Statistics and Machine Learning 1

Lecture 6: Regression II

Diego Perez Ruiz
Department of Social Statistics
School of Social Sciences
University of Manchester

9 November 2023

what today?

- indicator variables
- interaction effects
- polynomial regression
- multicollinearity
- multilevel modelling
- missing data

dummy variables

what are they?

the way to include categorical variables as explanatory variables

if there are only two categories to a variables,
then you assign one group to 0 and the other to 1

a dummy variable takes on two values:

$$x = \begin{cases} 0, & \text{if the observation does not belong to the category} \\ 1, & \text{if the observation belongs to the category} \end{cases}$$

dummy variable

what are they?

in a simple regression formula

$$y = \beta_0 + \beta_1 x + \epsilon$$

where x is a dummy variable, we get the following interpretation:

- β_0 is the mean of the first group
- β_1 is the difference between the group means

dummy variable

what are they?

in a simple regression formula

$$y = \beta_0 + \beta_1 x + \epsilon$$

where x is a dummy variable, we get the following interpretation:

- β_0 is the mean of the first group
- β_1 is the difference between the group means

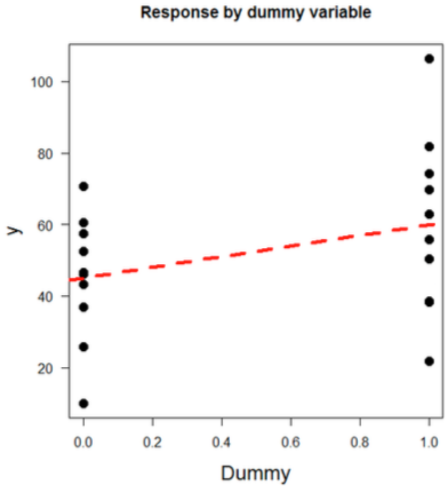
example.
in an estimated model with a dummy variable X (e.g. female/male)

$$\hat{y} = 45 + 15x$$

we predict a value for male=0 as $45 + 15(0) = 45$
and female=1 as $45 + 15(1) = 60$

category 1 (i.e. $x = \text{male} = 0$) is used as the 'baseline' for comparison

dummy variables



dummy variables

more categories than two

so what if the categorical variable has more than two categories?

dummy variables

more categories than two

so what if the categorical variable has more than two categories?

- we use more than one dummy
- one category is then baseline and called **reference variable**
- for reference variable, all dummies are 0
- for other categories, one of the dummies is 1, the rest 0

dummy variables

more categories than two

so what if the categorical variable has more than two categories?

- we use more than one dummy
- one category is then baseline and called **reference variable**
- for reference variable, all dummies are 0
- for other categories, one of the dummies is 1, the rest 0

generally
a variable with k categories requires $(k-1)$ dummy variables
(alternatively, use k dummy variables but drop intercept term)

dummy variables

example.
assume a variable with three categories ($k = 3$)
we need two ($k - 1$) dummy variables to describe it

y = car sales
x = colour of car (blue, red, green)

a linear model with sales predicted by colour includes two dummies:

$$y = \beta_0 + \beta_1 \underbrace{(x_{red})}_{\text{1 if red}} + \underbrace{\beta_2(x_{green})}_{\text{1 if green}} + \epsilon$$

blue is the baseline/reference variables (when red and green are 0)

- β_0 for blue cars
- $\beta_0 + \beta_1$ for red cars
- $\beta_0 + \beta_2$ for green cars

dummy variables

example.
assume a variable with three categories ($k = 3$)
we need two ($k - 1$) dummy variables to describe it

y = car sales
x = colour of car (blue, red, green)

a linear model with sales predicted by colour includes two dummies:

$$y = \beta_0 + \beta_1 \underbrace{(x_{red})}_{\text{1 if red}} + \underbrace{\beta_2(x_{green})}_{\text{1 if green}} + \epsilon$$

blue is the baseline/reference variables (when red and green are 0)

- β_0 for blue cars
- $\beta_0 + \beta_1$ for red cars
- $\beta_0 + \beta_2$ for green cars

- β_0: mean sale for blue
- β_1: difference in means green & blue
- β_2: diff. in means red & blue

dummy variables

advantages

- allows the inclusion of multiple categorical variables in model
- can show which means are significantly different from baseline

disadvantage

- any test is done in comparison to baseline

dummy variables

example from lab 6

life expectancy in 2007 predicted by 5 continents

$$\hat{y} = 54.8 + 18.8 I_{America} + 15.9 I_{Asia} + 22.8 I_{Europe} + 25.9 I_{Oceania}$$

(africa as reference variable)

```
Call:
lm(formula = lifeExp ~ continent, data = gapminder2007)

Residuals:
Min          1Q       Median    3Q      Max
-26.9005  -4.0399   0.2565   3.3840  21.6360

Coefficients:
                   Estimate   Std. Error t value Pr(>|t|)
(Intercept)         54.806      1.025    53.446  < 2e-16 ***
continentAmericas   18.802      1.800    10.448  < 2e-16 ***
continentAsia       15.922      1.646     9.675  < 2e-16 ***
continentEurope     22.843      1.695    13.474  < 2e-16 ***
continentOceania    25.913      5.328     4.863  3.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.395 on 137 degrees of freedom
Multiple R-squared:  0.6355,Adjusted R-squared:  0.6249
F-statistic: 59.71 on 4 and 137 DF,  p-value: < 2.2e-16
```

interaction effects

when combining variables to make new regression terms

the relationship between the primary predictor and outcome varies across levels of another predictor

interaction effects

when combining variables to make new regression terms

the relationship between the primary predictor and outcome varies across levels of another predictor

example. consider model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is salary, x_1 is experience, x_2 is gender

this model allows average salary to differ for men and women, but the difference in average salary between men and women is always the same regardless of experience

interaction effects

when combining variables to make new regression terms

the relationship between the primary predictor and outcome varies across levels of another predictor

example. consider model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is salary, x_1 is experience, x_2 is gender

this model allows average salary to differ for men and women, but the difference in average salary between men and women is always the same regardless of experience

effects of x_1 and x_2 are **additive**

interaction effects

we create a model that allows

- the average salary to differ for men and women
- the difference in average salary between men and women to change as experience increases

interaction effects

we create a model that allows

- the average salary to differ for men and women
- the difference in average salary between men and women to change as experience increases

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{(x_1 \times x_2)}_{\text{interaction}} + \epsilon$$

interaction effects

we create a model that allows

- the average salary to differ for men and women
- the difference in average salary between men and women to change as experience increases

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{(x_1 \times x_2)}_{\text{interaction}} + \epsilon$$

$(x_1 \times x_2) = (\text{gender} \times \text{experience})$
$= 0\times$ experience = 0 for men
$= 1\times$ experience = experience for women

interaction effects

we create a model that allows

- the average salary to differ for men and women
- the difference in average salary between men and women to change as experience increases

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{(x_1 \times x_2)}_{\text{interaction}} + \epsilon$$

$(x_1 \times x_2) = (\text{gender} \times \text{experience})$
$= 0\times$ experience = 0 for men
$= 1\times$ experience = experience for women

include **main effects** for any interaction terms being included

$\downarrow$

principle of hierarchy: required for interpretation of model

interaction effects

interactions of a categorical variable with more than two groups requires each dummy variable to have its own interaction term

this is interpreted as the increase/decrease in slope for each group compared to the baseline group

interaction effects

interactions of a categorical variable with more than two groups requires each dummy variable to have its own interaction term

this is interpreted as the increase/decrease in slope for each group compared to the baseline group

interaction terms can also be between two categorical variables but can be 'expensive'

interaction effects

interactions of a categorical variable with more than two groups requires each dummy variable to have its own interaction term

this is interpreted as the increase/decrease in slope for each group compared to the baseline group

interaction terms can also be between two categorical variables but can be 'expensive'

example.
an interaction between x_1 and x_2
where x_1 has 5 categories and x_2 has 6 categories would introduce $(5-1)(6-1) = 20$ interaction terms

interaction effects

interactions of a categorical variable with more than two groups requires each dummy variable to have its own interaction term

this is interpreted as the increase/decrease in slope for each group compared to the baseline group

interaction terms can also be between two categorical variables but can be ’expensive’

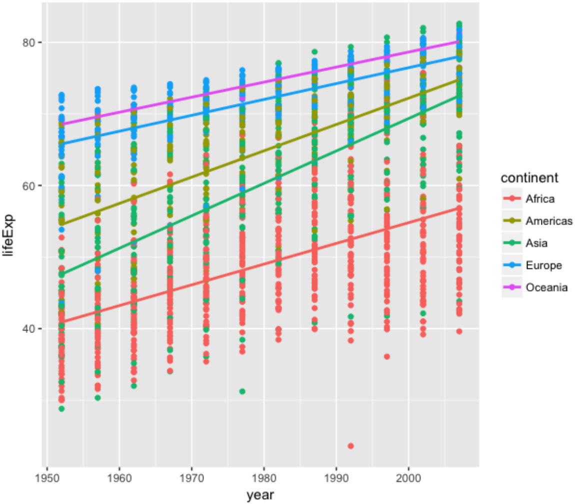
example.
an interaction between x_1 and x_2
where x_1 has 5 categories and x_2 has 6 categories would introduce $(5-1)(6-1) = 20$ interaction terms

third-order interactions can in practice also be in included but are generally not worth it

interaction effect

example from lab 6

life expectancy explained by year, grouped by the 5 continents



interaction effect

example from lab 6

we include interaction terms

```
Call:
lm(formula = lifeExp ~ year * continent, data = gapminder)

Residuals:
Min       1Q       Median    3Q      Max
-28.8854  -4.2696   0.3298   3.9835  21.1306

Coefficients:
                        Estimate   Std. Error t value Pr(>|t|)
(Intercept)           -524.25785   32.96343 -15.904  < 2e-16 ***
year                     0.28953    0.01665  17.387  < 2e-16 ***
continentAmericas     -138.84845   57.85058  -2.400  0.01650 *
continentAsia         -312.63305   52.90355  -5.909 4.14e-09 ***
continentEurope        156.84685   54.49776   2.878  0.00405 **
continentOceania       182.34988  171.28299   1.065  0.28720
year:continentAmericas   0.07812    0.02922   2.673  0.00758 **
year:continentAsia       0.16359    0.02672   6.121 1.15e-09 ***
year:continentEurope    -0.06760    0.02753  -2.455  0.01417 *
year:continentOceania   -0.07926    0.08653  -0.916  0.35980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.18 on 1694 degrees of freedom
Multiple R-squared:  0.6927,Adjusted R-squared:  0.6911
F-statistic: 424.3 on 9 and 1694 DF,  p-value: < 2.2e-16
```

interaction effect

example from lab 6

we include interaction terms

```
Call:
lm(formula = lifeExp ~ year * continent, data = gapminder)

Residuals:
Min       1Q        Median    3Q       Max
-28.8854  -4.2696   0.3298    3.9835   21.1306

Coefficients:
                        Estimate   Std. Error t value Pr(>|t|)
(Intercept)            -524.25785   32.96343 -15.904  < 2e-16 ***
year                      0.28953    0.01665  17.387  < 2e-16 ***
continentAmericas      -138.84845   57.85058  -2.400  0.01650 *
continentAsia          -312.63305   52.90355  -5.909 4.14e-09 ***
continentEurope         156.84685   54.49776   2.878  0.00405 **
continentOceania        182.34988  171.28299   1.065  0.28720
year:continentAmericas    0.07812    0.02922   2.673  0.00758 **
year:continentAsia        0.16359    0.02672   6.121 1.15e-09 ***
year:continentEurope     -0.06760    0.02753  -2.455  0.01417 *
year:continentOceania    -0.07926    0.08653  -0.916  0.35980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.18 on 1694 degrees of freedom
Multiple R-squared:  0.6927,Adjusted R-squared:  0.6911
F-statistic: 424.3 on 9 and 1694 DF,  p-value: < 2.2e-16
```

interaction effect

example from lab 6

we include interaction terms

```
Call:
lm(formula = lifeExp ~ year * continent, data = gapminder)

Residuals:
Min       1Q        Median    3Q      Max
-28.8854  -4.2696   0.3298   3.9835  21.1306

Coefficients:
                         Estimate  Std. Error t value Pr(>|t|)
(Intercept)            -524.25785   32.96343 -15.904  < 2e-16 ***
year                      0.28953    0.01665  17.387  < 2e-16 ***
continentAmericas      -138.84845   57.85058  -2.400  0.01650 *
continentAsia          -312.63305   52.90355  -5.909 4.14e-09 ***
continentEurope         156.84685   54.49776   2.878  0.00405 **
continentOceania        182.34988  171.28299   1.065  0.28720
year:continentAmericas    0.07812    0.02922   2.673  0.00758 **
year:continentAsia        0.16359    0.02672   6.121 1.15e-09 ***
year:continentEurope     -0.06760    0.02753  -2.455  0.01417 *
year:continentOceania    -0.07926    0.08653  -0.916  0.35980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.18 on 1694 degrees of freedom
Multiple R-squared:  0.6927,Adjusted R-squared:  0.6911
F-statistic: 424.3 on 9 and 1694 DF,  p-value: < 2.2e-16
```

interaction effect

example from lab 6

we include interaction terms

```
Call:
lm(formula = lifeExp ~ year * continent, data = gapminder)

Residuals:
Min       1Q        Median    3Q      Max
-28.8854  -4.2696   0.3298   3.9835  21.1306

Coefficients:
                         Estimate  Std. Error t value Pr(>|t|)
(Intercept)            -524.25785   32.96343 -15.904  < 2e-16 ***
year                      0.28953    0.01665  17.387  < 2e-16 ***
continentAmericas      -138.84845   57.85058  -2.400  0.01650 *
continentAsia          -312.63305   52.90355  -5.909 4.14e-09 ***
continentEurope         156.84685   54.49776   2.878  0.00405 **
continentOceania        182.34988  171.28299   1.065  0.28720
year:continentAmericas    0.07812    0.02922   2.673  0.00758 **
year:continentAsia        0.16359    0.02672   6.121 1.15e-09 ***
year:continentEurope     -0.06760    0.02753  -2.455  0.01417 *
year:continentOceania    -0.07926    0.08653  -0.916  0.35980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.18 on 1694 degrees of freedom
Multiple R-squared:  0.6927,Adjusted R-squared:  0.6911
F-statistic: 424.3 on 9 and 1694 DF,  p-value: < 2.2e-16
```

polynomial regression

regression allows for a tool to handle non-linearity:
the polynomial model

polynomial regression

regression allows for a tool to handle non-linearity:
the polynomial model

useful when transformations cannot linearise the relation between the predictors and the response

polynomial regression

regression allows for a tool to handle non-linearity:
the polynomial model

useful when transformations cannot linearise the relation between the predictors and the response

reminder: a polynomial function has the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots a_n x^n$$

where x_j is the j-th order polynomial term

- x is first order term, x^2 is second order term, etc.
- the degree of a polynomial is the highest order term

polynomial regression

a model is said to be linear when it is linear in parameters
so both below models are linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

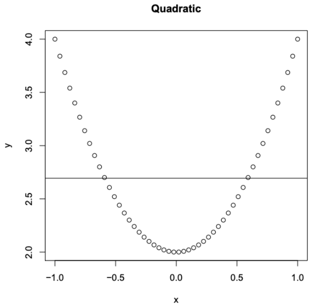
polynomial regression

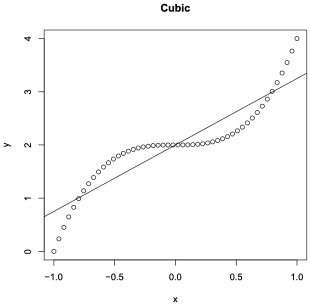
a model is said to be linear when it is linear in parameters
so both below models are linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

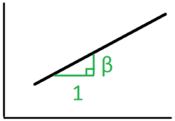
visual examples.





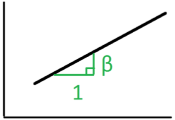
polynomial regression

in linear regression we assume that y increases/decreases with x but does so at the same rate for every x

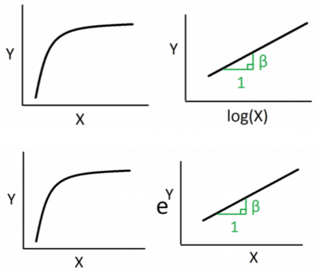


polynomial regression

in linear regression we assume that y increases/decreases with x but does so at the same rate for every x

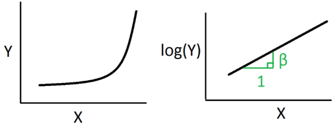


sometimes a transformation can fix a non-linear pattern



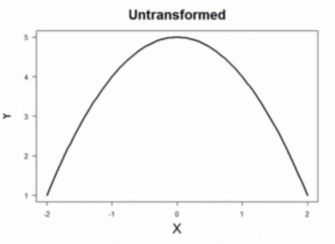
simple regression:
x or y can be transformed

multiple regression:
only y can be transformed



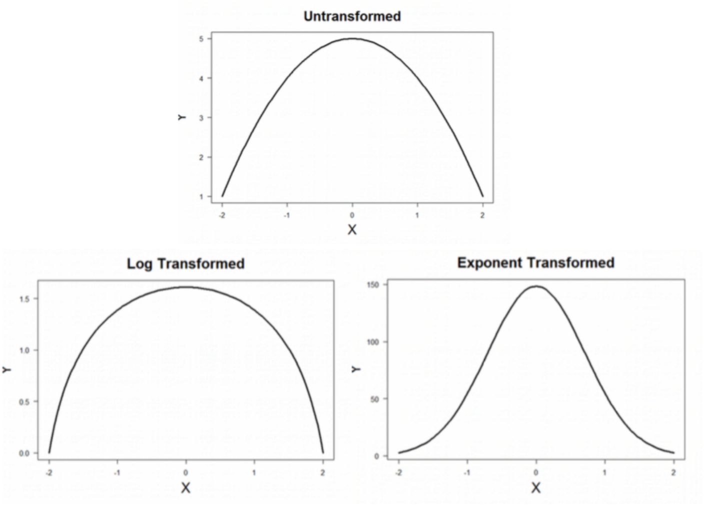
polynomial regression

transforms work when trend is monotonically increasing/decreasing
cases where trend reaches a maximum/minimum, transform will fail



polynomial regression

transforms work when trend is monotonically increasing/decreasing
cases where trend reaches a maximum/minimum, transform will fail



polynomial regression

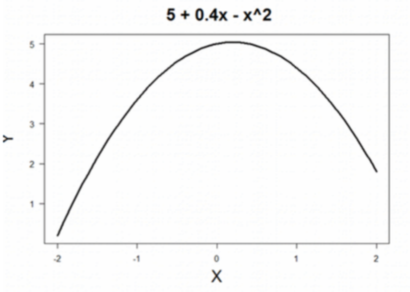
multiple terms used to describe a trend gives a model that fits well

polynomial regression

multiple terms used to describe a trend gives a model that fits well

$\downarrow$

polynomial model

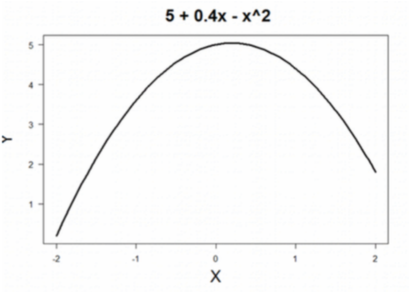


polynomial regression

multiple terms used to describe a trend gives a model that fits well

$\downarrow$

polynomial model



regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

with estimated parameters

$$b_0 = 5, \quad b_1 = 0.4, \quad b_2 = -1$$

the problem of multicollinearity

occurs when two or more **predictors** in the model are **correlated** and provide **redundant information** about the response

happens with interaction terms and in polynomial regression but also in standard regression models

examples.

- height and weight of a person
- years of education and income
- GDP and GNI

the problem of multicollinearity

occurs when two or more **predictors** in the model are **correlated** and provide **redundant information** about the response

happens with interaction terms and in polynomial regression but also in standard regression models

examples.

- height and weight of a person
- years of education and income
- GDP and GNI

consequences

- increased standard error of coefficient estimates (decreased reliability)
- often confusing and misleading results

the problem of multicollinearity

detecting multicollinearity

1. compute correlations between all pairs of predictors
   if close to -1 or1, remove one of the two correlated predictors

2. calculate **variance inflation factor** for each predictor

$$VIF_j = \frac{1}{1 - R_j^2}$$

   where R_j^2 is the coefficient of determination that includes all predictors except the j^{th}

   if $VIF_j \geq 10$, then problem with multicollinearity

the problem of multicollinearity

polynomial regression

- with terms of second or higher degree order
  e.g. x_1 and x_1^2 tend to be highly correlated

the problem of multicollinearity
polynomial regression

- with terms of second or higher degree order
  e.g. x_1 and x_1^2 tend to be highly correlated

- a solution is to use
  $$z_i = x_i - \overline{x}$$
  instead of just x_i

the problem of multicollinearity

polynomial regression

- with terms of second or higher degree order
  e.g. x_1 and x_1^2 tend to be highly correlated

- a solution is to use

  $$z_i = x_i - \overline{x}$$

  instead of just x_i

example.
assume the model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

define $z = x - \overline{x}$ and estimate model

$$\hat{y} = b_0 + b_1 z + b_2 z^2$$

the problem of multicollinearity*

useful resource

variables transformation

latent variable modelling (Principal Component Analysis)

Bayesian inference
https://avehtari.github.io/modelselection/collinear.html

multilevel modelling

data structure

assumption

response linearly related to covariates in an additive way

multilevel modelling

data structure

assumption

response linearly related to covariates in an additive way

example

final grade (**response**) of pupils explained by admission grade (**predictor**)

pupils nested and crossed from

- schools
- regions
- years

3 regions for each of 2 years

multilevel modelling

data structure

assumption

response linearly related to covariates in an additive way

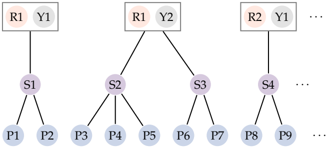
example

final grade (**response**) of pupils explained by admission grade (**predictor**)

pupils nested and crossed from

– schools
– regions
– years

$\Bigg\}$ 3 regions for each of 2 years



multilevel modelling

data structure

assumption

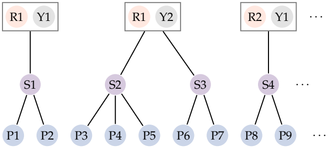
response linearly related to covariates in an additive way

example

final grade (**response**) of pupils explained by admission grade (**predictor**)

pupils nested and crossed from

$\left.\begin{array}{l}\text{– schools}\\ \text{– regions}\\ \text{– years}\end{array}\right\}$ 3 regions for each of 2 years



notations can get complicated

$y_{ijk\ell}$ = final grade of pupil ℓ from school k in region j and year i

$\ell = 1, \dots, n_{ijk}$ where n_{ijk}= # pupils from school k, region j, year i
$k = 1, \dots, S_{ij}$ where S_{ij}= # schools from region j, year i
$j = 1, 2, 3$
$i = 1, 2$

linear model

simplification: only one nested level

linear model

simplification: only one nested level

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, r$
r = total number of schools from the 3 regions and 2 years
$n = n_1 + n_2 + \cdots + n_r$ = total number of selected pupils

linear model

simplification: only one nested level

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \dots, n_j$
$j = 1, \dots, r$
r = total number of schools from the 3 regions and 2 years
$n = n_1 + n_2 + \cdots + n_r$ = total number of selected pupils

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

linear model

simplification: only one nested level

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, r$
r = total number of schools from the 3 regions and 2 years
$n = n_1 + n_2 + \cdots + n_r$ = total number of selected pupils

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

- how to estimate β_0 and β_1 from n observations
  $\Longrightarrow$ **simple OLS regression model**
- how to estimate β_{0j} and β_{1j} from n_j observations
  $\Longrightarrow$ **random intercept model**
  $\Longrightarrow$ **random slope model**

illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \dots, n_j$
$j = 1, \dots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$

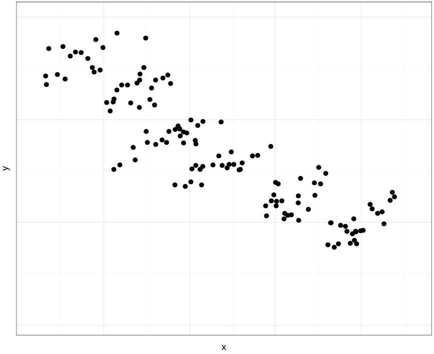
illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$

$$y_{ij} = \underbrace{(\beta_0 + u_{0j})}_{\beta_{0j}} + \underbrace{(\beta_1 + u_{1j})}_{\beta_{1j}} x_{ij} + \epsilon_{ij}$$

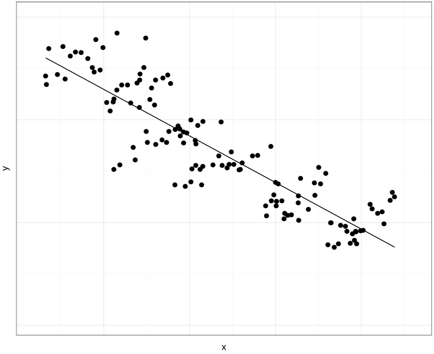


illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \dots, n_j$
$j = 1, \dots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$



simple OLS regression

$$y_{ij} = (\beta_0 + \underbrace{u_{0j}}_{=0}) + (\beta_1 + \underbrace{u_{1j}}_{=0})x_{ij} + \epsilon_{ij}$$
$$= \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

$$\beta_1 < 0$$

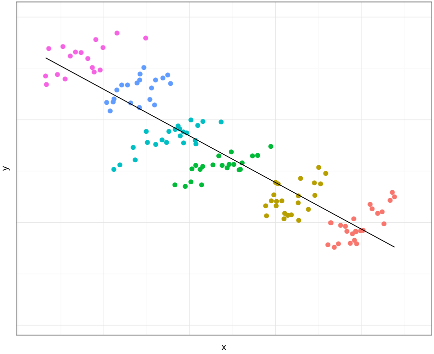
illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$

colour observations based on school



illustration

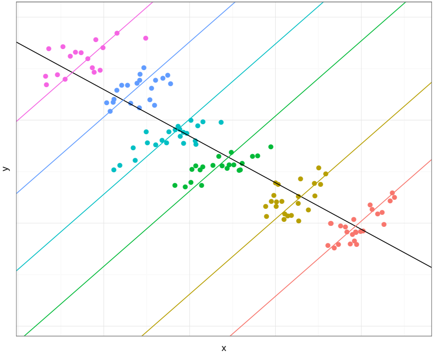
y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$

$j = 1, \ldots, 6$

$n = n_1 + n_2 + \cdots + n_6 = 130$



random intercept model

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + \underbrace{u_{1j}}_{=0})x_{ij} + \epsilon_{ij}$$

$$= \underbrace{\beta_0 + u_{0j}}_{\text{intercept}} + \beta_1 x_{ij} + \epsilon_{ij}$$

intercept is random with variance estimated from

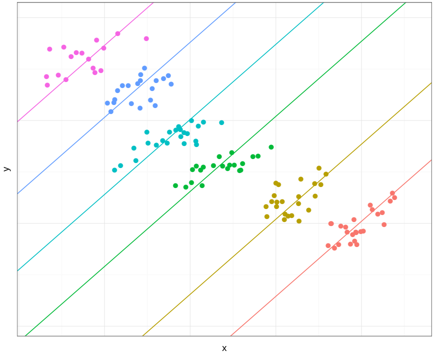
$u_{01}, u_{02}, \ldots, u_{06}$

illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$



random intercept model

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + \underbrace{u_{1j}}_{=0})x_{ij} + \epsilon_{ij}$$

$$= \underbrace{\beta_0 + u_{0j}}_{\text{intercept}} + \beta_1 x_{ij} + \epsilon_{ij}$$

intercept is random with variance estimated from
$u_{01}, u_{02}, \ldots, u_{06}$

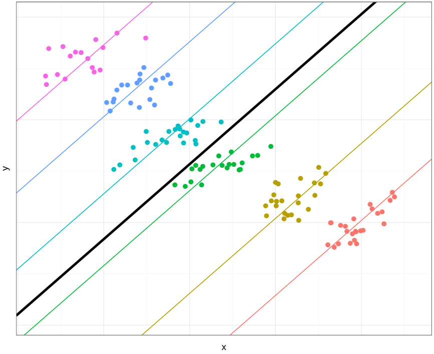
$\beta_1 > 0$

illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$



random intercept model

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + \underbrace{u_{1j}}_{=0})x_{ij} + \epsilon_{ij}$$

$$= \underbrace{\beta_0 + u_{0j}}_{\text{intercept}} + \beta_1 x_{ij} + \epsilon_{ij}$$

intercept is random with variance estimated from
$u_{01}, u_{02}, \ldots, u_{06}$

$\beta_1 > 0$

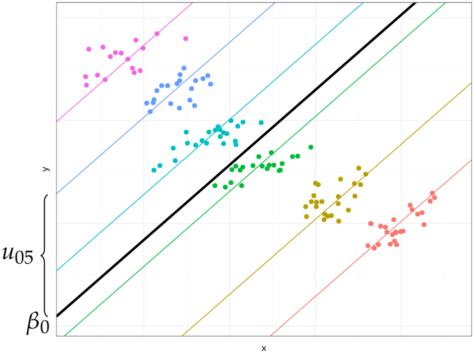
add overall population average line

illustration

y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$
$j = 1, \ldots, 6$
$n = n_1 + n_2 + \cdots + n_6 = 130$



random intercept model

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + \underbrace{u_{1j}}_{=0})x_{ij} + \epsilon_{ij}$$

$$= \underbrace{\beta_0 + u_{0j}}_{\text{intercept}} + \beta_1 x_{ij} + \epsilon_{ij}$$

intercept is random with variance estimated from

$u_{01}, u_{02}, \ldots, u_{06}$

$\beta_1 > 0$

add overall population average line

illustration

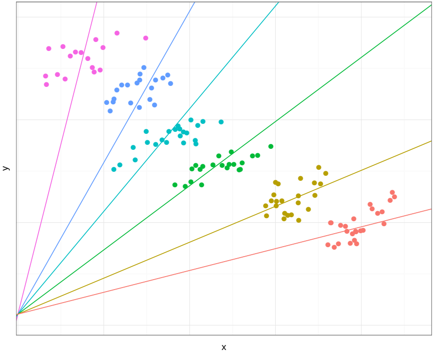
y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$

$j = 1, \ldots, 6$

$n = n_1 + n_2 + \cdots + n_6 = 130$



random slope model
(fixed intercept)

$$y_{ij} = (\beta_0 + \underbrace{u_{0j}}_{=0}) + (\beta_1 + u_{1j})x_{ij} + \epsilon_{ij}$$

$$= \beta_0 + \underbrace{(\beta_1 + u_{1j})}_{\text{slope}} x_{ij} + \epsilon_{ij}$$

slope is random with
variance estimated from
$u_{11}, u_{12} \ldots, u_{16}$
intercept is fixed

illustration

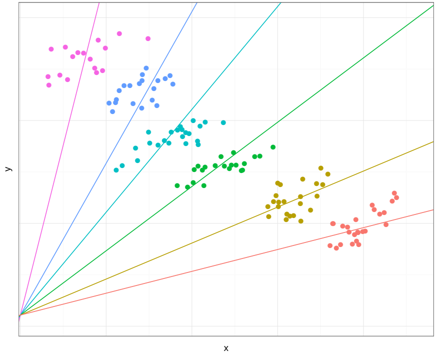
y_{ij} = final grade of pupil i from school j

x_{ij} = admission grade of pupil i from school j

$i = 1, \ldots, n_j$

$j = 1, \ldots, 6$

$n = n_1 + n_2 + \cdots + n_6 = 130$



random slope model
(fixed intercept)

$$y_{ij} = (\beta_0 + \underbrace{u_{0j}}_{=0}) + (\beta_1 + u_{1j})x_{ij} + \epsilon_{ij}$$

$$= \beta_0 + \underbrace{(\beta_1 + u_{1j})}_{\text{slope}} x_{ij} + \epsilon_{ij}$$

slope is random with
variance estimated from
$u_{11}, u_{12} \ldots, u_{16}$
intercept is fixed
what if we 'unfix' the intercept?
$\Longrightarrow$ **lab**

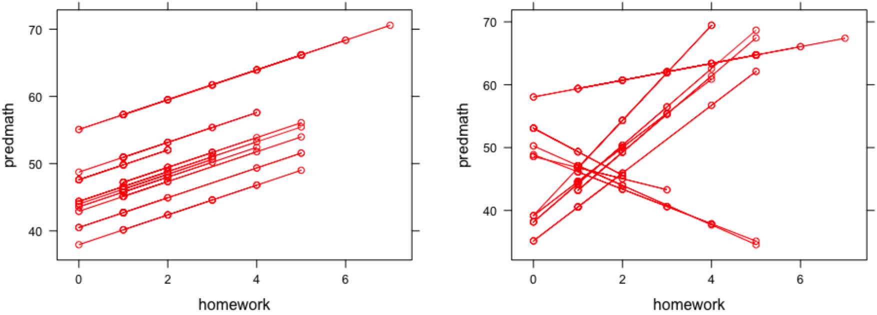
multilevel modelling

example from lab 6

random intercept and slope model

math score is predicted using hours spent on homework

data over 260 pupils in 10 schools



missing data

how is the data missing?

- **missing completely at random (MCAR)**
  data is completely missing at random:
  missingness is independent of the data values
  example: we run a taste study for 20 different drinks
  each subject was asked to rate only 4 drinks chosen at random

missing data

how is the data missing?

- **missing completely at random (MCAR)**
  data is completely missing at random:
  missingness is independent of the data values
  example: we run a taste study for 20 different drinks
  each subject was asked to rate only 4 drinks chosen at random

- **missing at random (MAR)**
  missingness depends only on the non-missing data
  (in principle the missing values can be predicted from them)
  example: in a survey, low income respondents were less likely to answer a question about drug use than wealthy subjects

missing data

how is the data missing?

- **missing completely at random (MCAR)**
  data is completely missing at random:
  missingness is independent of the data values
  example: we run a taste study for 20 different drinks
  each subject was asked to rate only 4 drinks chosen at random

- **missing at random (MAR)**
  missingness depends only on the non-missing data
  (in principle the missing values can be predicted from them)
  example: in a survey, low income respondents were less likely to answer a question about drug use than wealthy subjects

- **missing not at random (MNAR)**
  missingness depends on the missing and non-missing data
  example: respondents with high income less likely to report income

missing data

how is the data missing?

- **missing completely at random (MCAR)**
  data is completely missing at random:
  missingness is independent of the data values
  example: we run a taste study for 20 different drinks
  each subject was asked to rate only 4 drinks chosen at random

- **missing at random (MAR)**
  missingness depends only on the non-missing data
  (in principle the missing values can be predicted from them)
  example: in a survey, low income respondents were less likely to answer a question about drug use than wealthy subjects

- **missing not at random (MNAR)**
  missingness depends on the missing and non-missing data
  example: respondents with high income less likely to report income
  non-ignorable!

missing data

imputation

is the substitution of some value for a missing data point

missing data

imputation

is the substitution of some value for a missing data point

mean value imputation

replace missing values with the sample average for that item

- advantages
  - easy
  - does not affect estimates of the mean
- disadvantages
  - distorts the distribution (spike at the mean value)
  - can yield underestimation of standard errors (reduces variability)
  - weakens covariance and correlation estimates in the data (ignores relationship between variables)

missing data

imputation

is the substitution of some value for a missing data point

mean value imputation

replace missing values with the sample average for that item

- advantages
  - easy
  - does not affect estimates of the mean
- disadvantages
  - distorts the distribution (spike at the mean value)
  - can yield underestimation of standard errors (reduces variability)
  - weakens covariance and correlation estimates in the data (ignores relationship between variables)

imputation by subgroup can be more accurate

example: if data is an individual's height, better to use mean value after grouping by gender

missing data

listwise deletion

remove all data for observations that has one or more missing values

- advantages
  - simplicity
  - comparability across analyses
- disadvantages
  - reduces statistical power (lowers sample size)
  - information loss
  - estimates may be biased if data not MCAR

missing data

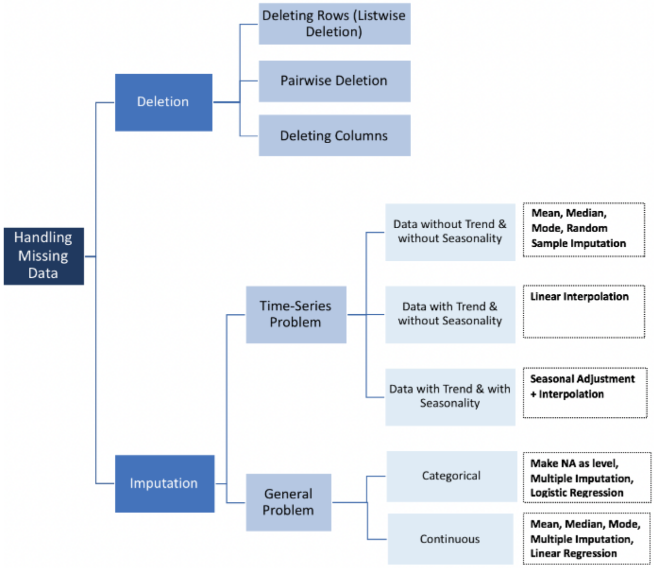


chart from towardsdatascience.com

final words: avoid too much analysis

- avoid complex models for small datasets
- try to obtain new data to validate your proposed model
- use theory and past experience with similar data to guide choice of model

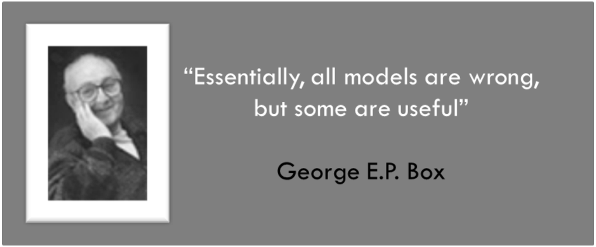
reading

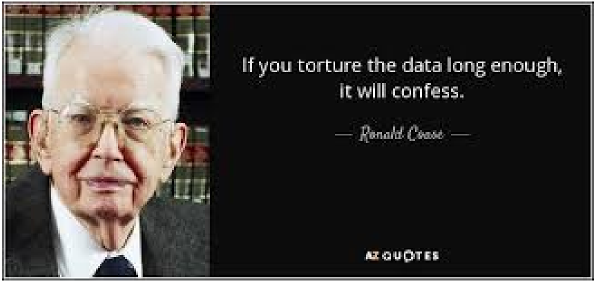
Agresti A., 2018, Statistical Methods for the Social Sciences, Fifth Edition, **Chapters 11.4, (14) 14.3, 14.5, 16.1-2**

link to the book via Manchester library

Multilevel models

final words


"Essentially, all models are wrong,
but some are useful"
George E.P. Box


If you torture the data long enough,
it will confess.
— Ronald Coase —
AZ QUOTES