Correlated Data Models

Big Data Summer Institute 2022

Irena Chen

Department of Biostatistics, University of Michigan

Table of contents

Intro

Learning Objectives

- Identify sources of correlation in datasets
- Understand intuition behind correlation and be comfortable with the mathematical definition
- Gain exposure to models for correlated data

Before we begin...

Access R files through:

- Github:

  https://github.com/realirena/bdsi_2022

- Option 1: clone the repository and work along in the ".Rmd" file on your local R
- Option 2: Download the ".html" file and follow along on your laptop

What is correlation?

Intuition behind Correlation

- Statistical models generally want to relate predictor variable X to an outcome variable Y
- Correlation is one summary that measures the degree of association (relationship) between two variables
- If X is not correlated with Y, this is not a particularly interesting question

Correlated Data Models

- **Correlated data models** focus on measuring other sources of correlation
- Want to reduce other variation (noise) sources in the data so that we can focus on X, Y
- For example: observations of X could be correlated

Example of Correlation

A researcher is interested in college students' dietary patterns. She picks a neighborhood at random and goes door-to-door interviewing any college students she encounters.

1. Are there any potential sources of correlation in the collected data?
2. What could improve this study design?

Example of Correlation

Your supervisor asks you to analyze some data to determine if individual mobility patterns can predict risk of contracting disease. You collect mobile phone data on individual travel patterns over time and whether or not the individual contracted the disease.

Your supervisor suggests that you run a simple linear regression using the distance traveled on each day to predict whether or not the individual contracted the disease on that day.

1. What are some potential sources of correlation in this data?
2. Based on your answer to Q1, would you fit the suggested linear regression model?

Mathematical Definition of Correlation

- Pearson Correlation Coefficient: common method of measuring correlation between two variables
- Mathematical definition: $\rho(X, Y) = \dfrac{E[(X - \mu_x) - (Y - \mu_y)]}{\sigma_X \sigma_y}$
    1. $E[(X - \mu_x) - (Y - \mu_y)]$: covariance between X, Y
    2. σ_x, σ_y : standard deviations of X, Y
    3. Correlation is a **standardized measurement** of covariance

Mathematical Definition of Correlation

- Pearson Correlation Coefficient: common method of measuring correlation between two variables
- Mathematical definition: $\rho(X, Y) = \dfrac{E[(X - \mu_x) - (Y - \mu_y)]}{\sigma_X \sigma_y}$
  1. $E[(X - \mu_x) - (Y - \mu_y)]$: covariance between X, Y
  2. σ_x, σ_y : standard deviations of X, Y
  3. Correlation is a **standardized measurement** of covariance
- Related to **independence**, but not equivalent
  - If X, Y are independent, $\rho(X, Y) = 0$
  - X, Y are independent **iff** $P(X|Y) = P(X)$

Mathematical Definition of Correlation

- Pearson Correlation Coefficient: common method of measuring correlation between two variables
- Mathematical definition: $\rho(X, Y) = \frac{E[(X - \mu_x) - (Y - \mu_y)]}{\sigma_x \sigma_y}$
  1. $E[(X - \mu_x) - (Y - \mu_y)]$: covariance between X, Y
  2. σ_x, σ_y : standard deviations of X, Y
  3. Correlation is a **standardized measurement** of covariance
- Related to **independence**, but not equivalent
  - If X, Y are independent, $\rho(X, Y) = 0$
  - X, Y are independent iff $P(X|Y) = P(X)$

Brain Teaser: What is the correlation between a random variable and itself? (i.e. $cor(X, X) =?$)

Identifying correlation in datasets

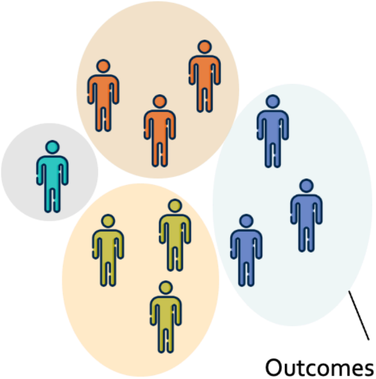
Three Examples of Correlation in Datasets

1. Clustered Data
2. Time Series Data
3. Longitudinal Data

Clustered Data

Idea: Dataset can be “clustered” into groups

- Correlation can be within group or between group
- **Hierarchical structure** to the data



Outcomes between members of the same cluster might be correlated

Clustered Data

Idea: Dataset can be "clustered" into groups

- Correlation can be within group or between group
- **Hierarchical structure** to the data

Things to consider:

1. What is the definition of a cluster?
2. What is the goal of clustering?

Examples of Clustered Data

- Spatial data
- Network data (e.g. social media communities)
- Genotyping clustering

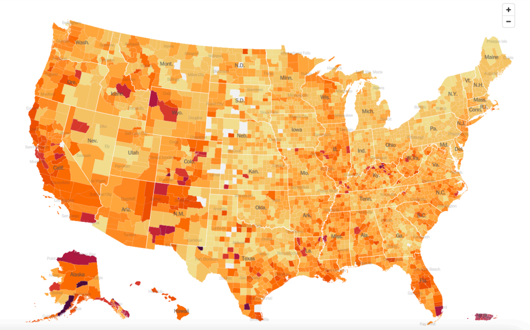


Figure 1: Originally from: https://www.nytimes.com/interactive/2021/us/covid-cases.html

K-means clustering algorithm

- Sometimes the clusters are predefined beforehand
- In other settings, the goal might be to identify clusters
- We can use **classification algorithms**

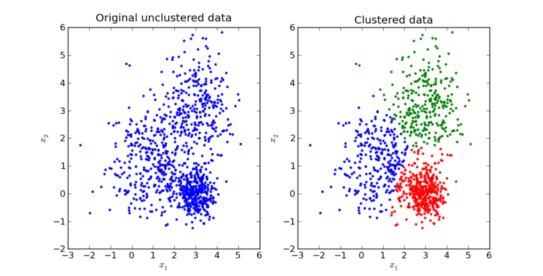


Figure 2: Example of data with k-means clustering, originally from:https://mubaris.com/posts/kmeans-clustering/

R Example

- We will use a subset of the WHO dataset (in github) for the k-means algorithm
- This dataset contains observations by country and year on health, economic, and social indicators
- **Question:** Can we detect clusters of countries that share similarities, based on these variables?

K-means clustering algorithm

- One of the most well-known algorithms
- Groups n observations into k clusters
- Specify k beforehand
- Requires continuous data
- Unsupervised algorithm

K-means clustering algorithm

For a set of observations $x_1, \ldots, x_n$ and k clusters $S_1, \ldots S_K$, we want to minimize:

$$\underset{S_i}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_i \in S_i} \|x_i - \mu_i\|^2 = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{K} \mid S_i \mid VarS_i$$

- μ_i = mean of the points in S_i
- Goal is to minimize the amount of variability in each S_i (cluster)

K-means clustering algorithm

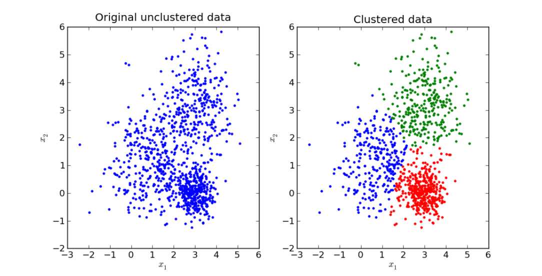
1. Pick K random points as cluster centers called centroids.
2. Assign each x_i to nearest cluster by calculating its distance to each centroid.
3. Find new cluster center by taking the average of the assigned points.
4. Repeat Step 2 and 3 until none of the cluster assignments change.

Visualization: http://shabal.in/visuals/kmeans/6.html

K-means clustering algorithm

Brain Teaser: Now that we have these clusters, it might be interesting to know if these are significant. Can we run a test or statistical model to see if these clusters are significant?

K-means clustering algorithm



- Traditional statistical methods require a hypothesis **before** data collection
- Using the same data twice will inflate Type 1 error
- The clusters are **likely significant** because we found them!
- We either need new data or novel statistical approaches
- Check out Dr. Daniela Witten’s research lab at UW for “double dipping” methods

Time Series Data

A data sequence taken at different **points in time**.

Example: Companies' stock market prices are time series, e.g. $X_{it} = X_{i,t-1} + \epsilon_{it}$

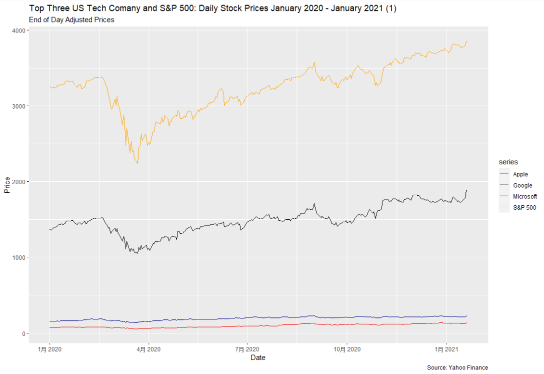


Figure 3: 2020 stock prices, originally from: https://medium.com/analytics-vidhya/plot-stock-prices-with-r-6bdbaebc8ec1

Time Series Data

A data sequence taken at different **points in time**.

Example: Companies' stock market prices are time series, e.g. $X_{it} = X_{i,t-1} + \epsilon_{it}$

- Data is not i.i.d.
- Time index (ordering) matters!
- Potential contamination: Using X_{it} to predict $X_{i,t-1}$

Time Series Data

Other examples of time series:

- Disease outbreak
- vehicle navigation prediction
- Signal processing

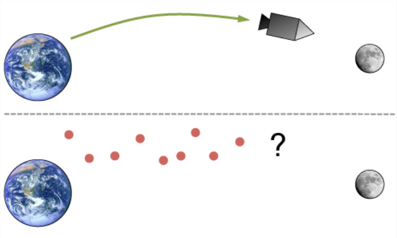


Figure 4: Image produced by https://plus.maths.org/

Time Series Data Example: AR(1) Model

Let’s return to our R script and dataset.

- We will analyze a univariate time series in this example
- Botswana has annual estimated life expectancy from 2000-2015
- Goal: Predict future life expectancy values based on past data on life expectancy

AR(1) Model

A simple **first order** AR model has the form:

$$\begin{aligned} y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t, t = 2, \cdots, T \\ \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned}$$

- y_{t-1} is the value of y at the previous timepoint
- y_0 usually taken to be a constant
- **Key difference:** The predictor, y_{t-1} is random (stochastic).
- In linear regression, the predictors are treated as fixed (deterministic).

Visualizing the AR(1) Model

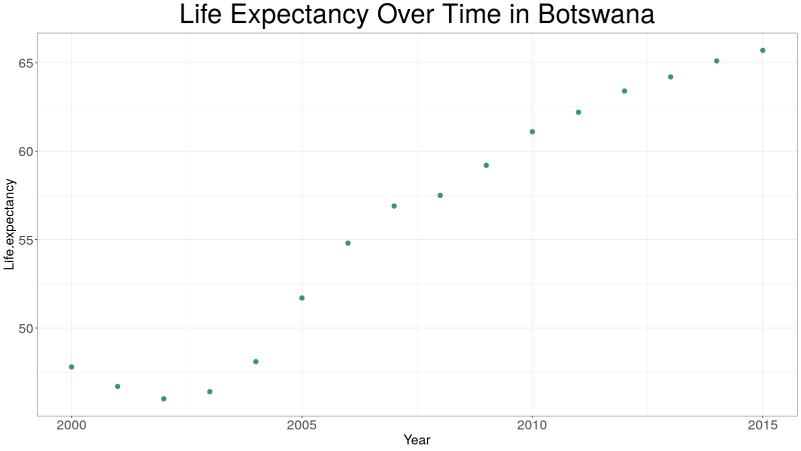


Figure 5: Scatterplot of life expectancy over time for Botswana.

Visualizing the AR(1) Model

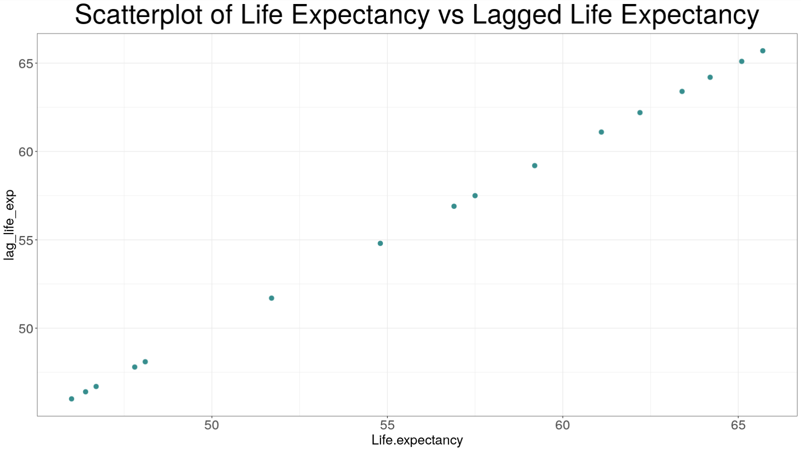
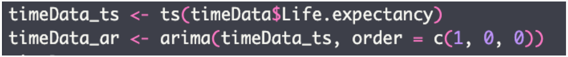


Figure 6: Scatterplot of lagged life expectancy and life expectancy. There appears to be evidence for a linear relationship.

AR(1) Model



```r
timeData_ts <- ts(timeData$Life.expectancy)
timeData_ar <- arima(timeData_ts, order = c(1, 0, 0))
```

1. Format the Life expectancy variable as a time series using ts()
2. Fit the AR(1) model using arima()

Linear Regression compared with AR(1) Model

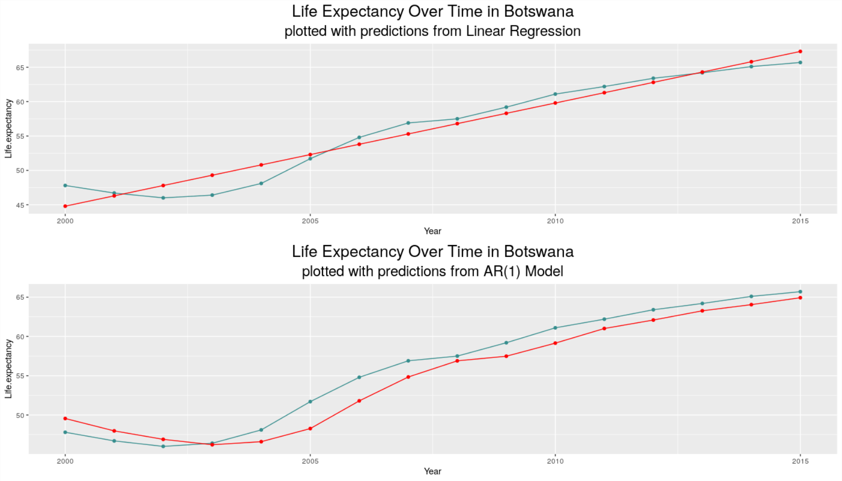


Figure 7: Plots of fitted vs observed values from a linear regression model and an AR(1) model.

Longitudinal Data

Data collected over time from the same subjects

- Sometimes called panel data because of its dimensions (time and subjects)
- Can think of this as combination of clustered data and time series data
- **Goal:** want to estimate population effects (like in linear regression) but also account for individual variations

Longitudinal Data Example

- Our WHO dataset contains Life expectancy for each country from 2000-2015
- Also includes covariates (health, economic, social factors)
- Question: Is alcohol consumption predictive of adult mortality? Should we assume that this varies by country?
- **Logical assumption:** Observations within a country are more likely to be correlated over time

Methods for Longitudinal Data

- Linear Mixed Models:
    - Accounts for individual variations in the dataset
    - Maintains interpretability for population and individual effects
- Generalized Estimating Equations:
    - alternative approach to LMMs
    - Integrates out the individual variations in the dataset
    - Robust to model misspecification

Linear Mixed Model

$$LifeExpectancy_{ij} = X\boldsymbol{\beta} + Z\mathbf{b_i} + \epsilon_{ij} \tag{1}$$

- $\boldsymbol{\beta}$ = coefficients for population effects
- b_i = coefficients for individual effects
- $\mathbf{b_i}$ measures how much each individual i deviates from the overall population trend
- $\mathbf{X}$: predictor variables for population effects
- $\mathbf{Z}$: subset of $\mathbf{X}$, covariates that could impact deviations from population trend

Linear Mixed Model: Visualization

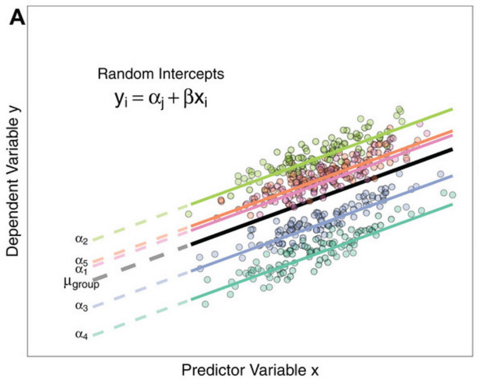


Figure 8: Linear Mixed Model with random intercept. Originally published in: https://peerj.com/articles/4794/

Linear Mixed Model: Visualization

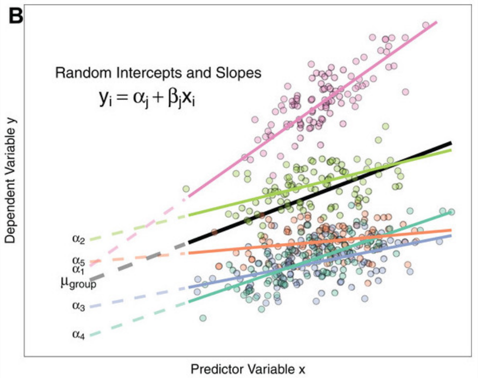
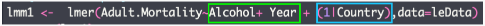


Figure 9: Linear Mixed Model with random intercept and slope. Originally published in: https://peerj.com/articles/4794/

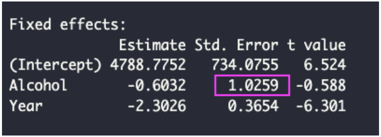
Linear Mixed Model

- We will fit an LMM using Adult mortality as the outcome variable and Alcohol consumption and Year as the covariates
- Give each country a random intercept b_{0i}
- This means that country is allowed to have a different **starting** value than the population average



```
lmm1 <-  lmer(Adult.Mortality~Alcohol+ Year + (1|Country),data=leData)
```

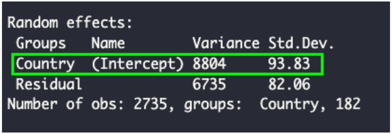
Coding Example: Linear Mixed Model



```
Fixed effects:
             Estimate Std. Error t value
(Intercept) 4788.7752   734.0755   6.524
Alcohol       -0.6032     1.0259  -0.588
Year          -2.3026     0.3654  -6.301
```

1. Alcohol consumption does not appear to be predictive of mortality
2. Time (Year) is significantly associated with mortality

Examining the Random Effects Variance



```
Random effects:
 Groups   Name        Variance Std.Dev.
 Country  (Intercept) 8804     93.83
 Residual             6735     82.06
Number of obs: 2735, groups:  Country, 182
```

1. If $var(b_{0i})$ were close to 0, we might be able to get away with linear regression
2. Estimated variance of b_{0i} is high, indicating that there is a lot of within-country variability
3. We could try to fit this model with a random slope as well (can try yourself in the next code block!)

Plotting the Best Linear Unbiased Predictions

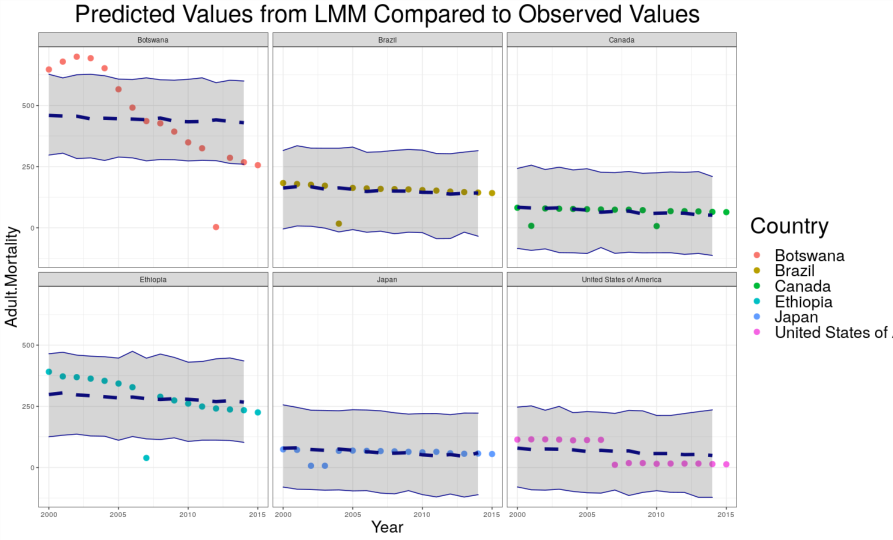


Figure 10: Predicted values for each country, compared to observed mortality

What happens if we ignore correlation?

Correlated Data Models

What happens if we ignore correlation in datasets?

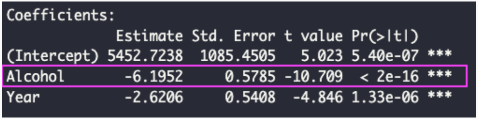
Correlated Data Models

What happens if we ignore correlation in datasets?

- Linear Regression assumes all observations are **independent**
- LMM (usually) assumes that between-individual observations are independent, but correlation can exist within-individuals

Coding Example: Linear Regression vs Linear Mixed Model

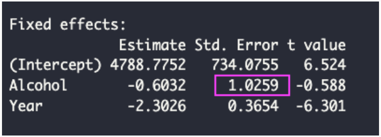
Coding Example: Linear Regression



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5452.7238  1085.4505   5.023 5.40e-07 ***
Alcohol       -6.1952     0.5785 -10.709  < 2e-16 ***
Year          -2.6206     0.5408  -4.846 1.33e-06 ***
```

1. Alcohol is significantly associated with Mortality
2. Time (Year) is also significantly associated with Mortality

Coding Example: Linear Mixed Model, accounting for within-country correlation



```
Fixed effects:
             Estimate Std. Error t value
(Intercept) 4788.7752   734.0755   6.524
Alcohol       -0.6032     1.0259  -0.588
Year          -2.3026     0.3654  -6.301
```

1. Estimate of Alcohol consumption coefficient is a lot lower
2. Standard error estimate for alcohol is almost double
3. Effect of alcohol is no longer significant

Coding Example: Linear Mixed Model, accounting for within-country correlation

- When we account for within-country correlations, the variability between observations decreases
- Overall “sample size” also decreases since observations are grouped by individuals
- Standard errors are not **artificially deflated** as a result
- This means chance of Type 1 error (false positive) is higher if you fit a linear regression

Conclusion

Recap

- Many potential sources of correlation in data
- Three types of correlated data
- Ignoring correlation can lead to biased statistical results

Questions?

License

Get the source of this theme and the demo presentation from

`github.com/matze/mtheme`

The theme *itself* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.