CS 559: Machine Learning Fundamentals and Applications
8^{th} Set of Notes

Instructor: Philippos Mordohai
Webpage: www.cs.stevens.edu/~mordohai
E-mail: Philippos.Mordohai@stevens.edu
Office: Lieb 215

Project Proposal

- Dataset
  - How many instances
- Classes (or what is being predicted)
- Inputs
  - Include feature extraction, if needed
  - If your inputs are images or financial data, this must be addressed
- Methods
  - At least one simple classifier (MLE with Gaussian model, Naïve Bayes, kNN)
  - At least one advanced classifier (SVM, Boosting, Random Forest, CNN)

Project Proposal

- Typical experiments
  - Measure benefits due to advanced classifier compared to simple classifier
  - Compare different classifier settings
    - k in kNN
    - Different SVM kernels
    - AdaBoost vs. cascade
    - Different CNN architectures
  - Measure effects of amount of training data available
  - Evaluate accuracy as a function of the degree of dimensionality reduction using PCA

Project Proposal

- Email me a pdf with all these
- I must say “approved” in my response, otherwise address my comments and resubmit

Overview

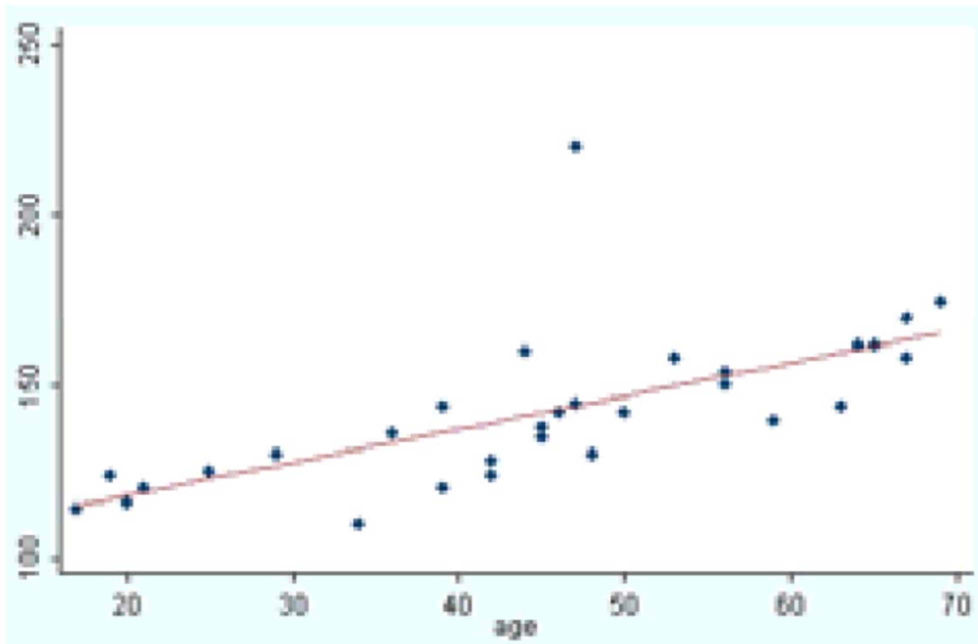
- Linear Regression
  - Barber Ch. 17
  - HTF Ch. 3

Simple Linear Regression

- How does a single variable of interest relate to another (single) variable?
  - Y = outcome variable (response, dependent...)
  - X = explanatory variable (predictor, feature, independent...)
- Data: n pairs of continuous observations $(X_1,Y_1) \dots (X_n,Y_n)$

Example

- How does systolic blood pressure (SBP) relate to age?

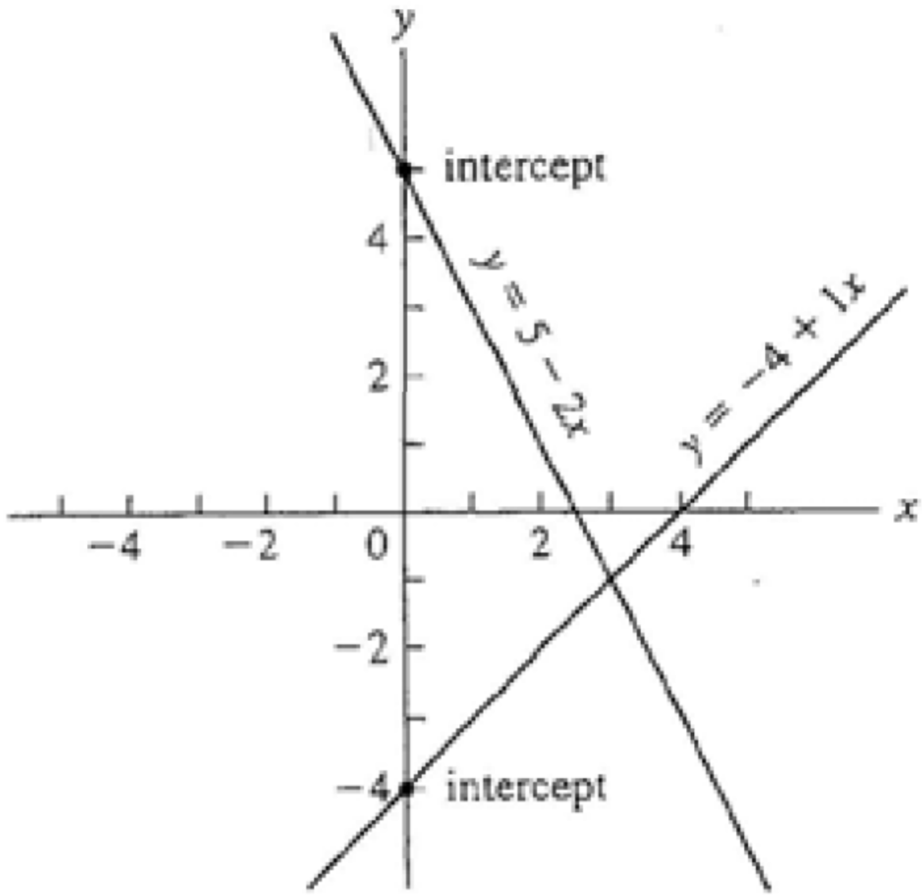


- Graph suggests that Y relates to X in an approximately linear way

Regression: Step by Step

1. Assume a linear model: $Y = \beta_0 + \beta_1 X$
2. Find the line which "best" fits the data, i.e. estimate parameters β_0 and β_1
3. Does variation in X help describe variation in Y ?
4. Check assumptions of model
5. Draw inferences and make predictions

Straight-line Plots



Assumptions of Linear Regression

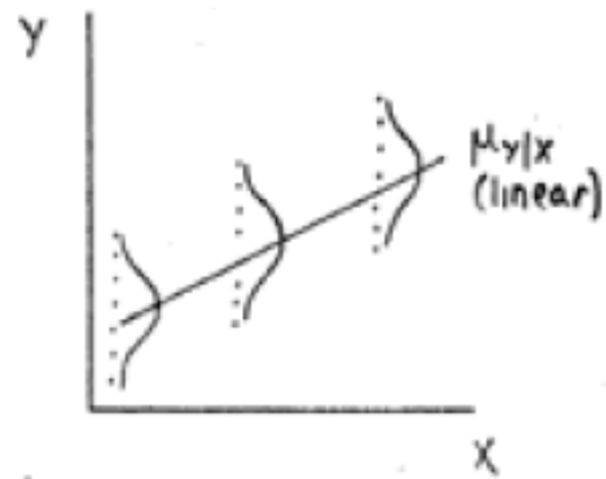
- Five basic assumptions

1. Existence: for each fixed value of X, Y is a random variable with finite mean and variance
2. Independence: the set of Y_i are independent random variables given X_i

Assumptions of Linear Regression

3. Linearity: the mean value of Y is a linear function of X

$$\mu_{Y|X} = E[Y|X] = \beta_0 + \beta_1 X$$



Assumptions of Linear Regression

4. Homoscedasticity: the variance of Y is the same for any X
5. Normality: For each fixed value of X, Y has a normal distribution (by assumption 4, σ^2 does not depend on X)

$$Y \sim N(\mu_{Y|X}, \sigma^2)$$

Formulation

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

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

$\Rightarrow Y_i$ are $N(\beta_0 + \beta_1 X_i, \sigma^2)$ given X_i

$\Rightarrow E(Y_i | X_i) = \beta_0 + \beta_1 X_i$

Var $(Y_i | X_i) = \sigma^2 =$ variability of Y_i about $\mu_{Y|X_i}$

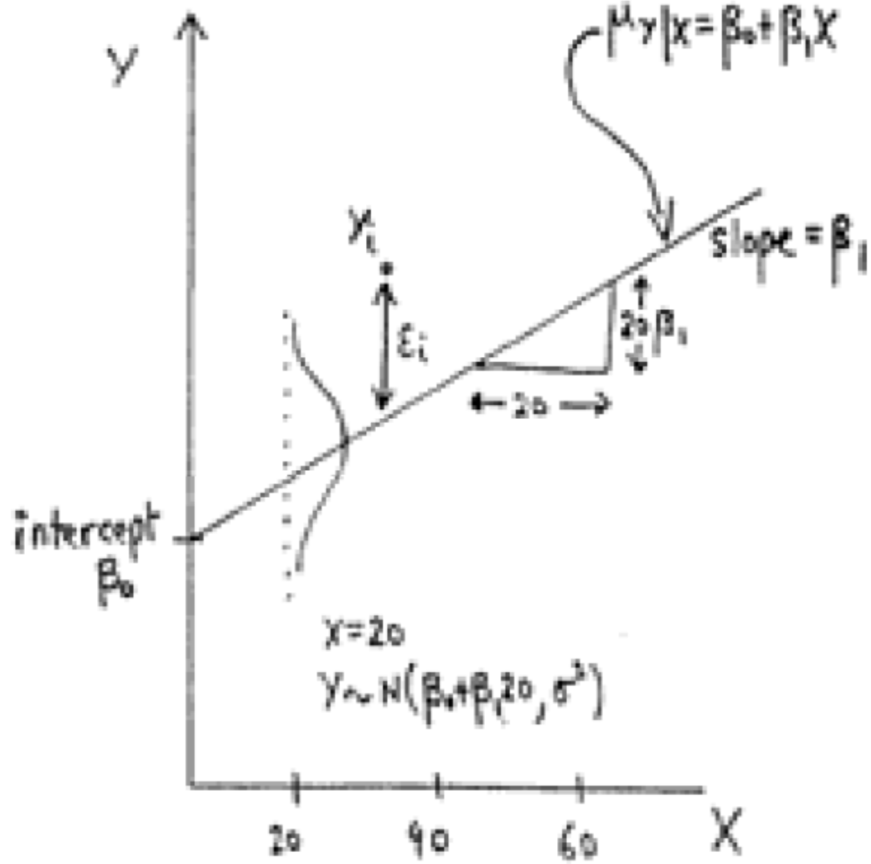
- Y_i are linear function of X_i plus some random error

$\varepsilon_i =$ error $\left(= Y_i - \mu_{Y|X_i}\right)$

$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ is called the "residual" for Y_i

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

Linear Regression



Estimating β_0 and β_1

- Find "best" line
- Criterion for "best": estimate β_0 and β_1 to minimize:

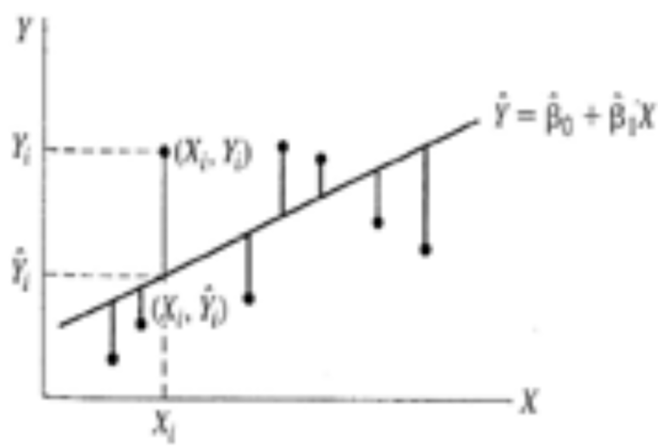
$$\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$$= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

- This is the residual sum of squares, sum of squares due to error, or sum of squares about regression line
- Least Squares estimator

Rationale for LS Estimates

- $\hat{\varepsilon}^2$ measures the "deviance" of Y_i from the estimated model
- The "best" model is the one from which the data deviate the least



Least Squares Estimators

- Taking derivatives with respect to β, we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{S_{xy}}{S_x^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- The residual variance is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$= \frac{1}{n-2} SSE$$

Example: SBP/age data

$$\bar{X} = 45.13 \quad \bar{Y} = 142.53$$

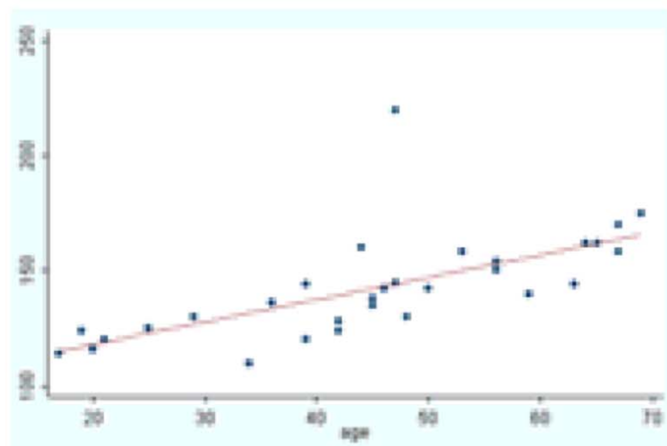
$$S_x^2 = 233.91 \quad S_y^2 = 509.91 \quad S_{xy} = 227.10$$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{227.10}{233.91} = 0.97$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = 142.53 - 0.97(45.13) = 98.71$$

$$\hat{y} = 98.71 + .97x$$

.97mm Hg ↑ for every
1 yr ↑ in age



Using the Model

- Using the parameter estimates, our best guess for any Y given X is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Hence at $\bar{X}$

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) + \hat{\beta}_1 \bar{X} = \bar{Y}$$

- Every regression line goes through $(\bar{X}, \bar{Y})$
- Also $\hat{Y} = \bar{Y} + \hat{\beta}_1 \left(X - \bar{X}\right)$

Correlation and Regression Coefficient

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} \text{ and } r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$\Rightarrow r_{xy} = \hat{\beta}_1 \cdot \frac{S_x}{S_y}$$

Example

Suppose we have:

x	y
1	2
2	4
-1	0
3	4
4	6

$$n = 5$$

$$\bar{x} = \frac{1+2-1+3+4}{5} = 1.8$$

$$\bar{y} = \frac{2+4+0+4+6}{5} = 3.2$$

Calculate r:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1-1.8 = -0.8	2-3.2 = -1.2	(-0.8)(-1.2) = 0.96
2-1.8 = 0.2	4-3.2 = 0.8	(0.2)(0.8) = 0.16
-1-1.8 = -2.8	0-3.2 = -3.2	(-2.8)(-3.2) = 8.96
3-1.8 = 1.2	4-3.2 = 0.8	(1.2)(0.8) = 0.96
4-1.8 = 2.2	6-3.2 = 2.8	(2.2)(2.8) = 6.16
0	0	17.2

Example

$$\sum (x_i - \bar{x})^2 = (-0.8)^2 + (0.2)^2 + (-2.8)^2 + (1.2)^2 + (2.2)^2 = 14.8$$

$$\sum (y_i - \bar{y})^2 = (-1.2)^2 + (0.8)^2 + (-3.2)^2 + (0.8)^2 + (2.8)^2 = 20.8$$

$$S_x = \sqrt{\frac{14.8}{4}} = 1.92$$

$$s_y = \sqrt{\frac{20.8}{4}} = 2.28$$

$$r = \frac{17.2}{\sqrt{(14.8)(20.8)}} = 0.98$$

$$\hat{\beta}_1 = r \cdot \frac{s_y}{S_x} = (0.98)\left(\frac{2.28}{1.92}\right) = 1.16$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.2 - (1.16)(1.8) = 1.11$$

Estimated regression line is $\hat{y}_i = 1.11 + 1.16 x_i$

Example

Fitted values are the $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{y}_1 = 1.11 + 1.16(1) = 2.27$ $\quad$ $\hat{y}_4 = 1.11 + 1.16(3) = 4.59$

$\hat{y}_2 = 1.11 + 1.16(2) = 3.43$ $\quad$ $\hat{y}_5 = 1.11 + 1.16(4) = 5.75$

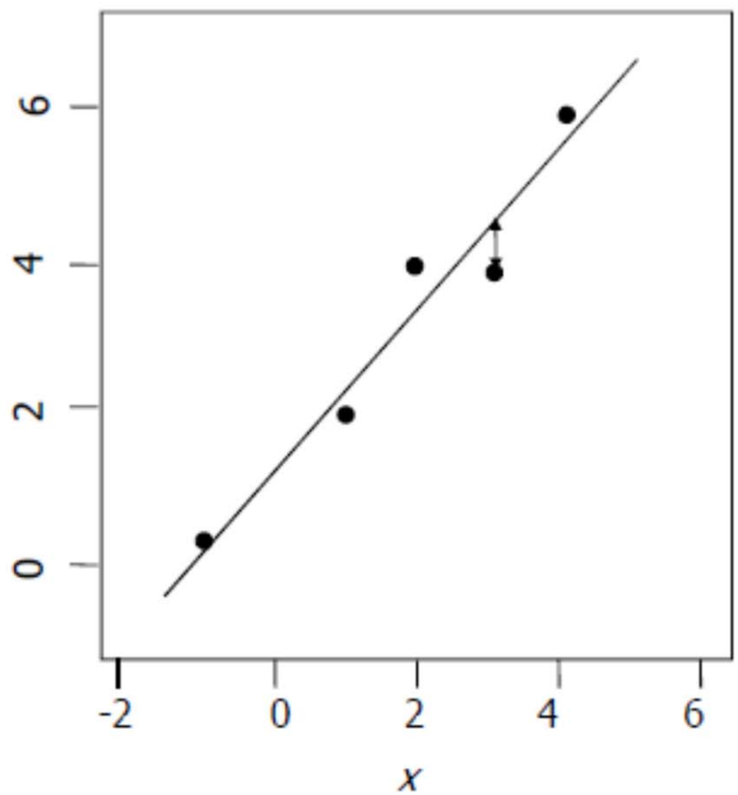
$\hat{y}_3 = 1.11 + 1.16(-1) = -0.05$

Residuals =

$y_i - \hat{y}_i = \hat{\varepsilon}_i$	$(y_i - \hat{y}_i)^2$
2-2.27 = -0.27	$(-0.27)^2 = 0.073$
4-3.43 = 0.57	$(0.57)^2 = 0.345$
0-(-0.05) = 0.05	$(0.05)^2 = 0.0025$
4-4.59 = -0.59	$(-0.59)^2 = 0.348$
6-5.75 = 0.25	$(0.25)^2 = 0.0625$
$\sum \hat{\varepsilon}_i = \sum (y_i - \hat{y}_i) = 0$	$\sum (y_i - \hat{y}_i)^2 = 0.811$

$$\hat{\sigma}^2 = \frac{0.811}{5-2} = 0.270 \qquad \hat{\sigma} = \sqrt{.270} = 0.520$$

Example



$$\hat{y} = 1.11 + 1.16x$$

At $x = 3$:

$$y_i - \hat{y}_i = 4 - 4.59 = -.59$$