Tufts

CS135
Introduction to Machine Learning

Lecture 4: Regression

Linear Regression

Two main objectives:

1. Establish relationship between 2 variables
   - Positive if ↑ in 1 causes ↑ in other
   - Negative if ↑ in 1 causes ↓ in other

   e.g., Income & Spending

2. Forecast new observations

   Use relationship to predict unobserved events

   e.g., Sales for next quarter

Linear Regression

Variable Roles:

- Dependent Variable
  - Value to explain or forecast
  - Depends on something else
  - Denoted as y
- Independent Variable
  - Variable to explain the other
  - Values independent
  - Denoted as x

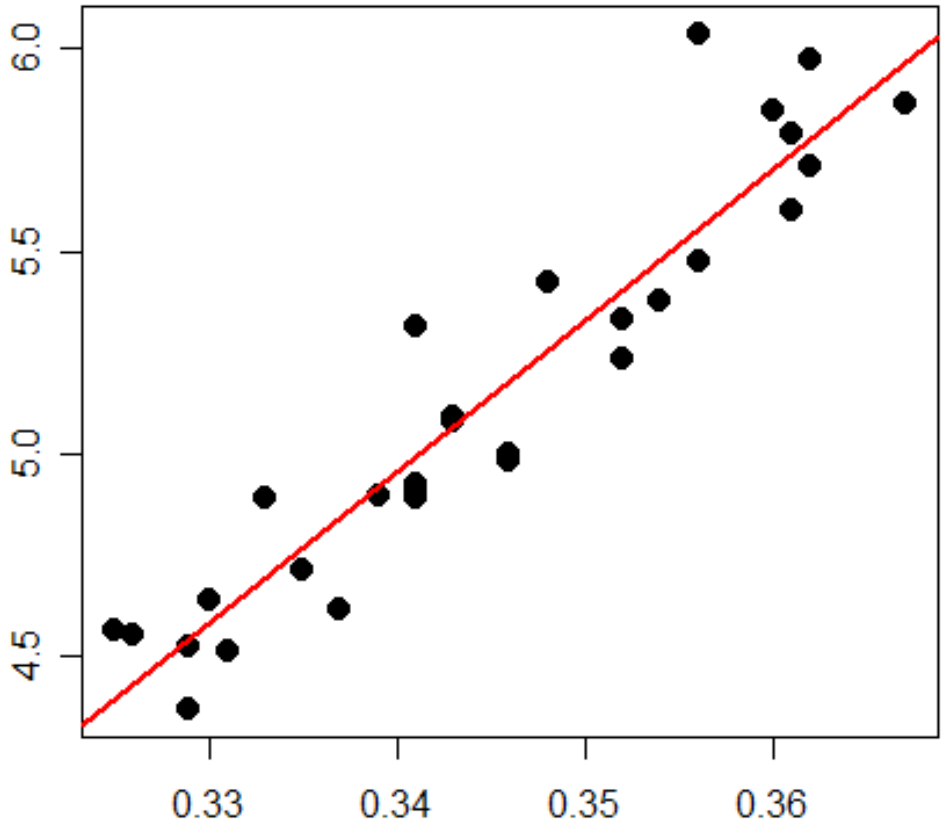
Linear Regression

Linear Equation

$$y = a + bx$$
$$y = mx + b$$

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

- β_0 is the intercept
- β_1 is the slope



- We will find these values via least squares

The Least Squares Criterion

The formulas for b_0 and b_1 that minimize the least squares criterion are:

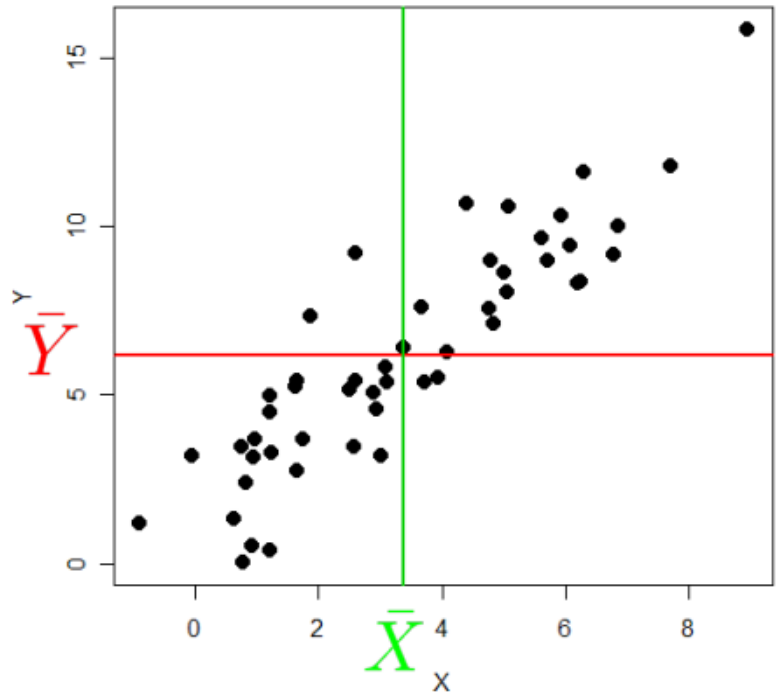
$$b_1 = corr(X, Y) \times \frac{s_Y}{s_X} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

where,

$$s_Y = \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \quad \text{and} \quad s_X = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Correlation and Covariance

Measure the *direction* and *strength* of the linear relationship between variables Y and X



$$Cov(Y, X) = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{n - 1}$$

Tufts

Correlation and Covariance

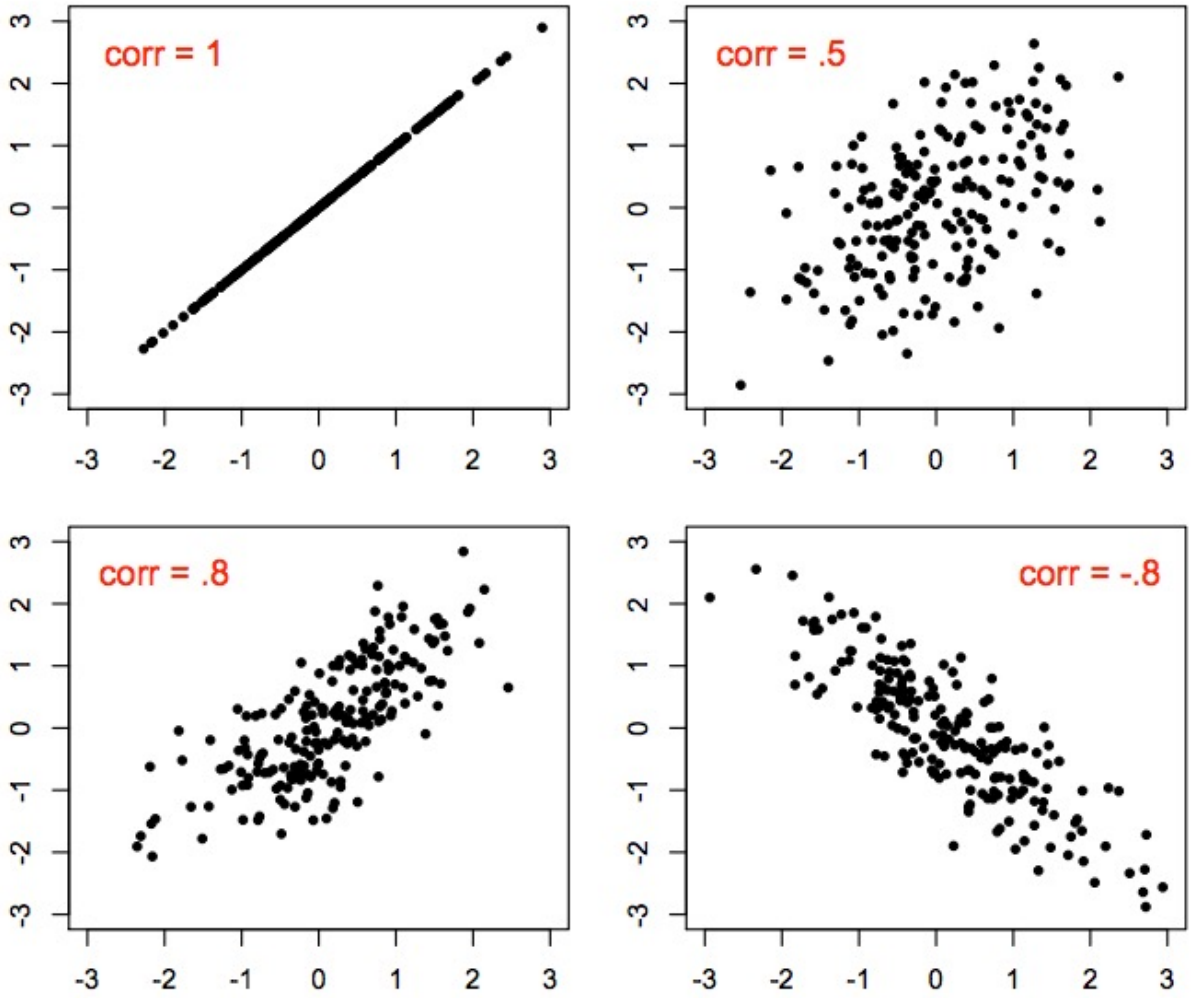
Correlation is the standardized covariance:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}$$

The correlation is scale invariant and the units of measurement don't matter: It is always true that $-1 \leq \text{corr}(X, Y) \leq 1$.

This gives the direction (- or +) and strength ($0 \rightarrow 1$) of the linear relationship between X and Y.

Correlation

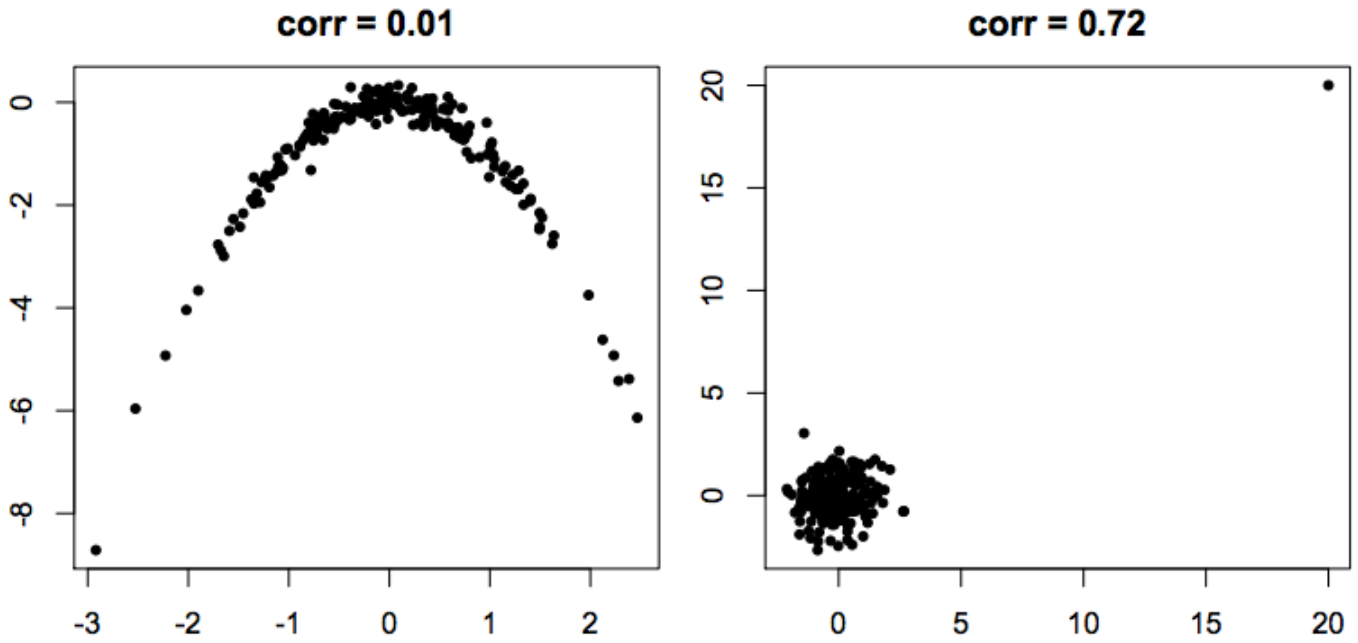


Tufts

Correlation

Only measures linear relationships:

$\text{corr}(X, Y) = 0$ does not mean the variables are not related!



Also be careful with influential observations.

Tufts

Back To Least Squares

1. Intercept:

$$b_0 = \bar{Y} - b_1\bar{X} \Rightarrow \bar{Y} = b_0 + b_1\bar{X}$$

   - The point $(\bar{X}, \bar{Y})$ is on the regression line!
   - Least squares finds the point of means and rotate the line through that point until getting the "right" slope

2. Slope:

$$b_1 = corr(X, Y) \times \frac{s_Y}{s_X}$$

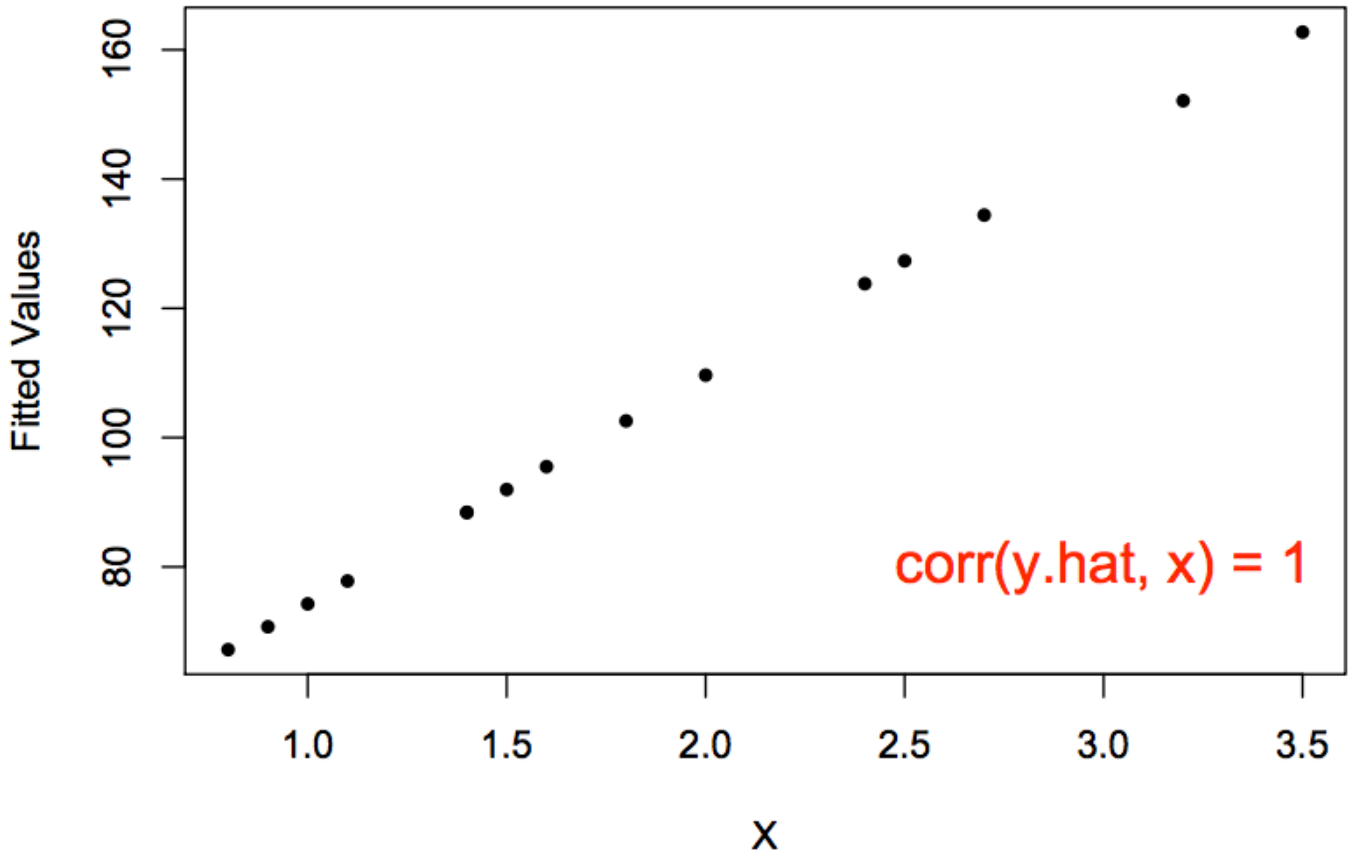
   - So, the right slope is the *correlation coefficient* times a *scaling factor* that ensures the proper units for b_1

More on Least Squares

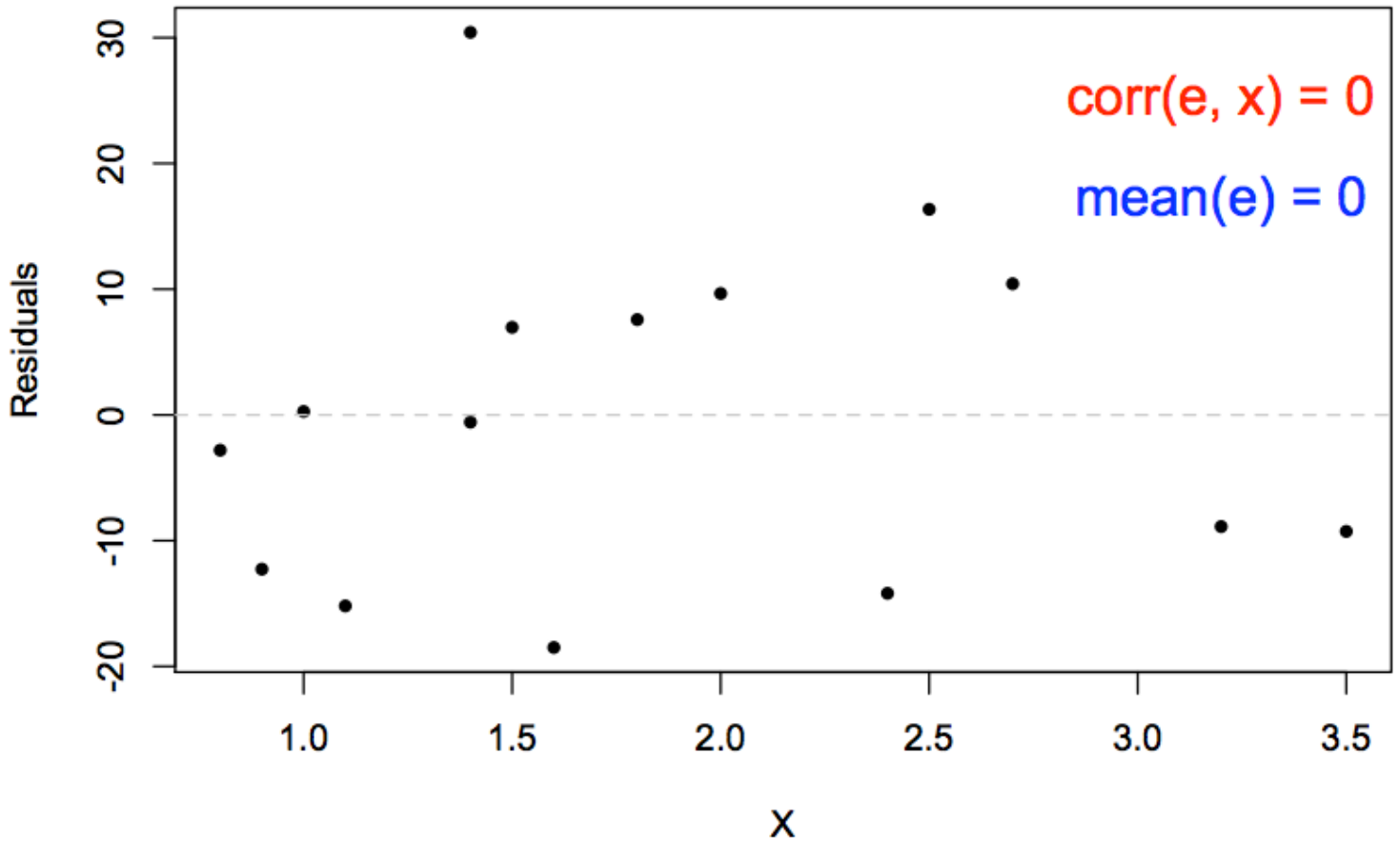
From now on, terms "fitted values" ($\hat{Y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties. Lets look at the housing data analysis to figure out what these properties are...

The Fitted Values & X



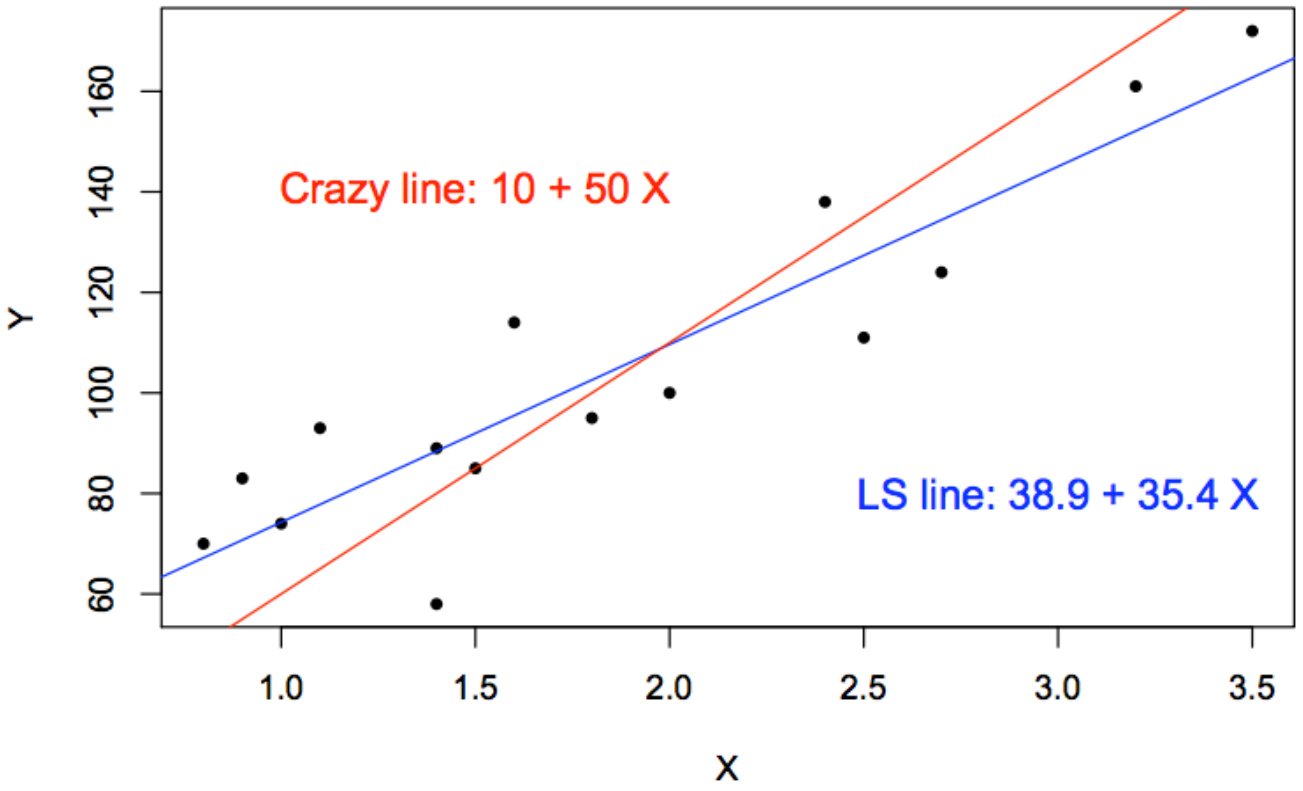
The Residuals of X



Why?

What is the intuition for the relationship between $\hat{Y}$ and e and X?

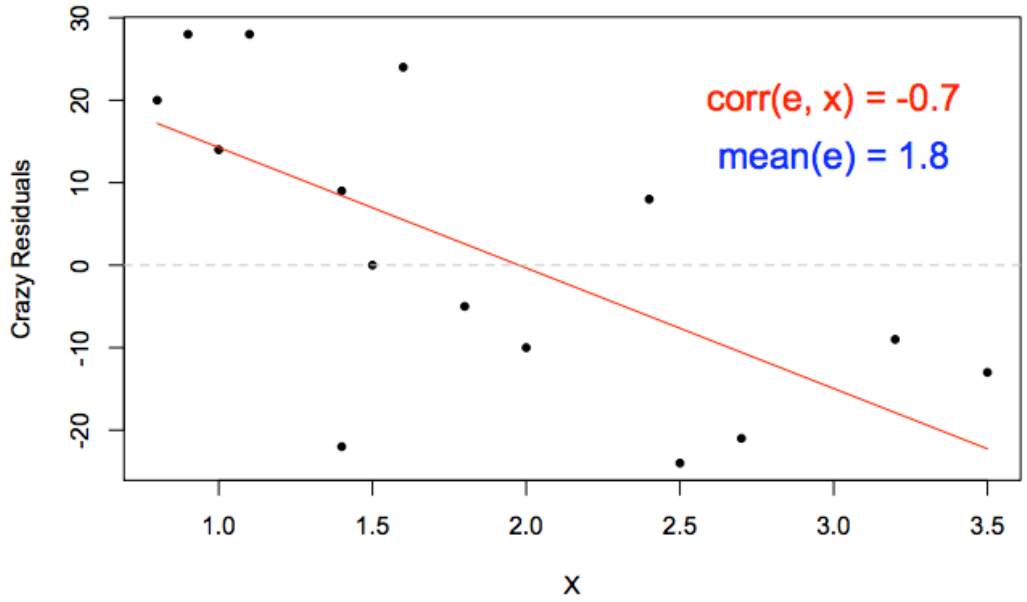
Lets consider some "crazy" alternative line:



Tufts

Fitted Values and Residuals

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses.



Clearly, we have left some predictive ability on the table!

Fitted Values and Residuals

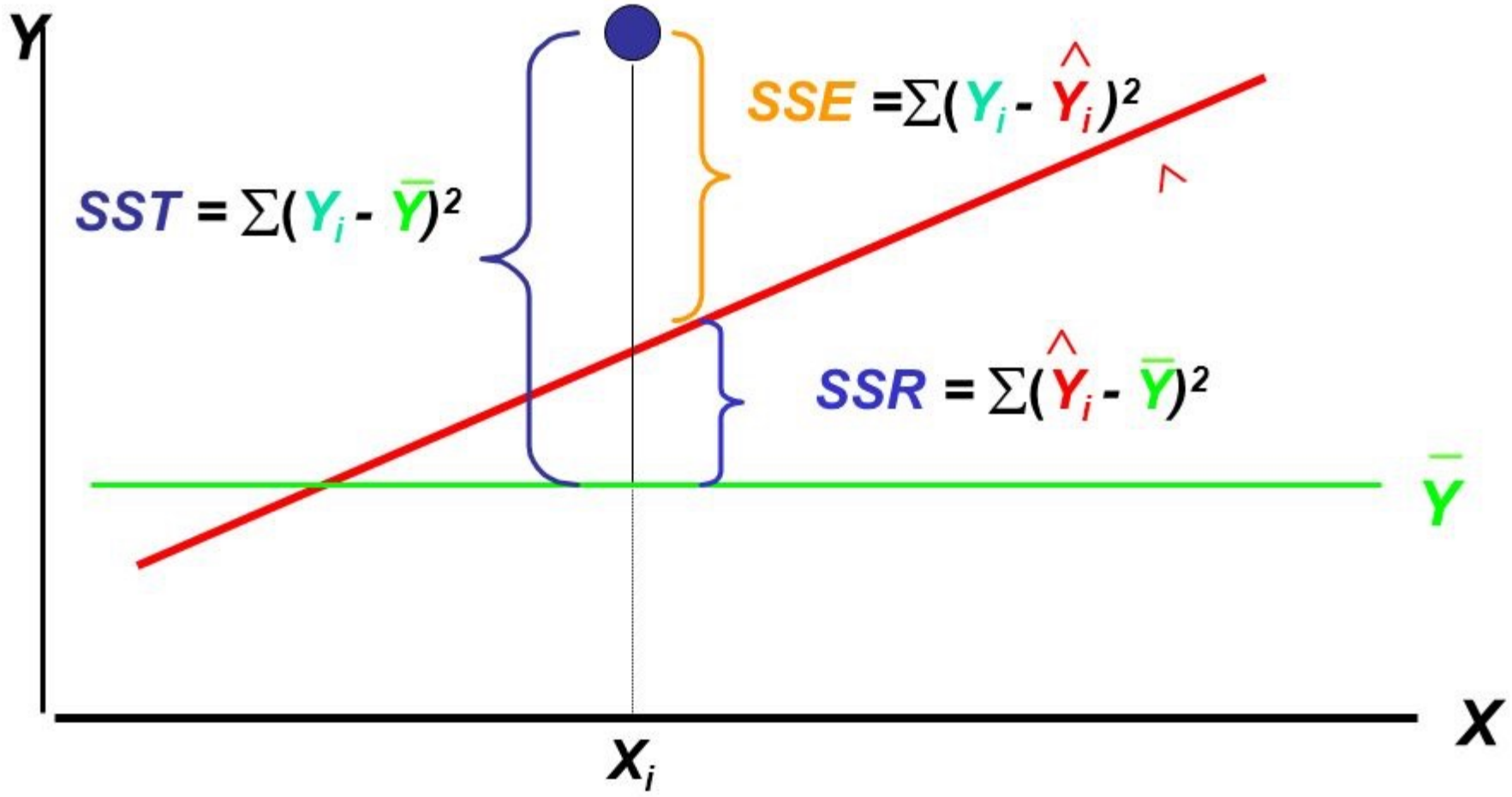
As long as the correlation between e and X is non-zero, we could always adjust our prediction rule to do better.

We need to exploit all of the predictive power in the X values and put this into $\hat{Y}$, leaving no "*Xness*" in the residuals.

In Summary: $Y = \hat{Y} + e$ where:

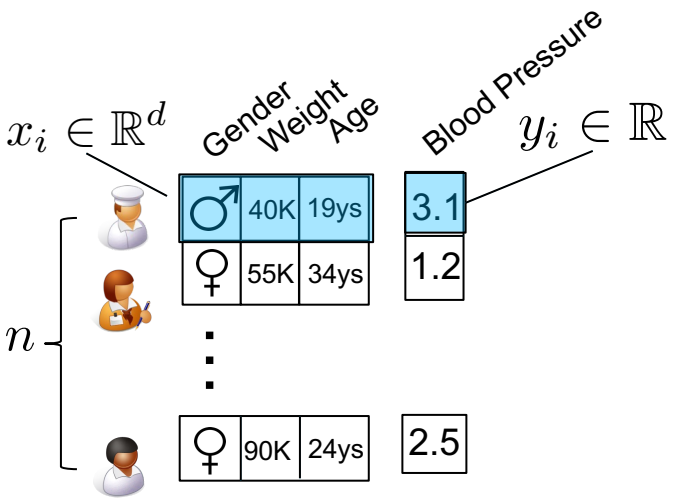
- $\hat{Y}$ is "made from X"; $\text{corr}(X, \hat{Y}) = 1$.
- e is unrelated to X; $\text{corr}(X, e) = 0$.

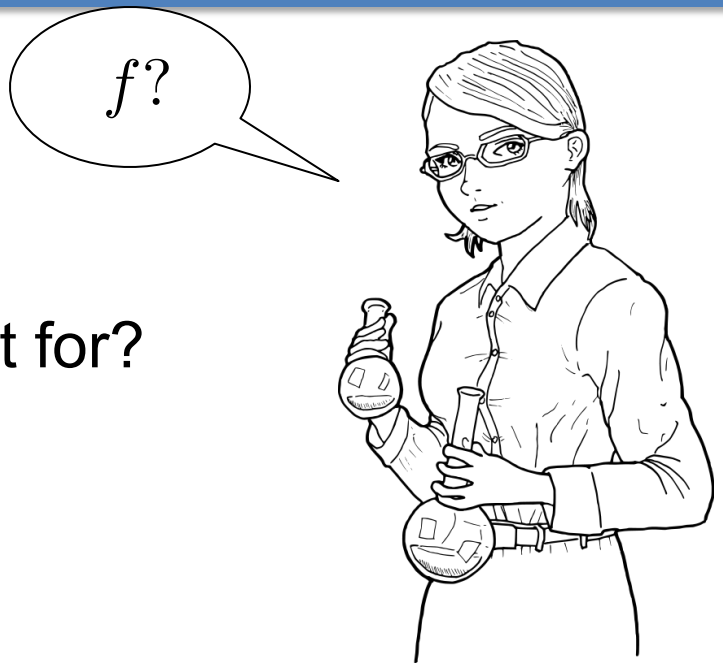
Measure Variation: The Sum of Squares



Tufts

Regression





What for?

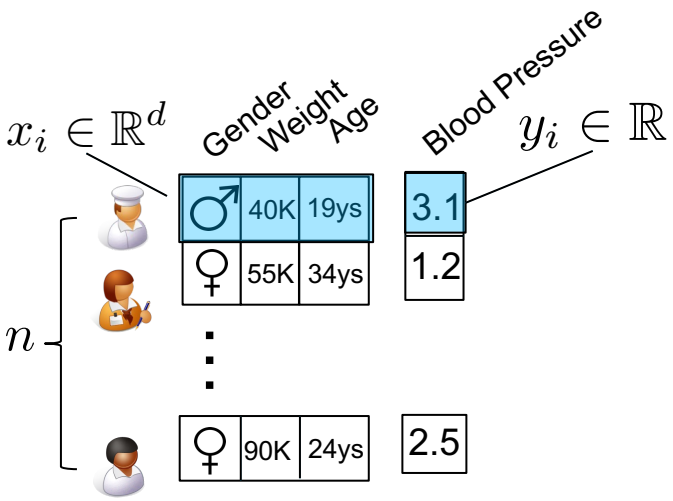
$$y_i \approx f(x_i), \quad i = 1, \ldots, n$$

for some $f : \mathbb{R}^d \to \mathbb{R}$

- Prediction: If $x =$ ♀ 70K 20ys then $y =$?
- Correlation: If Weight ↑ then y ↑

Tufts

Linear Regression



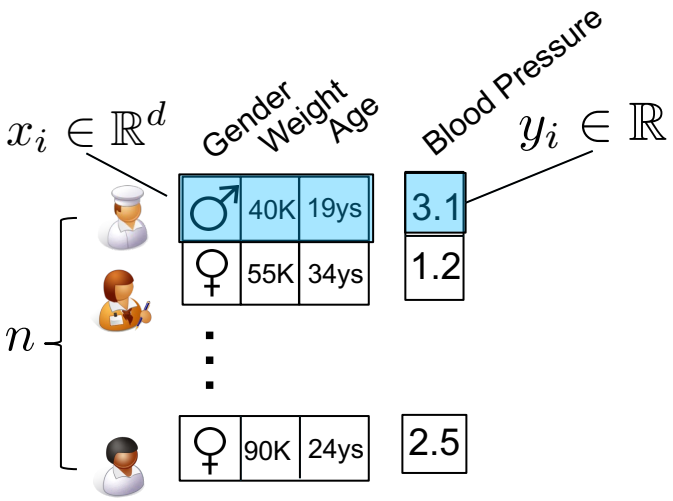
$$y_i \approx f(x_i) = \beta^\top x_i = \sum_{k=1}^{d} \beta_k x_{ik}$$

Assumption: There exists $\beta \in \mathbb{R}^d$ such that:

$$y_i = \langle \beta, x_i \rangle + \varepsilon_i, \quad i = 1, \ldots, n$$

where ε_i are i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$

Linear Regression



$$y_i \approx f(x_i), \quad i = 1, \ldots, n$$

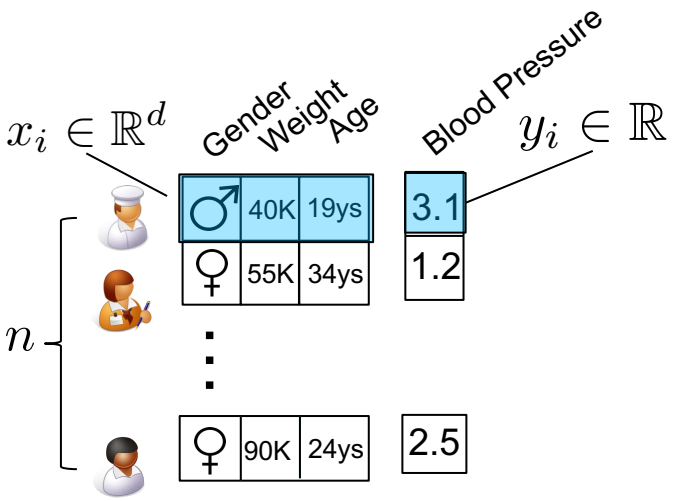
for some $f : \mathbb{R}^d \to \mathbb{R}$

$$y = f(X) + \varepsilon$$

Depends on statistical method

$$f(X) = \beta_0 + \beta_0 X$$

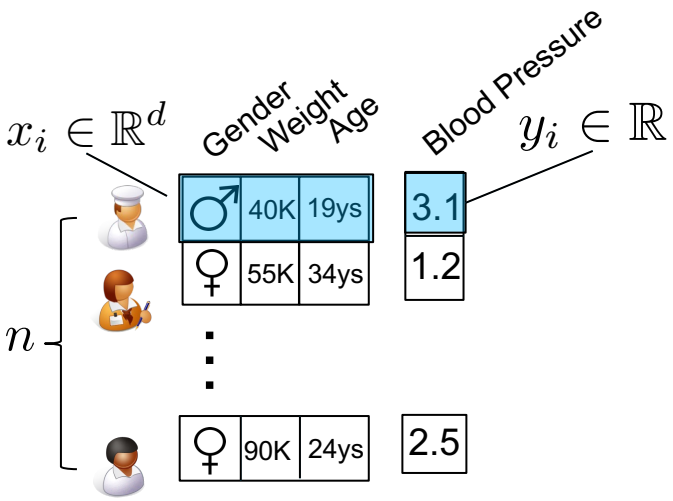
Regression: Terminology



$$y_i \approx f(x_i), \quad i = 1, \ldots, n$$

- $x_i \in \mathbb{R}^d$: features, independent variables, covariates, inputs,…
- $y_i \in \mathbb{R}$: label, dependent variable, outcome, response, output,…

Regression



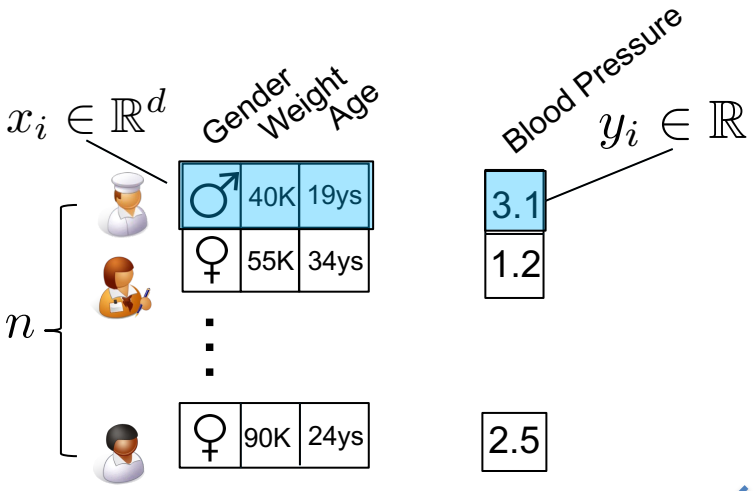
random "noise" variables

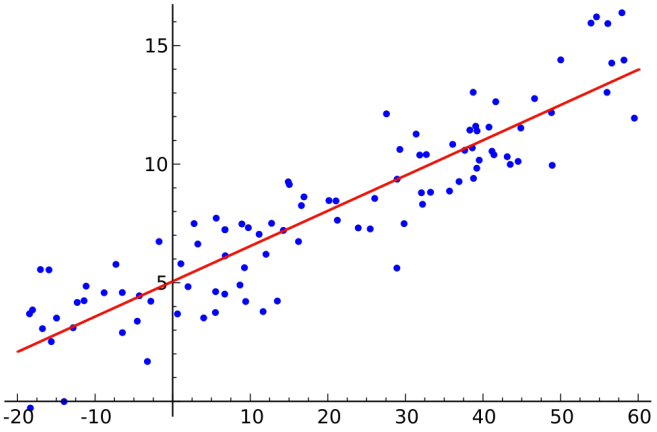
$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

where ε_i are **independent and identically distributed** (i.i.d), and

$$\mathbb{E}[\varepsilon_i] = 0 \qquad \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$$

Affine Can Be Written as Linear





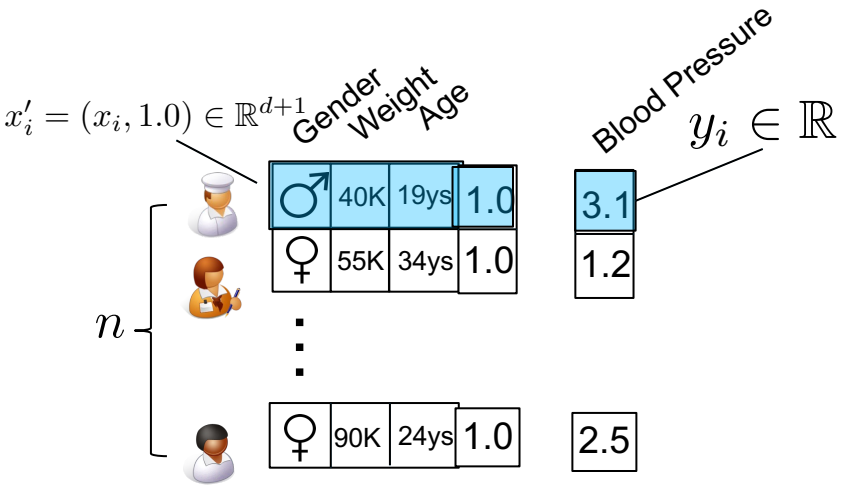
bias, offset

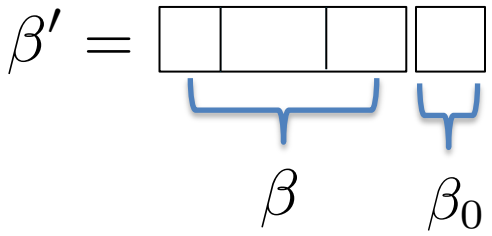
$$f(x_i) = \beta^\top x_i + \beta_0 \text{, where } \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}$$

$$= \beta'^\top x'_i, \qquad \text{where } \beta' = (\beta, \beta_0) \in \mathbb{R}^{d+1}$$

$$x'_i = (x_i, 1.0) \in \mathbb{R}^{d+1}$$

Affine Can Be Written as Linear



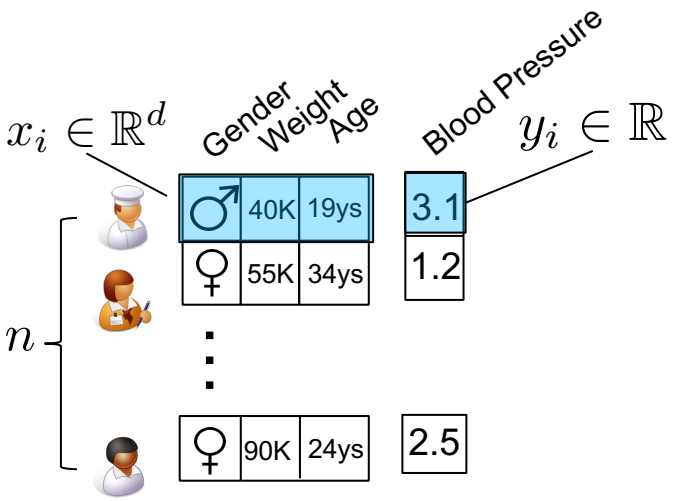


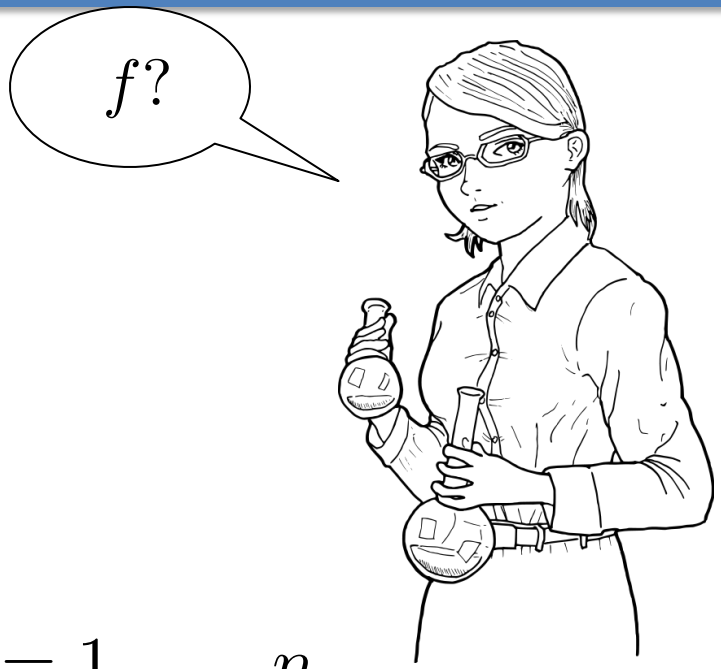
$$f(x_i) = \beta^\top x_i + \beta_0 \text{ , where } \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}$$

$$= \beta'^\top x_i', \qquad \text{where } \beta' = (\beta, \beta_0) \in \mathbb{R}^{d+1}$$

$$x_i' = (x_i, 1.0) \in \mathbb{R}^{d+1}$$

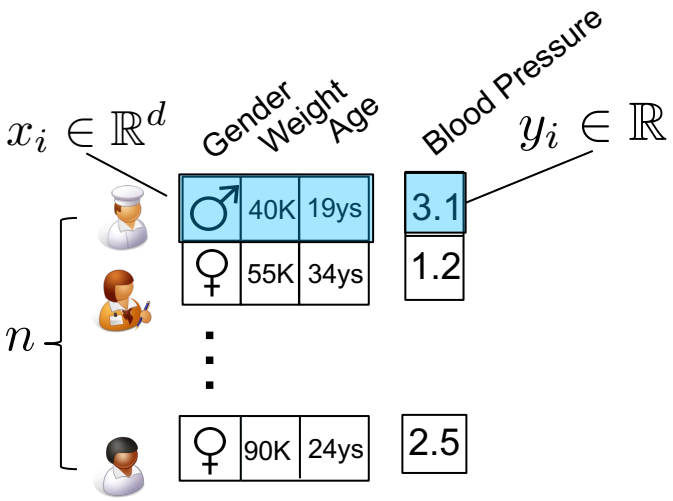
Regression: Noiseless Setting

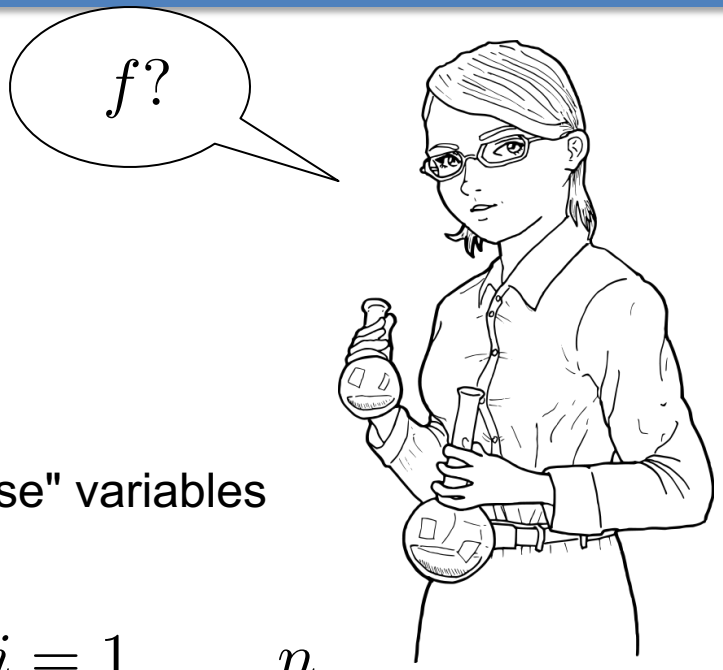




$$y_i = f(x_i), \qquad i = 1, \ldots, n$$

Regression: Noisy setting





random "noise" variables

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

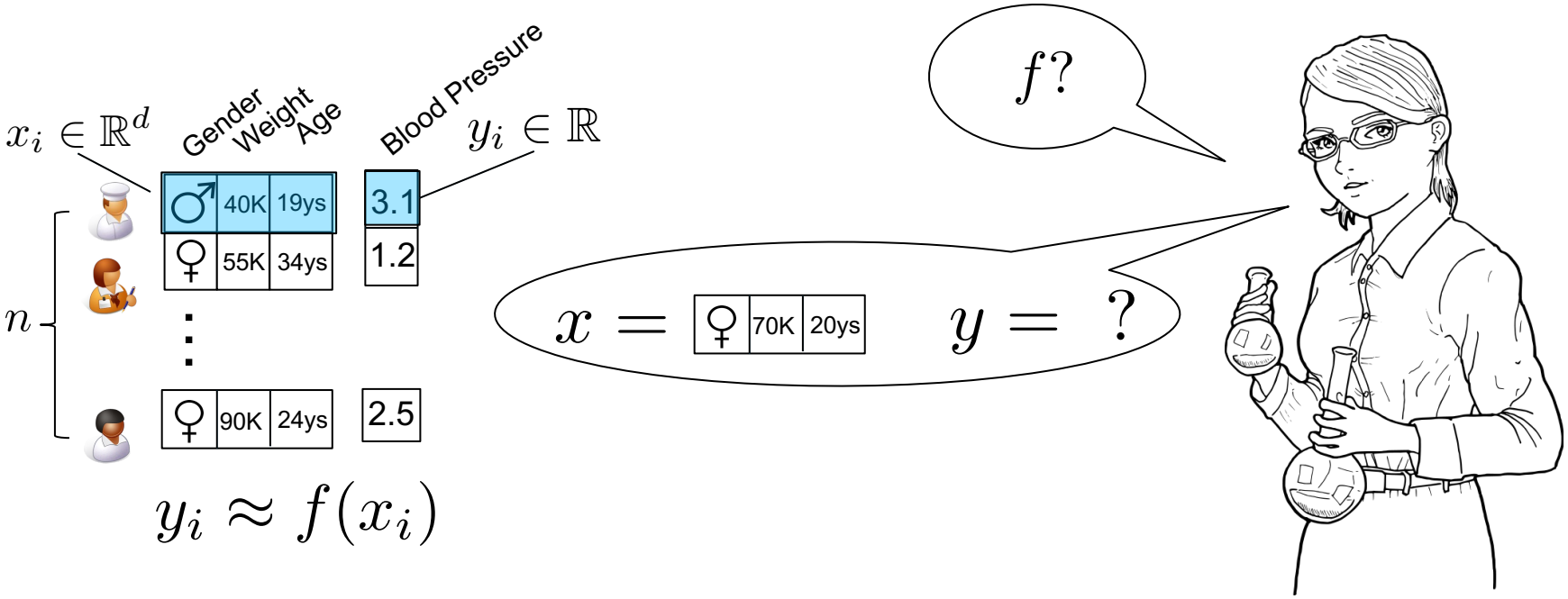
where ε_i are **independent and identically distributed** (i.i.d), and

$$\mathbb{E}[\varepsilon_i] = 0 \qquad \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$$

Note: This implies that $y_i, i = 1, \ldots, n,$ are **independent** random variables, where

$$\mathbb{E}[y_i] = f(x_i) \qquad \mathtt{Var}[y_i] = \mathbb{E}\left[(y_i - \mathbb{E}[y_i])^2\right] = \sigma^2$$

How would you solve this problem?



…you need to start making some assumptions on f!

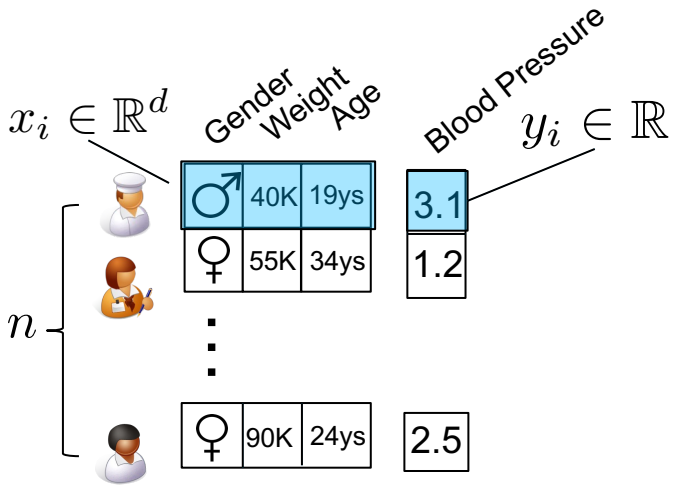
Assumption: Continuity!

- **Assumption**: Function $f : \mathbb{R}^d \to \mathbb{R}$ is **continuous**

$$\text{If } \lim_{k \to \infty} x_k = x \quad \text{then } \lim_{k \to \infty} f(x_k) = f(x)$$

- Values of f at points near x tell you something about $f(x)$!

Operations that Preserve Linearity



- Affine in $\mathbb{R}^d$ is linear in $\mathbb{R}^{d+1}$
- Linear transforms on features (i.e., rescaling):
  - E.g., from kilograms to pounds
- Affine transforms in features (i.e., rescaling and shifting):
  - E.g., from F° to C°.

Linear Regression (return)

- Linear Regression Estimate

$$\hat{y} = \widehat{f(X)}$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X$$

- Linear Regression Actual Response

$$y = f(X) + \epsilon$$

Linear Regression

- Residual Error

$$e = y - \hat{y}$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X$$

- Do not confuse residual error with irreproducible error

$$e = \beta_0 + \beta_1 X + \epsilon - (\widehat{\beta_0} + \widehat{\beta_1} X)$$

$$e = (\beta_0 - \widehat{\beta_0}) + (\beta_1 - \widehat{\beta_1}) X + \epsilon$$

Linear Regression

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$$

- Residual Error of i[th] sample

$$e_i = y_i - \hat{y}_i$$

- Sum of Squares Residual (RSS)

$$RSS = \sum_{i=1}^{n} e_i^2$$

Linear Regression

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1} X_i)\right)^2$$

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

Linear Regression

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

- Differentiating wrt β_0

$$2\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)(-1) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right) = 0$$

$$\widehat{\beta_0} = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta_1} X_i\right)}{n}$$

$$\widehat{\beta_0} = \sum_{i=1}^{n} \left(\frac{y_i}{n} - \widehat{\beta_1}\frac{X_i}{n}\right)$$

$$\widehat{\beta_0} = \sum_{i=1}^{n} \frac{y_i}{n} - \widehat{\beta_1} \sum_{i=1}^{n} \frac{X_i}{n}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Linear Regression

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

- Differentiating wrt β_1

$$2\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)(-X_i) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)(X_i) = 0$$

$$\sum_{i=1}^{n} X_i y_i - \widehat{\beta_0} \sum_{i=1}^{n} X_i - \widehat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i y_i - \left(\sum_{i=1}^{n} \frac{y_i}{n} - \widehat{\beta_1} \sum_{i=1}^{n} \frac{X_i}{n}\right) \sum_{i=1}^{n} X_i - \widehat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i y_i - \sum_{i=1}^{n} \frac{y_i}{n} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} \frac{X_i}{n} \sum_{i=1}^{n} X_i - \widehat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} X_i - \widehat{\beta_1} \times \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i - \widehat{\beta_1} \sum_{i=1}^{n} X_i^2 = 0$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Linear Regression

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$$

Least Squares Criteria

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X$$

Car Sales

Number of Ads Run

Multiple Regression

- General Parametric Equation

$$\hat{y} = f(\boldsymbol{X}) + \epsilon$$

Depends on statistical method

$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + + \beta_p X_p$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p$$

- Number of operations $=n\times(p-1)^2$, where n is the number of samples

Multiple Regression

$$n\ x\ 1 \qquad n\ x\ (p+1) \qquad (p+1)x1 \qquad n\ x\ 1$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & . & . & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & . & . & . \\ 1 & X_{3,1} & X_{3,2} & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & X_{n,1} & X_{n,2} & . & . & X_{n,p} \end{bmatrix} , \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} , \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$$y = X\beta + \epsilon \qquad \hat{y} = X\hat{\beta}$$

Multiple Regression

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \widehat{y_1} \\ y_2 - \widehat{y_2} \\ y_3 - \widehat{y_3} \\ . \\ . \\ . \\ y_n - \widehat{y_n} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} - \begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \widehat{y_3} \\ . \\ . \\ . \\ \widehat{y_n} \end{bmatrix} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

$$RSS = \sum_{i=1}^{n} e_i^2 \longrightarrow RSS = e^T e$$

Multiple Regression

$$RSS = e^T e$$

$$RSS = (y - \hat{y})^T (y - \hat{y})$$

$$RSS = (y - X\hat{\beta})^T (y - \widehat{X\hat{\beta}})$$

$$RSS = (y - \hat{\beta}^T X^T)\ (y - \widehat{X\hat{\beta}})$$

$$RSS = y^T y^T - y^T X\hat{\beta} - \hat{\beta} X^T y + \hat{\beta}^T X^T X\hat{\beta}$$

Multiple Regression

Matrix Differentiation

$x = m \times 1$ *matrix*
$A = n \times m$ *matrix;* $A \perp x$

$$\boldsymbol{y = A \rightarrow \frac{\delta y}{\delta x} = 0}$$

$$\boldsymbol{y = Ax \rightarrow \frac{\delta y}{\delta x} = A}$$

$$\boldsymbol{y = xA \rightarrow \frac{\delta y}{\delta x} = A^T}$$

$$\boldsymbol{y = x^T Ax \rightarrow \frac{\delta y}{\delta x} = 2x^T A}$$

$$RSS = \boldsymbol{y^T y - y^T X\hat{\beta} - \hat{\beta} X^T y + \hat{\beta}^T X^T X\hat{\beta}}$$

$$\frac{\delta(RSS)}{\delta\hat{\beta}} = \frac{\delta\left(\boldsymbol{y^T y - y^T X\hat{\beta} - \hat{\beta} X^T y + \hat{\beta}^T X^T X\hat{\beta}}\right)}{\delta\hat{\beta}} = 0$$

$$\frac{\delta(\boldsymbol{y^T y})}{\delta\hat{\beta}} - \frac{\delta(\boldsymbol{y^T X\hat{\beta}})}{\delta\hat{\beta}} - \frac{\delta(\boldsymbol{\hat{\beta} X^T y})}{\delta\hat{\beta}} + \frac{\delta(\boldsymbol{\hat{\beta}^T X^T X\hat{\beta}})}{\delta\hat{\beta}} = 0$$

$$\boldsymbol{0 - y^T X - (X^T y)^T + 2\hat{\beta}^T X^T X = 0}$$

$$\boldsymbol{2\hat{\beta}^T X^T X = 2y^T X} \rightarrow \boldsymbol{\hat{\beta}^T X^T X = y^T X}$$

$$\boldsymbol{\hat{\beta}^T = y^T X(X^T X)^{-1}} \rightarrow \boxed{\boldsymbol{\hat{\beta} = (X^T X)^{-1} X^T y}}$$

Multiple Regression

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$$

Least Squares Criteria

High Bias Problem

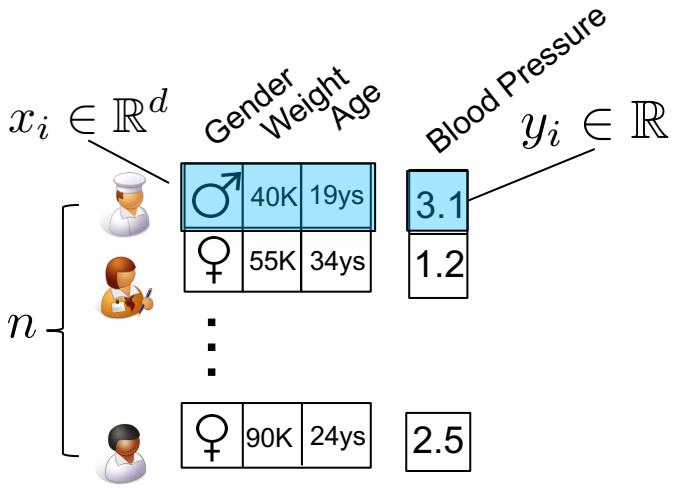
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{y} = \widehat{f(\boldsymbol{X})} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Car Sales

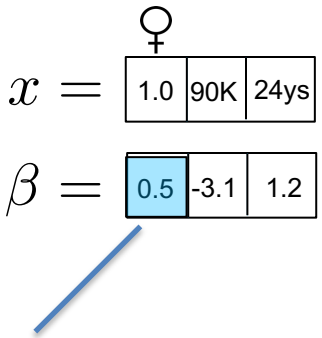
Number of Ads Run
Number of Salesman
.
:
feature$_p$

Binary Features



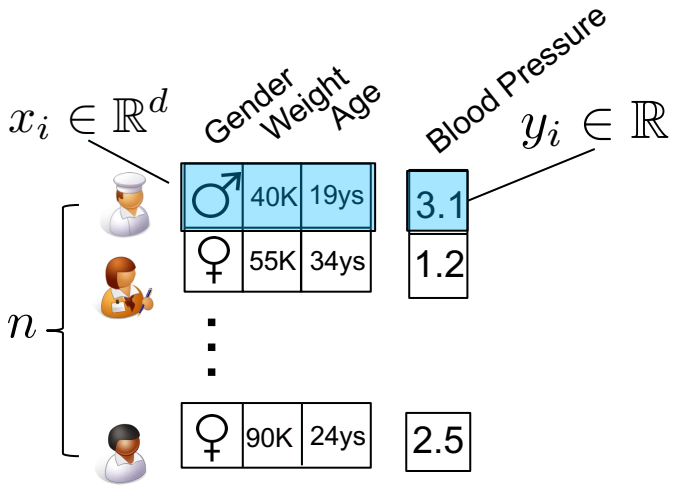
♂ = 0.0

♀ = 1.0



Tufts

Binary Features



♂ = 0.0

♀ = 1.0

$x =$ ♀ | 1.0 | 90K | 24ys

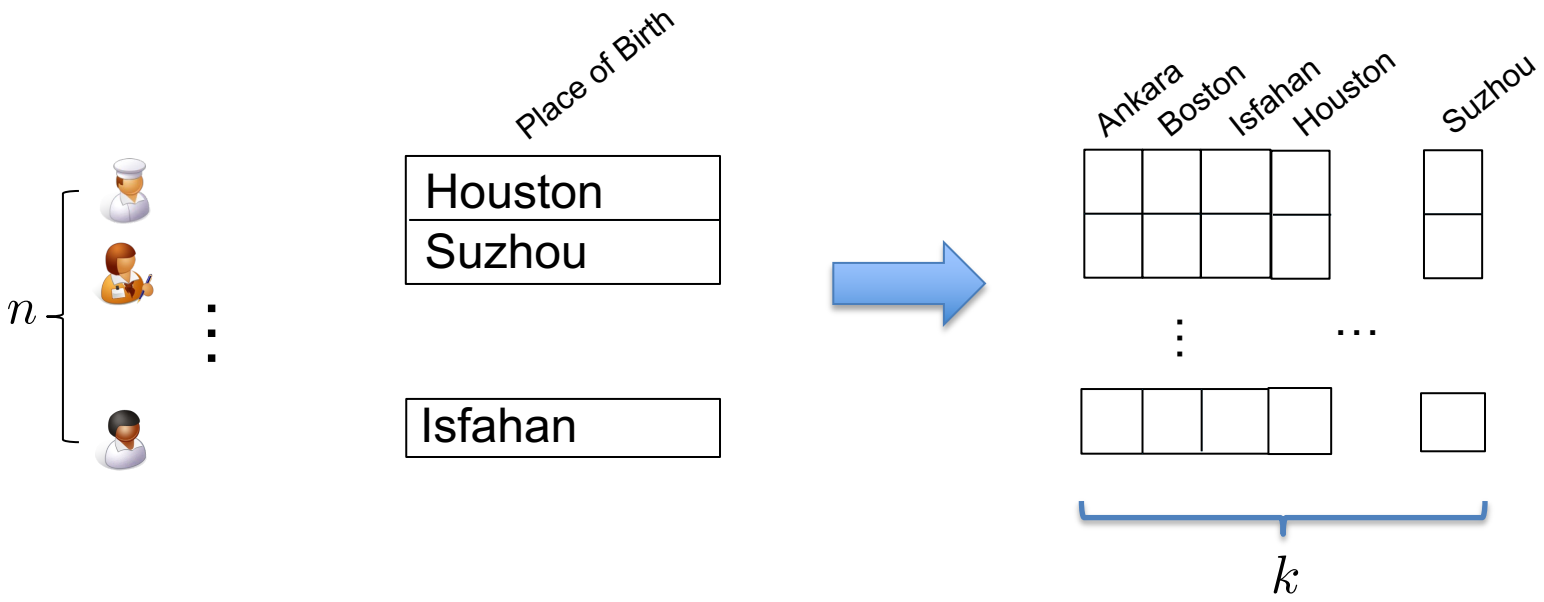
$x' =$ ♂ | 0.0 | 90K | 24ys

$\beta =$ | 0.5 | -3.1 | 1.2

gender bias

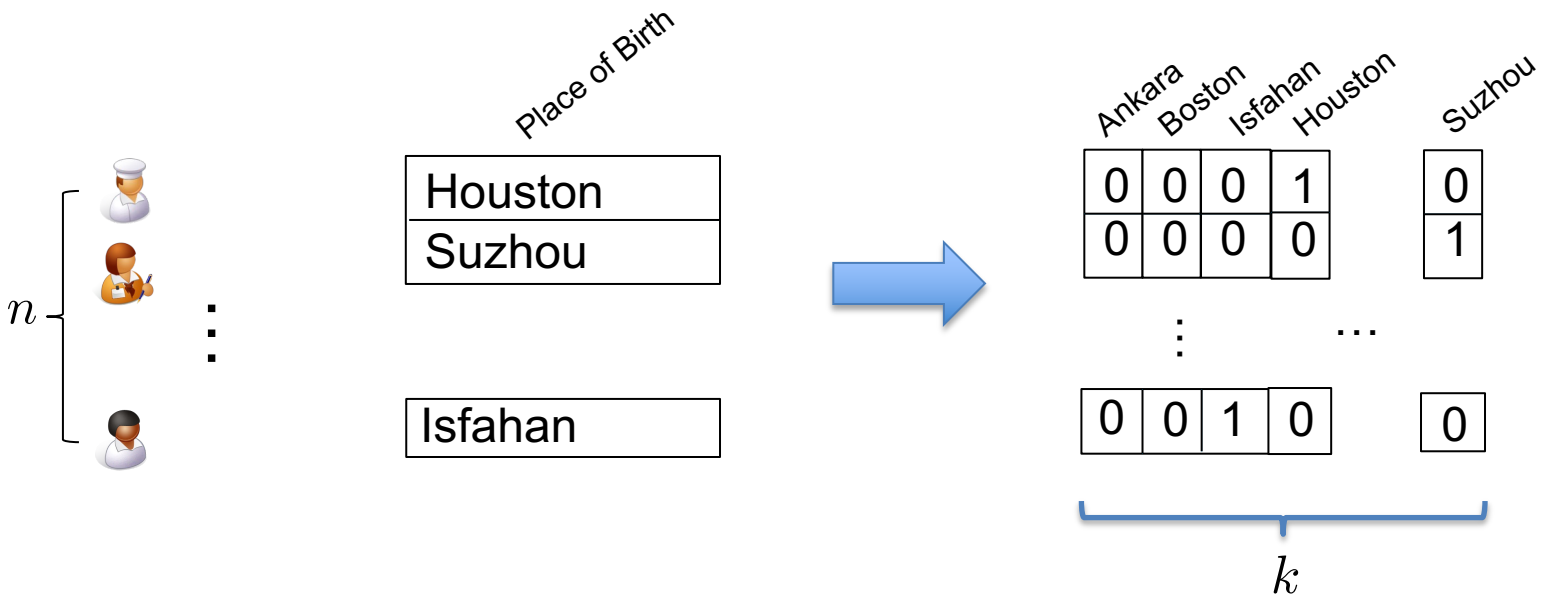
$$\mathbb{E}[y - y'] = \beta^\top x - \beta^\top x' = 0.5$$

Categorial Features: Binarization



Categories (cities) = k

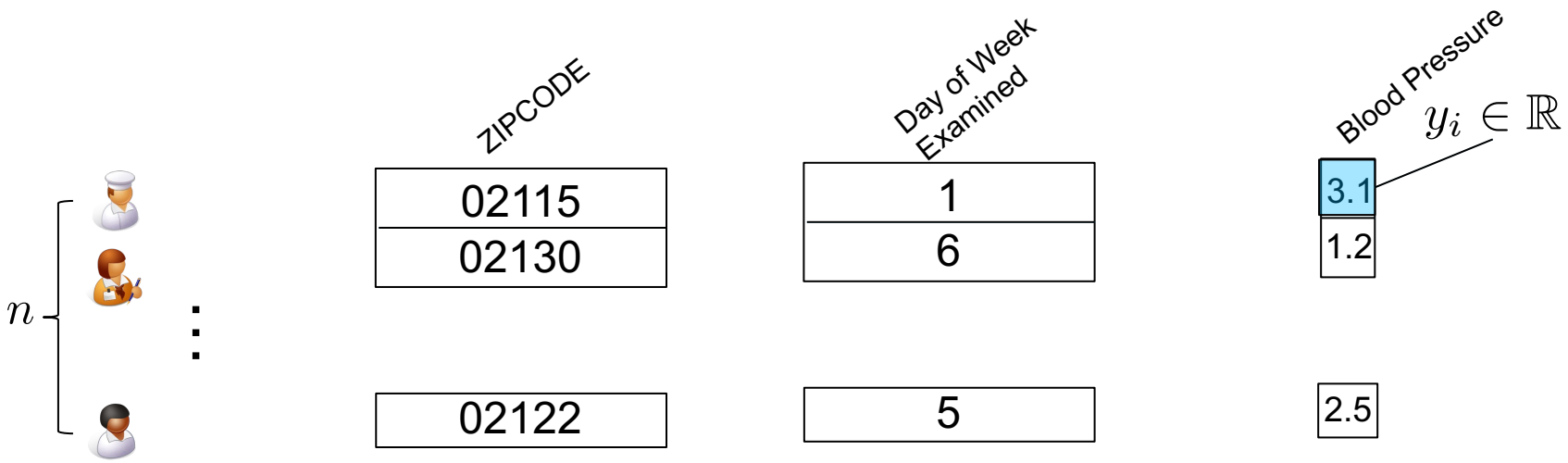
Categorial Features: Binarization



Categories (cities) = k

- Categorical features are very common: locations, genes, words in document
- Binarization leads to feature vectors that are **sparse:** most elements are 0!

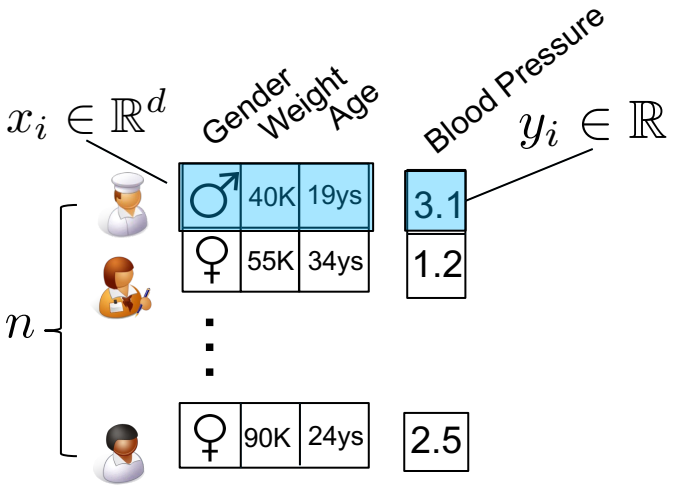
Numeric Features May Be Categorical!!!!!



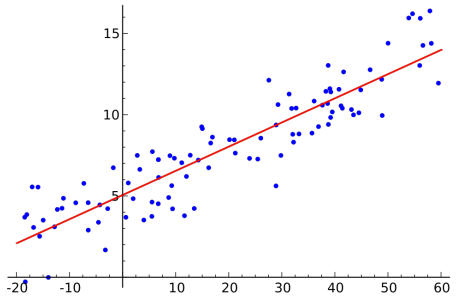
Mon: 1
Tue: 2
…
Sun: 7

Rule of thumb: if 2 does not mean "2 times" 1, treat it as categorical

Least Squares Estimator (LSE)



$$y_i \approx f(x_i) = \beta^\top x_i = \sum_{k=1}^{d} \beta_k x_{ik}$$

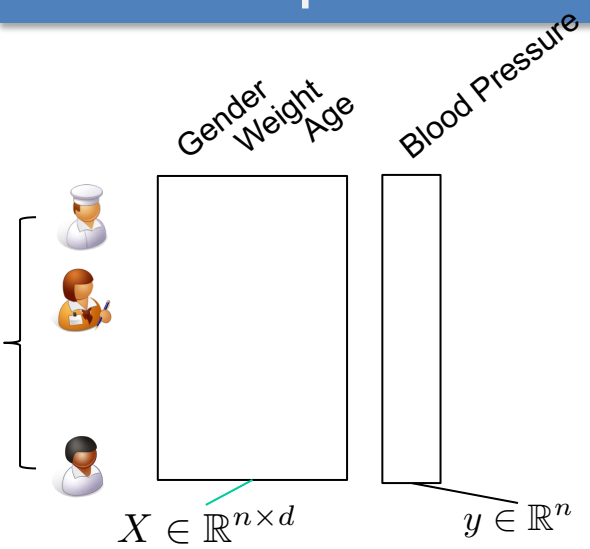


Estimate of β

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \langle \beta, x_i \rangle)^2$$

Tufts

Least Squares Estimator (LSE)

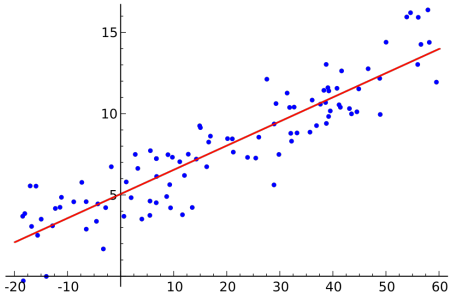


Why LSE?

$$y_i \approx f(x_i) = \beta^\top x_i = \sum_{k=1}^{d} \beta_k x_{ik}$$

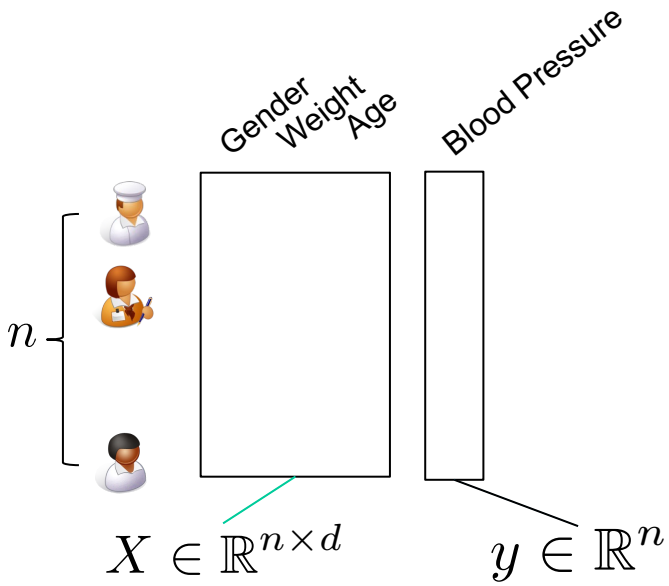
Estimate of β

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \langle \beta, x_i \rangle)^2$$
$$= \arg\min_{\beta \in \mathbb{R}^d} \|X\beta - y\|_2^2$$
$$= (X^T X)^{-1} X^T y$$



Tufts

Reason: If Noise is Gaussian, LSE is an MLE!



$$y_i = \beta^\top x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\varepsilon_i \text{ i.i.d., } \mathbb{E}[\varepsilon_i] = 0, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$$

- Suppose, in addition, that

$$\varepsilon_i \sim N(0, \sigma^2)$$

Then, the negative log-likelihood of the labels is:

$$-\log\left(P\left(y|\beta, X\right)\right) = -\log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \beta^\top x_i)^2/2\sigma^2}\right)$$

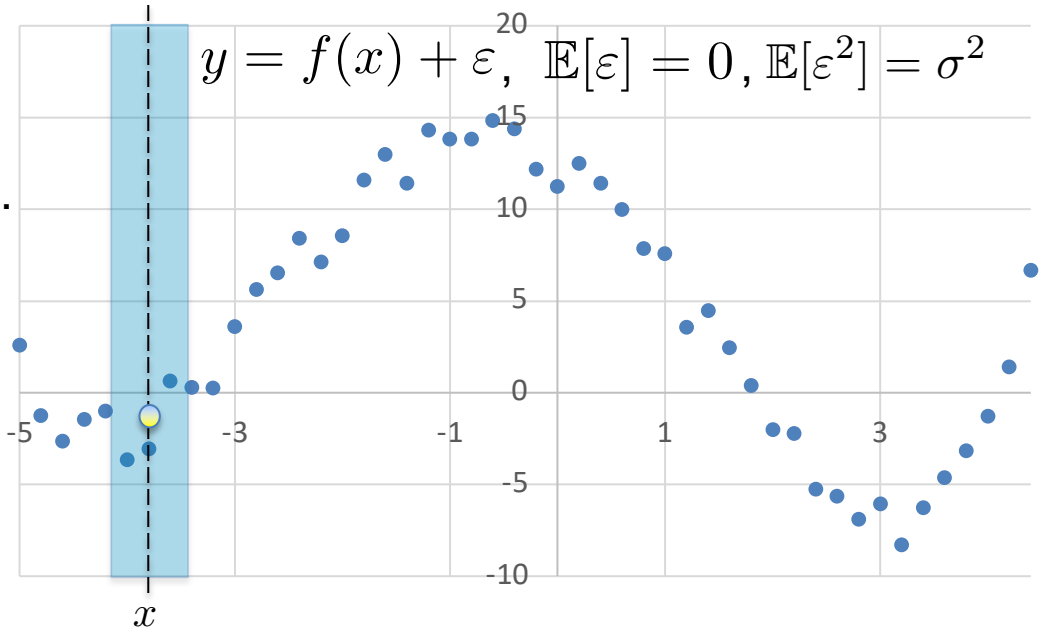
$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^\top x_i)^2 + C$$

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

$$\varepsilon_i \text{ i.i.d., } \mathbb{E}[\varepsilon_i] = 0\,, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$$

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

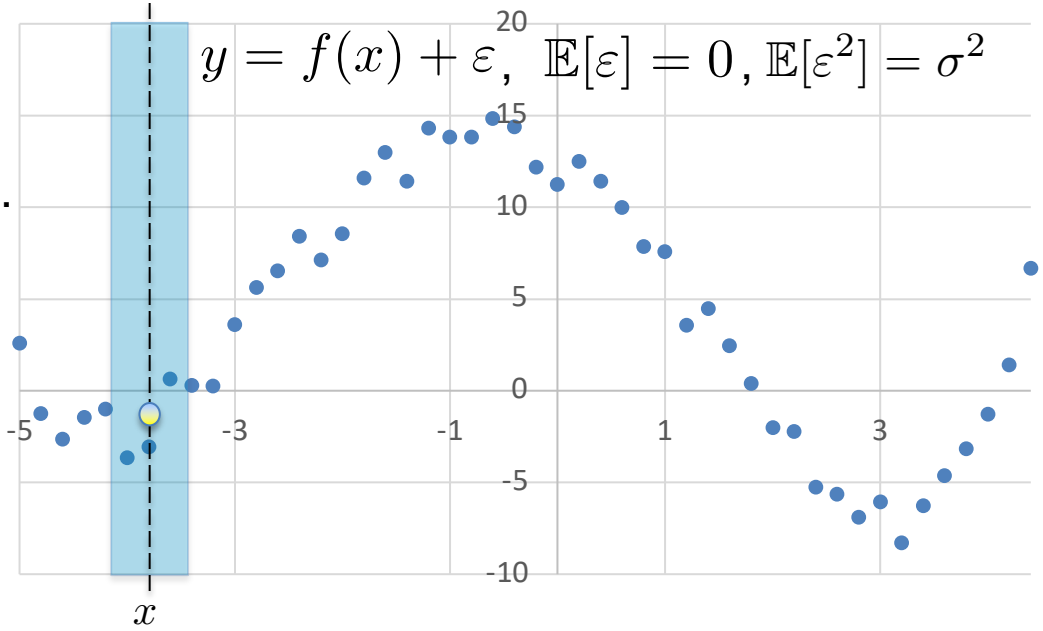
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$.

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

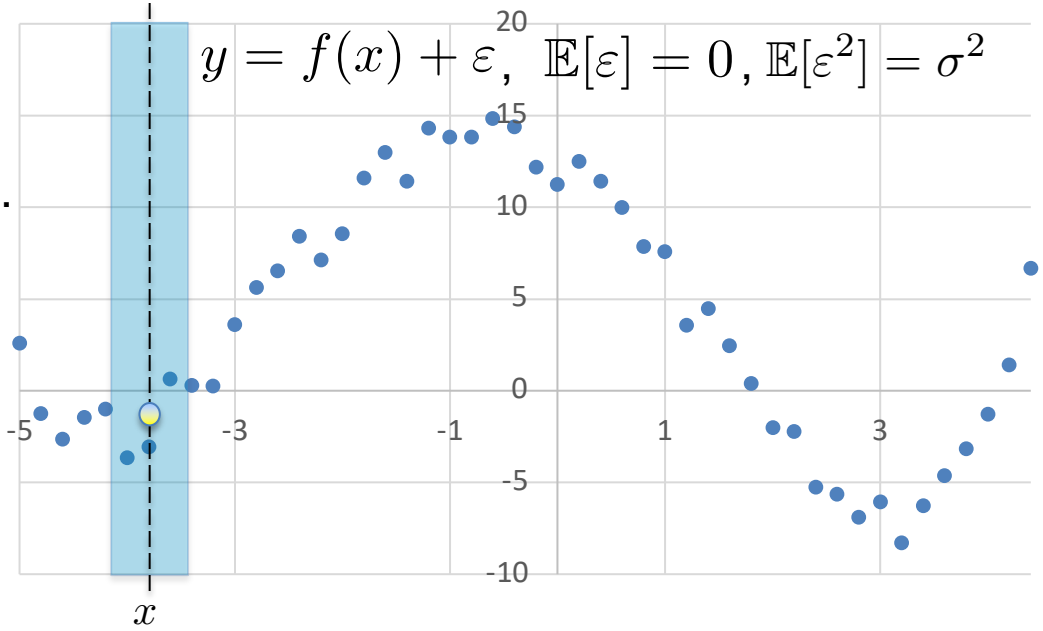
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \underbrace{\mathbb{E}\left[(y - \mathbb{E}[y])^2\right]}_{\text{inherent noise}} + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$.

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

estimator **bias**

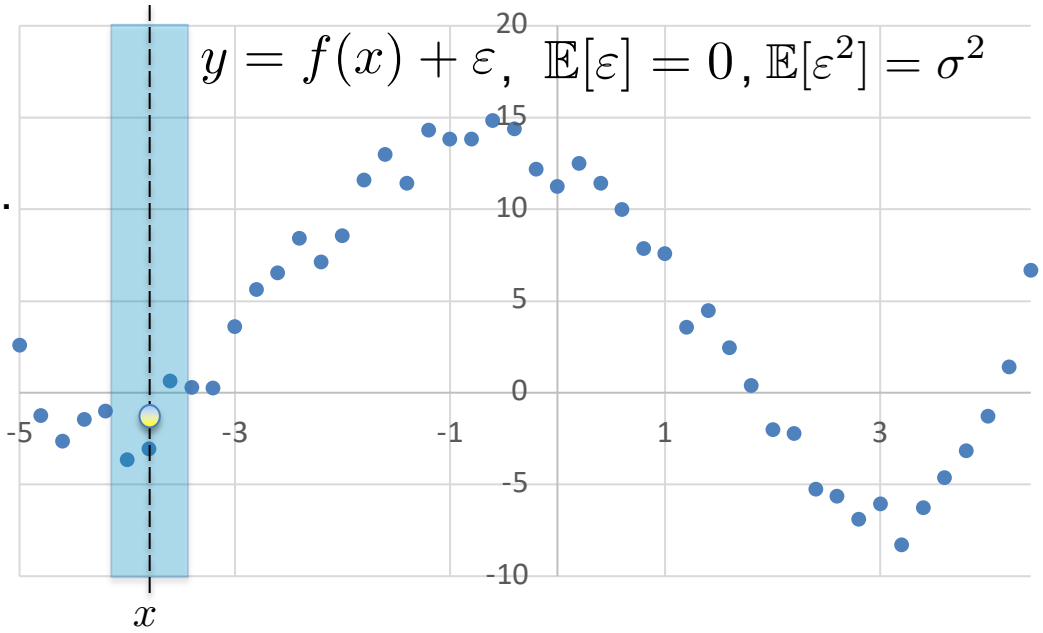
Tufts

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$.

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

estimator **variance**

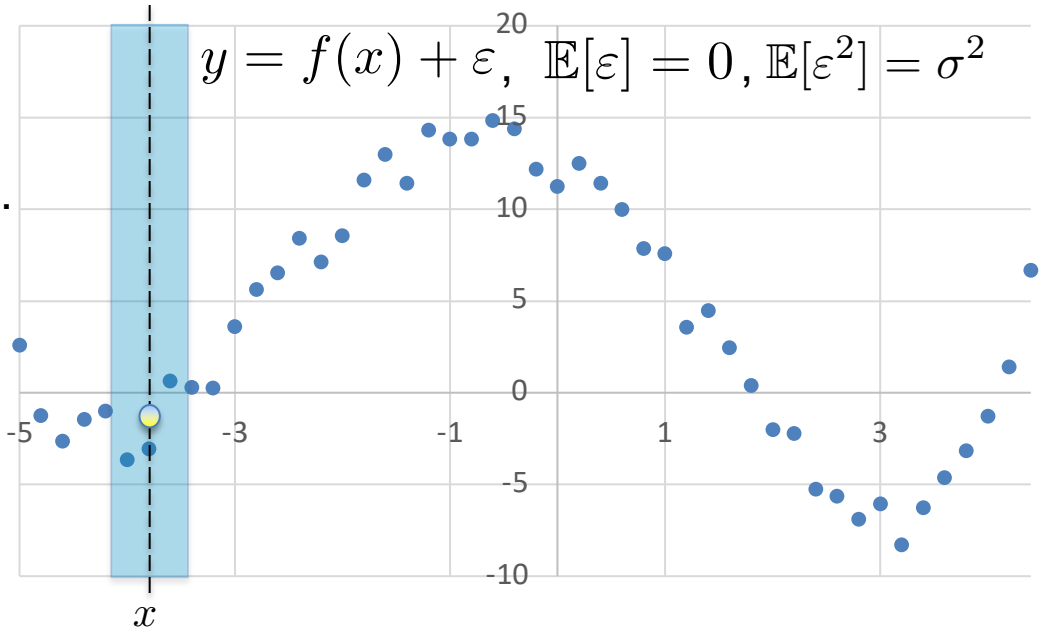
Tufts

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

$$\varepsilon_i \text{ i.i.d., } \mathbb{E}[\varepsilon_i] = 0\,, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$$

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

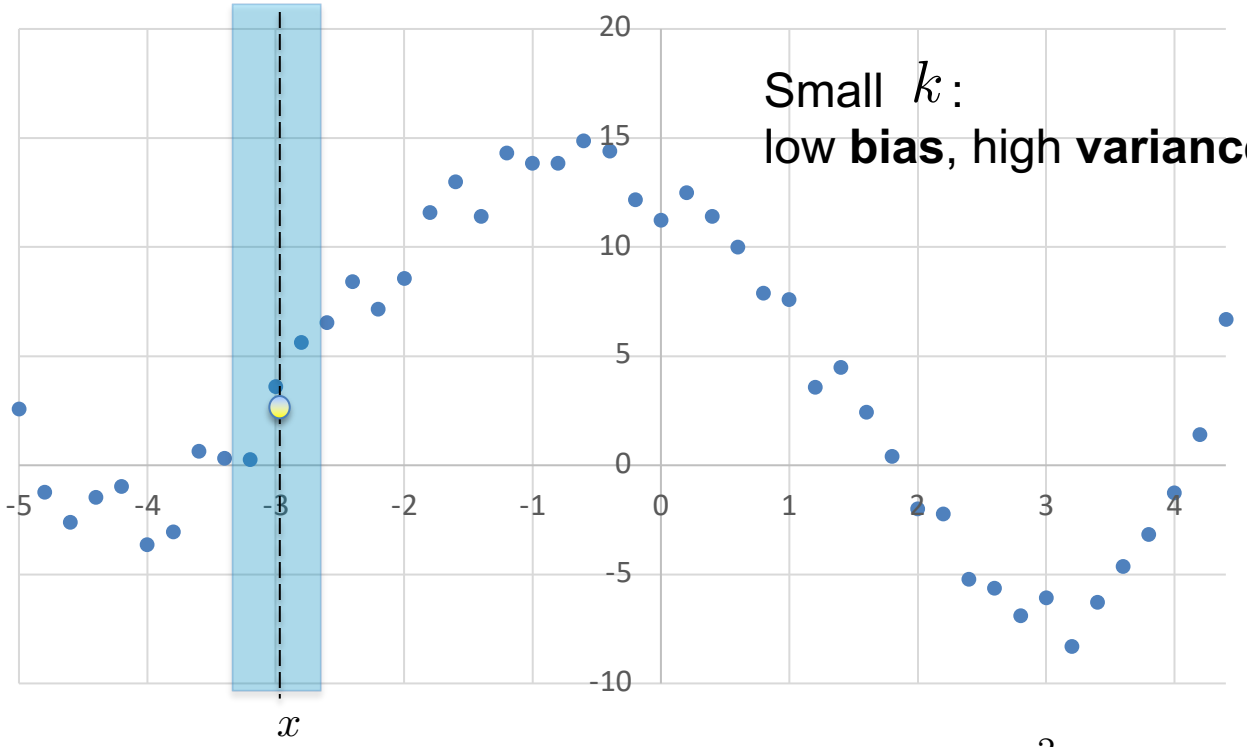


Expected Prediction Error (EPE):

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

$$= \sigma^2 + \left(f(x) - \frac{1}{k} \sum_{i \in N_k(x)} f(x_i)\right)^2 + \frac{\sigma^2}{k}$$

Bias vs. Variance Trade-off

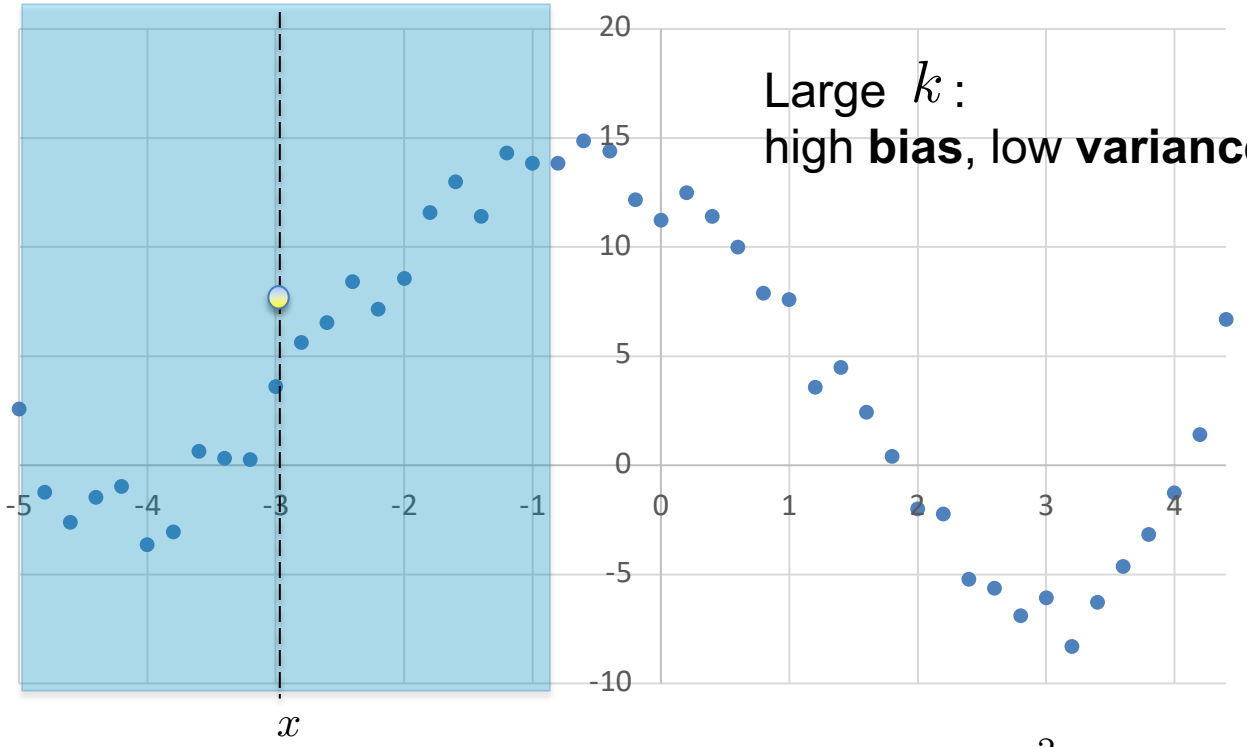


EPE: $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \sigma^2 + \left(f(x) - \frac{1}{k}\sum_{i \in N_k(x)} f(x_i)\right)^2 + \frac{\sigma^2}{k}$

estimator **bias** estimator **variance**

Tufts

Bias vs. Variance Trade-off



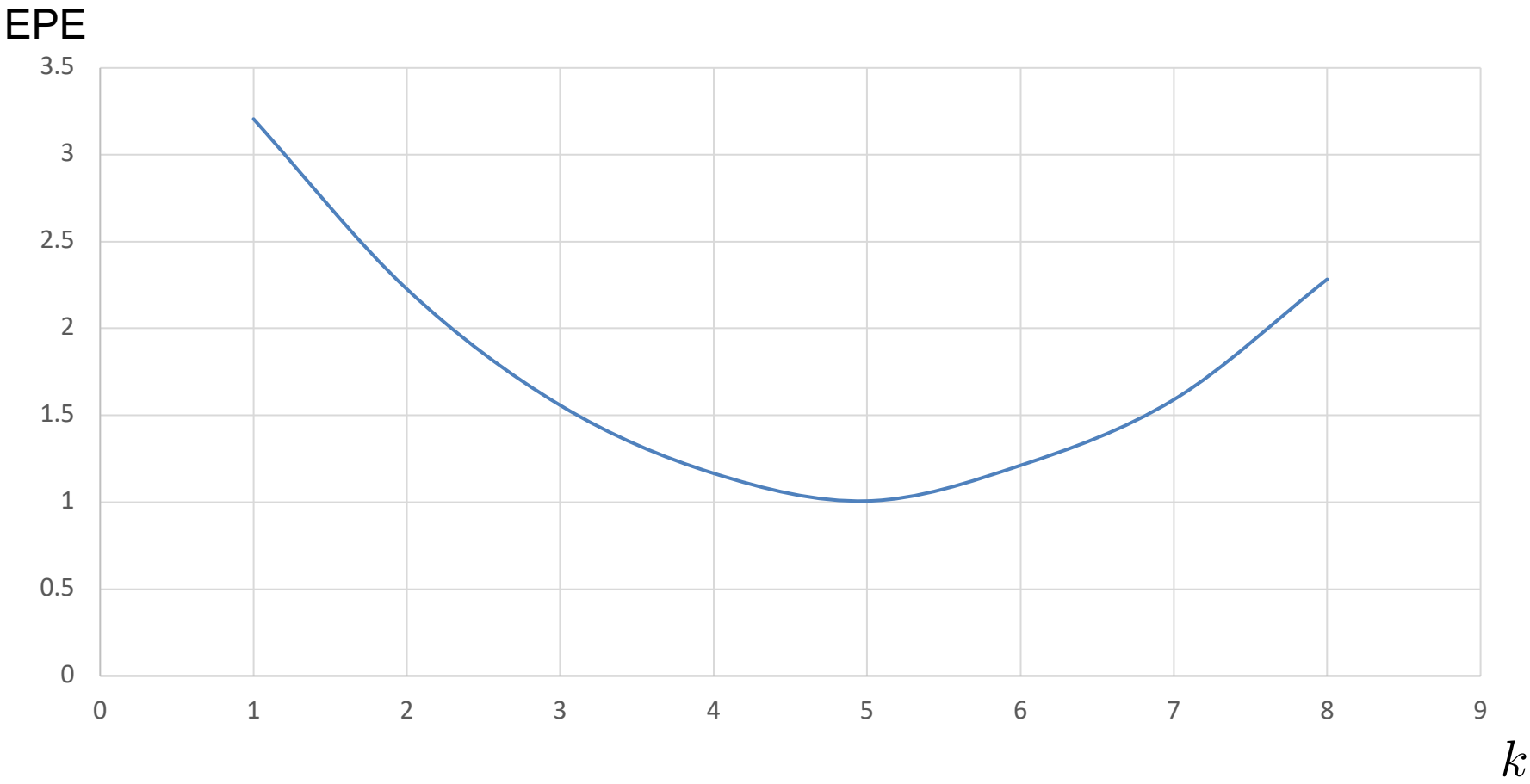
Large k:
high **bias**, low **variance**

EPE: $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \sigma^2 + \left(f(x) - \frac{1}{k}\sum_{i \in N_k(x)} f(x_i)\right)^2 + \frac{\sigma^2}{k}$

estimator **bias** estimator **variance**

Tufts

Bias vs. Variance Tradeoff



Tufts