PSTAT 126

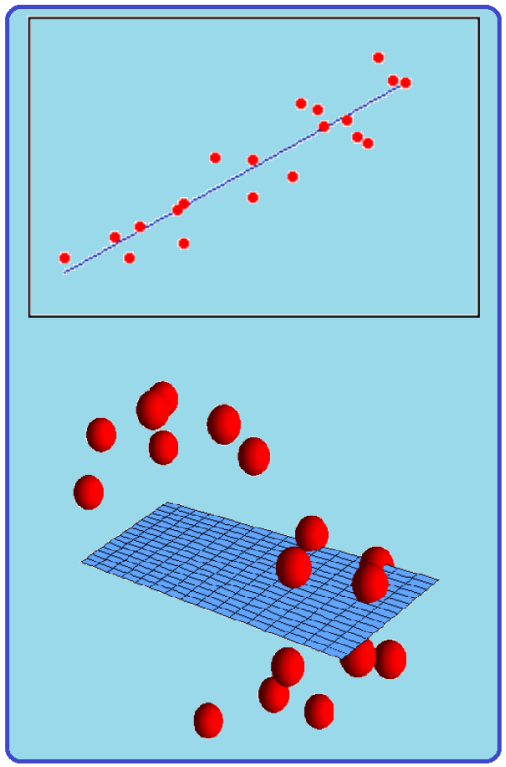
Regression Analysis

Laura Baracaldo

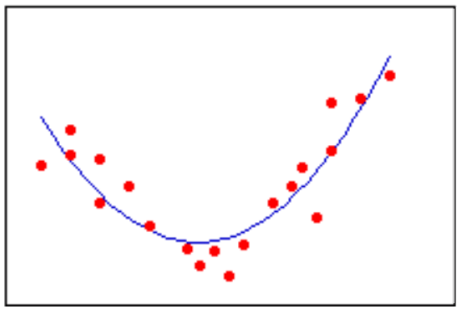
Lecture 1
Introduction to Linear Models

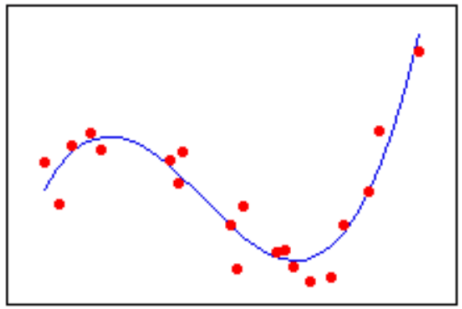
Introduction

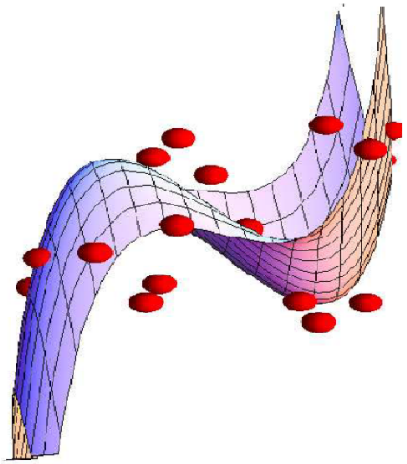
Regression Analysis: Statistical method that allows the exploration of the relationship between two or more variables.



Linear Regression Models







Introduction

- Y : Response, outcome, output, dependent variable. (Continuous)
- $(X_1, ...X_p)^T$: vector of predictors, inputs, independent or explanatory variables. (Can be of any kind: qualitative or quantitative)

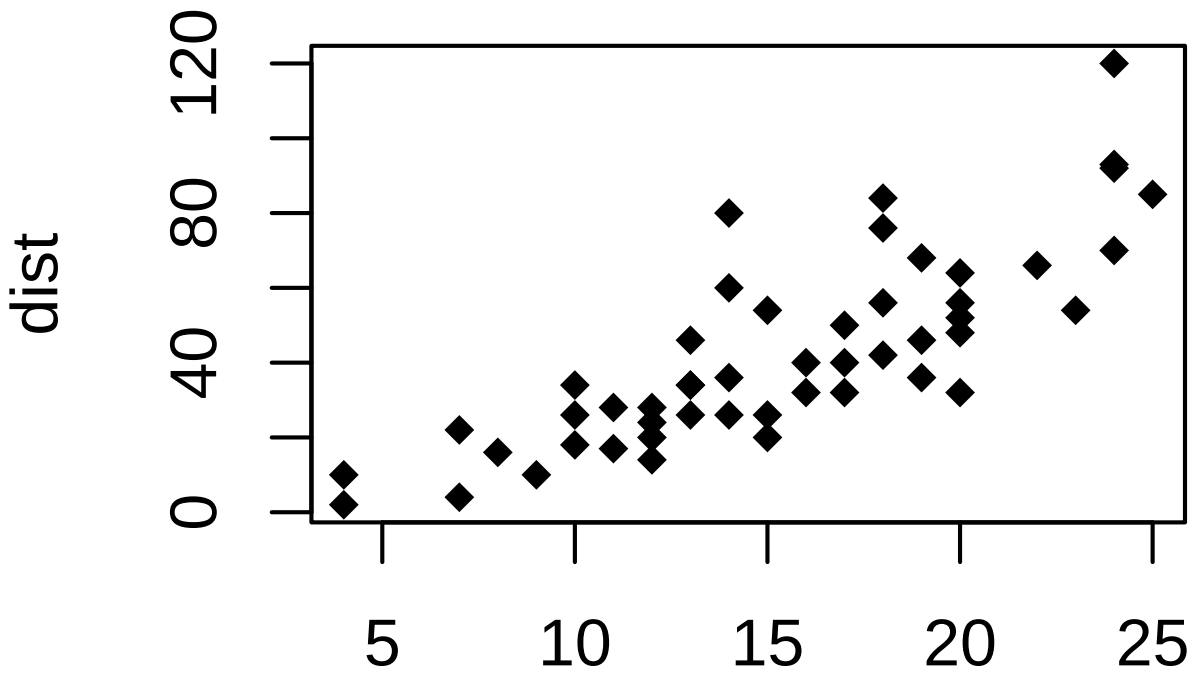
Goals:

1. Prediction of future or unmeasured responses given values of the predictors.
2. Assessment of the effect/relationship between explanatory variables and response. Existence of causality?

Cars example

X : Speed Y : Distance

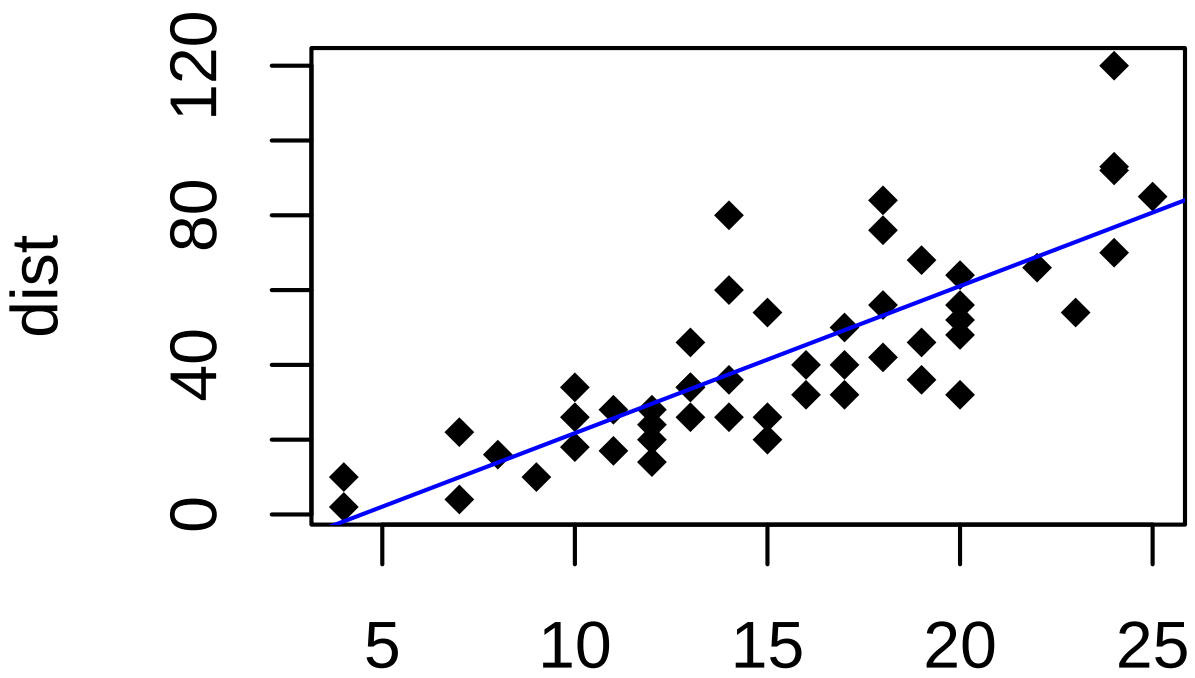
```
par(mar=c(3,4,0.5,1))
plot(cars, pch=18)
```



Cars example

X : Speed Y : Distance

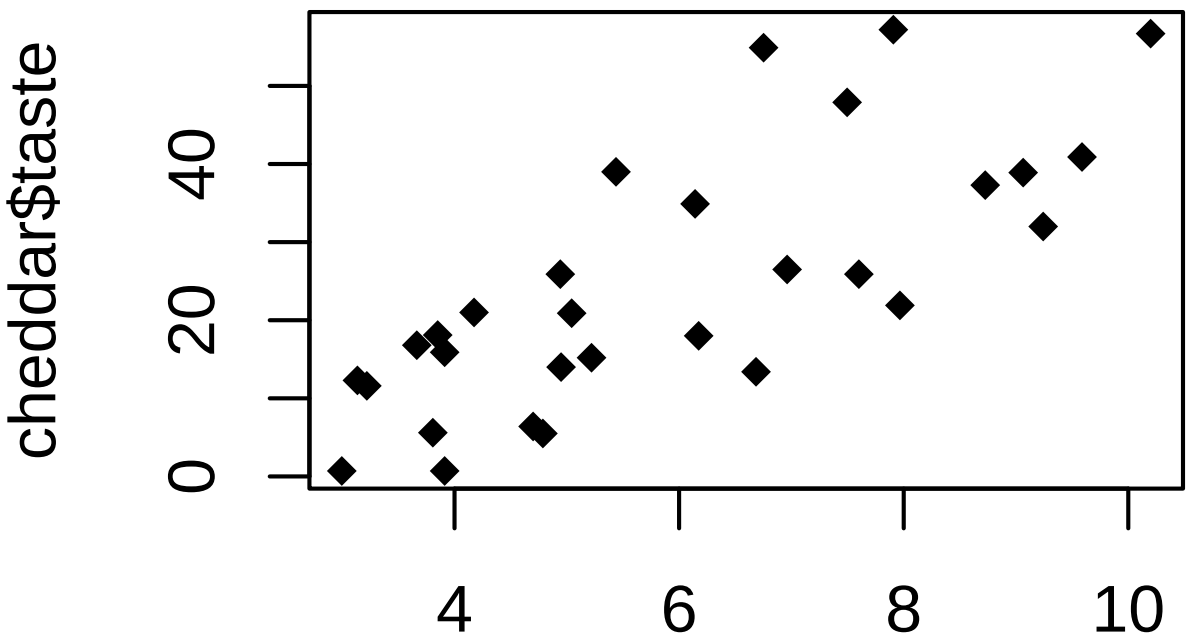
```
par(mar=c(3,4,0.5,1))
plot(cars, pch=18)
abline(lm( dist ~ speed, data=cars), col="blue")
```



Cheddar example

X : Hydrogen sulfide concentration Y : Taste score

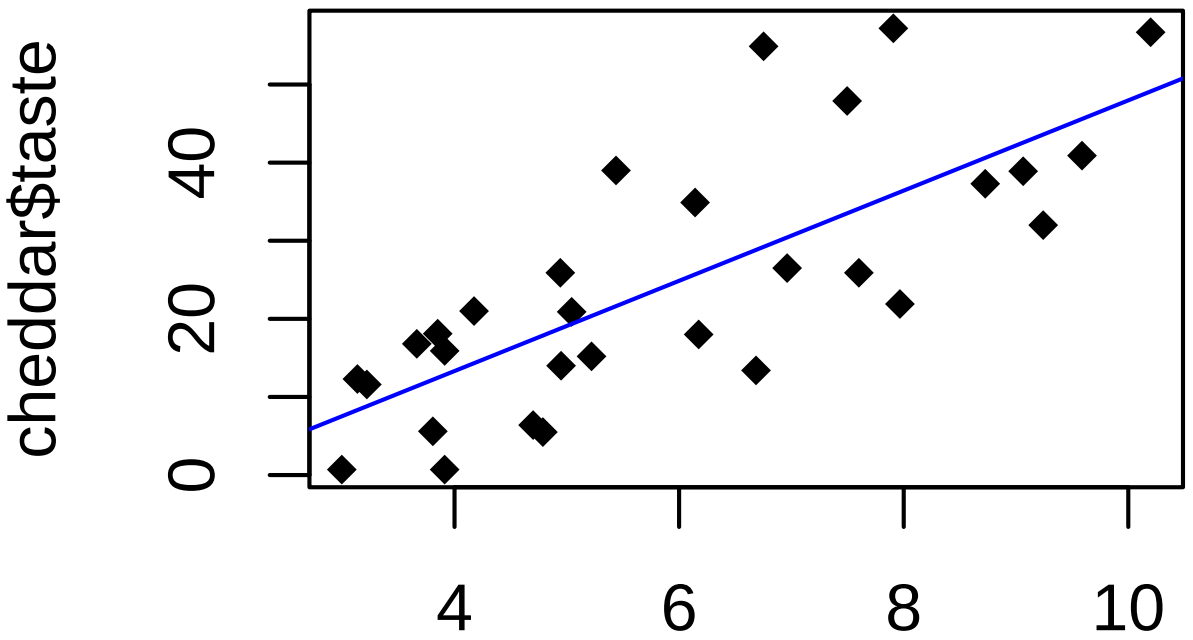
```
library(faraway)
par(mar=c(3,4,0.5,1))
plot(cheddar$H2S, cheddar$taste, pch=18)
```



Cheddar example

X : Hydrogen sulfide concentration Y : Taste score

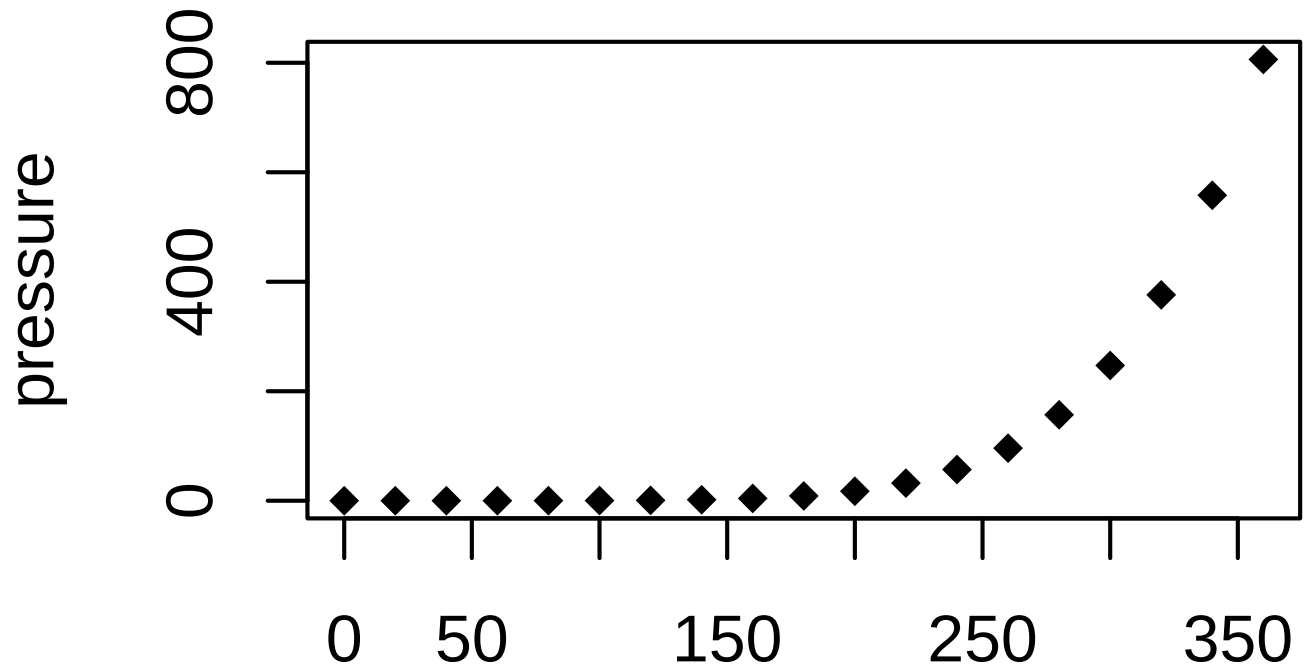
```
library(faraway)
par(mar=c(3,4,0.5,1))
plot(cheddar$H2S, cheddar$taste, pch=18)
abline(lm( taste ~ H2S, data=cheddar), col="blue")
```



Pressure example

X : Temperature Y : Pressure

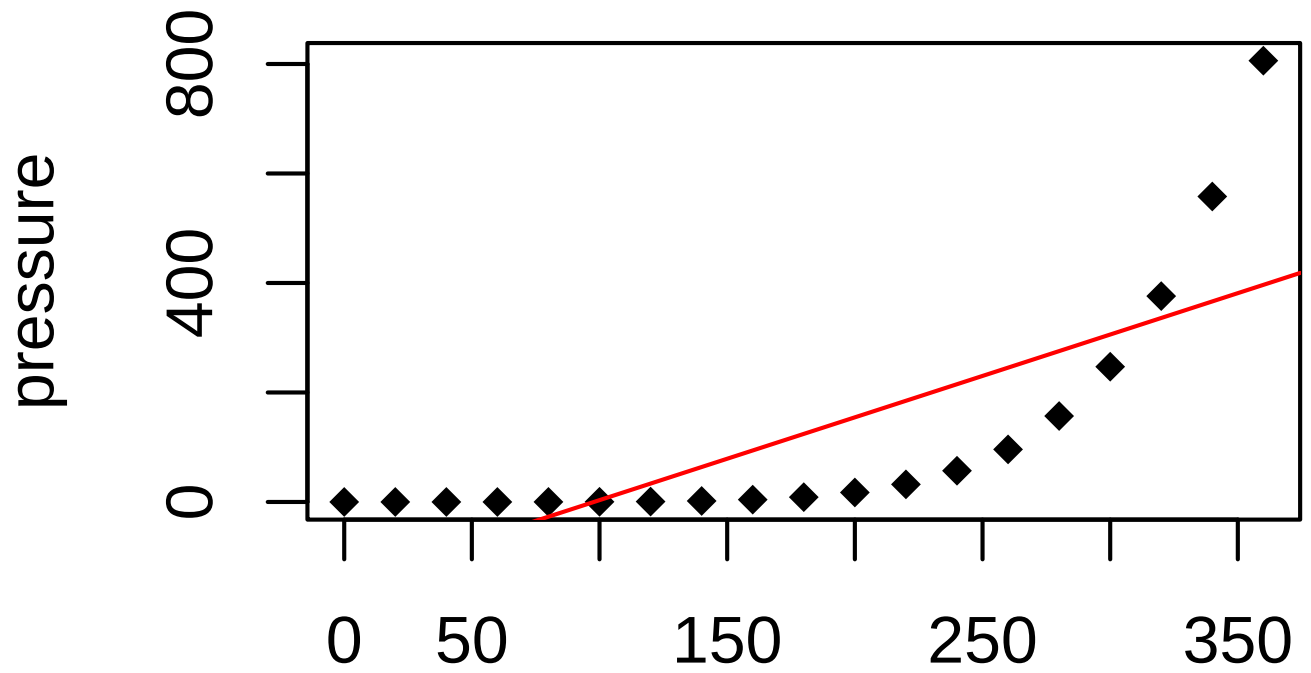
```
par(mar=c(3,4,0.5,1))
plot(pressure, pch=18)
```



Pressure example

X : Temperature Y : Pressure

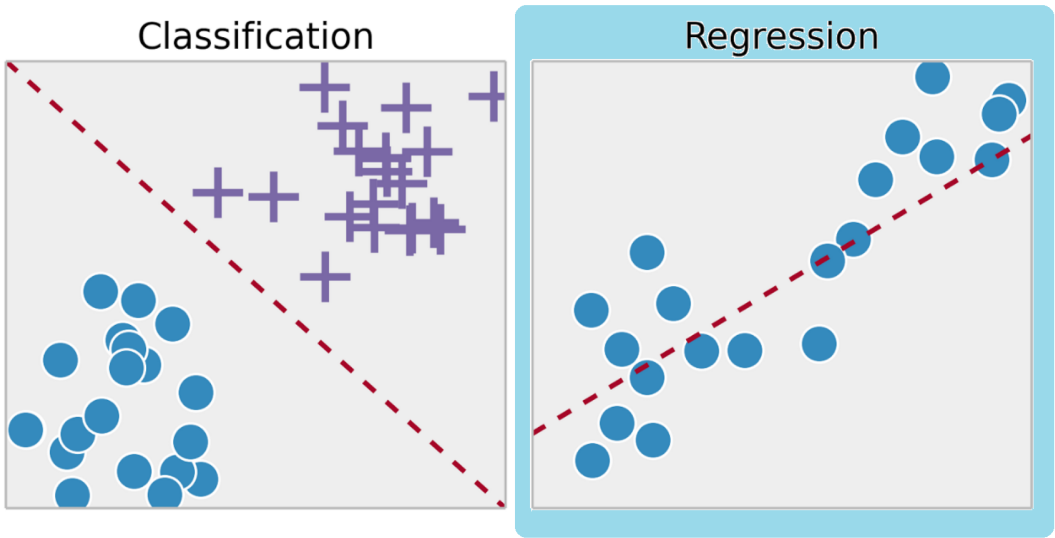
```
par(mar=c(3,4,0.5,1))
plot(pressure, pch=18)
abline(lm( pressure ~ temperature, data=pressure), col="red")
```



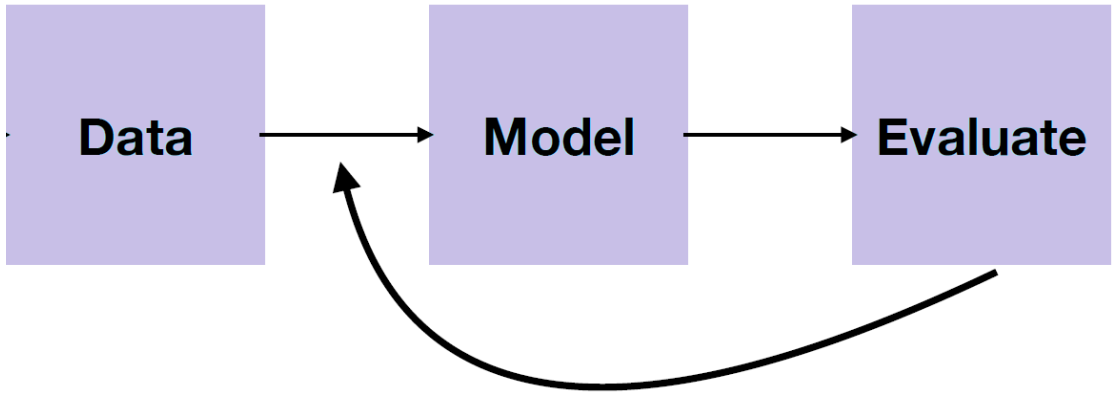
Not Linear!

Regression vs Classification

- Regression: Y is continuous. Numerical values (e.g. price, blood pressure, confirmed COVID-19 cases).
- Classification Y discrete labels. Categorical values. (e.g. survived/died, digit 0-9, if Bitcoin price is going up tomorrow).



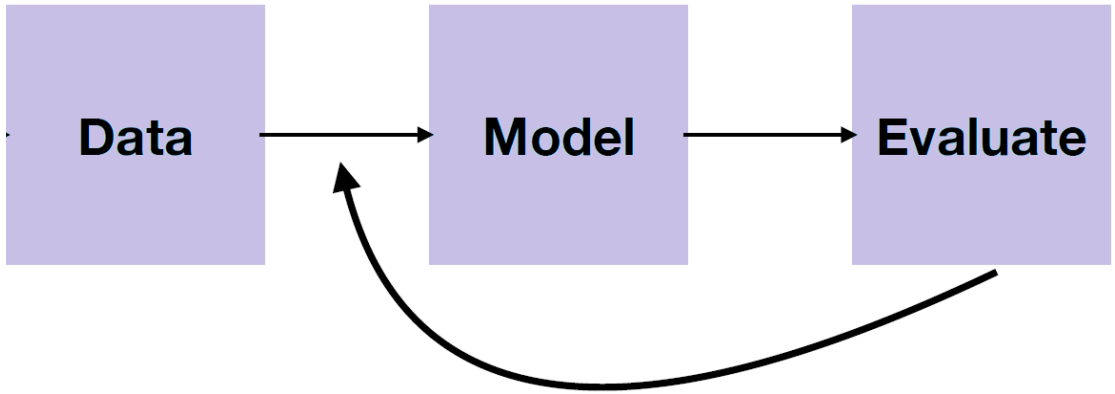
General Workflow



1. Data: We collect data for all n subjects in the study, and identify Y and X:

$$Y = \begin{bmatrix} y_1 \\ \\ \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ & & & \end{bmatrix}$$

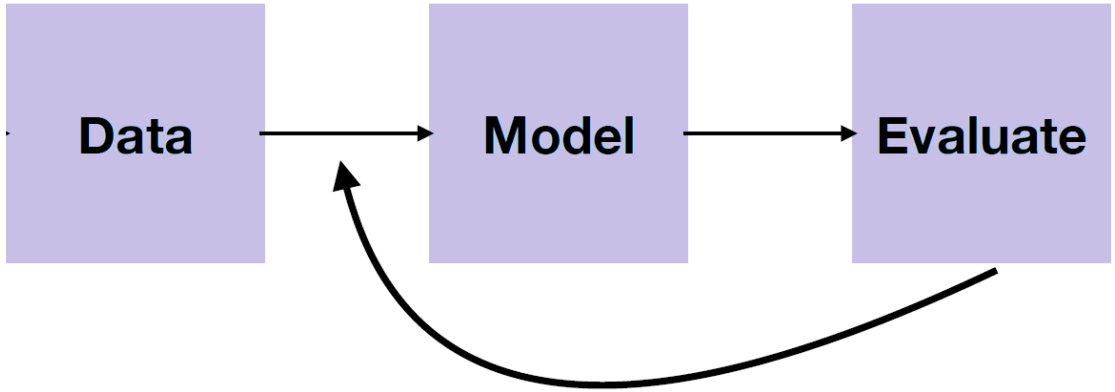
General Workflow



1. Data: We collect data for all n subjects in the study, and identify Y and X:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \\ \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ & & & \end{bmatrix}$$

General Workflow



1. Data: We collect data for all n subjects in the study, and identify Y and X:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$

Regression Model

$$Y = f(X_1, X_2, ...X_p) + \epsilon$$

- ϵ : Error/noise
- f is unknown and must be learned from observations $(Y, X_1, ..., X_p)$
- Questions about f. Continuous? Smooth?
- f must be restricted. For the most part of this course: We assume f is linear.

Linear Regression Model

We assume the response is written as linear representation of predictors:

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_p X_p + \epsilon$$

- **Estimation:** The goal is to estimate unknown parameters $\beta_0, \beta_1, ...\beta_p$.

Linear Regression Model

- **Linear:**

$$Y = \beta_0 + \beta_1 log(X_1) + ...\beta_p X_p^2 + \epsilon = \beta_0 + \beta_1 W_1 + ...\beta_p W_p + \epsilon$$

- **NOT Linear:**

$$Y = \beta_0 + X_1^{\beta_1} + ...\beta_p X_p + \epsilon$$

Matrix Representation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = X\beta + \epsilon$$

Model Assessment

1. Prediction: We seek to accurately predict future response given predictors.
2. Inference: We want to assess the quality of our predictions and (or) estimation.
3. Diagnostics: what if the linear model assumptions go wrong? how can we tell?
4. Model selection: We try to find the "best" linear model tha links the response and the set of predictors.