Introduction to Statistical Learning with Applications

CM1: Introduction, simple linear regression, and some multivariate statistics

Pedro L. C. Rodrigues

Structure of the course

Pedro L. C. Rodrigues
pedro.rodrigues@inria.fr

Isabella Costa Maia
isabella.costa-maia@grenoble-inp.fr

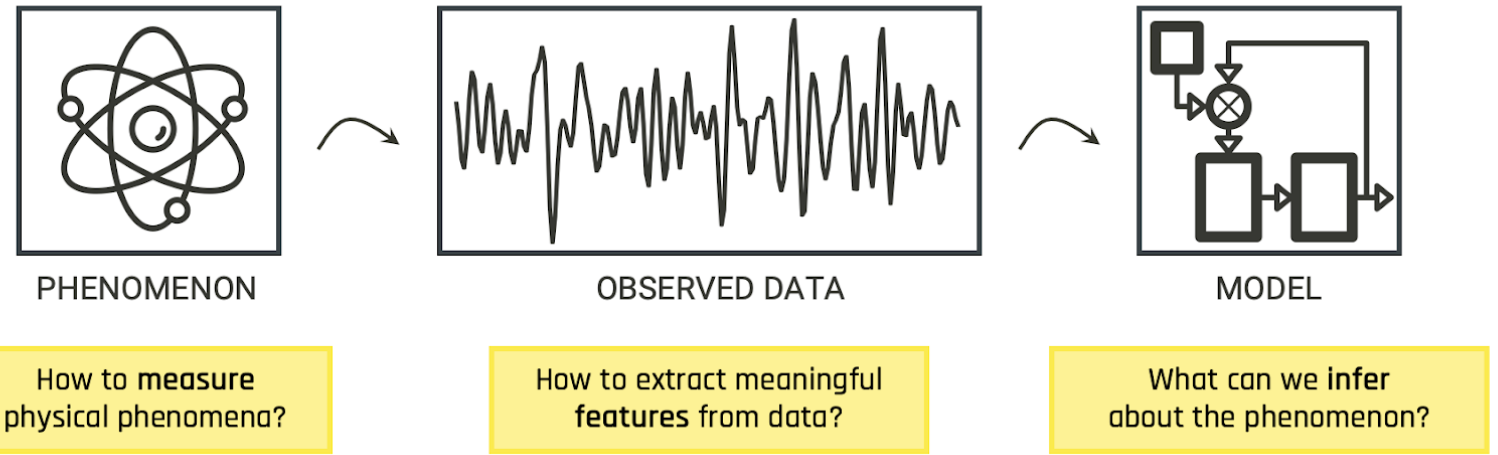
Complementary courses
for M1AM students

Razan Mhanna
razan.mhanna@inria.fr

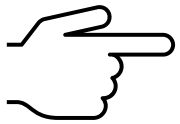
Pedro L. C. Rodrigues

- Researcher at Inria Grenoble since 2021
  (and teaching this class since 2018)

- Working mostly with **ML for experimental data**
  (neuro, climate, astro)



Keywords: (Deep) generative modelling, simulation-based inference, geometry-aware ML methods



- Administrative stuff
- Course overview
- Simple linear regression
- Multivariate statistics

Administrative stuff

Course evaluation:

- Your final grade will be **equally** split between:
  - (**Group**) Score on the TP: 50% report, 50% competition
  - (**Individual**) Score of the final exam

- The final exam is a **mix of theoretical and practical** questions and will be done in the TP rooms using Python

It will be a **3h exam similar to a TP session**
Be sure to use at least once during the semester the ENSIMAG computers

Administrative stuff

Textbooks:

- Our main reference is the book by James et al.

  "Introduction to statistical learning with applications to Python"

- Another excellent reference is Hastie et al.

  "Elements of Statistical Learning"

- Many of the examples in this course are taken from Cosma Shalizi's

  "The truth about linear regression"

The links to all these materials are available on the course website

Administrative stuff

About the **practical sessions** aka Travaux Pratique (TP):

- All sessions are to be done using Python -- 🙂 and 😕
- Teams of three students to be declared on Teide (if possible)
- Three TPs with reports to be sent on the dates informed in the website
- All reports are to be written in **English** and sent in `ipynb` format
- Suggestions:
  - Use as much as possible the computers from ENSIMAG
  - Avoid depending too much of fancy libraries and packages
  - Doing extra investigations and including references will be appreciated

Tour on the course website

Website: **github.com/ISLA-Grenoble/**

- Administrative stuff
- Course overview
- Simple linear regression
- Multivariate statistics

Course overview

We will be dealing with three main statistical tasks

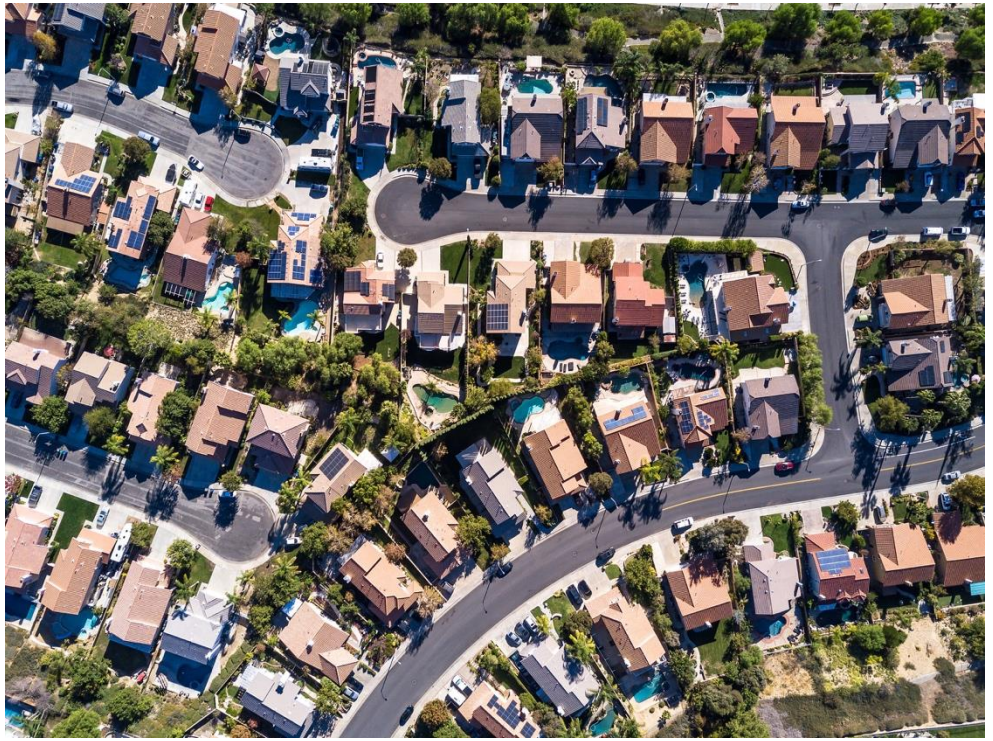
- **Regression:** how can we relate a continuous variable to another set of variables?
- **Classification:** how can we relate a discrete variable (i.e. a label) to another set of variables?
- **Clustering:** can we detect classes in a dataset and split its samples accordingly?

Regression

How can we relate a continuous variable to another set of variables?

Example

Can we predict the average price of houses in a district given the average income of people living there?



California housing dataset

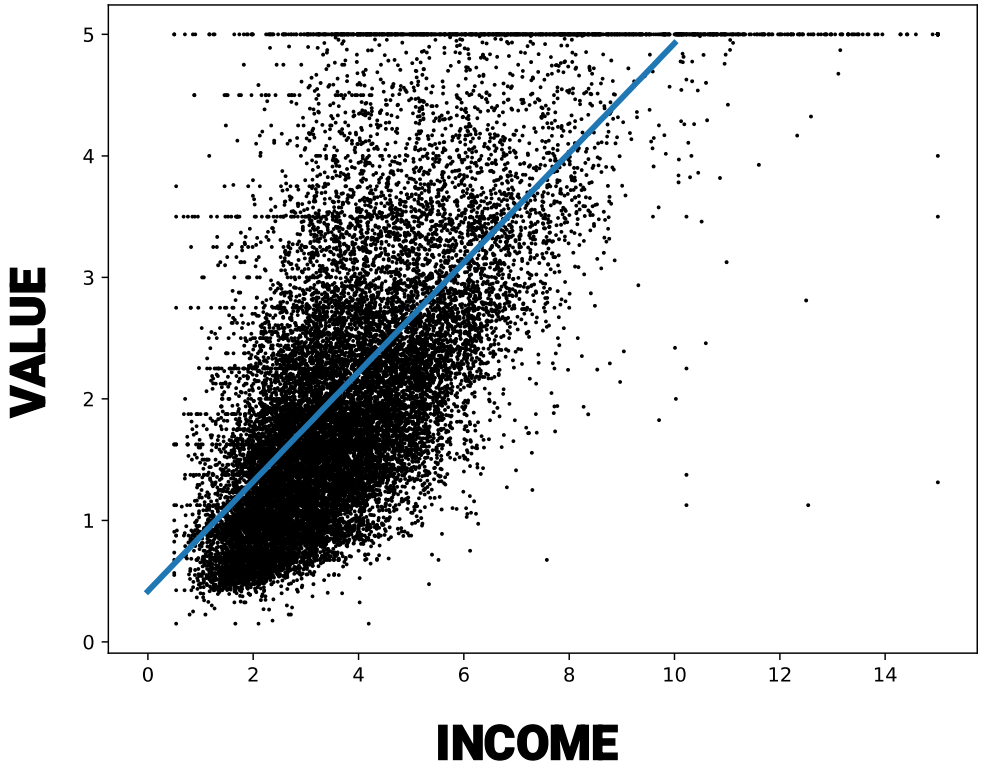
Regression

How can we relate a continuous variable to another set of variables?

Example

Can we predict the average price of houses in a district given the average income of people living there?

$$\text{VALUE} \approx \beta_1 \times \text{INCOME} + \beta_0$$



California housing dataset

Regression

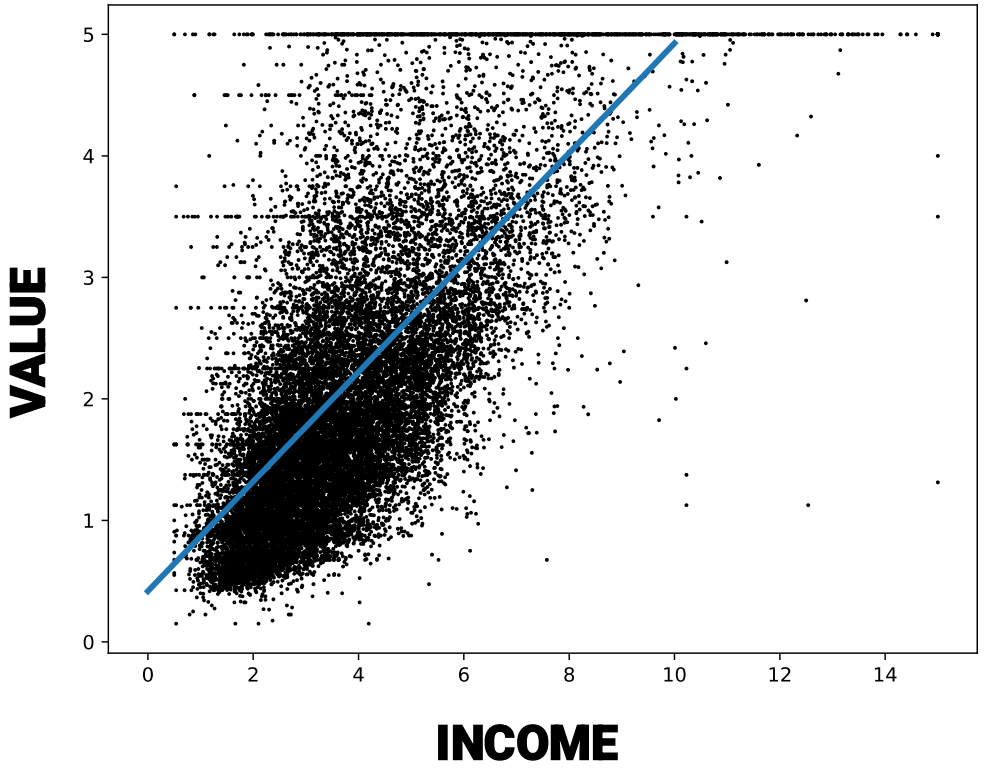
How can we relate a continuous variable to another set of variables?

Example

$$\text{VALUE} \approx \beta_1 \times \text{INCOME} + \beta_0$$

How significant are the parameters in the model? Can we define confidence intervals? Is the linear model a good one?

↳ Should we include more predictors?



California housing dataset

Out[69]:

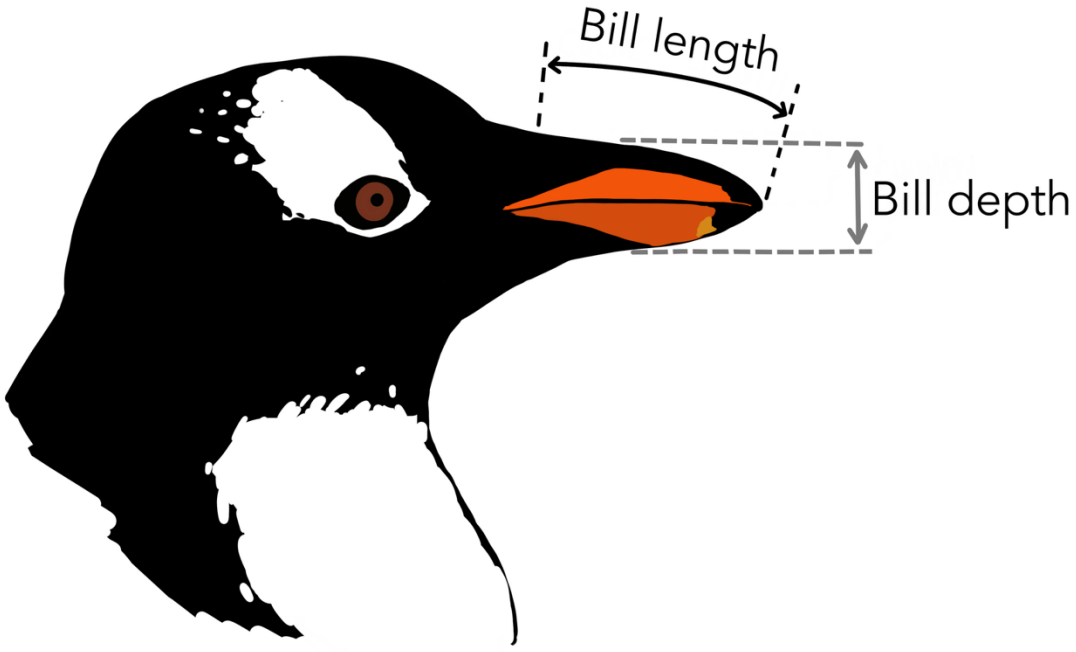
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:            MedHouseVal   R-squared:                       0.606
Model:                            OLS   Adj. R-squared:                  0.606
Method:                 Least Squares   F-statistic:                     3970.
Date:                Fri, 27 Dec 2024   Prob (F-statistic):               0.00
Time:                        10:49:44   Log-Likelihood:                -22624.
No. Observations:               20640   AIC:                         4.527e+04
Df Residuals:                   20631   BIC:                         4.534e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
MedInc         0.4367      0.004    104.054      0.000       0.428       0.445
HouseAge       0.0094      0.000     21.143      0.000       0.009       0.010
AveRooms      -0.1073      0.006    -18.235      0.000      -0.119      -0.096
AveBedrms      0.6451      0.028     22.928      0.000       0.590       0.700
Population -3.976e-06   4.75e-06     -0.837      0.402   -1.33e-05    5.33e-06
AveOccup      -0.0038      0.000     -7.769      0.000      -0.005      -0.003
Latitude      -0.4213      0.007    -58.541      0.000      -0.435      -0.407
Longitude     -0.4345      0.008    -57.682      0.000      -0.449      -0.420
intercept    -36.9419      0.659    -56.067      0.000     -38.233     -35.650
==============================================================================
Omnibus:                     4393.650   Durbin-Watson:                   0.885
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14087.596
Skew:                           1.082   Prob(JB):                         0.00
Kurtosis:                       6.420   Cond. No.                     2.38e+05
==============================================================================
```

Classification

How can we relate a discrete variable (i.e. a label) to another set of variables?

Example

Can we use the dimensions of a Penguin beak (aka bill) to classify him into one of three species?

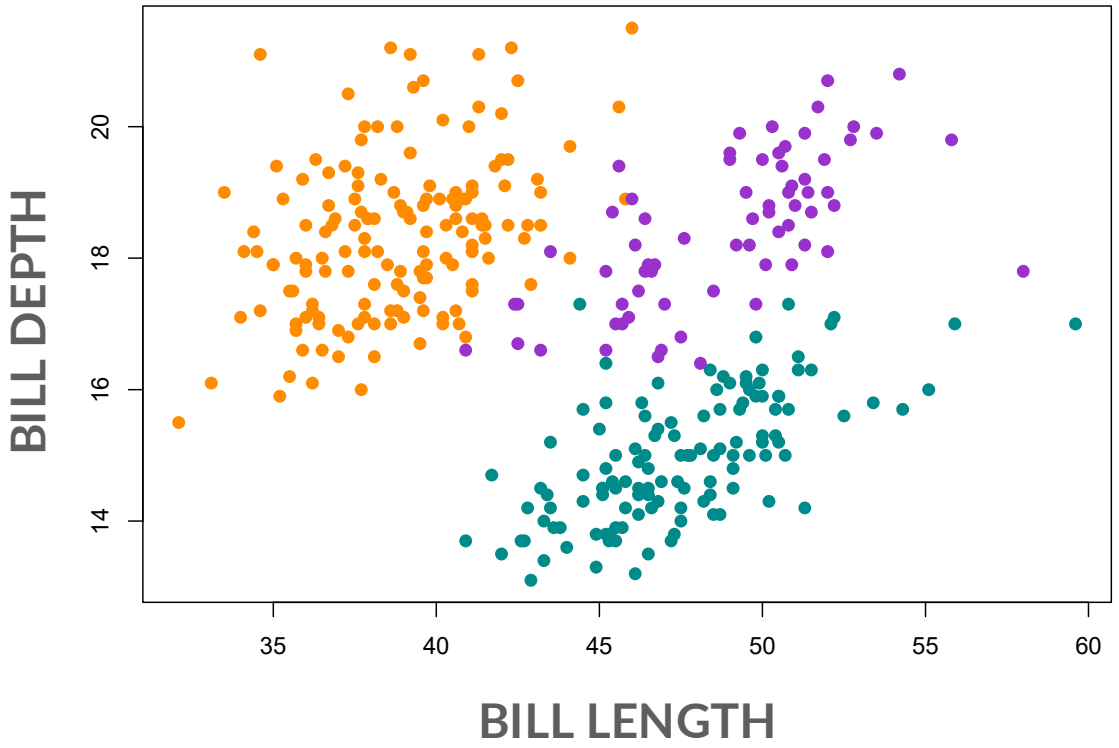


Classification

How can we relate a discrete variable (i.e. a label) to another set of variables?

Example

Can we use the dimensions of a Penguin beak (aka bill) to classify him into one of three species?



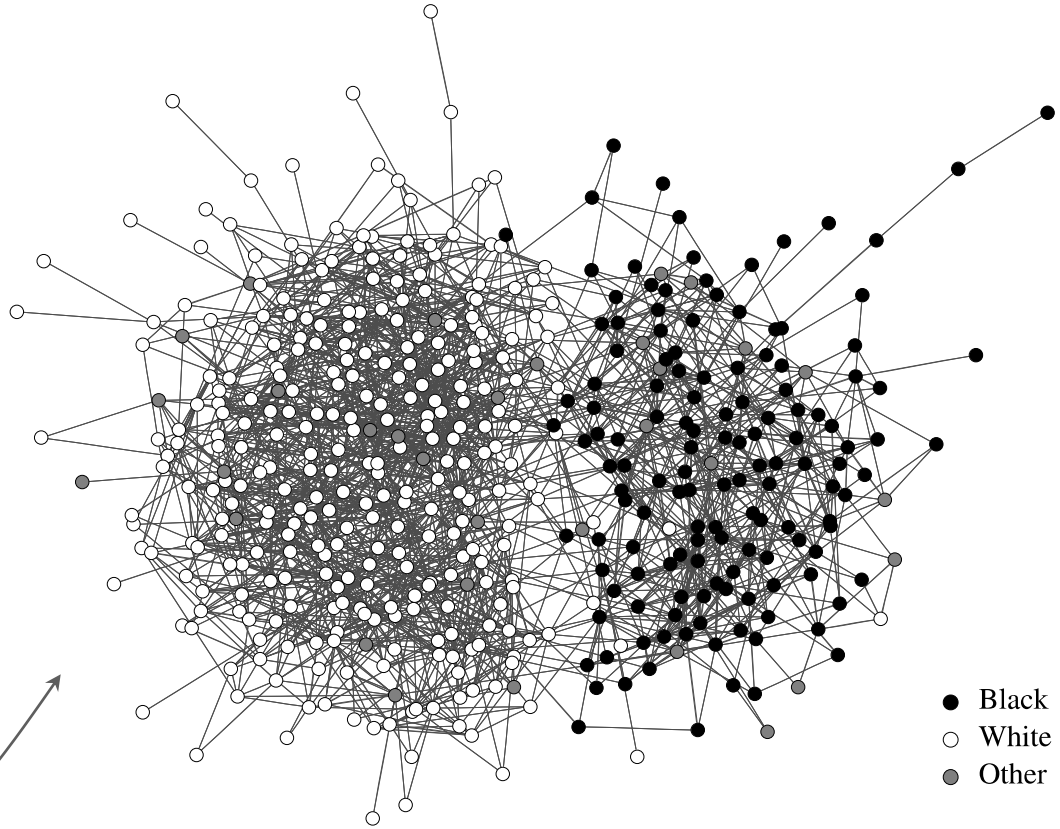
Clustering (and community detection)

Can we detect classes in a dataset and split its samples accordingly?

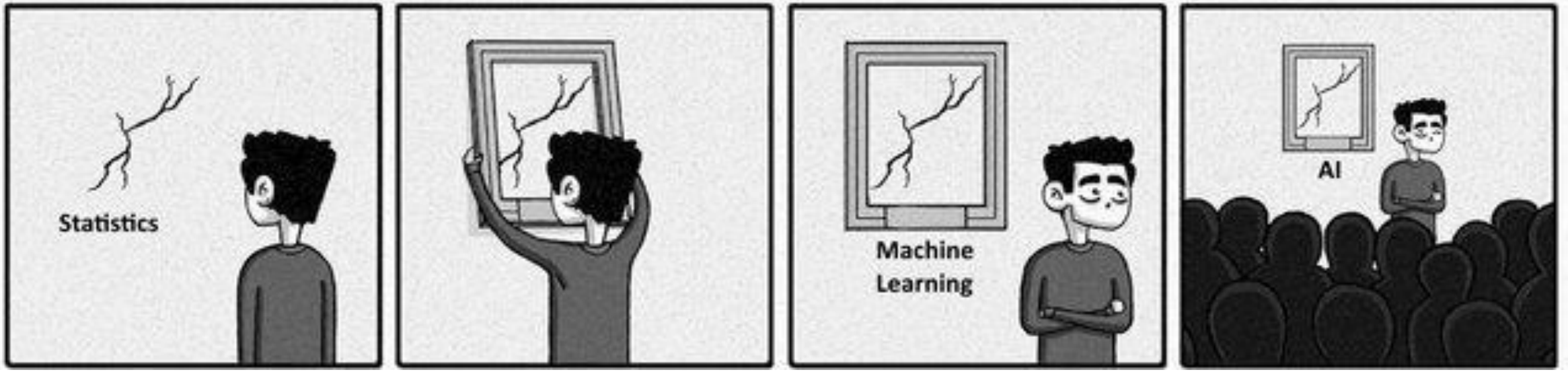
Example

Can we understand social relations using statistical analysis on graphs?

Nodes are students at a US high school and colors encode race. Using community detection in this situation helps understanding the social interactions in the school and evaluate which actions can be made for better integration.



Statistical Learning Vs Machine Learning



“When you’re fundraising, it’s AI.
When you’re hiring, it’s ML.
When you’re implementing, it’s logistic regression.”

Statistical Learning Vs Machine Learning

- We will be following more of a **statistician's** approach to machine learning
  - Under which **conditions** is the linear regression model reliable?
  - What sort of **errors** does the linear regression model can make ?
  - What can the linear regression model tell us when it works?
  - What are the signs that something has **gone wrong**?
- Flexibility (e.g. neural networks) Vs Interpretability (e.g. linear models)
- Understand **the inner workings** of building blocks for more complex models

Some tools we will be using all the time

and that you shold review on your own!

Probability & Statistics

- Expected value, variance, covariance
- Probability density function, CDF
- Conditional probability

Hypothesis tests

- Null hypothesis and p-value
- Confidence intervals
- t-test, F-test

Multivariate calculus

- Gradient and Hessian
- Multivariate integrals
- Change of variables, chain rule

Matrix analysis

- Eigenvalues and eigenvectors
- Symmetric positive definite matrices
- Trace, rank, norm

Concepts you will learn in this course

… and that you can mention during your internship interviews 😉

- Multiple linear regression in the statistician's way
- Principal component analysis aka PCA or ACP in FR
- Cross-validation and overfitting, model selection
- Linear classification with discriminative and generative models
- Decision trees and random forests
- Ensemble methods: stacking, bagging, and (gradient) boosting
- Introduction to social network analysis

- Administrative stuff
- Course overview
- Simple linear regression
- Multivariate statistics

Simple linear regression

"One of the things that people most often want to know about the world is **how different variables are related to each other**, and one of the central tools in statistics to tackle this question is regression."

Cosma Shalizi in "Advanced Data Analysis from an Elementary Point of View"

$$\begin{matrix} X & Y \end{matrix}$$
$$\begin{bmatrix} 1.1 & 0.5 \\ 1.9 & 1.1 \\ 3.1 & 1.4 \\ 4.0 & 2.0 \end{bmatrix}$$

$$X \sim p_X(x)$$

$$Y \sim p_Y(y)$$

Simple linear regression

"One of the things that people most often want to know about the world is **how different variables are related to each other**, and one of the central tools in statistics to tackle this question is regression."

Cosma Shalizi in "Advanced Data Analysis from an Elementary Point of View"

Q: What is the optimal point forecast for a random variable Y ?

Simple linear regression

"One of the things that people most often want to know about the world is **how different variables are related to each other**, and one of the central tools in statistics to tackle this question is regression."

Cosma Shalizi in "Advanced Data Analysis from an Elementary Point of View"

Q: What is the optimal point forecast for a random variable Y ?

A: One possible way is to minimize the mean squared error

$$\mu = \underset{m \in \mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}_Y \left[(Y - m)^2 \right] = \mathbb{E}_Y [Y]$$

Simple linear regression

Q: What if instead of just using one constant number to predict Y we used some function μ of another random variable X as our estimate?

Q: What if instead of just using one constant number to predict Y we used some function μ of another random variable X as our estimate?

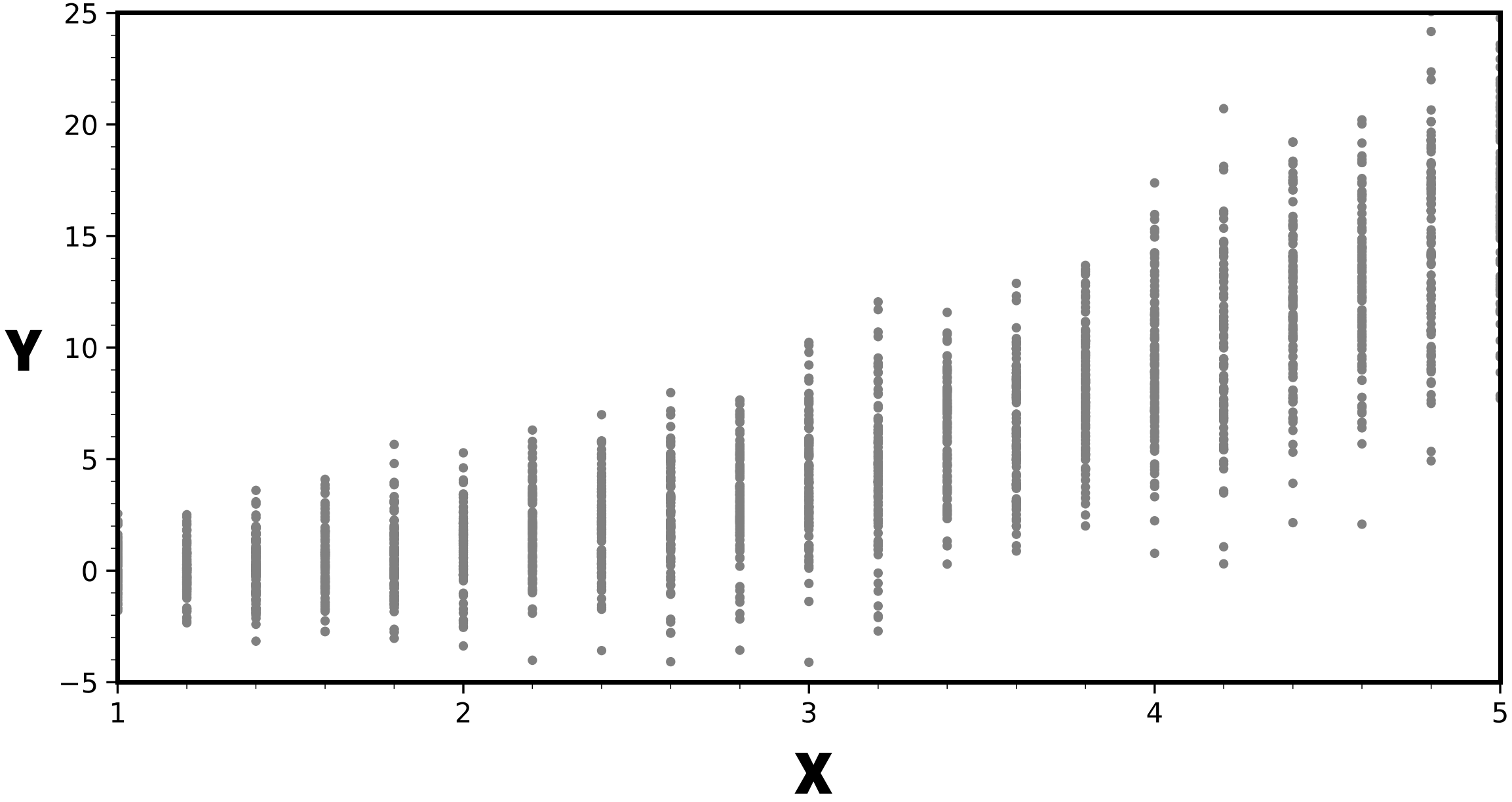
A: Our new optimization problem is

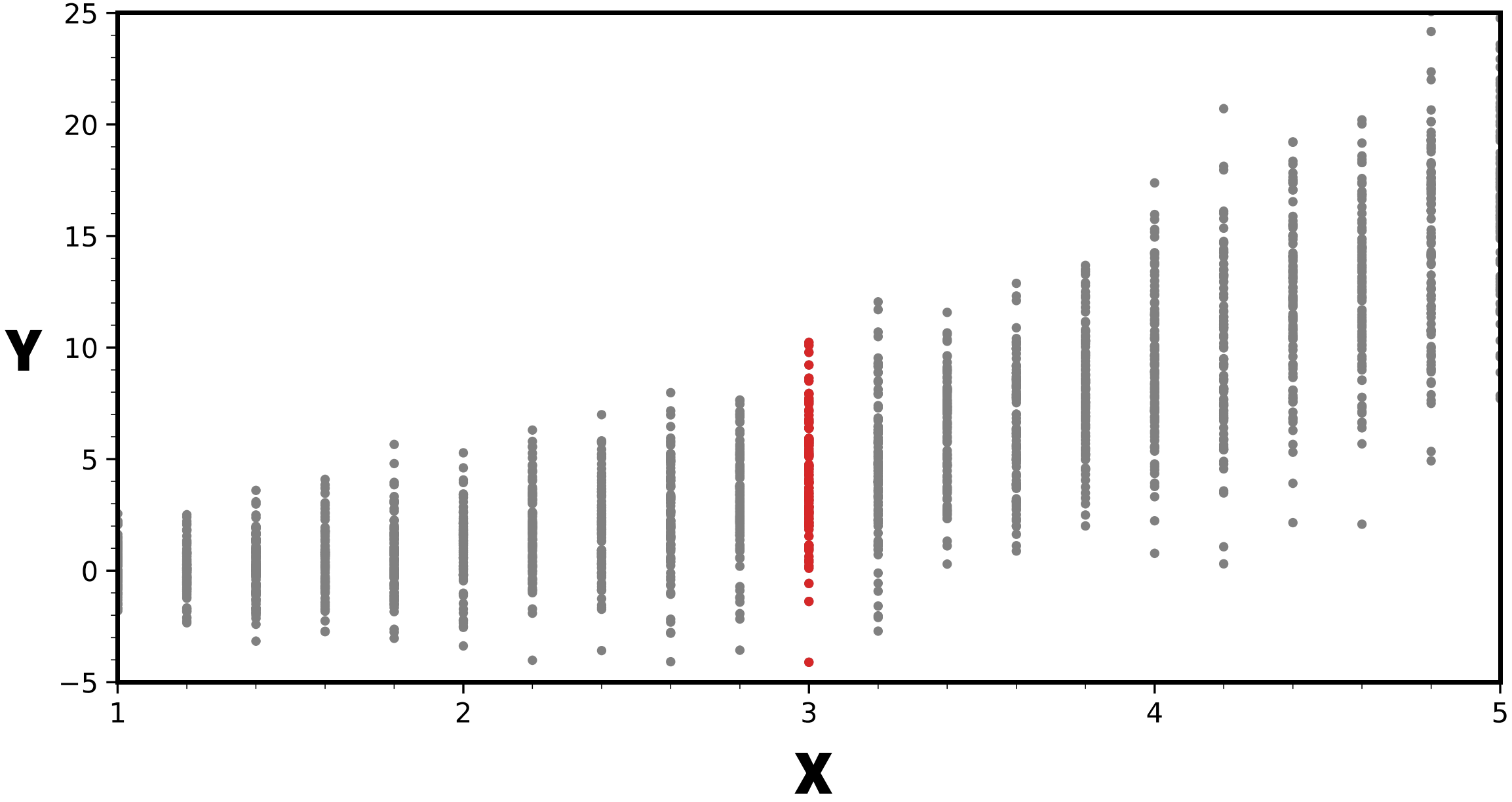
$$\mu = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}_{(X,Y)} \left[\Big(Y - f(X) \Big)^2 \right]$$

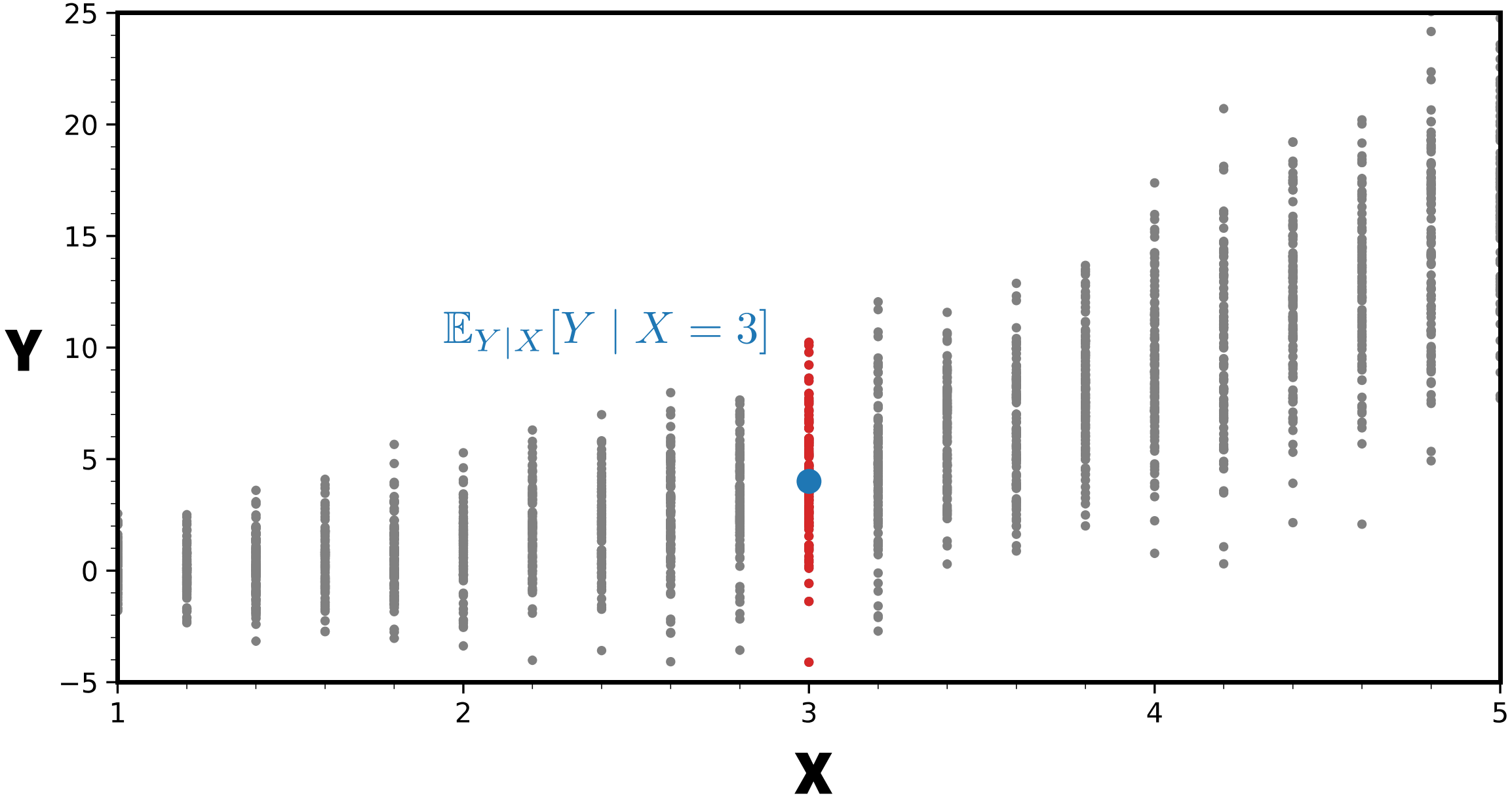
with solution

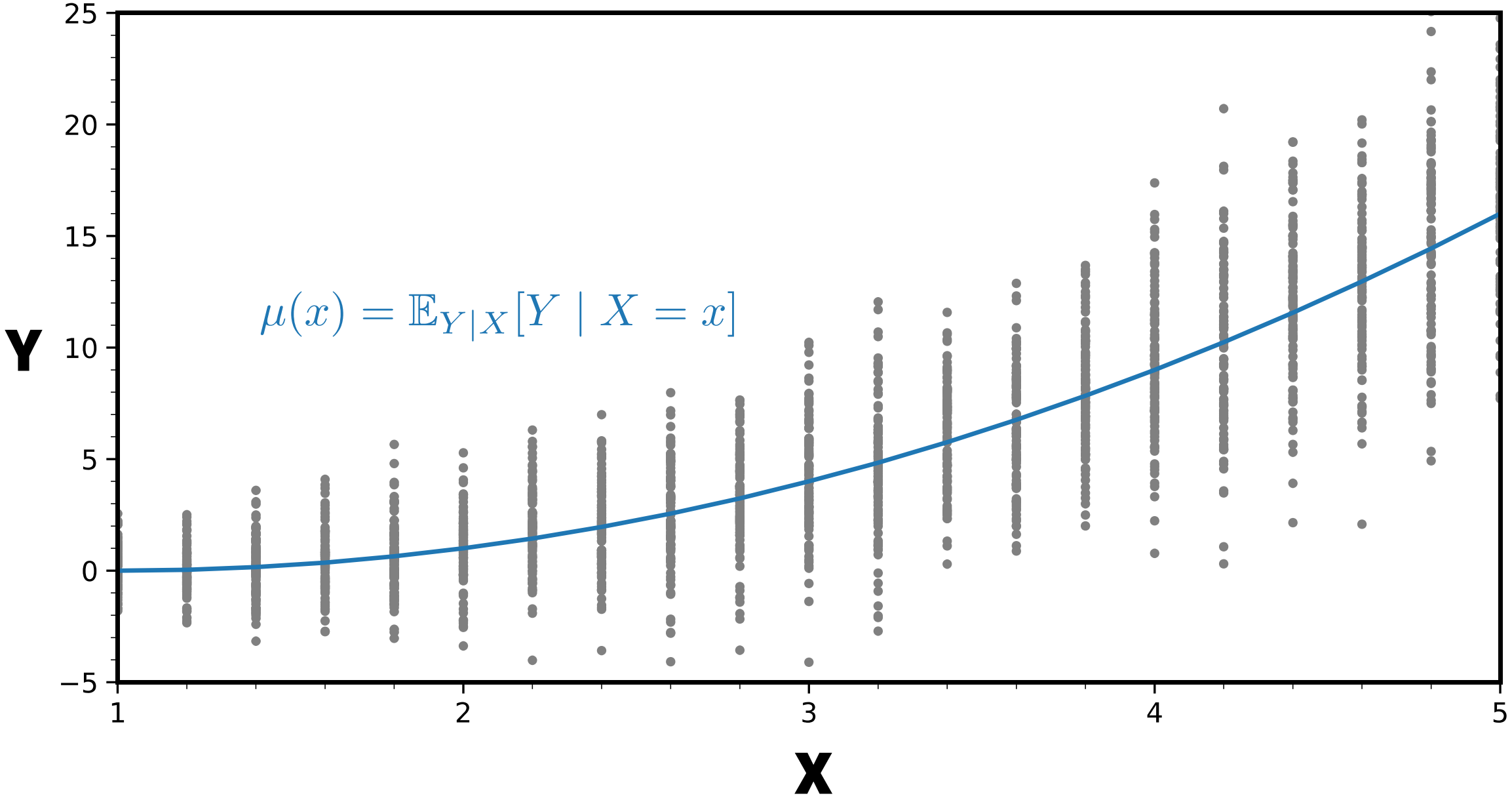
$$\mu(x) = \mathbb{E}_{Y|X}[Y \mid X = x]$$

Conclusion: the mean-squared optimal conditional prediction of Y in terms of X is the conditional expected value of Y for a fixed X


25
20
15
10
5
0
−5
Y
1
2
3
4
5
X


25
20
15
10
5
0
−5
Y
1
2
3
4
5
X


$\mathbb{E}_{Y|X}[Y \mid X = 3]$
Y
X
25
20
15
10
5
0
−5
1
2
3
4
5


$\mu(x) = \mathbb{E}_{Y|X}[Y \mid X = x]$
Y
X
25
20
15
10
5
0
−5
1
2
3
4
5

Q: What if instead of just using one constant number to predict Y we used some function μ of another random variable X as our estimate?

A: Our new optimization problem is

$$\mu = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}_{(X,Y)} \left[\Big(Y - f(X) \Big)^2 \right]$$

with solution

$$\mu(x) = \mathbb{E}_{Y|X}[Y \mid X = x]$$

Conclusion: the mean-squared optimal conditional prediction of Y in terms of X is the conditional expected value of Y for a fixed X

↳ **But how can we estimate the conditional expectation from observed data?**

Simple linear regression

The conditional expectation could be any function. **But this would be hard**.

We will restrict our regression function to have a linear form as in

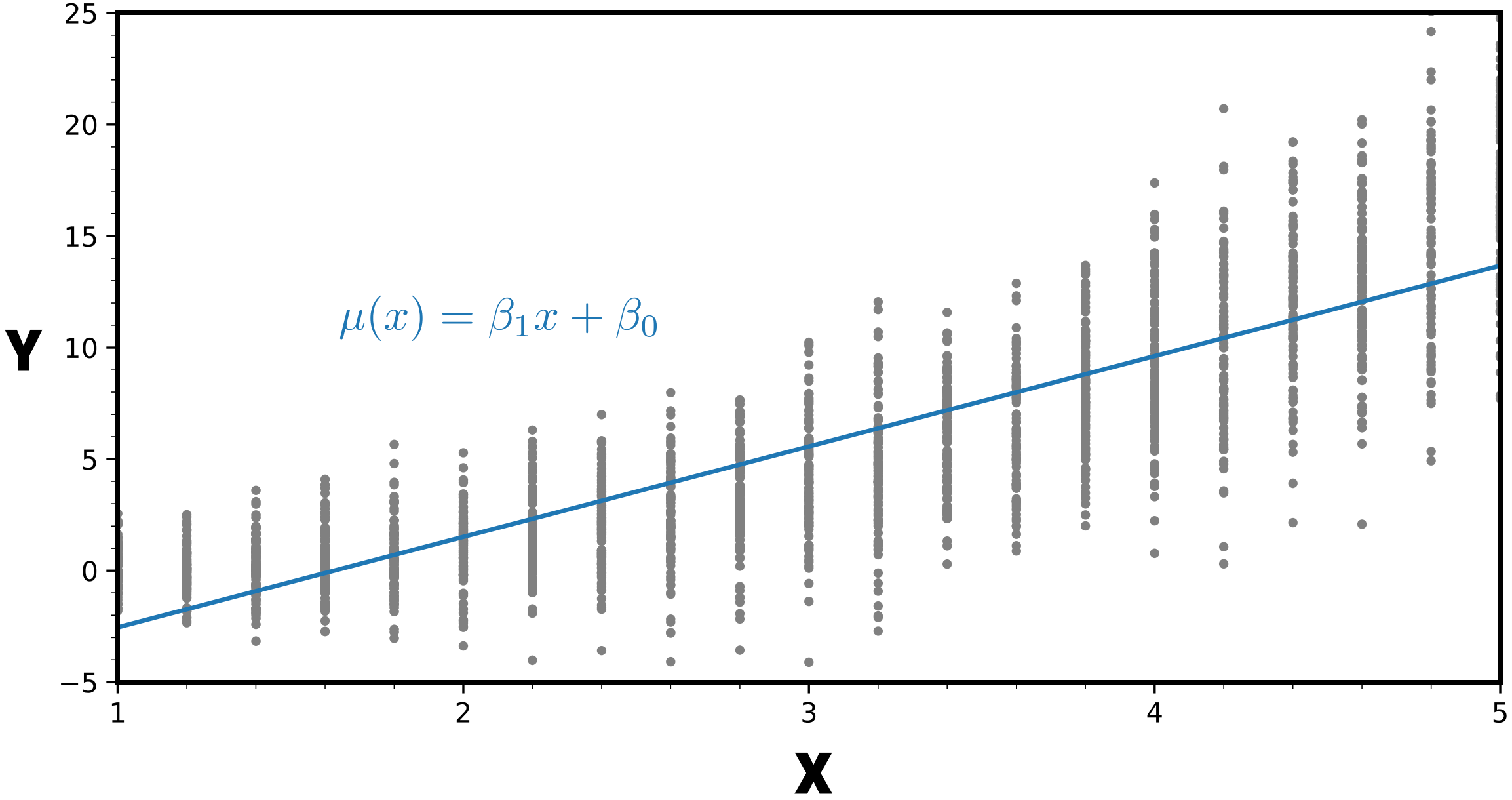
$$\mu(x) = \beta_1 x + \beta_0$$

The parameters of the model are thus estimated by minimizing the MSE

$$\mathcal{L}(\beta_0, \beta_1) = \mathbb{E}_{(Y,X)}\left[(Y - (\beta_0 + \beta_1 X))^2\right]$$

for which the minimizers are

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \qquad \beta_0 = \mathbb{E}_Y[Y] - \beta_1 \mathbb{E}_X[X]$$


$\mu(x) = \beta_1 x + \beta_0$
Y
X
25
20
15
10
5
0
−5
1
2
3
4
5

$$\mu(x) = \mathbb{E}_Y[Y] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\left(x - \mathbb{E}_X[X]\right)$$

Some insights and comments:

- The **optimal slope increases** the more X and Y tend to fluctuate together and gets pulled towards zero the more X fluctuates by itself.
- The expected values of X and Y play no role in the slope (only the variances and covariances). Therefore, the **optimal slope does not change if we re-center the data**

Simple linear regression

How do we estimate the coefficients from **observed data**?

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \quad \Rightarrow \quad \hat{\beta}_1 = \frac{c_{XY}}{s_X^2}$$

$$\beta_0 = \mathbb{E}_Y[Y] - \beta_1 \mathbb{E}_X[X] \quad \Rightarrow \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$c_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y}) \qquad s_X^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2 \qquad \bar{Z} = \frac{1}{N} \sum_{i=1}^{N} z_i$$

OK, we could stop here… like most machine learning courses do.

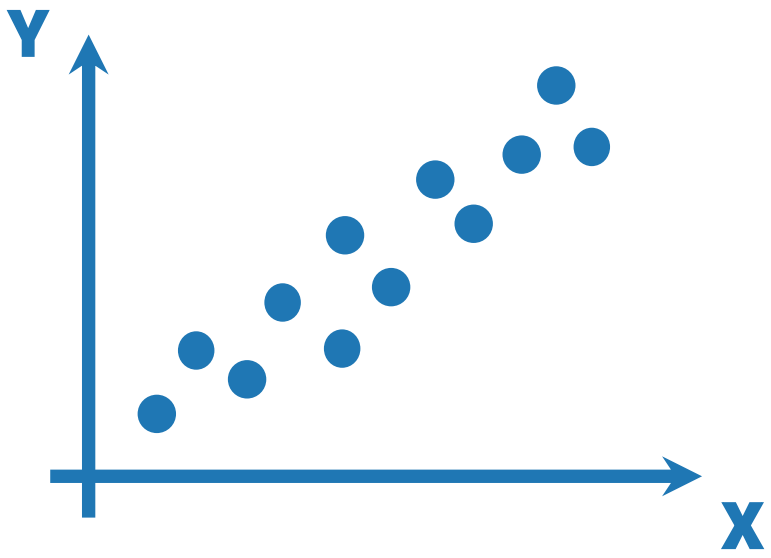
But how can we answer those questions that I mentioned previously?

- Under which **conditions** is the linear regression model reliable?
- What sort of **errors** does the linear regression model can make ?
- What can the linear regression model tell us when it works?
- What are the signs that something has **gone wrong**?

↳ **We need to make some assumptions on the data generative model**

Simple linear regression

DATA GENERATIVE MODEL

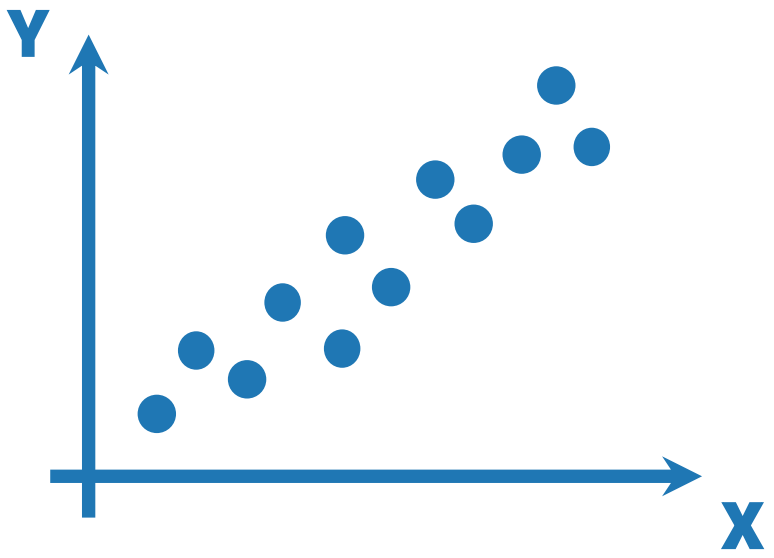


REGRESSION MODEL

$$Y = f(X)$$

Simple linear regression

DATA GENERATIVE MODEL



Assumption: "Quantity Y depends linearly on the values of X and random noise"

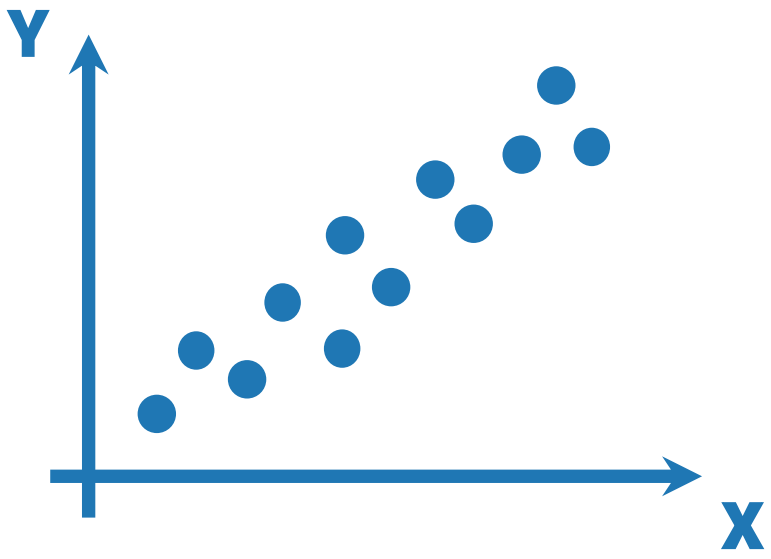
$$Y = \beta_1 X + \beta_0 + \varepsilon$$

REGRESSION MODEL

$$Y = f(X)$$

Simple linear regression

DATA GENERATIVE MODEL



Assumption: "Quantity Y depends linearly on the values of X and random noise"

$$Y = \beta_1 X + \beta_0 + \varepsilon$$

REGRESSION MODEL

$$Y = f(X)$$

$$f(X) = \hat{\beta}_1 X + \hat{\beta}_0$$

The hat-parameters are estimated from observed data and expected to converge to the true parameters

Simple linear regression

Assumption: "Quantity Y depends linearly on the values of X and random noise"

$$Y = \beta_1 X + \beta_0 + \varepsilon$$

The additive noise is assumed to have **three important properties**:

- Zero mean
- Constant variance
- No correlation with the values of X

Note that we make **no assumption about p(X)** (what does this imply?)

The assumption of additive noise is non-trivial, but a very common one.

↳ **Ideally, we should always check these assumptions…**

Gaussian simple linear regression

It is also common (though not trivial) to assume a Gaussian additive noise, i.e.

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

which is the same as saying that the generative model of the data is

$$Y \mid X = x \sim \mathcal{N}(\beta_1 x + \beta_0, \sigma^2)$$

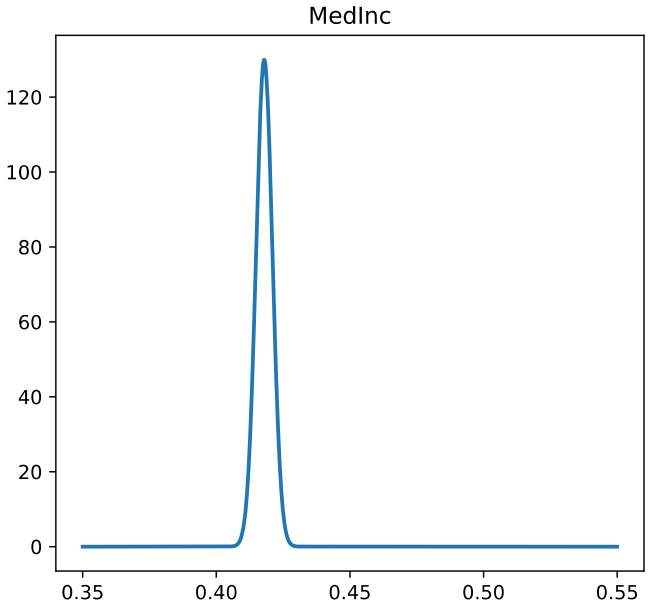
We can even write the statistical distribution of the estimated parameters

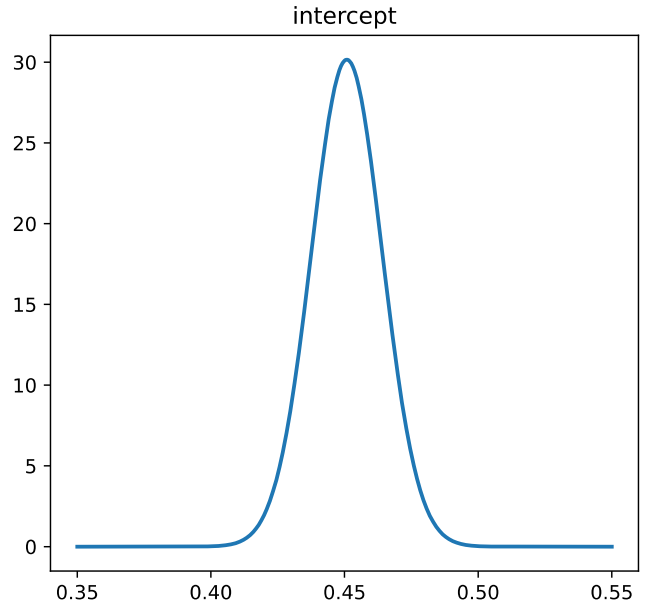
$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{N} \frac{1}{s_X^2}\right) \qquad \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{N}\left(1 + \frac{\bar{X}^2}{s_X^2}\right)\right)$$

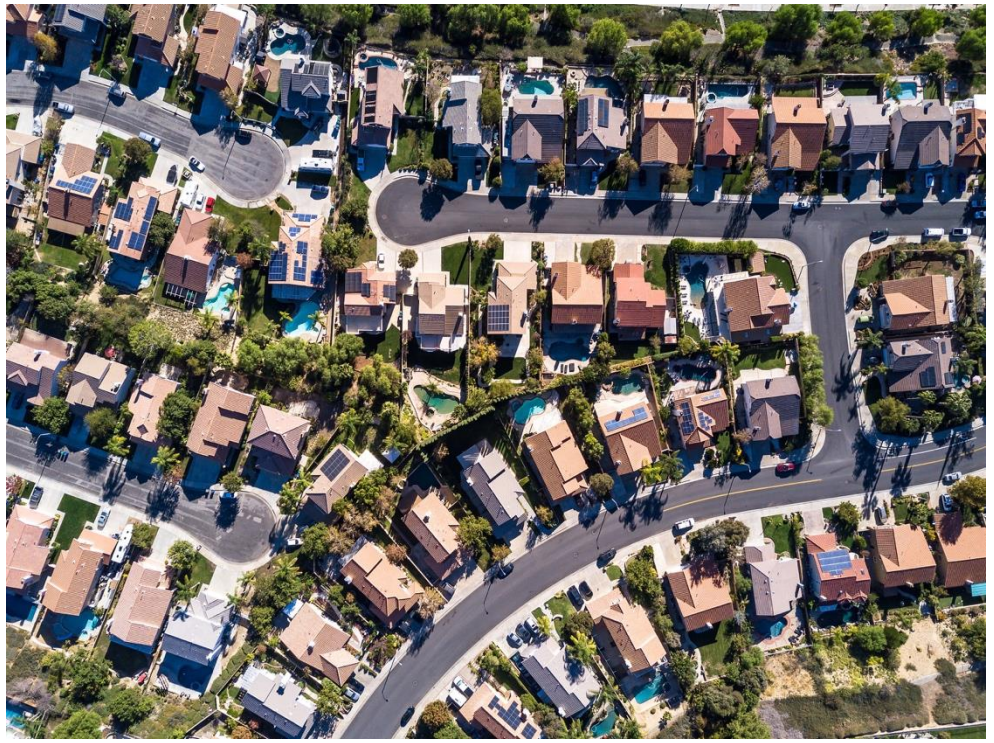
Coming back to our example with the California housing dataset

Example

Can we predict the average price of houses in a district given the average income of people living there?







California housing dataset

Q: But how can we **ensure** that the **assumptions** of the model are **respected**?

We can rewrite the Gaussian regression model as $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \sim \mathcal{N}(0, \sigma^2)$

and we define the **residuals** of our estimation as in $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

With a little bit of ~~magic~~ algebra we can obtain an expression for the residuals

$$e_i = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i$$

$$\mathbb{E}[e \mid X] = 0 \qquad \mathrm{Var}(e) = \frac{n-2}{n}\sigma^2 \qquad e \sim \mathcal{N}\left(0, \frac{n-2}{n}\sigma^2\right)$$

Q: But how can we **ensure** that the **assumptions** of the model are **respected**?

$$\mathbb{E}[e \mid X] = 0 \qquad \mathrm{Var}(e) = \frac{n-2}{n}\sigma^2 \qquad e \sim \mathcal{N}\left(0, \frac{n-2}{n}\sigma^2\right)$$

So if the assumptions of the Gaussian simple linear regression are satisfied, then:

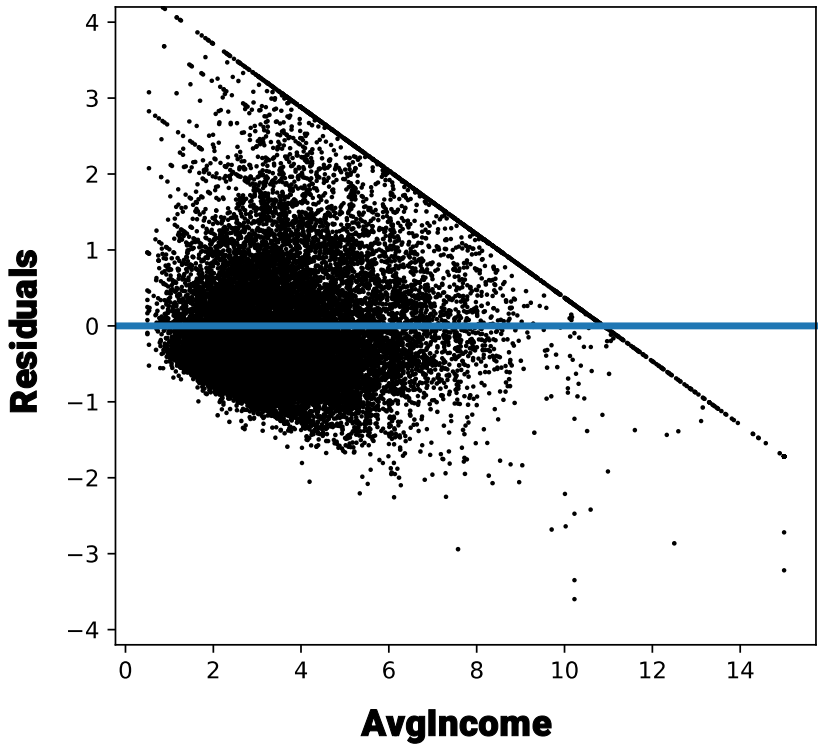
- The residuals should have **expectation zero** conditioned on X
- The residuals should show a nearly-**constant variance**
- The residuals should have a **Gaussian** distribution

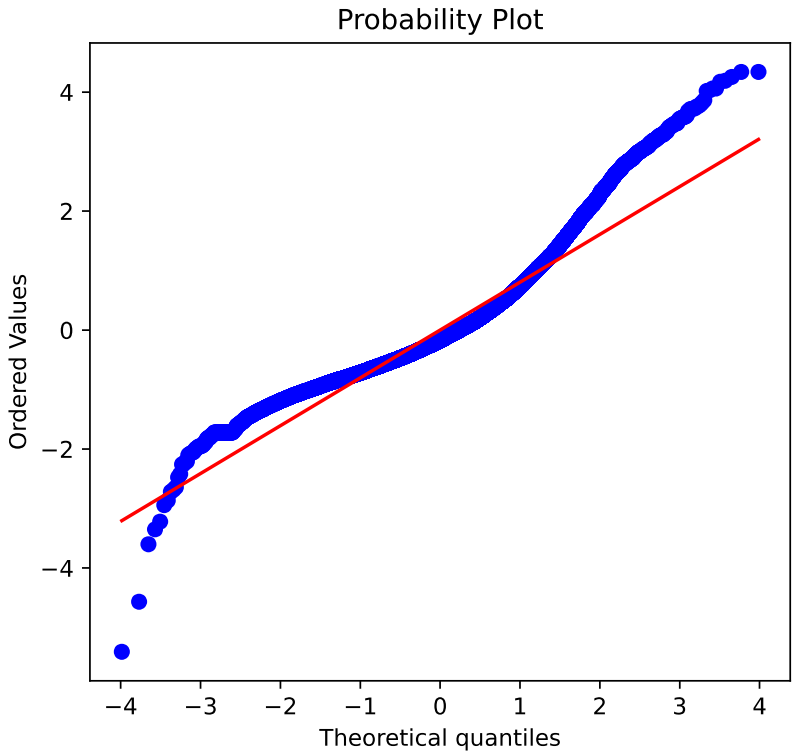
↳ **In practice, they never are... so it should be "almost" Gaussian**

Gaussian simple linear regression

So if the assumptions of the Gaussian simple linear regression are satisfied, then:

- The residuals should have **expectation zero** conditioned on X
- The residuals should show a nearly-**constant variance**
- The residuals should have a **Gaussian** distribution

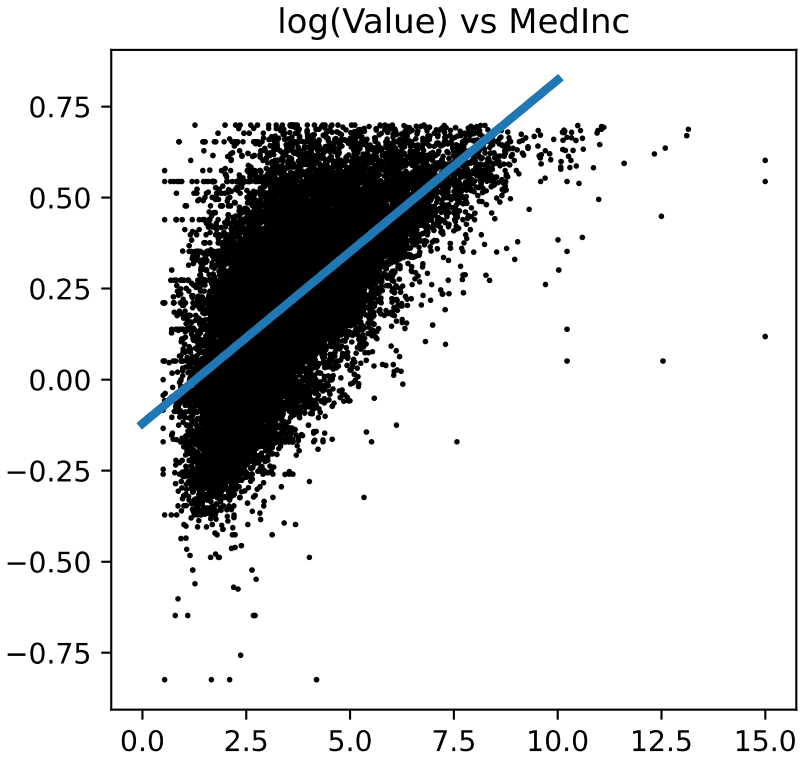


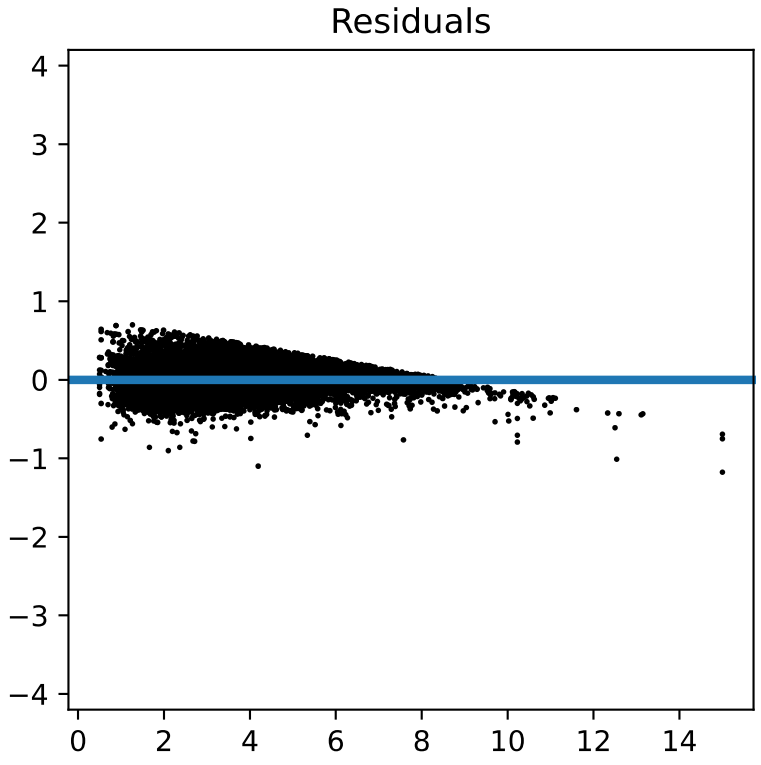


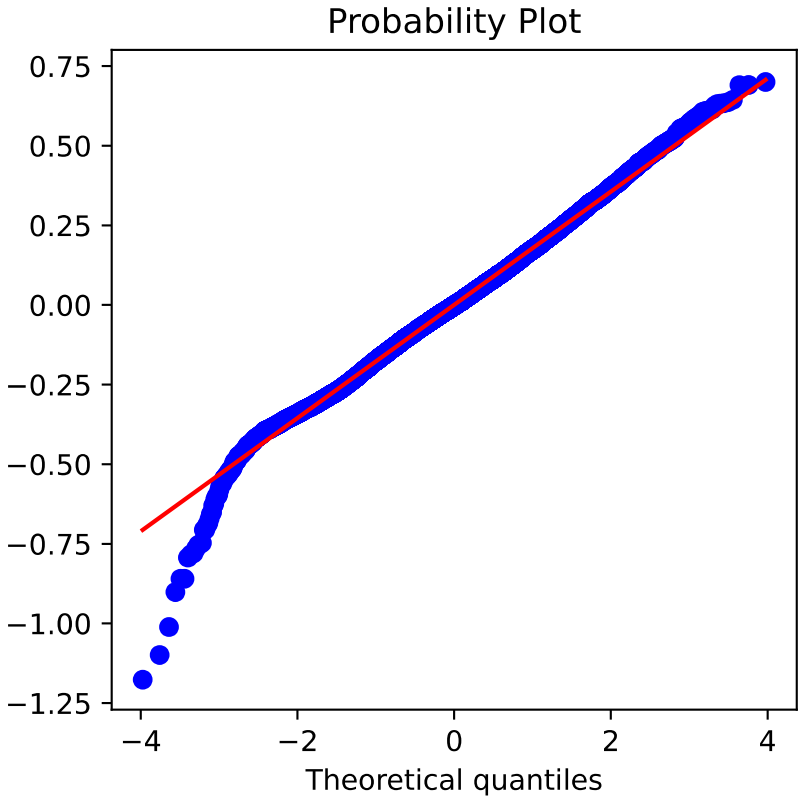
Gaussian simple linear regression

After using some pre-processing tricks…

- Filter the thresholded house values
- Apply a logarithmic transformation to the target values







Gaussian ~~simple~~ multiple linear regression

The next natural step would be to model the data using *p* predictors

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

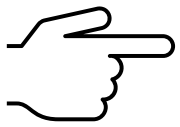
where the **p+1**-dimensional vector of parameters can be estimated via

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix}^T$$

↳ **How to describe the statistics of the vector of estimated parameters?**

- Administrative stuff
- Course overview
- Simple linear regression
- Multivariate statistics



Basic concepts of multivariate statistics

Consider a random vector defined in $\mathbb{R}^n$

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$ where each X_i is a random variable

Expectation

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

Covariance Matrix

$$\mathrm{Cov}(X) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$$

$$\sigma_{ij} = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\ \mathbb{E}[X_j]$$

Gaussian vectors

A random vector X is a **Gaussian vector** if, and only if, every linear combination

$$\sum_{i=1}^{n} a_i X_i$$

of its components has a normal distribution

We denote it as $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = \mathbb{E}[X]$ and $\Sigma = \mathrm{Cov}(X)$

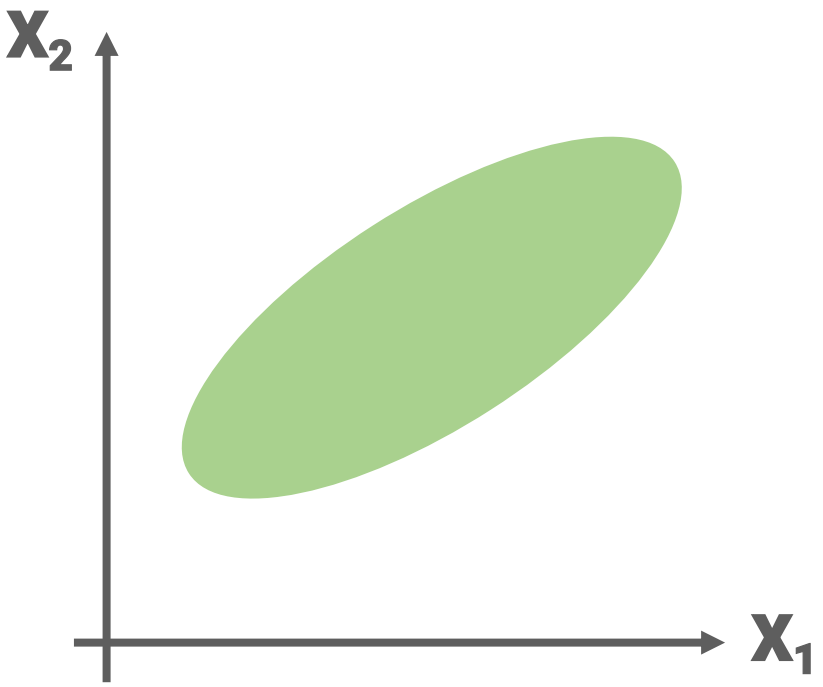
and the pdf is

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Gaussian vectors

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

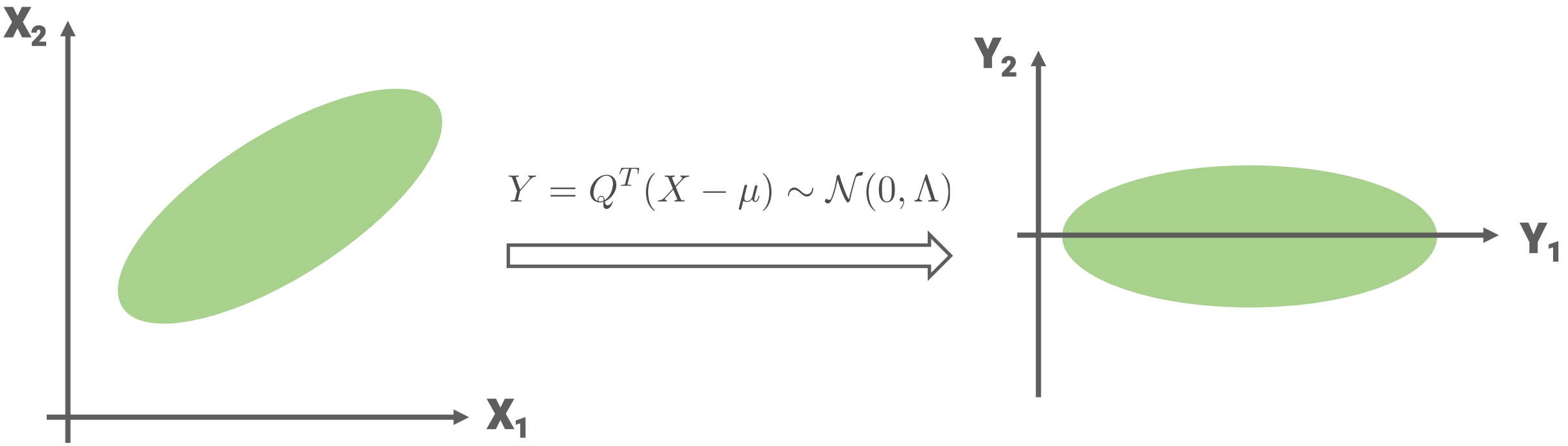
↳ **The equidensity contours are ellipsoids centered at the mean**



Gaussian vectors

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

↳ **The equidensity contours are ellipsoids centered at the mean**



Gaussian ~~simple~~ multiple linear regression

Coming back to our problem of Gaussian multiple linear regression

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 & \dots & y_N \end{pmatrix}^T$$

we can write that

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$