

# Regression

Machine Learning for Behavioral Data

March 7, 2022

# Today's Topic

Week	Lecture/Lab
1	Introduction
2	Data Exploration
<b>3</b>	<b>Regression</b>
4	Classification
5	Model Evaluation
6	Time Series Prediction
7	Time Series Prediction
8	Time Series Prediction

Complete pipeline for one use case:

- Data exploration
- Prediction
- Model evaluation

# Getting ready for today's lecture...

- **If not done yet:** clone the repository containing the Jupyter notebook and data for today's lecture into your Noto workspace.
- SpeakUp room for today's lecture:

<https://go.epfl.ch/mlbd-lecture>

---

# Short quiz about the past...

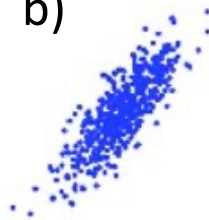
- Which of the four graphs have the following properties:

**High Pearson's Correlation, High Mutual Information**

a)



b)

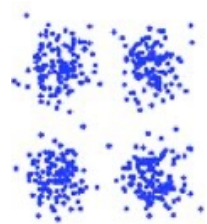


c) None

d)



e)



SpeakUp Poll!

0  
0 votes

14/02/2022 21:26, by me

0 comments

# Short quiz about the past...

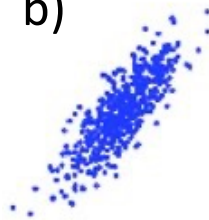
- Which of the four graphs have the following properties:

**High Pearson's Correlation, Low Mutual Information**

a)



b)

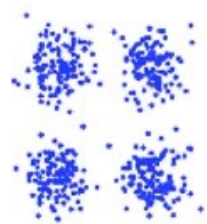


c) None

d)



e)



SpeakUp Poll!

0  
0 votes

14/02/2022 21:26, by me

0 comments

# Short quiz about the past...

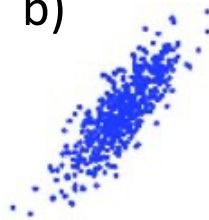
- Which of the four graphs have the following properties:

**Low Pearson's Correlation, Low Mutual Information**

a)



b)

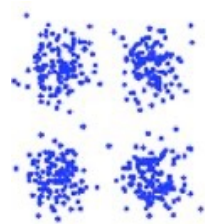


c) None

d)



e)



SpeakUp Poll!

0  
0 votes

14/02/2022 21:26, by me

0 comments

# Today's Use Case: Flipped Classroom Course

- Participants: 288 EPFL students of a course taught in *flipped classroom* mode with a duration of 10 weeks
  - Structure:
    - Preparation: watch videos (and solve simple quizzes) on **new content** at home as a preparation for the lecture
    - Lecture: discuss open questions and solve more complex tasks
    - Lab session: solve paper-an-pen assignments
  - Data: clickstream data (all interactions of the student with the system)
-

# Agenda

- **Linear Regresssion**
  - Generalized Linear Models
  - Mixed-Effect Models
  - Performance Metrics
  - Regression for Time-Series
-



# Idea

- In regression, a single aspect of the data (output variable) is modeled by some combination of other aspects of the data (input variables)
-

# More formal

- In regression, a single aspect of the data (output variable) is modeled by some combination of other aspects of the data (input variables)
  - Given:  $N$  data points  $(y_n, \mathbf{x}_n)$ , where  $y_n$  is the  $n$ 'th output variable and  $\mathbf{x}_n$  is a  $D$ -dimensional vector of input variables
  - Goal:  $y_n \approx f(\mathbf{x}_n)$
-

# Usage

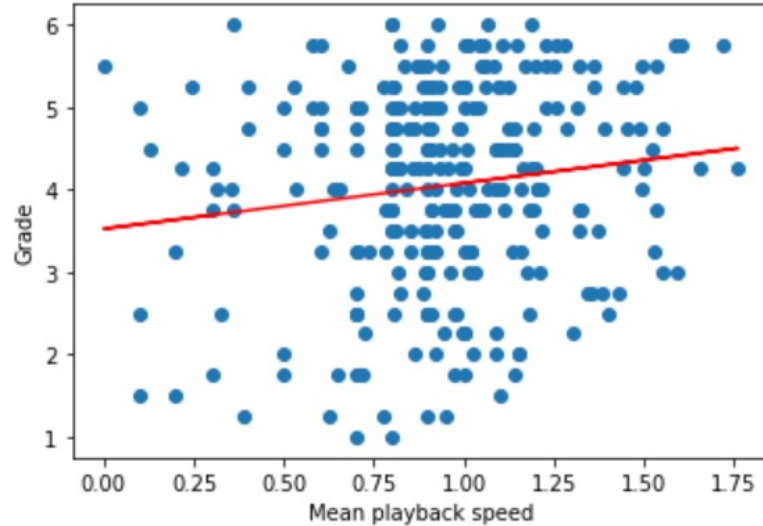
- *Prediction*: predict the output for a new (unseen) input vector  $x$
  - *Interpretation*: analyze the relationships between the variables (what effect the input variables have on the output variable)
-

# Example | Mean playback speed

**x-axis:** Mean playback speed of videos

**y-axis:** Course grade

Each point is one student



*Students who watch the videos faster tend to have better grades.*

# Linear Regression

The output variable  $y_n$  with  $n = 1, \dots, N$  is modeled by a **linear** combination of the input variables  $x_{n,d}$  with  $d = 1, \dots, D$ .

$$y_n = \beta_0 + \beta_1 x_{n,1} + \dots + \beta_D x_{n,D} + \epsilon_n$$

where  $\epsilon_n$  are error terms that should be as small as possible and  $\epsilon_n \sim N(0, \sigma^2)$ .

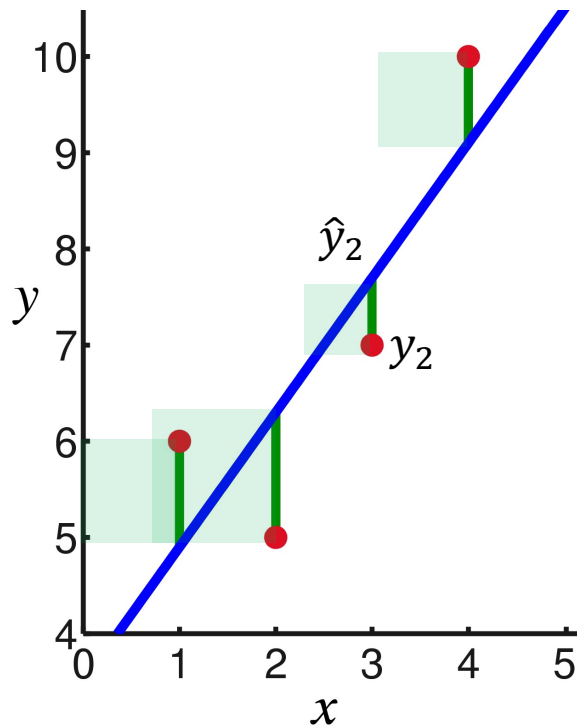
---

# Goal: find optimal parameters

Find parameters  $\hat{\beta}$  that minimize

$$\sum_{n=1}^N (y_n - \tilde{\mathbf{x}}_n^T \cdot \hat{\boldsymbol{\beta}})^2$$

$$\text{with } \tilde{\mathbf{x}}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ \dots \\ x_{n,D} \end{bmatrix} \text{ and } \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_D \end{bmatrix}$$



# Fitting the parameters

$$grade = \beta_0 + \beta_1 \cdot time\_in\_problem + \beta_2 \cdot percentage\_correct$$

Formula: `grade~time_in_problem+percentage_correct`

Family: gaussian                  Estimator: OLS

Std-errors: non-robust    CIs: standard 95%                  Inference: parametric

Number of observations: 288                  R<sup>2</sup>: 0.123                  R<sup>2</sup>\_adj: 0.117

Log-likelihood: -447.384                  AIC: 900.769                  BIC: 911.758

Fixed effects:

	Estimate	2.5_ci	97.5_ci	SE	DF	T-stat	P-val	Sig
Intercept	3.205159	2.881827	3.528491	0.164268	285	19.511809	0.000000	***
time_in_problem	0.000015	0.000009	0.000021	0.000003	285	4.681046	0.000004	***
percentage_correct	0.892133	0.293503	1.490764	0.304132	285	2.933372	0.003625	**

---

# Fitting the parameters

$$grade = \beta_0 + \beta_1 \cdot time\_in\_problem + \beta_2 \cdot percentage\_correct$$

**R-squared:**

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

**R-squared** adjusted  
for number of  
parameters

Std-errors: non-restricted OLS      CIs: standard 95%      Inference: parametric

Number of observations: 288

R<sup>2</sup>: 0.123

R<sup>2</sup> adj: 0.117

Log-likelihood: -447.384

Fixed effects:

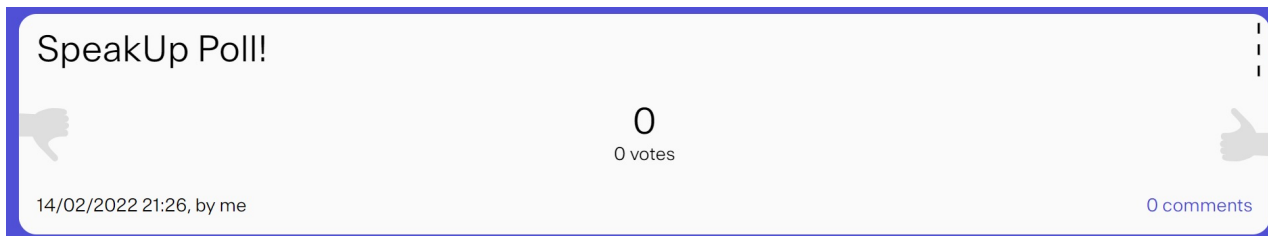
**p-value:** probability of  
getting this estimate if true  
coefficient value was 0

	Estimate	Std. Error	95% CI Lower	95% CI Upper	z	P-val	Sig		
Intercept	3.205159	2.000000	-0.794841	6.999999	1.602539	0.109809	***		
time_in_problem	0.000015	0.000009	-0.000021	0.000051	0.000003	285	4.681046	0.000004	***
percentage_correct	0.892133	0.293503	1.490764	0.304132	285	2.933372	0.003625	**	



# Influence of input variables

$$grade = 3.4 + 0.000016 \cdot time\_in\_problem + 0.72 \cdot percentage\_correct$$



Which of the input variables has the largest impact on *grade*?

- a) *time\_in\_problem*
  - b) *percentage\_correct*
  - c) I don't know
-

# Different units of measurements

$$\text{grade} = 3.4 + 0.000016 \cdot \text{time\_in\_problem} + 0.72 \cdot \text{percentage\_correct}$$

increase in  
*time\_in\_problem*  
by 1s -> increase of  
grade by 0.000015

increase in  
*percentage\_correct* by 1  
percentage point -> increase of  
grade by 0.89

# Transformation: Z-Scores

$$\tilde{x}_{n,d} = \frac{x_{n,d} - \bar{x}_d}{\sigma(\mathbf{x}_d)}$$

$$d = 1, \dots, D$$

$$n = 1, \dots, N$$

- ➡ Standardization via z-score:  $\tilde{x}_{n,d}$  denotes the distance between the raw feature  $x_{n,d}$  and the sample mean  $\bar{x}_d$  (in units of the standard deviation)
-

# Transformation: Example

$$grade = 4.05 + 0.35 \cdot time\_in\_problem + 0.15 \cdot percentage\_correct$$

Example in Jupyter Notebook

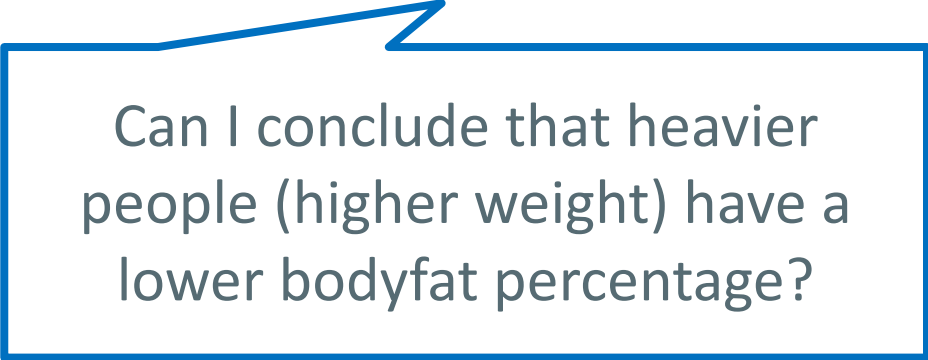
---

# Transformation: Summary

- Lets us compare the impact of input variables with different scales/units of measurements (e.g., time in problem in *seconds* and percentage correct)
  - Reduces interpretability of individual input variables
-

# Interpretation: Caveat

$$\text{bodyfat} = -45.95 + 0,99 \cdot \text{abdomen} - 0,33 \cdot \text{weight}$$



Can I conclude that heavier people (higher weight) have a lower bodyfat percentage?

# Interpretation: Caveat

$$\text{bodyfat} = -45.95 + 0,99 \cdot \text{abdomen} - 0,33 \cdot \text{weight}$$

Can I conclude that heavier people (higher weight) have a lower bodyfat percentage?

- a) Yes
- b) No
- c) I don't know

SpeakUp Poll!



0  
0 votes

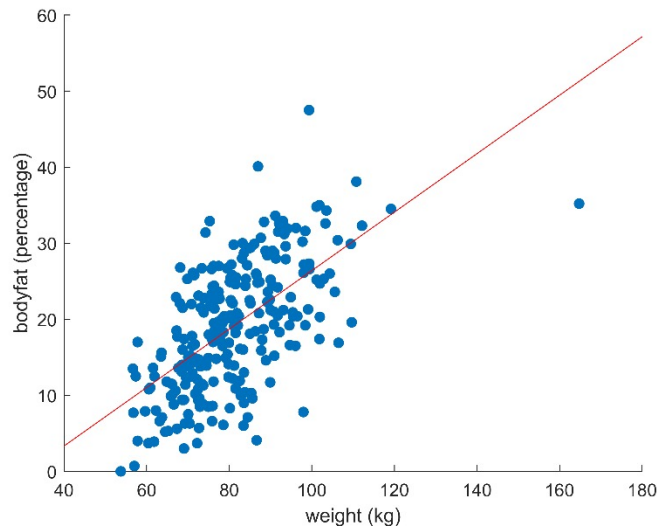


14/02/2022 21:26, by me

0 comments

# Interpretation: Caveat

- There is a positive correlation between *weight* and *bodyfat* ( $r = 0.61, p < .001$ ).





# Interpretation: Caveat

- There is a positive correlation between *weight* and *bodyfat* ( $r = 0.61, p < .001$ ).
    - *weight* only has a negative coefficient  $\beta$  in the context of *abdomen*, i.e. for fixed *abdomen* predictor
    - a predictor can only be interpreted **in the context** of the other predictors in the model
-

# What means linear?

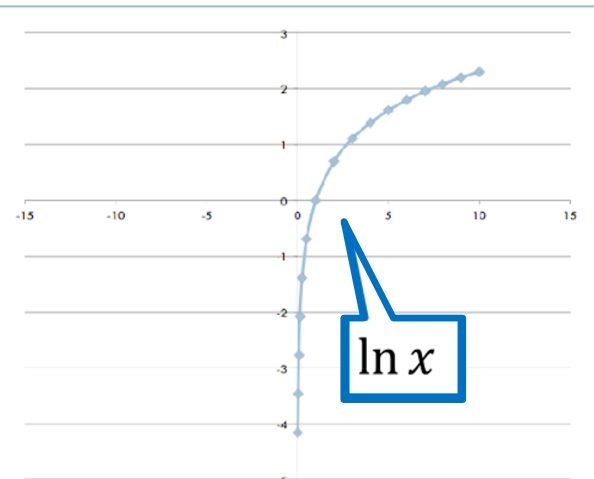
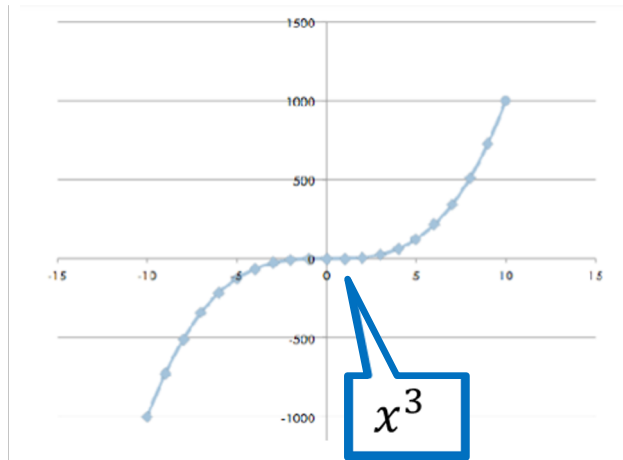
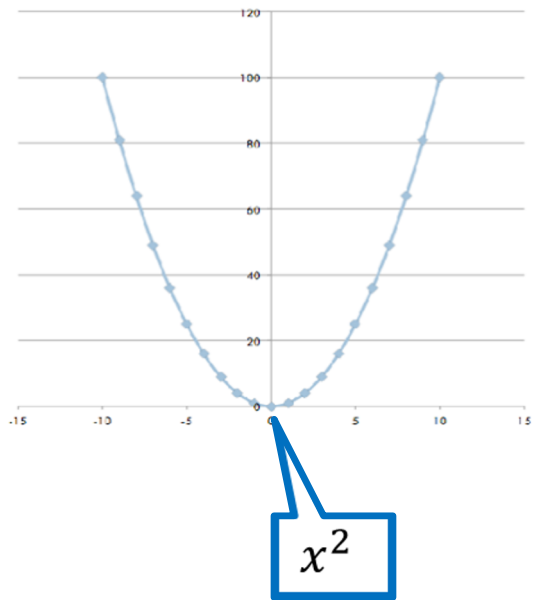
$$y_n = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_D x_{n,D} + \epsilon_n$$

*Linear* in the **parameters** -> we can apply arbitrary functions to the raw input variables, e.g.,

- logarithms, exponentials
  - polynomials
  - inverse
-

# What means linear

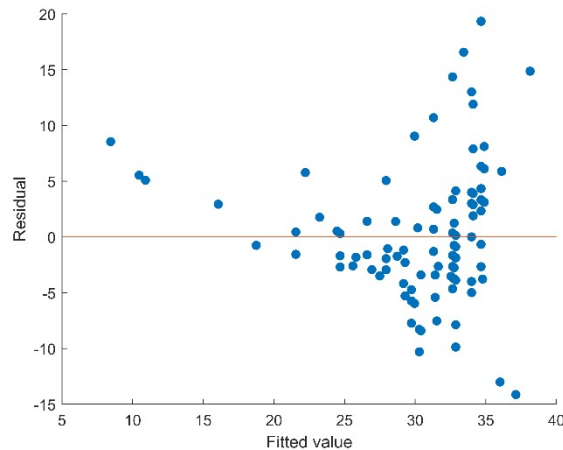
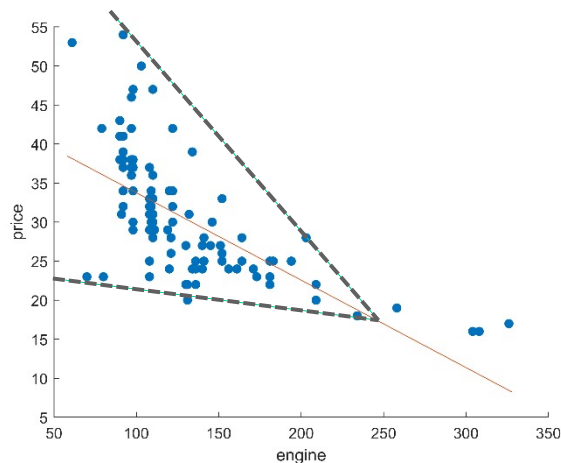
Different transformations



# Restrictions of linear models

For some cases, linear regression models are not appropriate:

- the variance of  $y$  depends on the mean



Assumption for statistics (t-test, chi, etc.):  
 $\epsilon \sim N(0, \sigma^2)$

# Restrictions of linear models

For some cases, linear regression models are not appropriate:

- the variance of  $\mathbf{y}$  depends on the mean
- the range of  $\mathbf{y}$  is restricted

$$\#bicycles = -2291 + 83 \cdot \text{maxTemp} - 13 \cdot \text{minTemp} - 890 \cdot \text{precipitation}$$

➡ prediction  $\hat{y}$  can be negative...

---

# Agenda

- Linear Regression
  - **Generalized Linear Models**
  - Mixed-Effect Models
  - Performance Metrics
  - Regression for Time-Series
-

# Generalized Linear Models

A generalized linear model is composed of a **linear predictor**

$$\pi_n = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_D x_{n,D}$$

and a **link function**

$$g(\mu_n) = \pi_n$$

with  $\mu_n = \mathbb{E}[Y|X = \mathbf{x}_n]$

---

# Generalized Linear Models

Conditional expectation: the mean  $\mu_n$  depends on the values of independent variables  $\mathbf{x}_n$

is composed of a **linear predictor**

$\mathbf{x}_{n,1}$

Each  $y_n$  represents the realization of the random variable  $Y$ , which is distributed according to a specific probability distribution

and a

**function**

$$g(\mu_n) = \pi_n$$

with  $\mu_n = \mathbb{E}[Y|X = \mathbf{x}_n]$



# Generalized Linear Models

A generalized linear model is composed of a **linear predictor**

$$\pi_n = \beta_0 + \beta_1 x_{n,1} + \dots + \beta_p x_{n,p}$$

and a **link function**

In practice (for parameter fitting): observed values  $y_n$  are assumed to represent  $\mu_n$

$$g(\mu_n) = \pi_n$$

with  $\mu_n = E[Y|X = \mathbf{x}_n]$

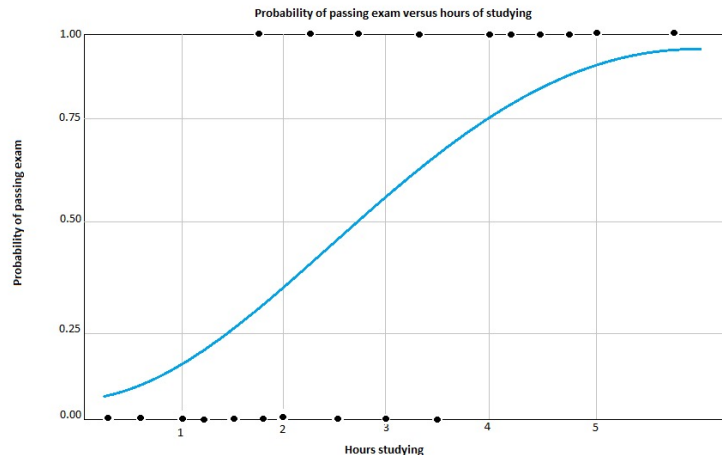
# Logistic Regression

In logistic regression, the link function is

$$g(\mu_n) = \log\left(\frac{\mu_n}{1 - \mu_n}\right)$$

and therefore (for fitting)

$$\log\left(\frac{y_n}{1 - y_n}\right) = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_D x_{n,D}$$



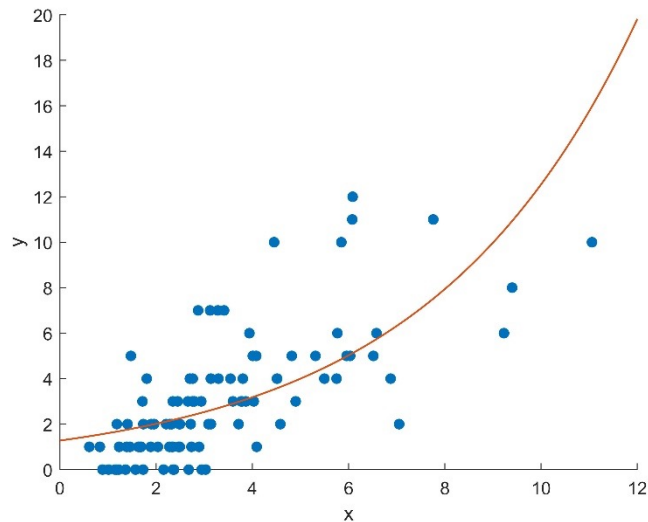
# Poisson Regression

In Poisson regression, the link function is

$$g(\mu_n) = \log(\mu_n)$$

and therefore (for fitting)

$$\log(\mu_n) = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_D x_{n,D}$$



# Linear Regression as a special case

For the linear regression, the link function is

$$g(\mu_n) = \mu_n$$

and therefore (for fitting)

$$y_n = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_D x_{n,D}$$

---

# Example

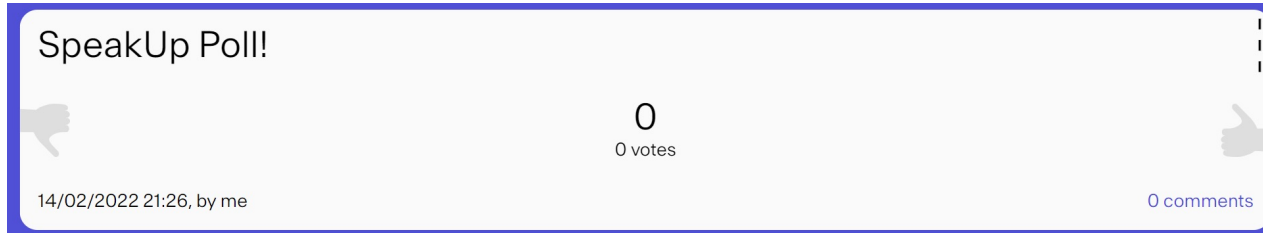
What type of model would you use for the following tasks?

1. Predict the **number of awards** earned by students at one high school. Predictors include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math.

(a) Generalized Linear Model

(b) Logistic Regression

(c) Poisson Regression



# Example

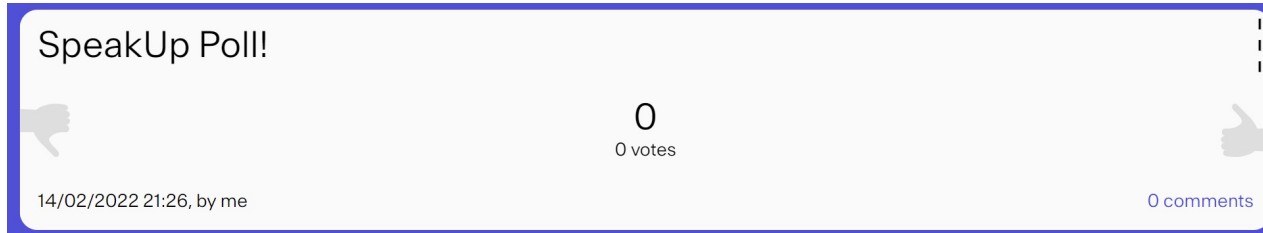
What type of model would you use for the following tasks?

2. Predict whether a student will **solve a task correctly**. Predictors include the difficulty of the task and the number of tasks the student has already solved.

(a) Generalized Linear Model

(b) Logistic Regression

(c) Poisson Regression



# Example

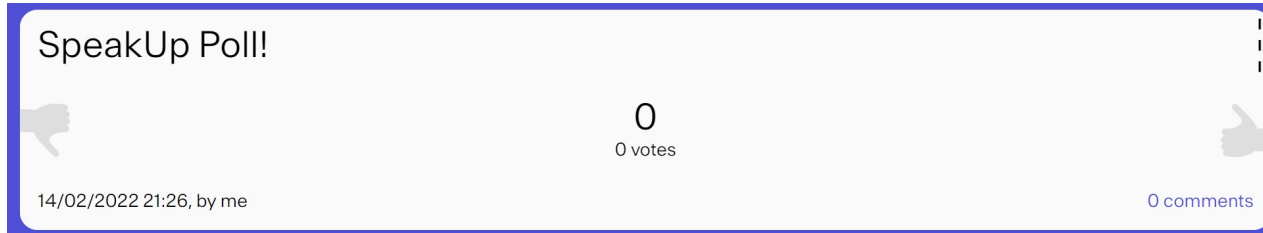
What type of model would you use for the following tasks?

3. Predict the **sales (in \$)** of a company based on their advertising budget on Youtube.

(a) Generalized Linear Model

(b) Logistic Regression

(c) Poisson Regression



# Agenda

- Linear Regression
  - Generalized Linear Models
  - **Mixed-Effect Models**
  - Performance Metrics
  - Regression for Time-Series
-

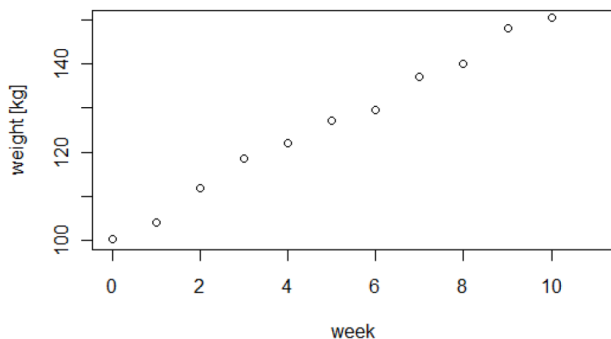


# Why mixed-effect models?

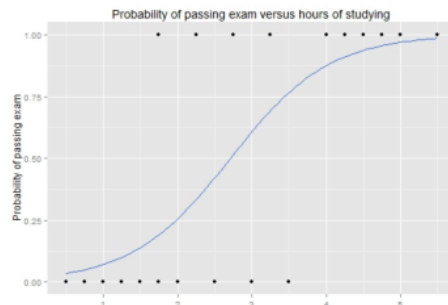
- Useful when we are dealing with **correlated** samples
    - **Grouping** of subjects (e.g., students within a classroom)
    - **Repeated measurements** on each subject over time (e.g., student in flipped classroom course over 10 weeks)
-

# Generalized Linear Models

- Example 1: strength gain by weight training
- Example 2: probability of passing exam of a course  $c$  depending on the hours studied

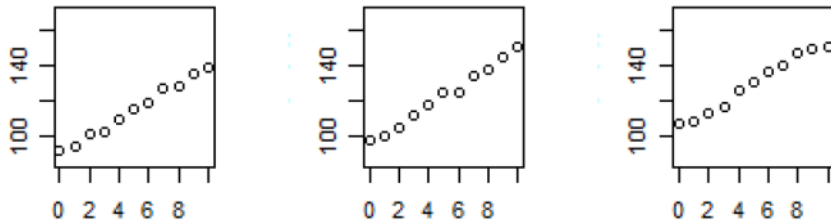


$$y_n = \beta_0 + \beta_1 x_{n,1}$$



# Generalized Linear Mixed Effects Model

- Example 1: strength gain by weight training
  - Each person has individual starting strength



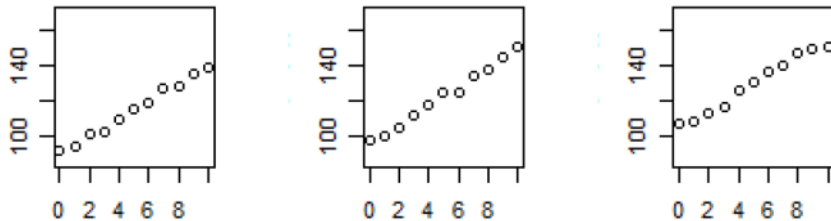
$$y_n = \beta_0 + u_n + \beta_1 x_{n,1} \quad u_n \sim N(0, \sigma_u^2)$$

“Fixed” Effects

“Random” Effect

# Generalized Linear Mixed Effects Model

- Example 1: strength gain by weight training
  - Each person has individual starting strength



$$y_n = \beta_0 + u_n + \beta_1 x_{n,1} \quad u_n \sim N(0, \sigma_u^2)$$

“Fixed” Effects

+

“Random” Effect

=

“Mixed” Effects

Fitting the parameters:

- Fixed effects only: linear least squares
- Mixed effects: maximum likelihood estimation

# Generalized Linear Mixed Effects Model

- In our case, students come from different origins and we assume that students from the same origin are more similar (same education system)
- We therefore use origin (*category*) as a proxy for prior knowledge and add a random intercept to the model

$$passed \sim 1 | category + percentage\_correct$$

---

# Agenda

- Linear Regression
  - Generalized Linear Models
  - Mixed-Effect Models
  - **Performance Metrics**
  - Regression for Time-Series
-

# Usage

- *Interpretation*: analyze the relationships between the variables (what effect the input variables have on the output variable)
  - *Prediction*: predict the output for a new (unseen) input vector  $x$
-

# Regression: $R^2$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where  $SS_{res} = \sum_i (y_i - f(x_i))^2$  and  $SS_{tot} = \sum_i (y_i - \bar{y})^2$

- Can be interpreted as the fraction of explained variability of the data
  - Often used when the goal is *interpretation*
  - Often used in the fields of Psychometrics, Learning Sciences, Psychology, etc.
-



# Regression: MAE and RMSE

- Mean absolute error:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - f(x_i)|$$

- Root mean squared error:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2}$$

- Often used when the goal is *prediction*
  - **RMSE is largely preferred to MAE**
-

# Hypothetical Example

- Given:
    - Student giving correct answers 70% of the time
    - Model A: predicts correct 70% of the time
    - Model B: predicts 100% correctness
-

# MAE: Model B is better

- 70% of the time the student gives a correct answer (response = 1)
    - Model A: absolute error = 0.3
    - Model B: absolute error = 0.0
  - 30% of the time the student answers wrong (response = 0)
    - Model A: absolute error = 0.7
    - Model B: absolute error = 1.0
  - $MAE_A = 0.42$ ,  $MAE_B = 0.30$
-

# RMSE: Model A is better

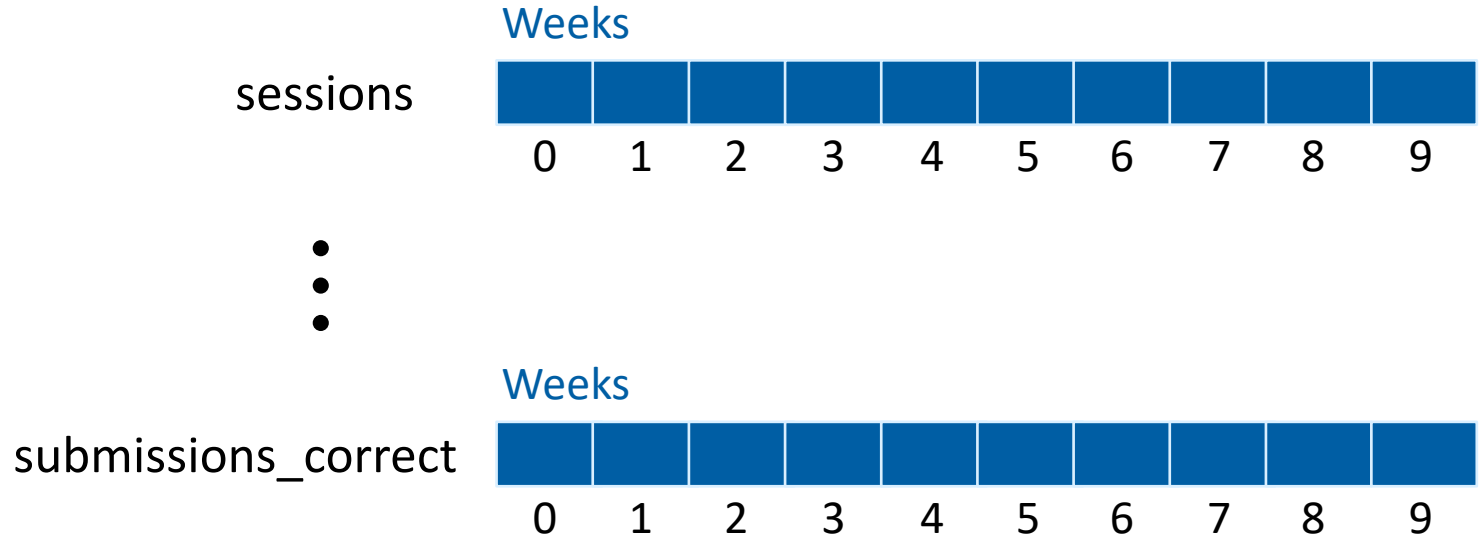
- $RMSE_A = 0.21$
  - $RMSE_B = 0.30$
  - **$RMSE$  penalizes large errors heavier**
-

# Agenda

- Linear Regresssion
  - Generalized Linear Models
  - Mixed-Effect Models
  - Performance Metrics
  - **Regression for Time-Series**
-

# Time Series – Our flipped classroom case

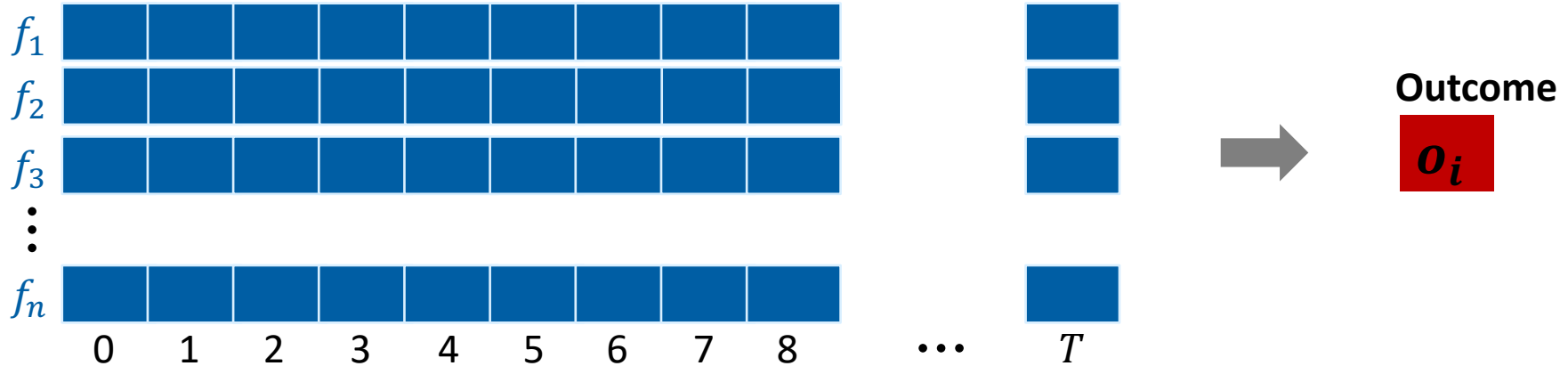
Student  $i$



# Time Series – Possible Tasks

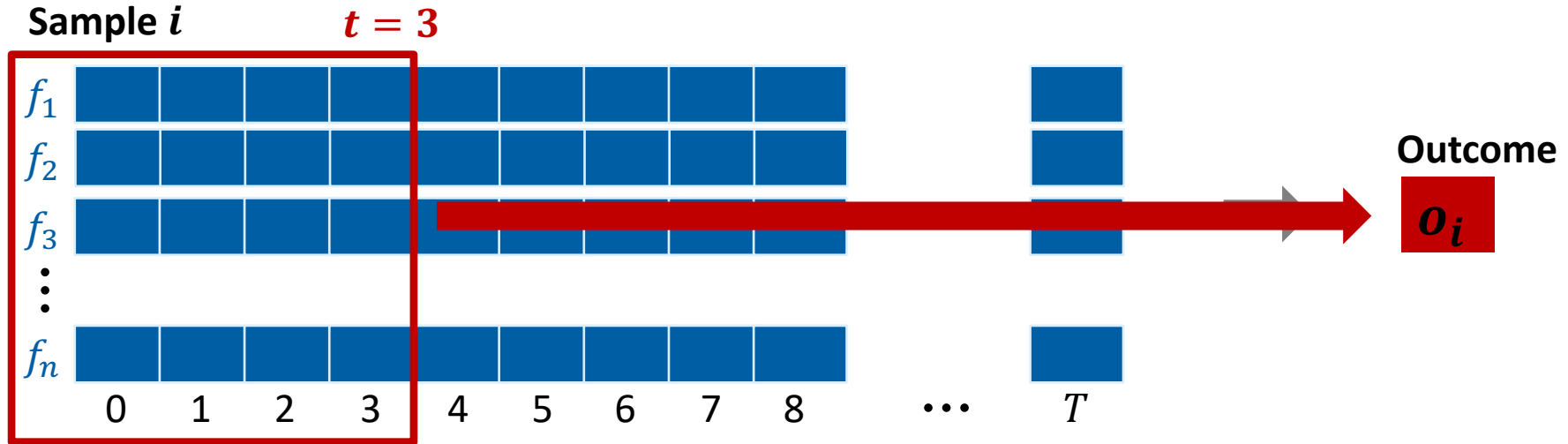
- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps

Sample  $i$



# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps





# Handling Time Series Data

- Flattening
    - The number of parameters of the model depends on the number of time steps of the model
  - Aggregation
    - Averaging across weeks
    - Accumulating across weeks
-

# Example – Prediction of Pass/Fail

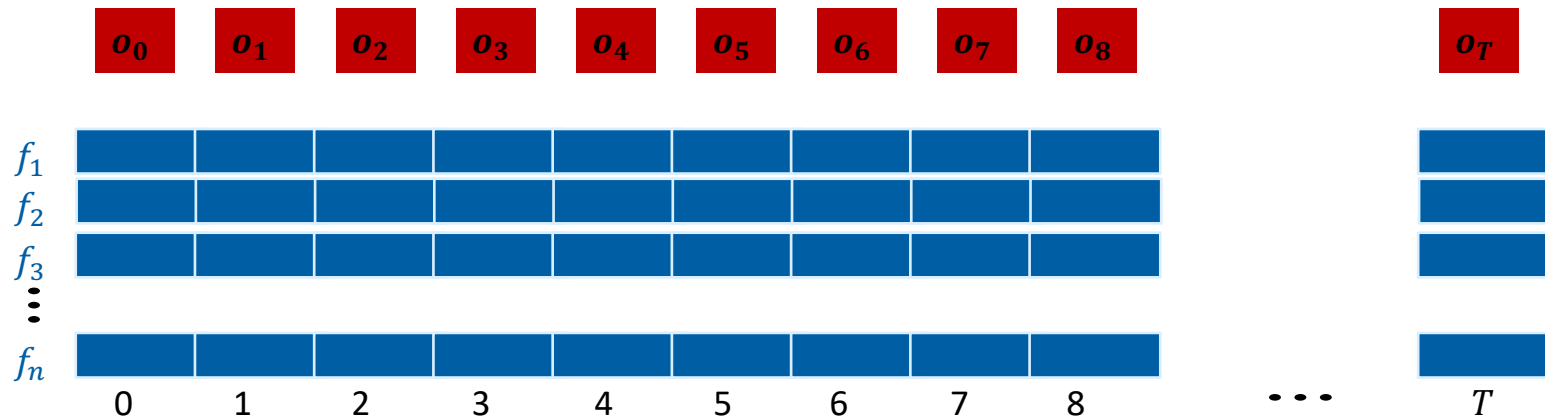
- Prediction of grade after  $t < T$  weeks
- We will try to predict after 5 weeks and after 10 weeks

$passed \sim (1|category) + average\_percentage\_correct [week\ n]$

---

# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .

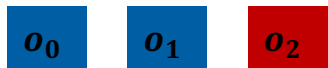
$o_0$

$o_1$

---

# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



# Time Series – Possible Tasks

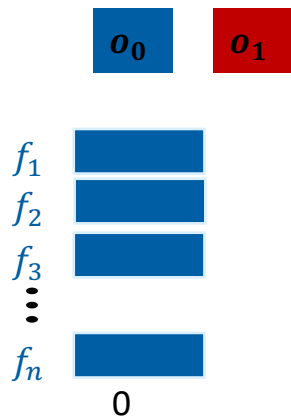
- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .





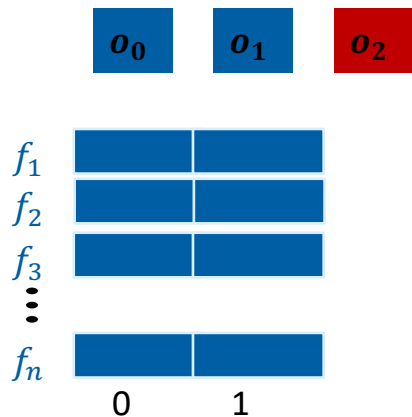
# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



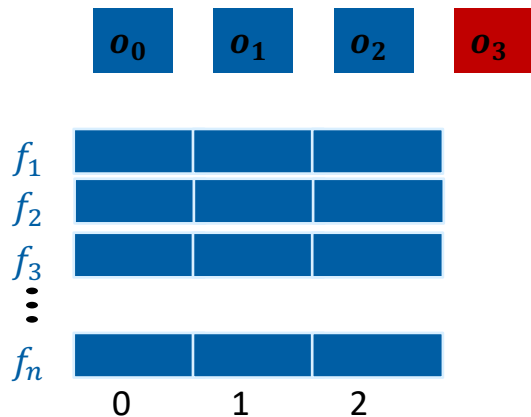
# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



# Time Series – Possible Tasks

- Prediction of a target variable after  $t < T$  time steps, where  $T$  is the total number of time steps
- Prediction of a variable in time step  $t + 1$ , based on time steps  $0, \dots, t$ .



# Your turn: prediction of quiz performance

- Prediction of quiz performance in week  $t + 1$ , based on features calculated for weeks  $0, \dots, t$
  - We provide an example equation with one random and one fixed effect already in the model
  - We provide a train/test split
-

# Your Turn – Feedback

Do you want feedback or have questions?

Upload your Jupyter Notebook here:

<https://go.epfl.ch/mlbd-activities>

---

# Summary

- Linear regression is a useful framework for interpreting data and making predictions
  - Caveat: be careful when interpreting the models
  - Linear regression is flexible, i.e. arbitrary functions can be applied to the raw input data
  - Generalized linear models are a more general framework appropriate for response variables from exponential family distributions
  - Mixed models allow for capturing correlation in the data
  - Modeling time series data requires some type of aggregation
-