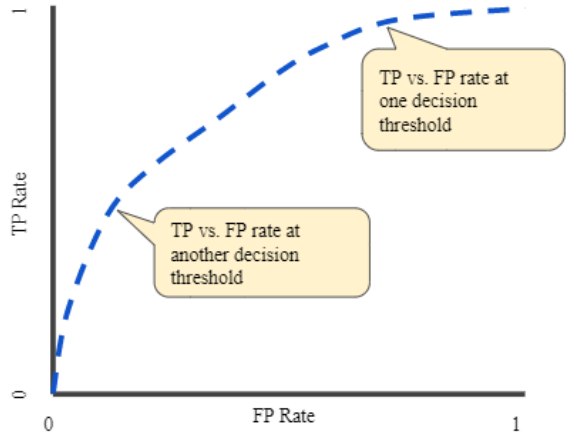
Difference in the probability of death between those who had stroke and those who did not

$$\frac{1}{1 + e^{-(-2+2*1)}} - \frac{1}{1 + e^{-(-2+2*0)}} = \frac{1}{1 + e^{0}} - \frac{1}{1 + e^{2}} = 0.38$$

Model assessment

Logistic regression can be assessed differently compared to the linear regression.

- Akaike information criteria - estimator of prediction error. Therefore – the smaller the better
- Area under Curve (AUC). Once we estimated the model and predicted the outcome values, we can estimate true positive and false positives rate and draw ROC curve (receiver operating characteristic curve)
  The close AUC to 1 the better (the more true positives)



- Visually estimate the quality of prediction – e.g. using boxplots.
- Estimate ratio of correctly predicted outcomes.
- Pseudo $R^2 = 1 - \frac{ln(L)}{var(L0)}$ where L is the regression we have just built and L0 is the regression with intercept only
- To compare models - Likelihood ratio test (similar to restricted VS unrestricted)

Exercise: interpret regression

Research question: impact of economic crisis on the risk of abortion. Outcome variable – risk of abortion.

Table 6 – main regression

Table 7 – interaction terms

Table 7. Marginal effects of economic crises on the risk of abortion for different subsamples of women

		Crisis 1990's	Crisis 2008-2009	Crisis 2014-2015
Marital status	Single	0.416*** (0.103)	-0.136 (0.158)	-0.329 (0.234)
	Married	0.241*** (0.060)	-0.012 (0.084)	-0.136 (0.092)
Area of living	Urban	0.357*** (0.063)	-0.069 (0.088)	-0.180* (0.099)
	Rural	0.287*** (0.106)	-0.026 (0.140)	-0.177 (0.173)
Region of living	West	0.451*** (0.097)	-0.120 (0.149)	-0.286 (0.163)
	East	0.201 (0.225)	-0.15 (0.090)	-0.136 (0.102)

**Significant at 10% level*
***Significant at 5% level*
****Significant at 1% level*

Table 6. Marginal effects of the determinants of abortion for model with regional dummies

Variables	Average marginal effects	Standard errors
Crisis 2014-2015	-0.079**	0.057
Crisis 2008-2009	-0.044	0.077
Crisis 1990's	0.363***	0.155
Age	0.058***	0.024
Age^2	-0.002***	0.000
Marital status	0.070***	0.006
Area	-0.135***	0.058
Number of children	0.043***	0.007
Number of children^2	-0.074***	0.017
Employment	-0.056	0.052
Log of Income	-0.031***	0.003
Concern to afford essentials	-0.073	0.047
High education	-0.170***	0.059
Good health	-0.061	0.045
Bad health	0.061	0.106
Disability	0.106	0.150
Alcohol more than 4 times per week	0.211***	0.082
Alcohol 1-3 times per week	0.165***	0.060
Alcohol 1-2 times per month	0.100***	0.010
Smoking	0.086***	0.021
Poor life satisfaction	0.111***	0.049
Regional dummies		
Northern and North Western	0.031***	0.010
Central and Central Black-Earth	0.018*	0.009
Volga-Vaytski and Volga Basin	0.022**	0.010
North Caucasian	0.059***	0.010
Ural	0.051***	0.010
Western Siberian	0.045***	0.009
Eastern Siberian and Far Eastern	0.036***	0.011

**Significant at 10% level*
***Significant at 5% level*
****Significant at 1% level*

Ordinal regression

We might have a model when dependent variable of interest has more than 2 categories. For example,

- Risk categories (high risk/medium risk/low risk)
- Cost categories
- Rating of a GP practice

We can use the principles of logistic regression.

$$Pr(Rating \geq r) = F(GPs, Receptionist) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 GPs + \beta_2 Receptionist + error)}}$$

Alternatively, we can use probit-model. In logit model we assumed that the Probability is a logistic distribution. In probit model, we assume that it is a normal distribution.

Ordinal regression works very much like binominal logistic regression and odds ratios will show the odds of the changing category

We are not going to go through ordinal regressions today, but might do it in future sessions if it is something you have to work to in your organisation.

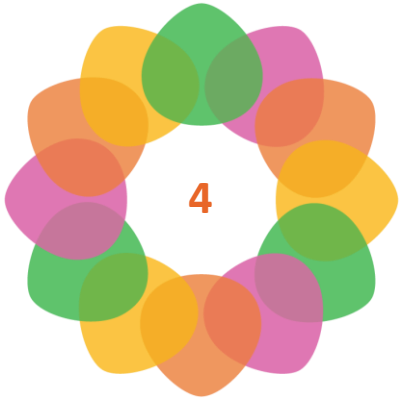
Exercise: interpret regression

Four cost groups: Very High (highest 2% costing service users), High (2%-10%), Medium (10%-50%) and Low (lowest 50% costing service users).

Patient-level data. Research question: how different services affect the total cost of the services for a patient?

	odds ratios	standard errors
AE	1.276***	-0.026
Amb	3.301***	-0.034
IpNonEmerg	17.046***	-0.025
IpEmerg	19.052***	-0.031
Mat	34.114***	-0.06
OpFirst	2.719***	-0.019
OpFoll	2.872***	-0.018
SCAssess	1.436***	-0.055
SCPackage	364.564***	-0.071
CommContact	2.515***	-0.019
CommIP	8.391***	-0.064
CommOP	2.025***	-0.023
MHContact	12.437***	-0.037
MHOP	3.464***	-0.043
MHIP	92.298***	-0.135
Note:	*p<0.1; **p<0.05; ***p<0.01	

SECTION



Time-Series Analysis

Time-Series data

On Tuesday, we looked at **cross-sectional data** (many observations at one point of time).

The other common type of data is **time-series data** - change in the variable over time. For example, time trend of A&E attendances, Delayed Transfers of Care (DTOC), number of appointments/admissions.

Why do we use time-series analysis?

- Forecasting future activity
- Explore causal relationship in time. If we increase number of beds now, how will it affect waiting times in a year?
- We don't have any other data. E.g. we don't have access to SUS/HES, but want to explore A&E activity

The other type of data is **panel data** - cross-sectional time series data – which looks at characteristics of observations over time. Most survey data (Understanding Society) and SUS/HES data is panel data. Sometimes it is also considered to be a quasi-experimental approach itself, so we will look into it at a separate session.

Univariate Time Series Models

Time-Series analysis can be univariate (the change in **one variable** over time) and multivariate (changes in **many variables**).

Univariate models:

- Autoregressive models (AR)
- Moving average models (MA)
- Autoregression with moving average (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)
- Seasonal Autoregressive Integrated Moving Average (SARIMA)
- White noise

Multivariate models:

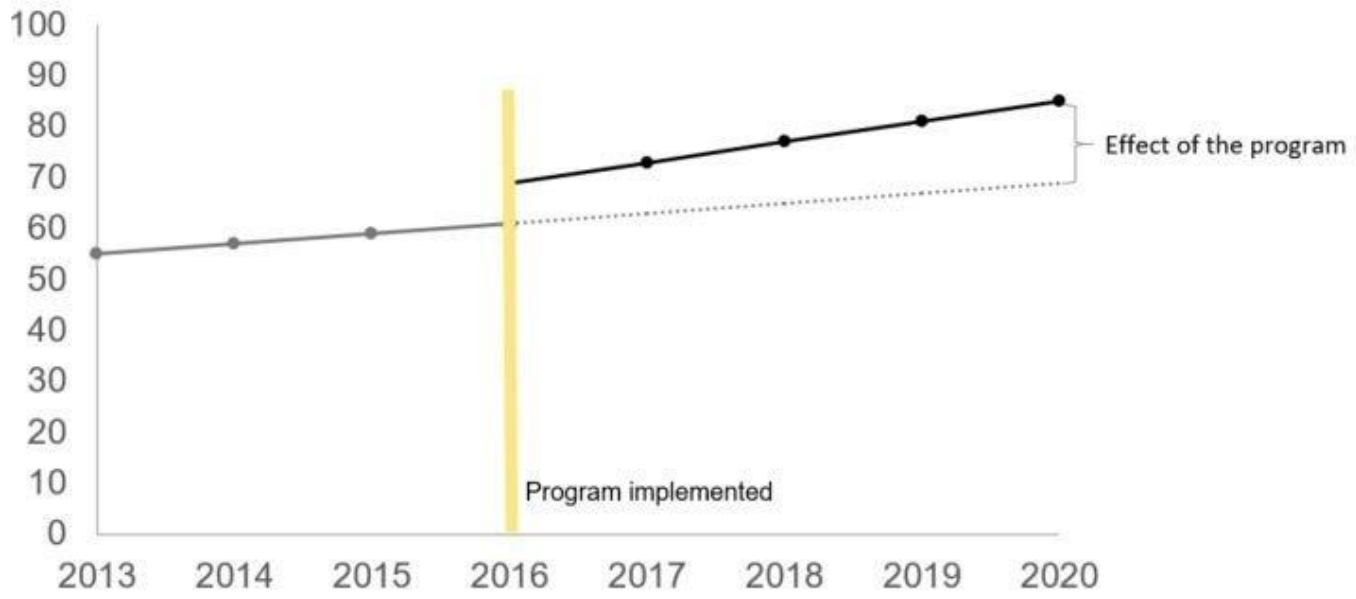
- Bayesian structural time series
- Vector Autoregressive models

Today we will look at univariate Time Series.

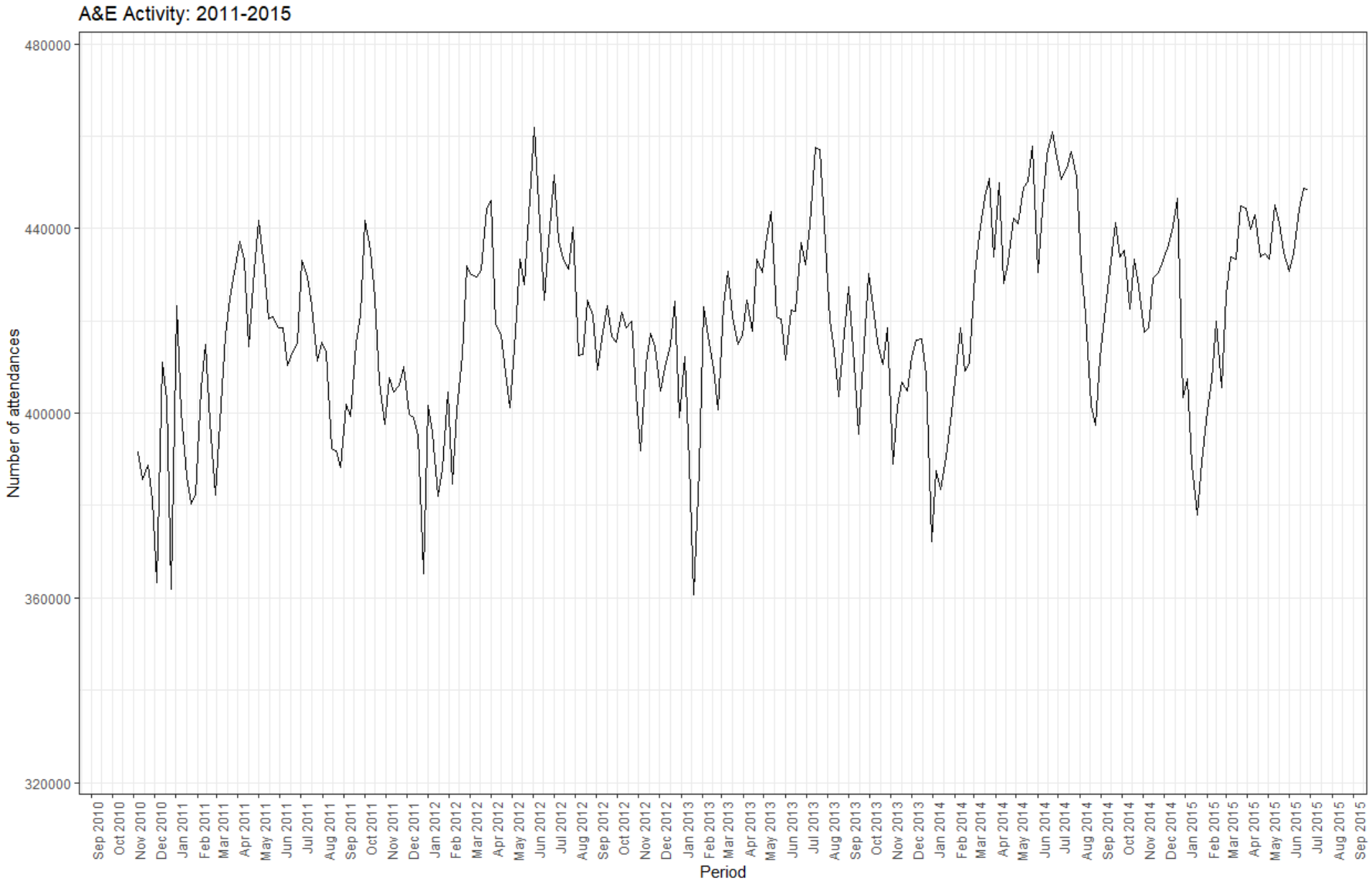
Year	Period	Total DTOC
2010-11	August	4,940
2010-11	September	5,004
2010-11	October	4,588
2010-11	November	4,409
2010-11	December	3,861
2010-11	January	4,597
2010-11	February	4,404
2010-11	March	4,170
2011-12	April	3,910
2011-12	May	4,056
2011-12	June	4,137
2011-12	July	4,228
2011-12	August	4,144
2011-12	September	4,165

Time-Series models

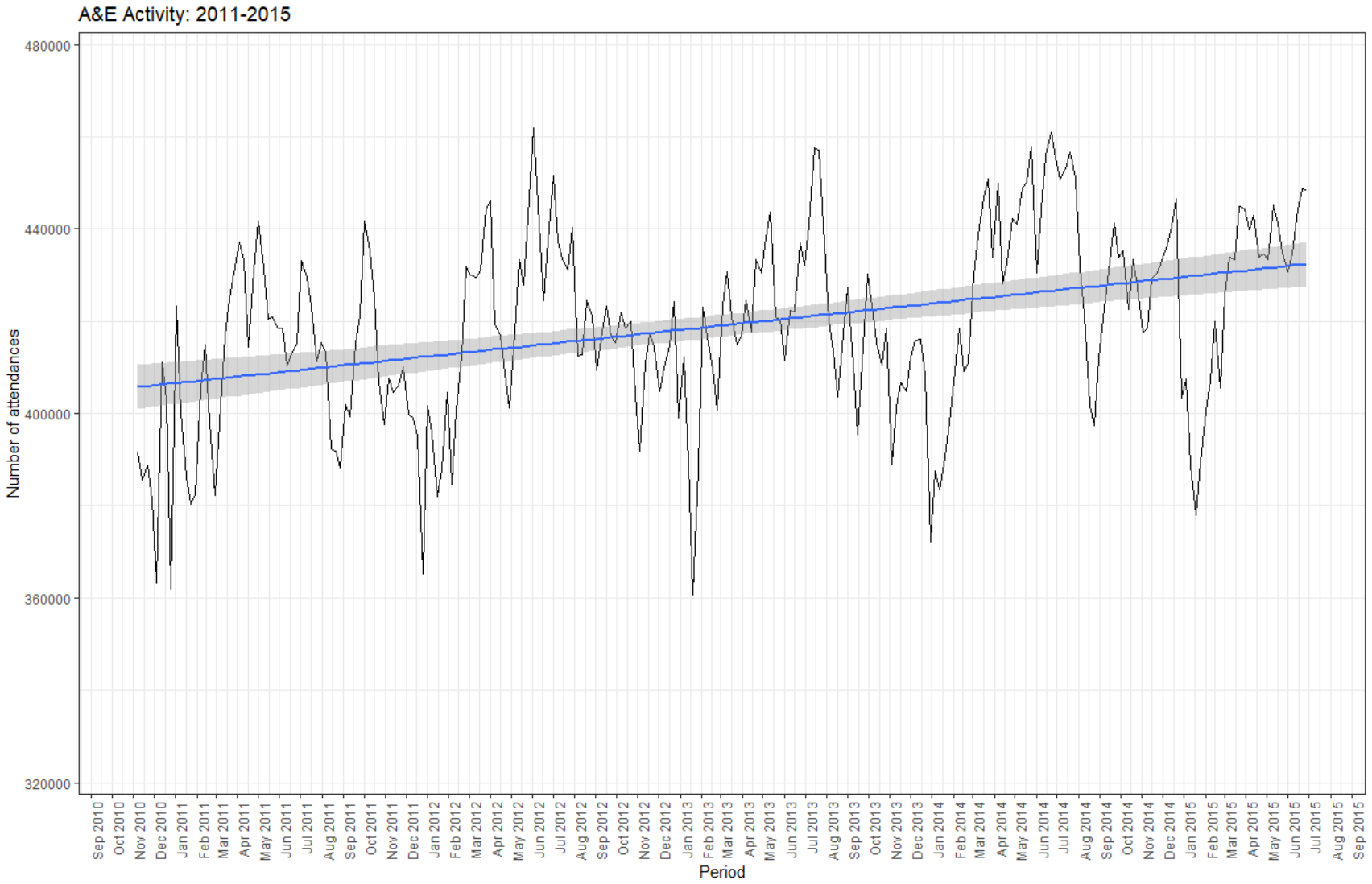
Univariate Time Series are widely used in quasi-experimental approaches, for example, in the interrupted time series analysis.



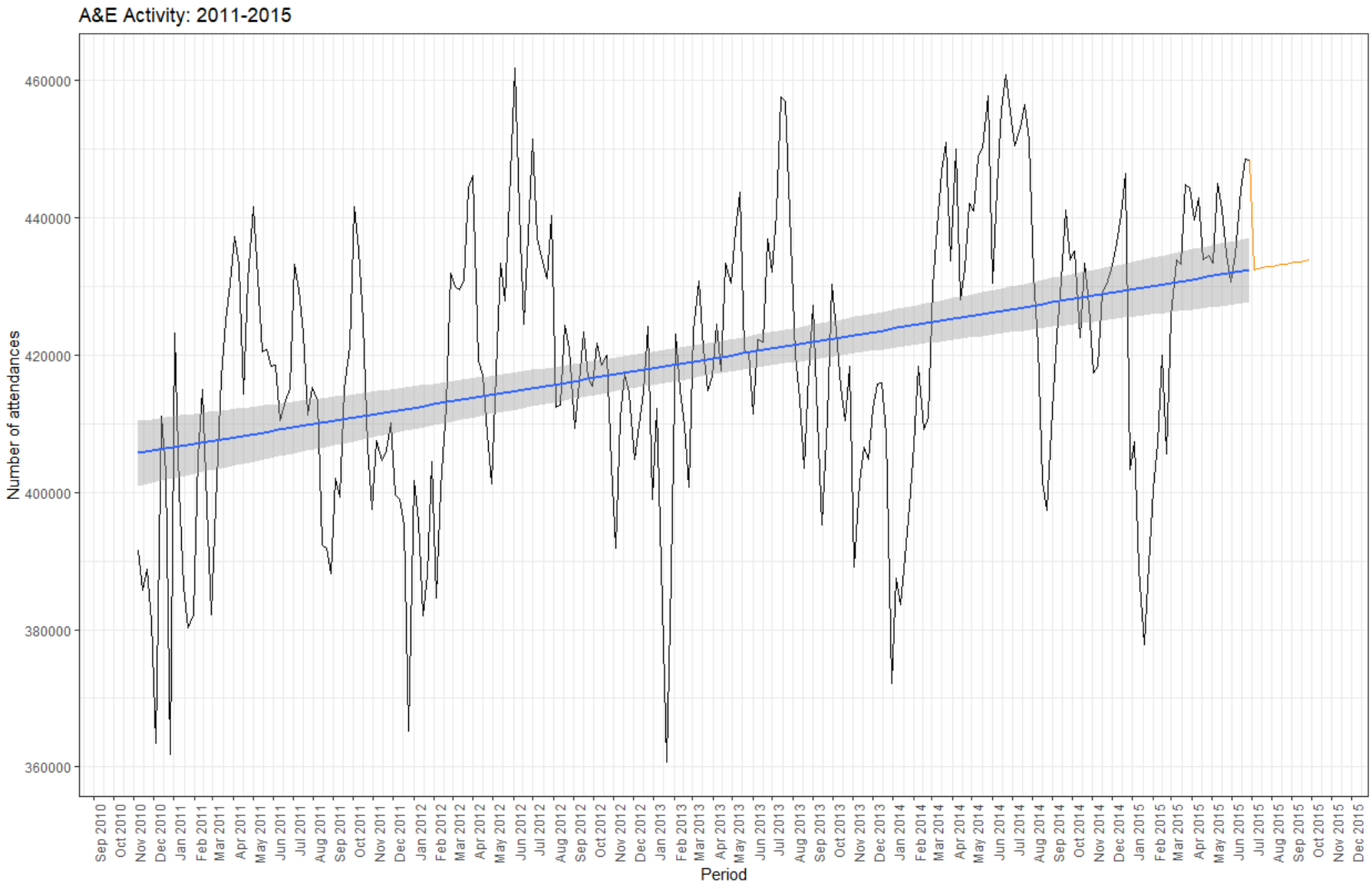
If we only have Number of A&E attendances over time, what can we use to predict future attendances?



We can say that A&E attendances is the function of time and use linear model, but it will cause problems



Knowing linear trend is helpful, but it is never accurate enough



Autoregressive (AR) and Moving average (MA) models

We can assume that the activity now (at moment t) depends on activity in the past

- In previous period: last year for annual data/last month for monthly data (at moment t-1) – AR(1)
- Or activity 2 periods ago (t-2) – AR(2)
- Or activity p periods ago – AR(p)

$$y_t = \beta_0 + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + \cdots + \beta_p * y_{t-p} + e$$
$$DTOC_{2019} = \beta_0 + \beta_1 * DTOC_{2018} + \beta_2 * DTOC_{2017} + e$$

On the other hand, activity now (at moment t) can be some function of average.

- Function of average and some error in last period – MA(1)
- Function of average and some error 2 periods ago – MA(2)
- Function of average and some error in the period p – MA(p)

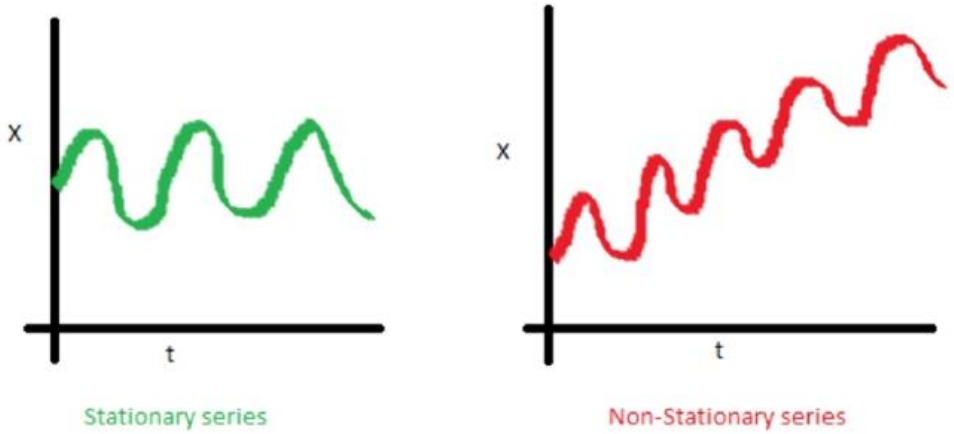
$$y_t = \mu + e_t + \alpha_1 * e_{t-1} + \cdots + \alpha_p * e_{t-p}$$
$$DTOC_{2019} = \mu + e_t + \alpha_1 * e_{t-1} + \cdots + \alpha_p * e_{t-p}$$

Or it can be both – ARMA(p,q)

$$y_t = \mu + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + \cdots + \beta_p * y_{t-p} + e_t + \alpha_1 * e_{t-1} + \cdots + \alpha_q * e_{t-q}$$

Stationarity

Both AR, MA and ARMA models imply the time series will be **stationary**. A stationary time series is one whose properties do not depend on the time at which the series is observed. Or, in other words, that the future will be like past.

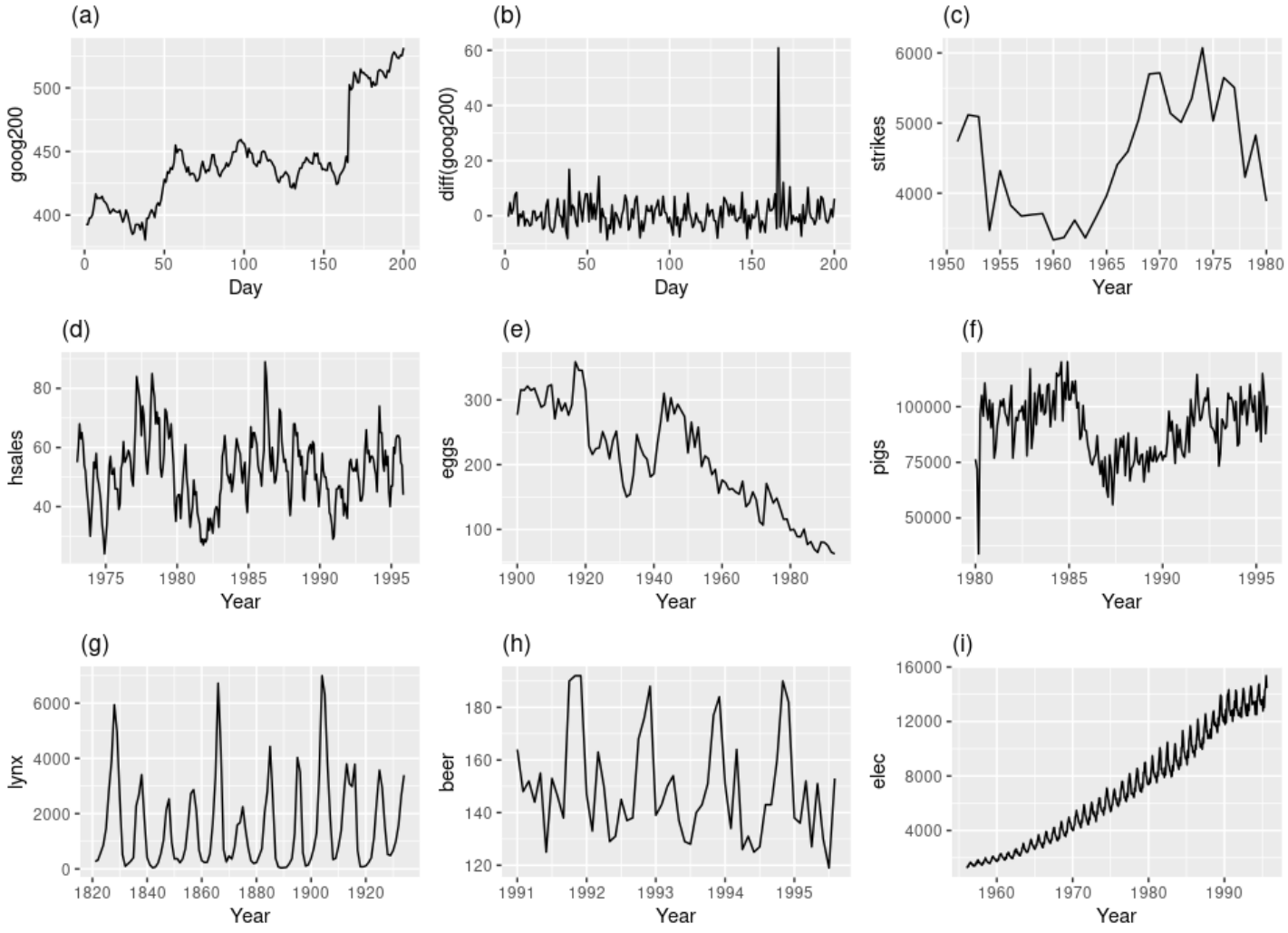


Visually, non-stationary time series can be distinguished from stationary, because they have a clear trend and a seasonality.

*Forecasting: principles and practice - https://otexts.com/fpp2/stationarity.html

Stationarity

Sometimes we can identify non-stationary time series visually



Stationarity

We cannot model non-stationary time-series. But how to formally test if the data stationary or not?

1. Augmented Dickey–Fuller (ADF) t-statistic test
   - Testing β_1=1 (unit root) in $y_t = \beta_0 + \beta_1 * y_{t-1} + e$
   - Different scenarios for different specifications can be tested: e.g. with or without intercept
   - If $\beta_1 = 1$, the time series is non-stationary

   In R, ADF test can be run using the tseries package, the function adf.test ()

2. Ljung-Box test. Testing the hypothesis that lags of the variable correlate. H0: the time series is non-stationary. Statistics value is being calculated using correlations values between lags.

   In R, Ljung-Box test can be performed using the base stats package, the function box.test ()

3. Kwiatkowski-Phillips-Schmidt-Shin (KPSS). H0: the time series is stationary.
   - Estimating $y_t = \beta_0 + \beta_1 * y_{t-1} + e$
   - Calculating residuals e and comparing KPSS statistics with the critical value

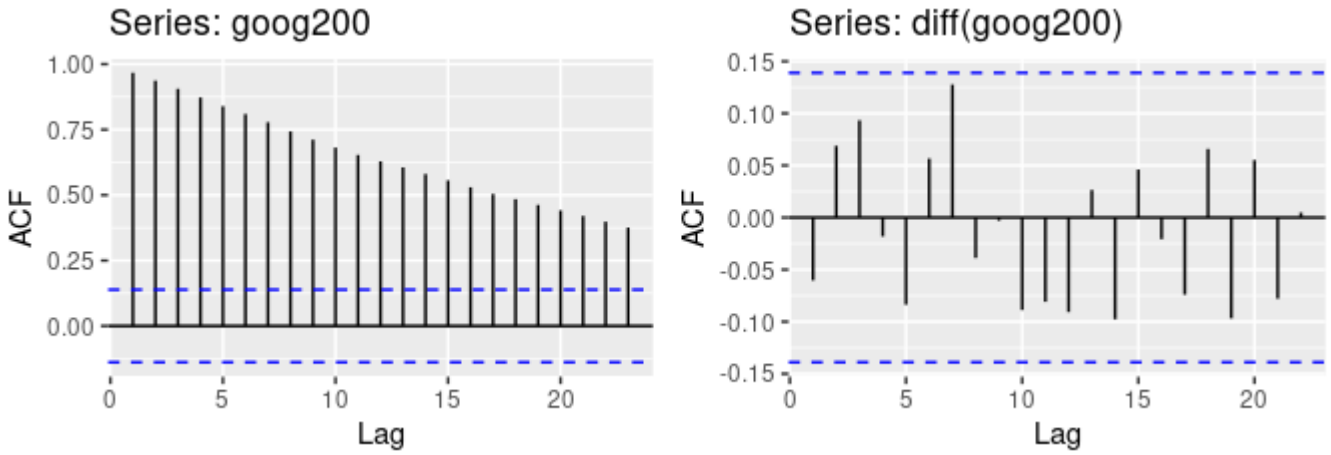
   In R, KPSS test can be run using using the tseries package, the function kpss.test ()

Stationarity

We can also check stationarity via looking at ACF and PACF plots.

ACF is an (complete) auto-correlation function - describes how well the present value of the series is related with its past values

PACF is a partial auto-correlation function - finds correlation of the residuals with the next lag value



ACF on the left is for non-stationary time series. ACF on the right is for stationary time series (no autocorrelations lying outside the 95% limits)

In R, ACF and PACF can be created using acf() and pacf() functions in the base stats package.

*Forecasting: principles and practice - https://otexts.com/fpp2/stationarity.html

Stationarity

If we have to deal with non-stationary series, we need to make them stationary first. How to make non-stationary data stationary?

We can perform data transformations such as taking logarithms to exclude the trend and seasonality from the data. However, the simplest way is to perform differencing.

If $y_t = \beta_0 + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + error$ is non-stationary (β_1 is very close to 1), we need to subtract y_{t-1} from the both sides of equation

$$y_t - y_{t-1} = \beta_0 + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} - y_{t-1} + error$$

$\Delta y_t = \beta_0 + (1 - \beta_1) * y_{t-1} + \beta_2 * y_{t-2} + error$ will be stationary and $(1 - \beta_1)$ will be close to 0. Hence no trend

In R, we can apply diff() function.

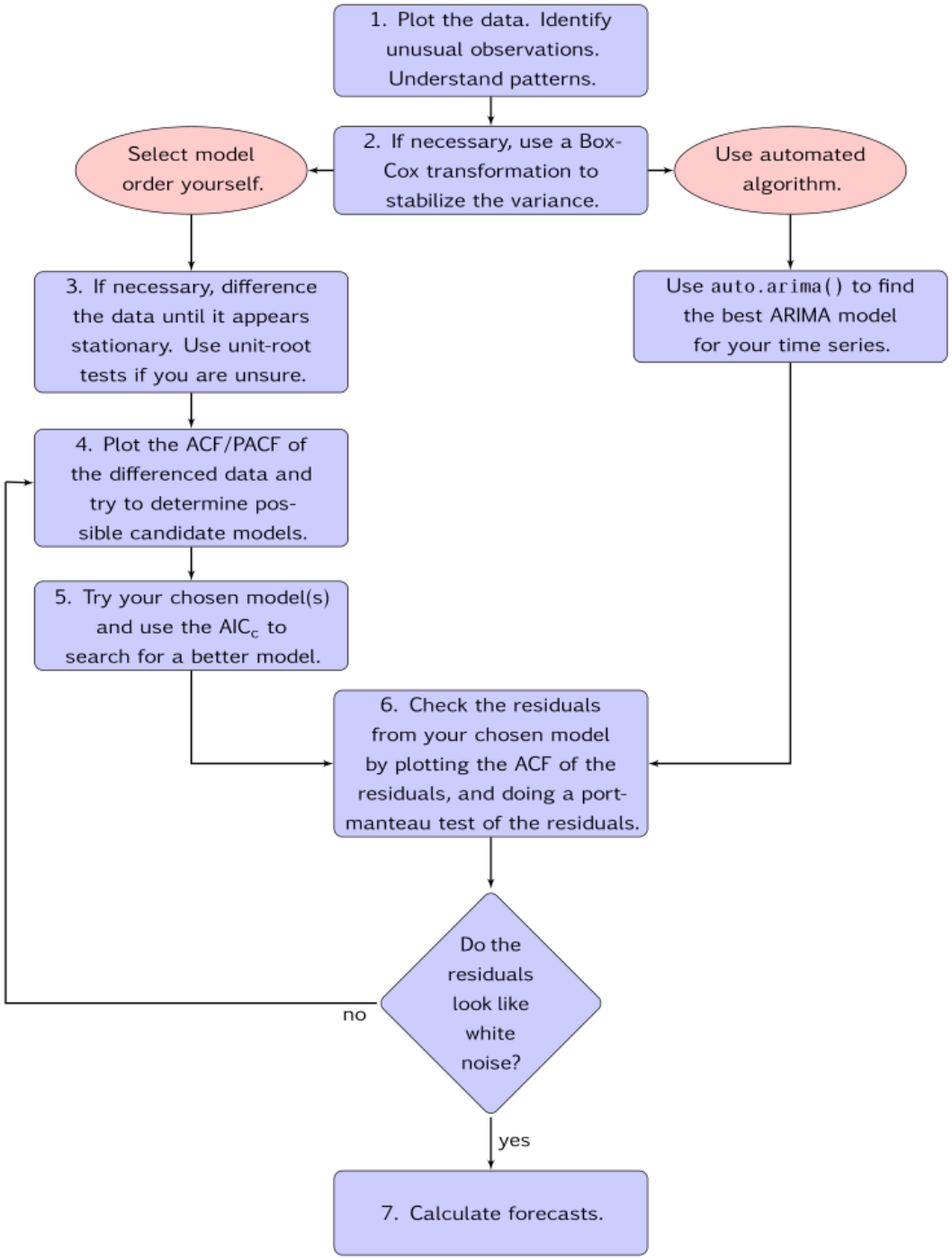
ARIMA models

ARIMA(p,d,q) model is the time series model for non-stationary time series. It has the same structure as ARMA model (p,q), but it has I element – integration – hence d value shows the order of differencing.

1. Test if the time series is stationary
   - If time series is stationary, we can start building the model
   - If time series is non-stationary, we need to perform differencing and test again. The order of differencing at which time series becomes stationary is the d value
     For example, if $y_t - y_{t-1}$ is a stationary time series, we are looking at ARIMA (p,1,q)
2. Analyse ACF and PACF charts to find optimal parametres.
   - If ACF is high and is close to 0 only for very high lags, it is AR(p). PACF needs to be close to 0 for lags>p
   - If PACF is high and is close to 0 only for very high lags, it is MA(q). ACF needs to be close to 0 for lags>q
   - Otherwise, we are having both autoregression and moving average. We can find the order of autoregression and moving average via looking at charts.
3. Build ARIMA Model. After we identified orders of p,q and d, we can call arima() function. Alternatively, we can use auto.arima function in R, which will find parameters automatically. It is not as straightforward to interpret coefficients in the ARIMA model because he had to do differencing before. But we still can use the model to predict future values.

ARIMA model

4. Assess the quality of the model
   - Check the significance of the coefficients. R does not calculate p values automatically, but we can
     - Check the t statistics similarly to linear models. $t = \frac{\beta}{se(\beta)}$. If t-statistics >=1.96, the parameters of the model are significant.
     - Calculate p-value (probability that the coefficients are not significant). Coeftest() function from the lmtest package will give us p-values or we can calculate them manually from the normal distribution (1-pnorm(abs(model1$coef)/sqrt(diag(model1$var.coef))))*2.
   - Information criteria: Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The smaller the better.
   - Residuals analysis and the model variance σ^2. We want both to be as low as possible
5. Forecast.
   - Predicting future values will help identify if the forecast is adequate
   - Pseudo Out-of-Sample Forecasts. We might want to predict past values to see how far it is from the actual true values. To do it, we need to take 10-15% of the sample, predict the values, calculate errors and calculate the average error. Ideally, it should be as close to 0 as possible.



Forecasting principles and practice
https://otexts.com/fpp2/arima-r.html

Seasonality

What if there are factors other then previous periods and average, that we are aware of?

- Season – winter pressures in the NHS
- Day of a week – less or more discharges depending on the day of a week
- Time – more activity in the A&E during out of office hours
- ‘Moving’ seasons – Easter Break, School holidays – not so easy to incorporate, but possible

Various time series models deal with seasonality. For example, we can build seasonal ARIMA models– SARIMA. SARIMA models have the same parameters as ARIMA (p,d,q), but it also has a seasonal component. SARIMA(p,d,q)x(P,D,Q,s):

- p and seasonal P: indicate number of lags
- d and seasonal D: indicate differencing that must be done to stationarize series
- q and seasonal Q: indicate number of lags of the forecast errors
- s: indicates seasonal length in the data

The process of fitting sarima model is very similar to ARIMA models. We can either try to find parameters manually using sarima function, or use automated function auto.arima and state that there is a seasonality (seasonal=T).

SECTION

Next Steps

What is next?

Next full-day session – Directed Acyclic Graphs (DAGs) – week commencing 28th of June AM

Optional homework

If you want (and have capacity) to test your regression modelling knowledge, you can either

- Use your own data and research question
- You have a question but no data – we can try to find some open data or create a synthetic data for you
- You don't any regression modelling question you want to test, but you want to practice - I can send you some synthetic data

You can then attend optional 'clinic' sessions on the week commencing 21st of June to talk about your model of choice and ask any questions.

If you have any questions or comments, please email me! Anastasiia.zharinova@nhs.net

Extra Reading

General regression modelling:

1. "Mostly harmless econometrics" by Angrist and Pischke – available online
2. DataCarpentry have a great online page with materials/exercises - I wish I saw it when I prepared my materials
3. Some NHS/public health examples:
   - NHS England - Understanding drivers of emergency admissions for ambulatory care-sensitive conditions
   - NHS Digital – Dentists working patterns, motivation and morale
   - What is driving obesity trends? – some methods we have not learnt yet, but the article might be interesting

Multivariate regressions:

1. Neus Valveny and Stephen Gilliver: How to interpret and report the results from multivariable analyses
2. Getting started with multivariate multiple regression
3. Logistic Regression in R
4. ROC/AUC to measure the quality of logistic model

Time Series Analysis:

1. Forecasting: principles and practice - Brilliant blog/book with step-by-step ARIMA chapter
2. Applied Time Series in R - similar to the previous one, but harder coding
3. Forecasting R training by Bahman Rostami-Tabar and NHS-R Community