Modern Machine Learning
Linear Regression

Kenneth E. Barner
Department of Electrical and Computer Engineering
University of Delaware

Linear Regression

Problem: We are given n observations of variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \boldsymbol{x}_p$ and output $\mathbf{y}$

Objective: Predict $\mathbf{y}$ using $\mathbf{x}_1, \mathbf{x}_2, \dots, \boldsymbol{x}_p$

Model: Linear model of the form:

$$\mathbf{y} = \beta_0 + \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_p\beta_p = \beta_0 + \sum_{i=1}^{p} \mathbf{x}_i\beta_i$$

- p: $=$number of variables
- n: $=$ number of observations (observations indexed in next slide)
- Classical setting: $n \gg p$. Given sufficient observations, build a prediction model for Y

Assumption: Y is statistically related to X

Matrix notation:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad \boldsymbol{X} = \begin{pmatrix} | & | & | & & | \\ 1 & \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p \\ | & | & | & & | \end{pmatrix}_{n\times(p+1)} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}$$

The goal is to solve

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

Note: the system is overdetermined and has no solution since $n > p$

Solution: Solve the system in the least squares sense

Least squares formulation:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{(p+1)}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$$

Compute solution using matrix derivatives

The residual sum of squares (RSS) is:

$$RSS(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0 \qquad (*)$$

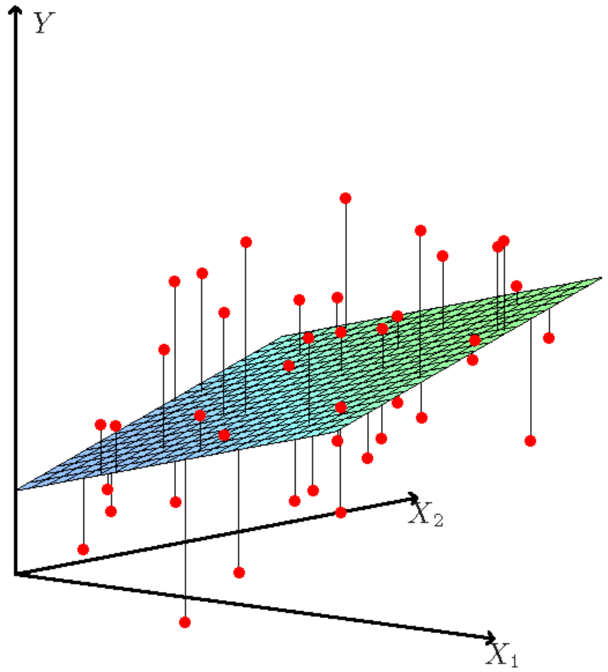
Assuming that $\mathbf{X}$ is full rank, $\mathbf{X}^T\mathbf{X}$ is invertible, which leads to

$$\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T\mathbf{y}$$

Therefore the predicted value for an input vector $\mathbf{x}_0 \in \mathbb{R}^p$ is calculated as

$$\hat{y}_0 = (1, \quad \mathbf{x}_0)\widehat{\boldsymbol{\beta}}_{LS}$$

Note: prediction function is a hyperplane.



Linear least squares fitting with $\boldsymbol{X} \in \mathbb{R}^2$

Orthogonality Principal & Geometric Interpretation

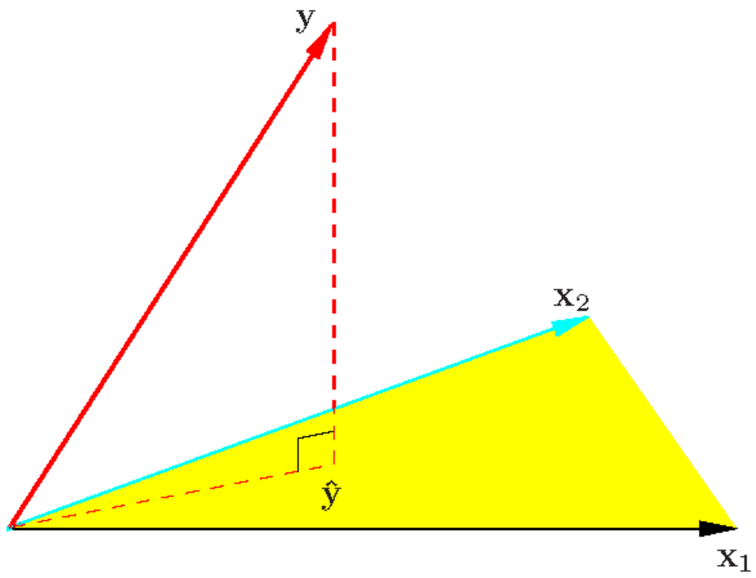
Rearranging $(*)$

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$$

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

$$\mathbf{X}^T(\mathbf{y} - \widehat{\boldsymbol{y}}) = 0$$

Interpretation: estimate error is orthogonal to the observation



Optimal estimate $\widehat{\boldsymbol{y}}$ is achieved by projecting $\boldsymbol{y}$ on to $\mathbf{X}$. Estimate error, $(\mathbf{y} - \widehat{\boldsymbol{y}})$, is orthogonal to the observation $\mathbf{X}$.

Properties of the least squares solution:

$E(\widehat{\boldsymbol{\beta}}_{LS}) = \boldsymbol{\beta}$ [unbiased]

$Var\left(\widehat{\boldsymbol{\beta}}_{LS}\right) = (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\sigma$

Assumptions:

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

where $\mathbf{X}$ is non-random; $\boldsymbol{\epsilon}$ elements are iid $N(0, \sigma^2)$

Proof: For $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}$ iid $N(0, \sigma^2)$ we have

$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ [Normally distributed]

Substituting within the estimate

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{LS} &= (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T\boldsymbol{\epsilon}\end{aligned}$$

Therefore

$$\widehat{\boldsymbol{\beta}}_{LS} \sim N\left(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\sigma^2\right) \quad \text{[Normally distributed]}$$

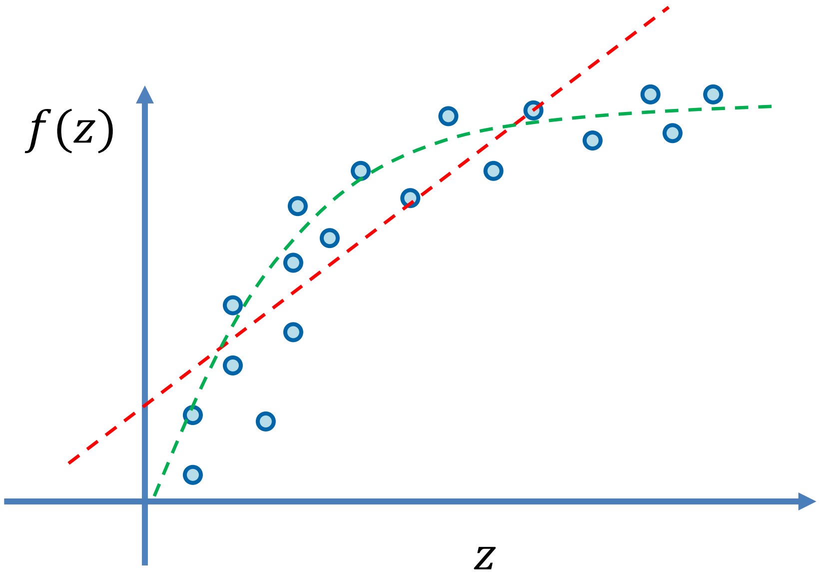
with $E(\widehat{\boldsymbol{\beta}}_{LS}) = \boldsymbol{\beta}$ and $Var\left(\widehat{\boldsymbol{\beta}}_{LS}\right) = (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\sigma^2$

Linear Regression

Note: Observation samples may be

- Quantitative inputs
- Transformations of quantitative inputs, e.g., log, square root or square
- Basis expansions, such as $X_2 = X_1^2, X_3 = X_1^3$ (polynomial expansion)

Note: appropriate observation (feature) transformation may improve performance



$f(z) = \beta_0 + \beta_1 z$

$f(z) = \beta_0 + \beta_1 z + \beta_2 \sqrt{z}$

New feature vectors $\mathbf{z}_i = [1, z_i, \sqrt{z_i}]$

Example: Prostate Cancer

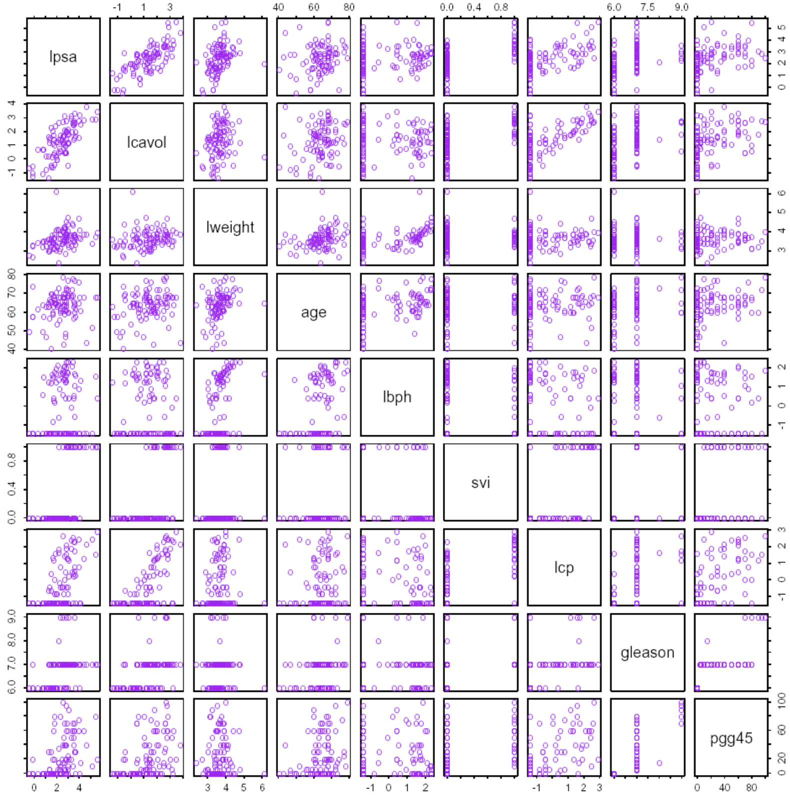
Example: Stanley et al. (1989) examined the correlation between prostate-specific antigen and multiple clinical measures in prostatectomy patients.

Data: Given variables (clinical measures):

- `lcavol`: log cancer volume
- `lweight`: log prostate weight
- `age`
- `lbph`: log benign hyperplasia amount
- `svi`: seminal vesicle invasion
- `lcp`: log capsular penetration
- `gleason`: gleason score
- `pgg45`: percent gleason scores 4 or 5

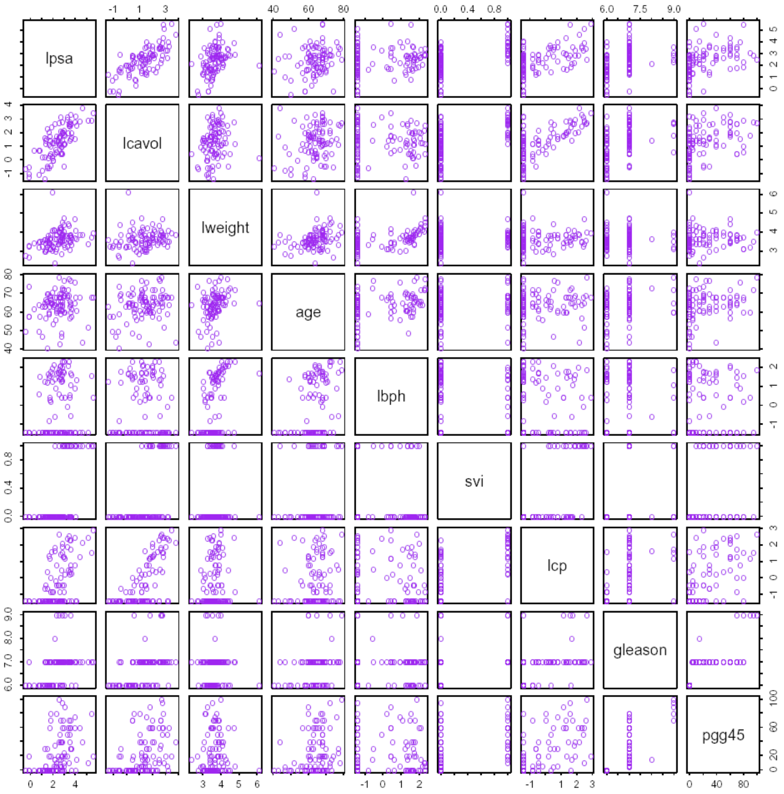
Objective: Predict `lpsa` (log of prostate specific antigen) level

Data size: Measurements from 97 men



Pairwise scatterplot matrix of data. First row is `lpsa` versus each clinical measure. Note `svi` and `gleason` are categorical.

Example: Prostate Cancer



Pairwise scatterplot matrix of data. First row is `lpsa` versus each clinical measure. Note `svi` and `gleason` are categorical.

Observations:

- `lcavol` and `lcp` have a strong relationship with `lpsa` (and each other)
- Apply LS linear regression prediction of utilizing all clinical measure observations to untangle relationships between predictors and response

Correlations of predictors in the prostate cancer data

	lvavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	-0.139			
lcp	0.692	0.157	0.173	-0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	-0.030	0.481	0.663	0.757

Example: Prostate Cancer

Methodology:

- Fit a linear model to `lpsa`
- Training/testing split: 67/30
- Optimization loss function: least squares
- Z-Scores measure the effect of dropping that variable from the model (based on null hypothesis testing)
  - A Z-score >2 in absolute value is considered significant, meaning that the coefficient is relevant to the model and should be kept

Result & Observations:

- `intercept` is the bias term
- `lcavol` has the strongest effect
- `lweight` and `svi` also have significant effect
- `lcp` is not significant given `lcavol` in the model
  - `lcp` is significant without `lcavol` in the model
- For comparison, consider the base error rate: mean value of `lpsa` (in the training set)
- The linear model mean prediction error on the test data is 0.521, a 50% reduction compared to the 1.057 base error rate

Linear model fit to the prostate cancer data

Term	Coefficient (β)	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	-0.14	0.10	-1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	-0.29	0.15	-1.87
`gleason`	-0.02	0.15	-0.15
`pgg45`	0.27	0.15	1.74