

Applied Data Science

Machine Learning Lecture 1

John Tsitsiklis

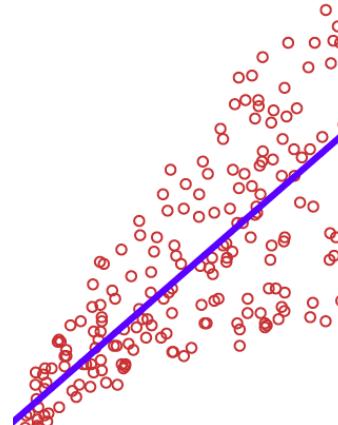
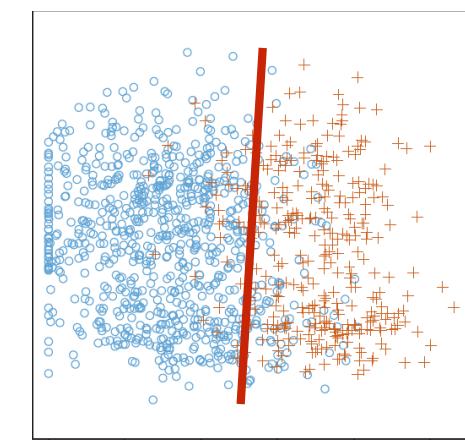
August 30, 2021

Introductions



<http://www.mit.edu/~jnt/home.html>

Overview of this week/module

- Central methods in Machine Learning
 - we will only discuss “supervised learning”: learn from examples
- Predict the value of an unobserved variable
Regression (linear)A scatter plot with red data points forming a positive linear trend. A solid blue line represents the linear regression fit.
- Predict the type of an individual
classificationA scatter plot with two classes of data points: blue circles and orange crosses. A vertical red line represents a decision boundary or classification threshold.
- Assessment
 - How good is our method, our model, and our prediction?
 - Testing, validation

Today's agenda

- Regression
 - formulation
 - solution
 - interpretation
 - (classical) performance assessment
- Further topics (*next session*)
 - what can go wrong
 - using nonlinear features of the data
 - overfitting and regularization
 - ridge regression
 - sparse regression and lasso

MACHINE LEARNING AND STATISTICS

A conceptual big picture



Model?

X_1	Y_1
:	:
:	:
X_n	Y_n
X	$Y?$

X : symptoms, test results, etc.

Y : state of health

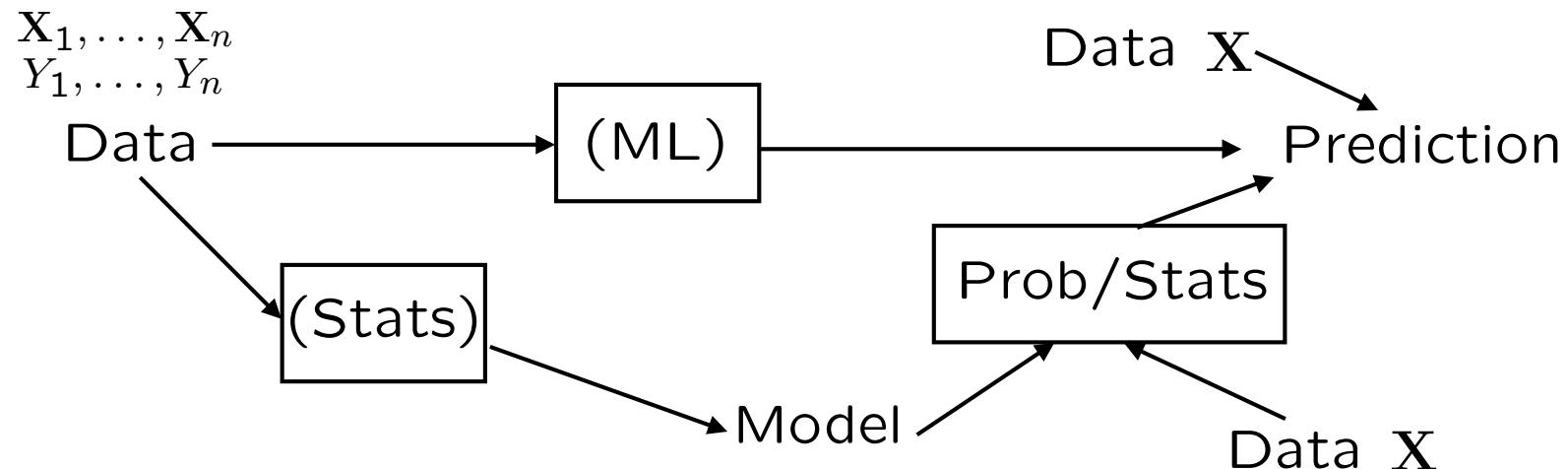
- “Predict” Y based on X

Y : sick or not (binary) [classification]

Y : life expectancy (any real number)
[regression]

- **Understand**

- build a model, a theory, a narrative, a mechanism



“All models are wrong, some are useful” (George E.P. Box)

Statistics versus machine learning — a caricature of a spectrum

Statistics

Machine learning

Deep understanding
of simple models/methods

Theory does not
always explain success

Deep theorems
on complex models

Fearless methods
Algorithmic emphasis

- Need a language: probability
- Build on two centuries of statistical knowledge

A basic statistical framework

- Unknown parameter vector θ , to estimate $\theta \in \mathbb{R}^m \quad \theta \in \{0, 1\}$
- Data $\mathbf{X}_1, \dots, \mathbf{X}_n$ drawn from distribution \mathbb{P}^θ i.i.d.
independent
- Estimator: $\widehat{\Theta} = g(\mathbf{X}_1, \dots, \mathbf{X}_n)$
to ↑
design
- Various approaches
 - plugin
 - maximum likelihood
 - empirical risk minimization
 - Bayesian
- Assessment of $\widehat{\Theta}$:
$$\mathbb{E}\left[\|\widehat{\Theta} - \theta\|^2\right]$$
$$\mathbb{P}(\widehat{\Theta} \neq \theta)$$

Notation key

- vectors: boldface
scalars: normal font
 - \mathbf{X}_2 : second data record
 - X_2 : second component of a vector \mathbf{X}
- number of data records: n
- “star” for true quantity, e.g., θ^*
- “hat” for estimates, e.g., $\widehat{\Theta}$
- Upper case: random variables, e.g., Y
Lower case: numbers and constants, e.g., y for realized value of Y

(LINEAR) REGRESSION

formulation

solution

interpretation

An example: Advertising and Sales

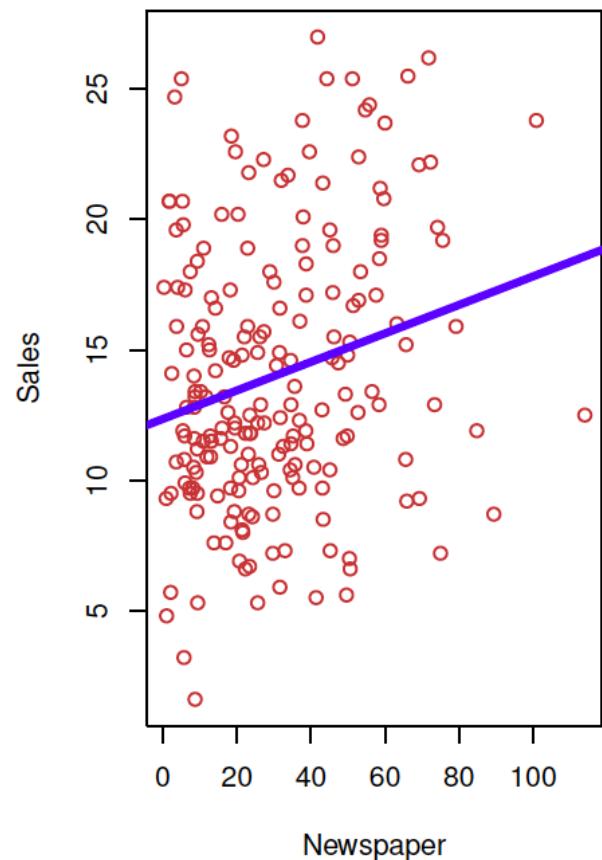
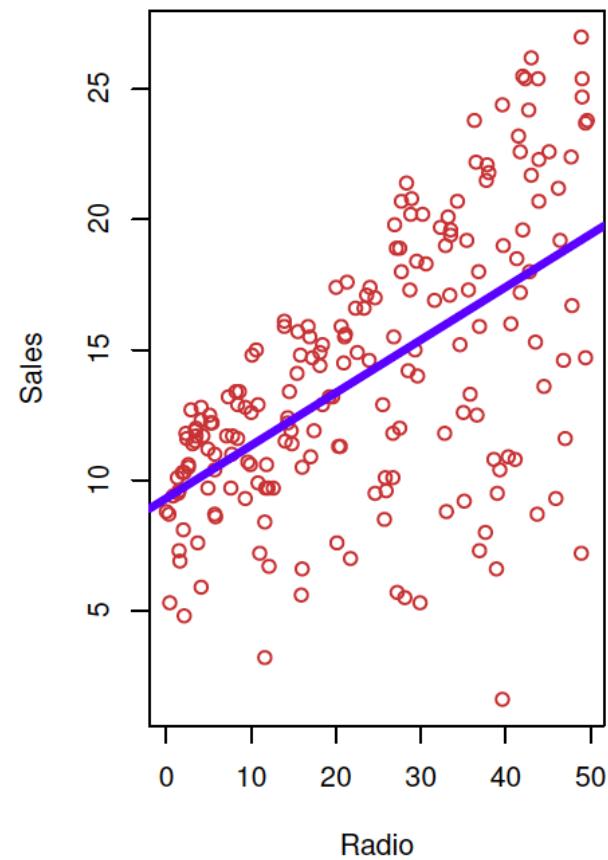
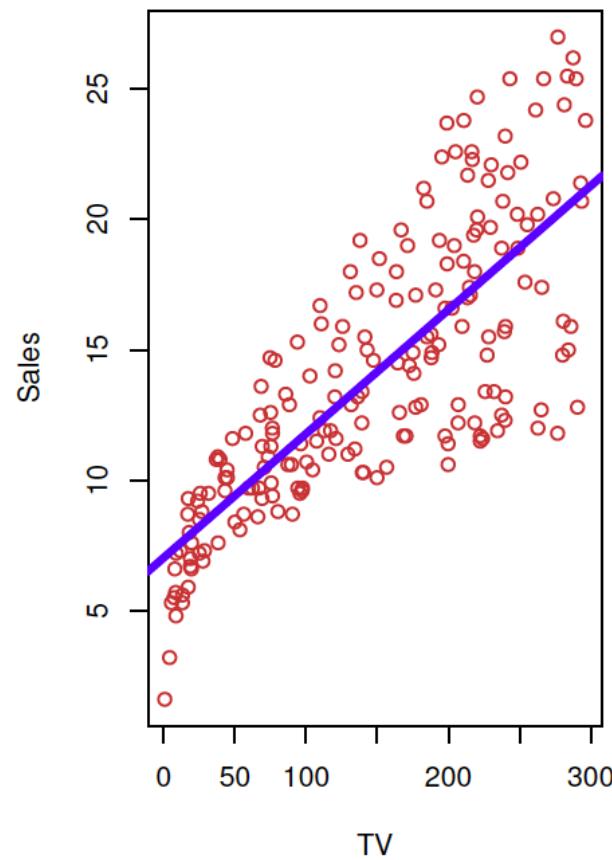
- Data across 200 Markets
 - Spending for TV, Radio, NewsPaper
 - Resulting Sales

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9
...
195	38.2	3.7	13.8	7.6
196	94.2	4.9	8.1	9.7
197	177.0	9.3	6.4	12.8
198	283.6	42.0	66.2	25.5
199	232.1	8.6	8.7	13.4

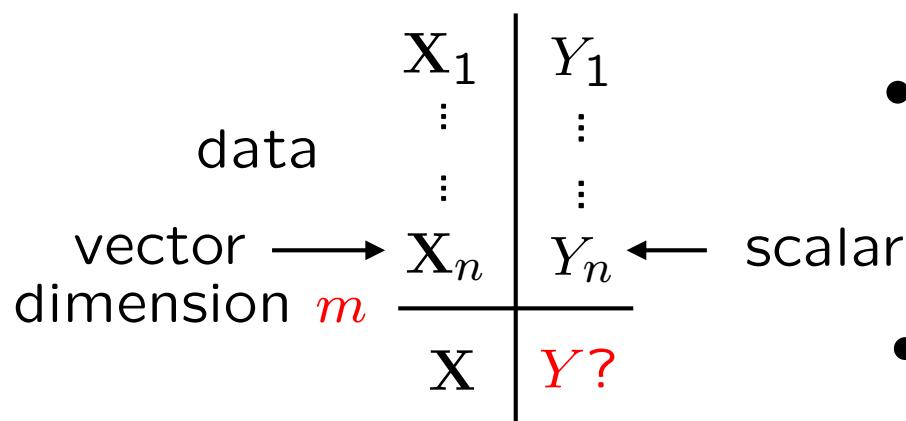
200 rows × 4 columns

- Questions
 - Is there a relation between Advertising Channel Budgets and Sales?
 - If yes, can we “predict” Sales given the Channel Budgets?

An example: Advertising and Sales



Regression



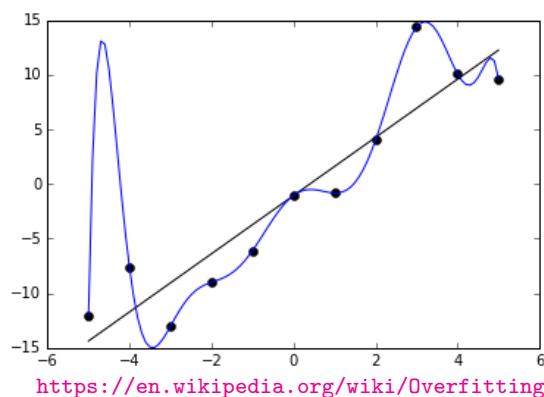
- **Regressor/predictor:** $\hat{Y} = g(\mathbf{X})$
- “Learn” a “good” g from the data

objective: $\mathbb{E}\left[(g(\mathbf{X}) - Y)^2\right]$
(risk)

proxy: $\frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2$

turns out: $g^*(\mathbf{X}) = \mathbb{E}[Y | \mathbf{X}]$
unavailable

“empirical risk minimization”



- Restrict to limited class of predictors

Linear regression

data	X_1 ⋮ X_n	Y_1 ⋮ Y_n
vector dimension m	X	$Y?$

$$\frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2$$

sum of over data points

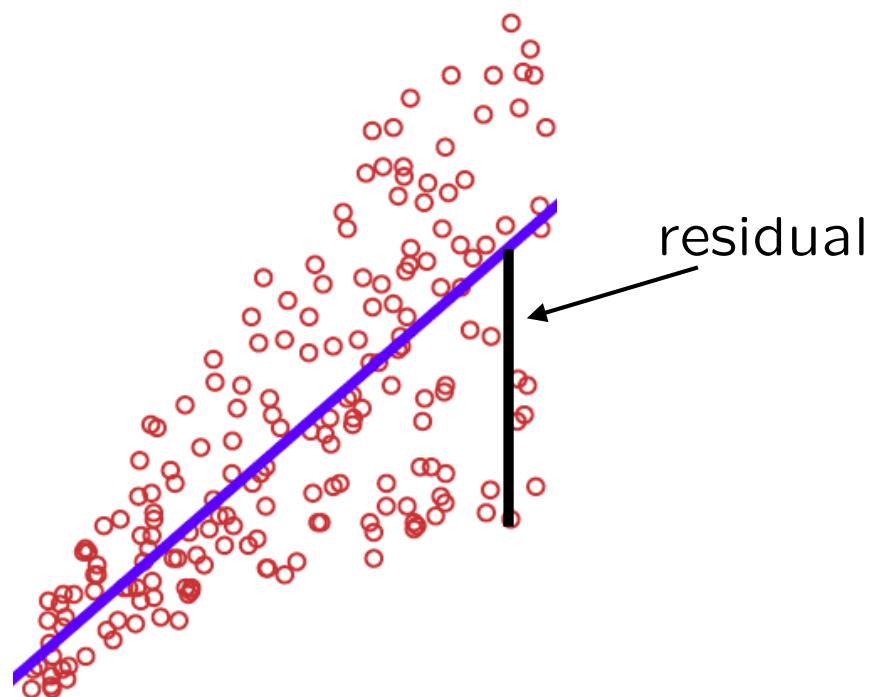
$$\min_{\theta} \sum_{i=1}^n (\theta^T \mathbf{X}_i - Y_i)^2$$

- Restrict to limited class of predictors:

$$\hat{Y} = \theta_0 + \theta_1 X_1 + \cdots + \theta_m X_m$$

$$\text{let } \mathbf{X} = (1, X_1, \dots, X_m)$$

$$\hat{Y} = g(\mathbf{X}) = \theta^T \mathbf{X}$$



Solution to the regression problem

$$\min_{\theta} \sum_{i=1}^n (\theta^T \mathbf{X}_i - Y_i)^2 \quad \begin{array}{l} n \text{ data points} \\ \mathbf{X}_i \text{ and } \theta \text{ have dimension } m+1 \end{array}$$

$$\min_{\theta} H(\theta) \quad \text{quadratic in } \theta$$

optimality conditions: $\nabla H(\theta) = 0$ $\frac{\partial H}{\partial \theta_j} = 0, \quad j = 0, 1, \dots, m$

linear system of $m + 1$ equations

• Formulas:

$$n \begin{bmatrix} 1 & \mathbf{X}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{X}_n^T \end{bmatrix}_{m+1} = \mathbb{X} \quad \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \mathbf{Y}$$

$$\hat{\theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}$$

Results for our example

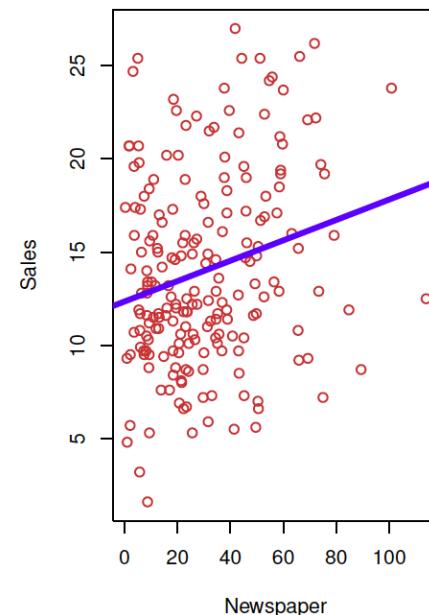
$$n = 200 \quad m + 1 = 4$$

$$\hat{\theta} = \begin{bmatrix} 2.94 \\ 0.046 \\ 0.19 \\ -0.001 \end{bmatrix}$$

$$\widehat{\text{Sales}} = 2.94 + 0.046 \cdot (\text{TV}) + 0.19 \cdot (\text{Radio}) - 0.001 \cdot (\text{NewsP})$$

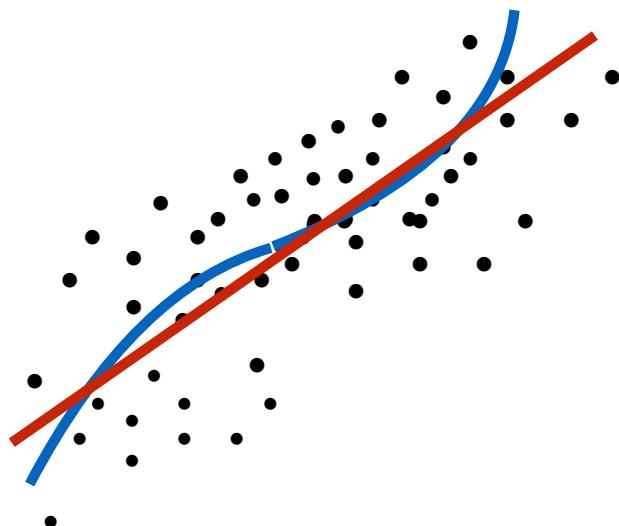
- Compare with **simple** linear regression

$$\widehat{\text{Sales}} = 12.35 + 0.055 \cdot (\text{NewsP})$$

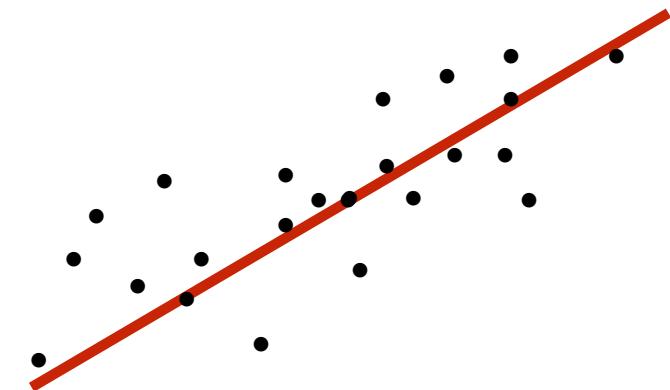


Interpretation and justification: empirical risk minimization

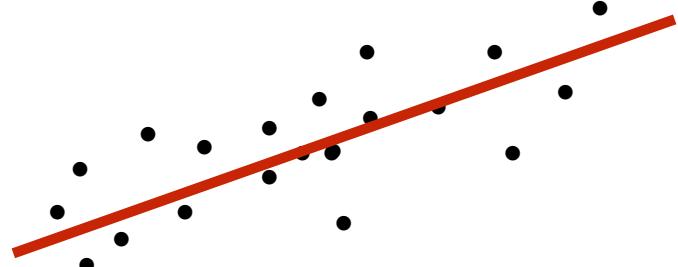
- Large true population



- true relation may be complex
- interested in best linear predictor



- Finite sample: find best linear fit
 $n \rightarrow \infty$: recover “population best”
(as long as samples are drawn representatively)



- Another finite sample: different results
how much variation do we expect?

Interpretation and justification: maximum likelihood

- Illustrate for $m = 1$

- assume structural model: $Y_i = \theta_0^* + \theta_1^* X_i + W_i$
- conditioned on all the X_i : all the W_i are independent and $\text{Normal}(0, \sigma^2)$

$$\max_{\boldsymbol{\theta}} \mathbb{P}(\mathbf{Y} | \mathbb{X}; \boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(Y_i - \theta_0^* - \theta_1^* X_i)^2}{2\sigma^2} \right\}$$

$$\prod \exp = \exp \sum$$

Maximizing the likelihood function = minimizing the empirical risk

Maximum likelihood estimates have appealing theoretical guarantees

PERFORMANCE ASSESSMENT

R^2 (R-squared)

- Prediction if no regression: $\bar{Y} = \frac{1}{n} Y_i$
- Total sum of squares: $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$
“initial” variation in Y
- Residual sum of squares: $RSS = \sum_{i=1}^n (Y_i - \hat{\theta}^T \mathbf{X}_i)^2$
unexplained variation in Y , after taking into account X
- $R^2 = 1 - \frac{RSS}{TSS}$ fraction of variation in Y that has been explained
 $0 \leq R^2 \leq 1$ high R^2 is preferred

in simple regression R^2 is an estimate of
the squared correlation coefficient between X and Y

R^2 for our example

- $\widehat{\text{Sales}} = 2.94 + 0.046 \cdot (\text{TV}) + 0.19 \cdot (\text{Radio}) - 0.001 \cdot (\text{NewsP})$
 $R^2 = 0.897$ All the budgets together explain a lot
- $\widehat{\text{Sales}} = 12.35 + 0.055 \cdot (\text{NewsP})$
 $R^2 = 0.05$ Newspaper budget explains little
For TV alone: $R^2 = 0.61$
For Radio alone: $R^2 = 0.33$
- More variables: R^2 can only go up (or stay the same)
 - but this may be a mirage
 - adjusted R^2 : $1 - \frac{\text{RSS}/(n - m - 1)}{\text{TSS}/(n - 1)}$ 0.897 → 0.896

How noisy/reliable are my estimates of θ^*

- Assume structural model



(If not, need to resort to simulation/bootstrap methods)

next
session

$$Y_i = (\theta^*)^T \mathbf{X}_i + W_i$$

W_i : independent,
zero mean, variance σ^2

- Regression:

$$\widehat{\Theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}$$

$\widehat{\Theta}$ is a random variable
(depends on random data)

$$\mathbb{E}[(\widehat{\Theta}_j - \theta_j^*)^2] = (\mathbb{E}[\widehat{\Theta}_j] - \theta_j^*)^2 + \text{var}(\widehat{\Theta}_j)$$

$$\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \text{var}(X)$$

bias

variance

$$\mathbb{E}[\widehat{\Theta}_j] = \theta_j^*$$

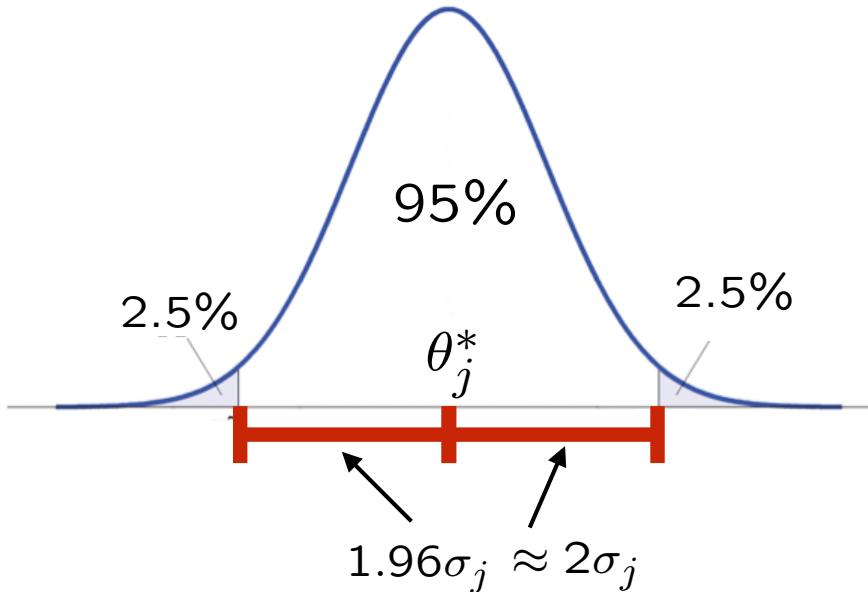
(unbiased)

- Hence focus on the variance of $\widehat{\Theta}_j$

The distribution of $\widehat{\Theta}$

- $\widehat{\Theta}$ “is” multivariate normal
- Each $\widehat{\Theta}_j$ is normal

$$\widehat{\Theta}_j \sim \mathcal{N}(\theta_j^*, \sigma_j^2)$$



$$\sigma_j = \sqrt{\text{var}(\widehat{\Theta}_j)} = \text{se}(\widehat{\Theta}_j)$$

standard error

$$Y_i = (\boldsymbol{\theta}^*)^T \mathbf{X}_i + W_i$$

W_i : independent,
zero mean, variance σ^2

$$\widehat{\boldsymbol{\Theta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}$$

- approximately: large n , central limit theorem
- exactly: if W_i are normal
- linear combinations of the $\widehat{\Theta}_j$ are normal

- Agenda:

calculate/approximate standard error
use it (confidence intervals, hypothesis testing)

The covariance matrix of $\widehat{\Theta}$

$$\mathbb{E}[(\widehat{\Theta} - \theta^*)(\widehat{\Theta} - \theta^*)^T]$$

dimensions $(m + 1) \times (m + 1)$

diagonal entries: $\text{Var}(\widehat{\Theta}_j)$

off-diagonal entries: $\text{Cov}(\widehat{\Theta}_i, \widehat{\Theta}_j)$

formula: $\sigma^2(\mathbb{X}^T \mathbb{X})^{-1}$

use $\widehat{\sigma}^2$

$$Y_i = (\theta^*)^T \mathbf{X}_i + W_i$$

W_i : independent,
zero mean, variance σ^2

$$\widehat{\Theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}$$

	const	TV	Radio	Newspaper
const	9.72867479E-02	-2.65727337E-04	-1.11548946E-03	-5.91021239E-04
TV	2.65727337E-04	1.9457371E-06	-4.47039463E-07	-3.26595026E-07
Radio	-1.11548946E-03	-4.47039463E-07	7.41533504E-05	-1.78006245E-05
Newspaper	-5.91021239E-04	-3.26595026E-07	-1.78006245E-05	3.44687543E-05

- Estimate σ^2 by

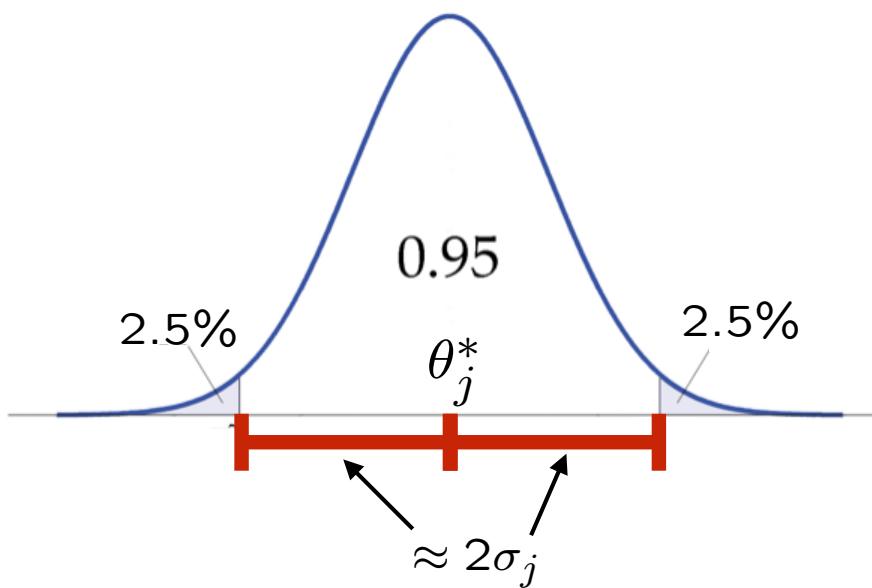
$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{\theta}^T \mathbf{X}_i)^2$$

slight downwards bias
negligible bias if $m \ll n$

Why? For large samples, $\widehat{\Theta} \approx \theta^*$, and

$$\sigma^2 = \mathbb{E}[W_i^2] \approx \frac{1}{n} \sum_{i=1}^n (Y_i - (\theta^*)^T \mathbf{X}_i)^2 \approx \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{\Theta}^T \mathbf{X}_i)^2$$

Confidence Interval (CI)



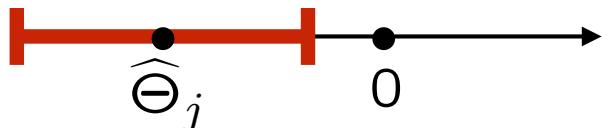
$$\widehat{\Theta}_j \sim \mathcal{N}(\theta_j^*, \sigma_j^2)$$

- With probability 95%: $|\text{error}| = |\widehat{\Theta}_j - \theta_j^*| \leq 2\sigma_j$
- 95%-CI:
$$[\widehat{\Theta}_j - 2\widehat{\sigma}_j, \quad \widehat{\Theta}_j + 2\widehat{\sigma}_j]$$
- $$\mathbb{P}(\theta_j^* \in \text{CI}) \approx 0.95$$
- Needs careful (“frequentist”) interpretation

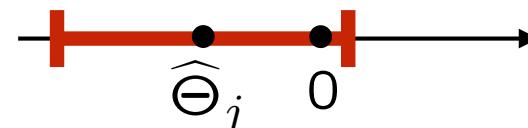
Testing the hypothesis $\theta_j^* = 0$

- Are the data compatible with the null hypothesis $\theta_j^* = 0$?

Wald test:



reject the null



do not reject the null

(data compatible with $\theta_j^* = 0$)

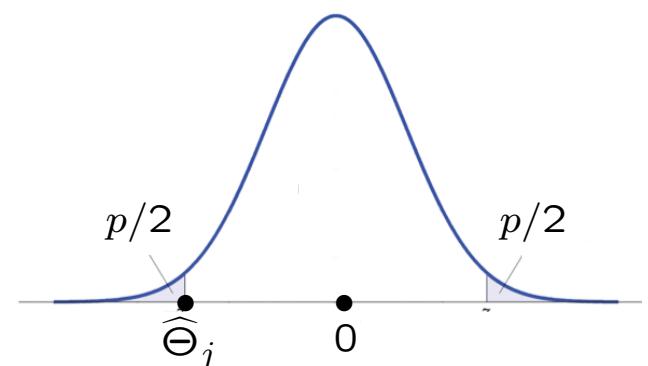
$$\mathbb{P}(\text{reject} \mid \theta_j^* = 0)$$

$$= \mathbb{P}(\text{the CI "misses" } 0 \mid \theta_j^* = 0)$$

$\approx 5\%$

p-value: probability of seeing something at least as extreme as the observed $\widehat{\Theta}_j$, under θ_j^*

- reject if p-value < 0.05



Back to our example

Error covariance matrix:

	const	TV	Radio	Newspaper
const	9.72867479E-02	-2.65727337E-04	-1.11548946E-03	-5.91021239E-04
TV	2.65727337E-04	1.9457371E-06	-4.47039463E-07	-3.26595026E-07
Radio	-1.11548946E-03	-4.47039463E-07	7.41533504E-05	-1.78006245E-05
Newspaper	-5.91021239E-04	-3.26595026E-07	-1.78006245E-05	3.44687543E-05

Confidence intervals

[0.025 0.975]

	coef	std err		
Intercept	2.9389	0.312		2.324 3.554
TV	0.0458	0.001		0.043 0.049
Radio	0.1885	0.009		0.172 0.206
Newspaper	-0.0010	0.006		-0.013 0.011

- Wald test: intercept, TV, Radio are “significant”
(reject the hypothesis that they are zero)

Newspaper: the hypothesis that $\theta_{\text{NewsP}}^* = 0$ “survives”
(not rejected)

Making new predictions

- After running the regression given some new \mathbf{X} , predict $\hat{Y} = \hat{\boldsymbol{\theta}}^T \mathbf{X}$
- Keep assuming structural model: $Y = (\boldsymbol{\theta}^*)^T \mathbf{X} + W$
- $\hat{\boldsymbol{\theta}}$ is unbiased estimate of $\boldsymbol{\theta}^*$
 $\Rightarrow \hat{\boldsymbol{\theta}}^T \mathbf{X}$ is unbiased estimate of $(\boldsymbol{\theta}^*)^T \mathbf{X}$ (and of Y)
- Two sources of error:
 - unavoidable, from W ; variance σ^2
 - variance of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{X}$
 - Total prediction error variance:

$$\begin{array}{c|c} \mathbf{X}_1 & Y_1 \\ \vdots & \vdots \\ \mathbf{X}_n & Y_n \\ \hline \mathbf{X} & \textcolor{red}{Y?} \end{array}$$

$$\sigma^2 \mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}$$

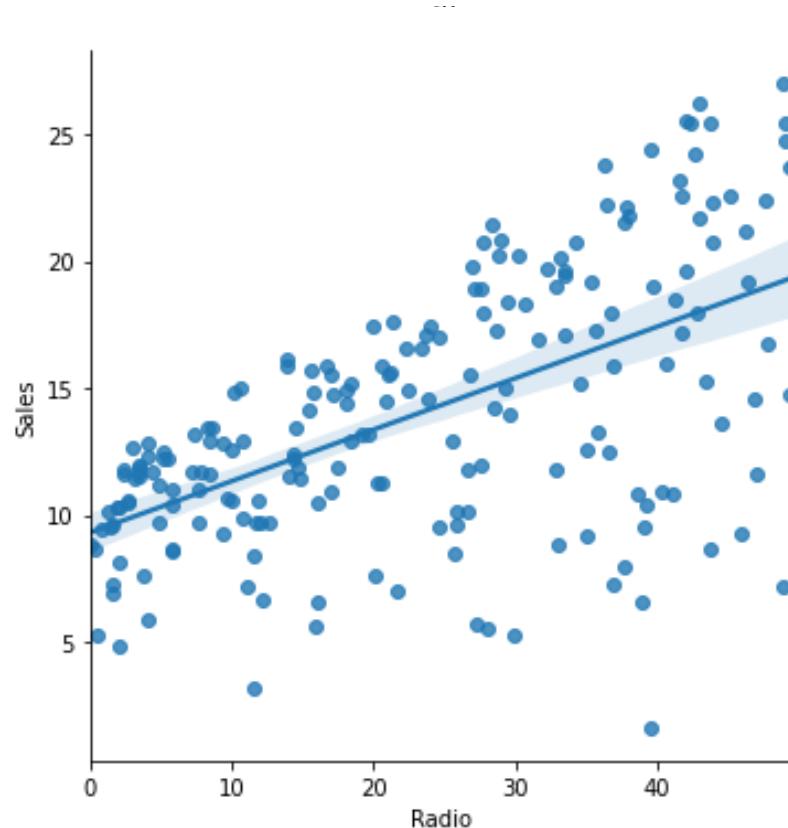
$$\sigma^2 + \sigma^2 \mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}$$

Confidence bands

- 95% confidence interval about the value of $(\theta^*)^T \mathbf{X}$:

$$(\hat{\theta})^T \mathbf{X} \text{ plus or minus } 2 \cdot \hat{\sigma} \sqrt{\mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}}$$

- confidence interval width changes with \mathbf{X}
- in simple regression, this gives a **confidence band**



Summary

- Linear regression
 - formulation
 - underlying assumptions
 - formulas
 - results: their interpretation and usage
- Still, many things can go wrong or be misinterpreted
- New issues when θ has high dimension
- Next session...