Applied Data Science

Machine Learning
Lecture 1

John Tsitsiklis

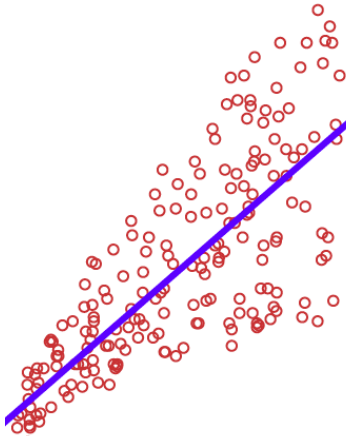
February 7, 2022

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 labeled examples

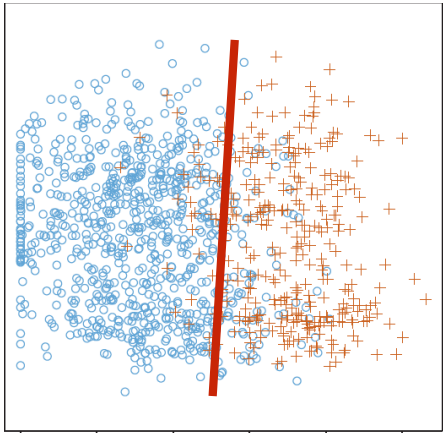


- Predict the value of an unobserved variable

  Regression (linear)

- Predict the type of an individual

  classification



- 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?

$\mathbf{X}_1$	Y_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\mathbf{X}_n$	Y_n
$\mathbf{X}$	Y?

$\mathbf{X}$: symptoms, test results, etc.

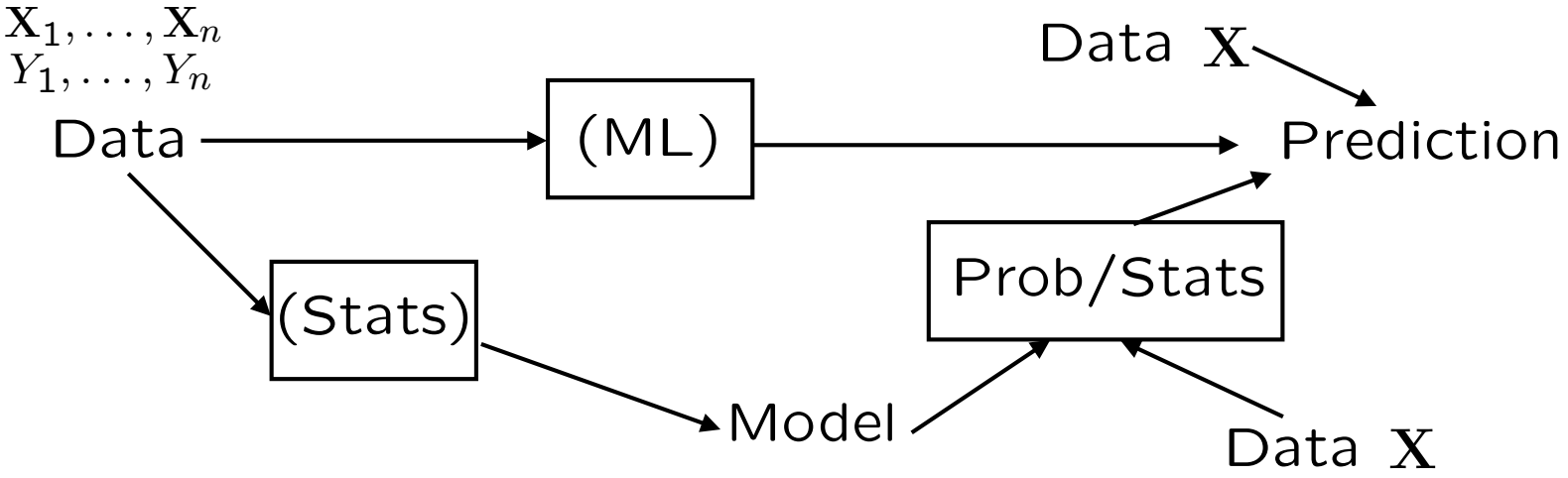
Y: state of health

- **"Predict"** Y based on $\mathbf{X}$

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

Deep theorems
on complex models

Theory does not
always explain success

Fearless methods

Algorithmic emphasis

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

Notation key

- vectors: boldface
  scalars: normal font
  - $\mathbf{X}_2$: second data record
  - X_2: second component of a vector $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$$\mathbf{X}^T = [X_1 \; X_2 \; X_3]$$

$$\mathbf{X}^T\mathbf{Y} = X_1Y_1 + X_2Y_2 + X_3Y_3$$

- 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

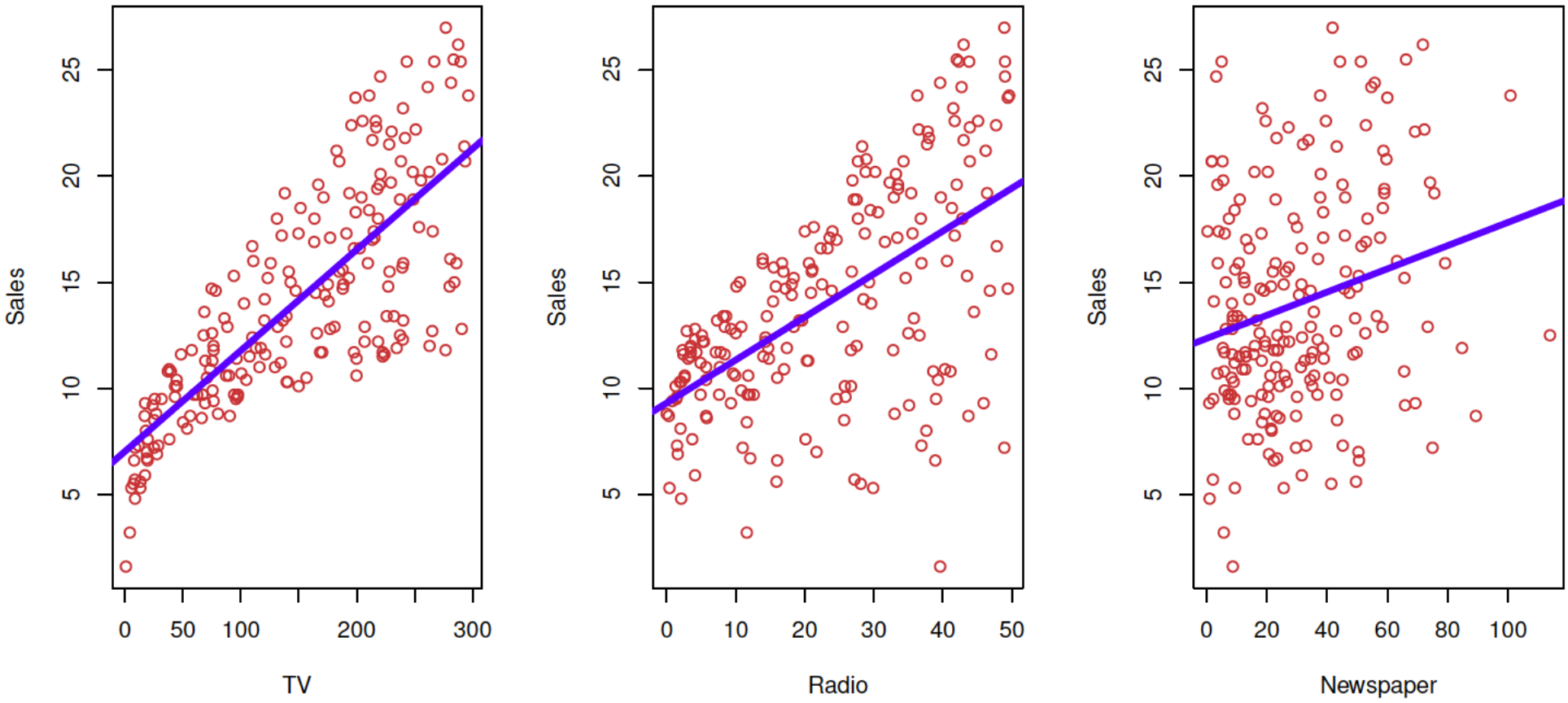
200 rows × 4 columns

- 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

data

$\mathbf{X}_1$	Y_1
$\vdots$	$\vdots$
$\mathbf{X}_n$	Y_n
$\mathbf{X}$	$Y?$

vector $\longrightarrow$ dimension m

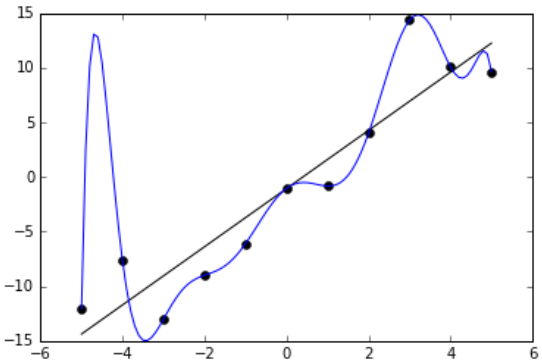
$\longleftarrow$ scalar

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

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

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

proxy: $\dfrac{1}{n}\sum_{i=1}^{n}\big(g(\mathbf{X}_i) - Y_i\big)^2$ "empirical risk minimization"



https://en.wikipedia.org/wiki/Overfitting

- Restrict to limited class of predictors

Linear regression

data

vector $\longrightarrow$ dimension m

$\mathbf{X}_1$	Y_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\mathbf{X}_n$	Y_n
$\mathbf{X}$	$Y?$

- Restrict to limited class of predictors:

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

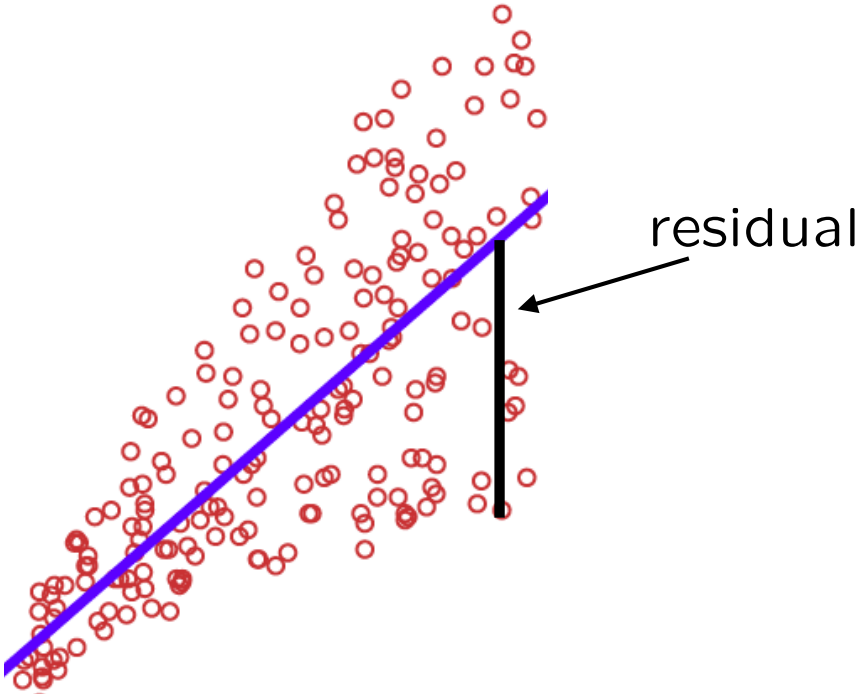
let $\mathbf{X} = (1, X_1, \ldots, X_m)$

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

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

sum of over data points

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



Solution to the regression problem

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

n data points

$\mathbf{X}_i$ and $\boldsymbol{\theta}$ have dimension $m+1$

- Formulas:

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

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

Math details:

$\min_{\boldsymbol{\theta}} H(\boldsymbol{\theta})$ quadratic in $\boldsymbol{\theta}$

optimality conditions: $\nabla H(\boldsymbol{\theta}) = 0$ $\qquad \dfrac{\partial H}{\partial \theta_j} = 0, \qquad j = 0, 1, \ldots, m$

linear system of $m+1$ equations

Results for our example

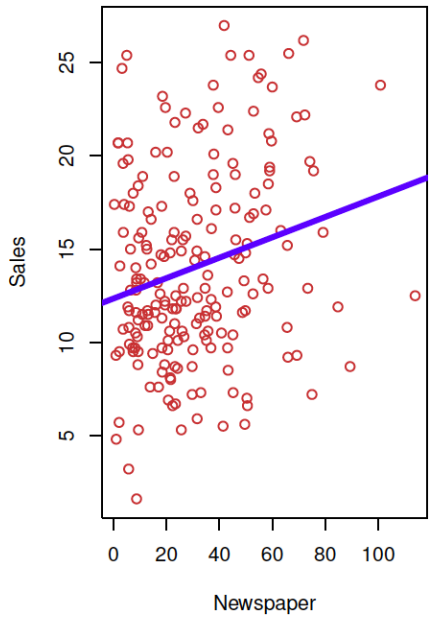
$n = 200 \qquad m + 1 = 4$

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

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

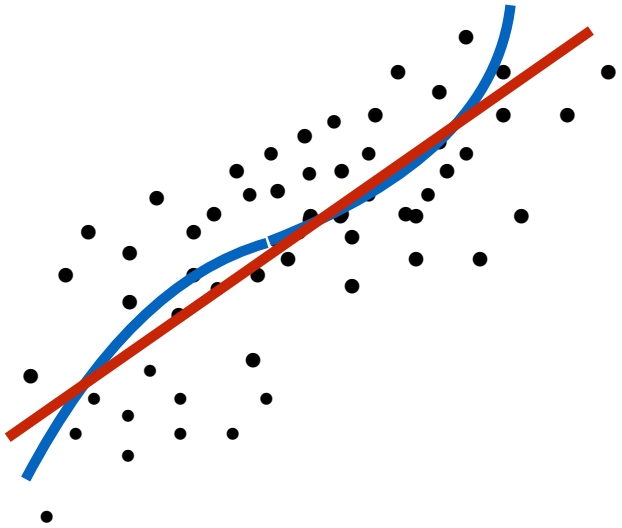
- Compare with simple linear regression

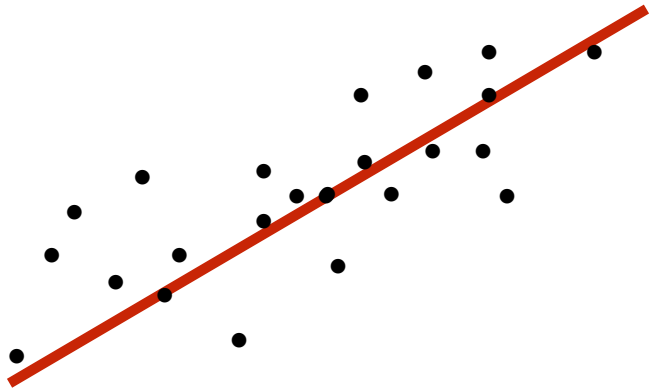
$$\widehat{\text{Sales}} = 12.35 + 0.055 \cdot (\text{NewsP})$$



Interpretation and justification: empirical risk minimization

- Large true population





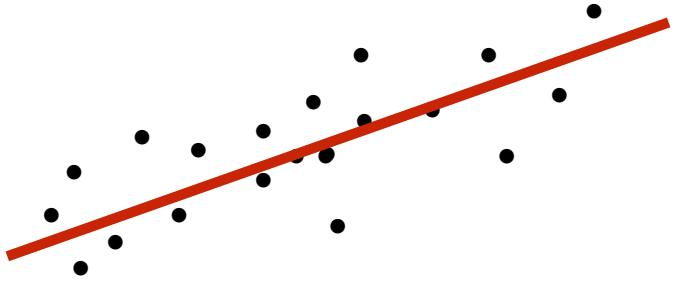
- Finite sample: find best linear fit

$n \to \infty$: recover "population best"

(as long as samples are drawn representatively)

– true relation may be complex

– interested in best linear predictor



- 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 Normal$(0, \sigma^2)$

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

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

Maximizing the likelihood function = minimizing the empirical risk

Maximum likelihood estimates have appealing theoretical guarantees

Summary of the two interpretations

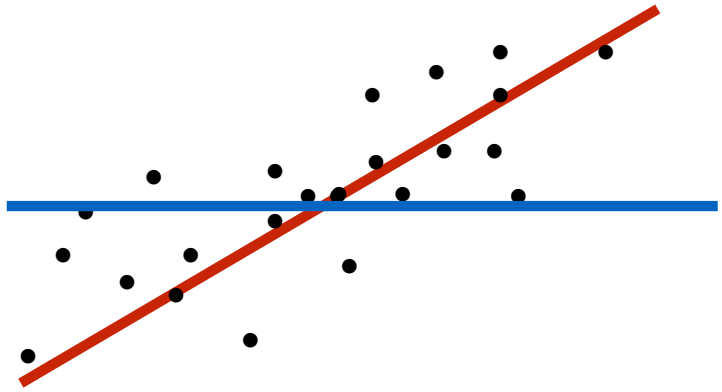
- $(\mathbf{X}, Y)$ (data, and new examples) come from some distribution
  we learn best linear **predictor**

- The world is linear; we know the structure of the relation
  we learn the **coefficients of the structural relation**

PERFORMANCE ASSESSMENT

R^2 (R-squared)

- Prediction if no regression: $\overline{Y} = \frac{1}{n} Y_i$



- Total sum of squares:

  $$\text{TSS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

  "initial" variation in Y

- Residual sum of squares: $\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{\boldsymbol{\theta}}^T \mathbf{X}_i)^2$

  unexplained variation in Y, after taking into account X

- $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$ 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: $1 - \dfrac{\text{RSS}/(n-m-1)}{\text{TSS}/(n-1)}$ $0.897 \to 0.896$

How noisy/reliable are my estimates of θ^*

- Assume structural model

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

W_i: independent, zero mean, variance σ^2

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

next session

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

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

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

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

bias variance

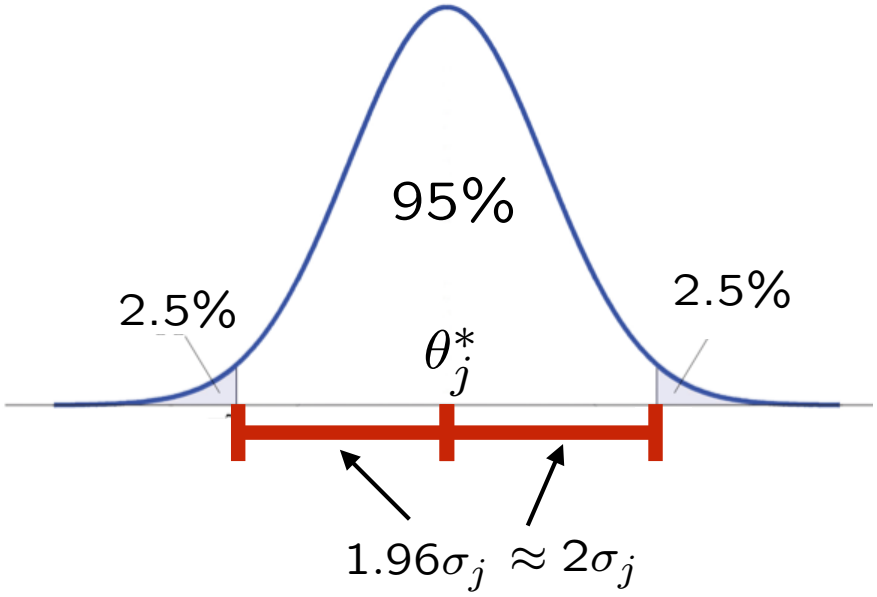
$\mathbb{E}[\widehat{\Theta}_j] = \theta_j^*$ (unbiased)

- 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

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

W_i: independent,
zero mean, variance σ^2

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

- 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 $\widehat{\Theta}$

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

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

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

W_i: independent, zero mean, variance σ^2

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

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

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

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

use $\hat{\sigma}^2$

	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
News paper	-5.91021239E-04	-3.26595026E-07	-1.78006245E-05	**3.44687543E-05**

- Estimate σ^2 by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\boldsymbol{\theta}}^T\mathbf{X}_i\right)^2$

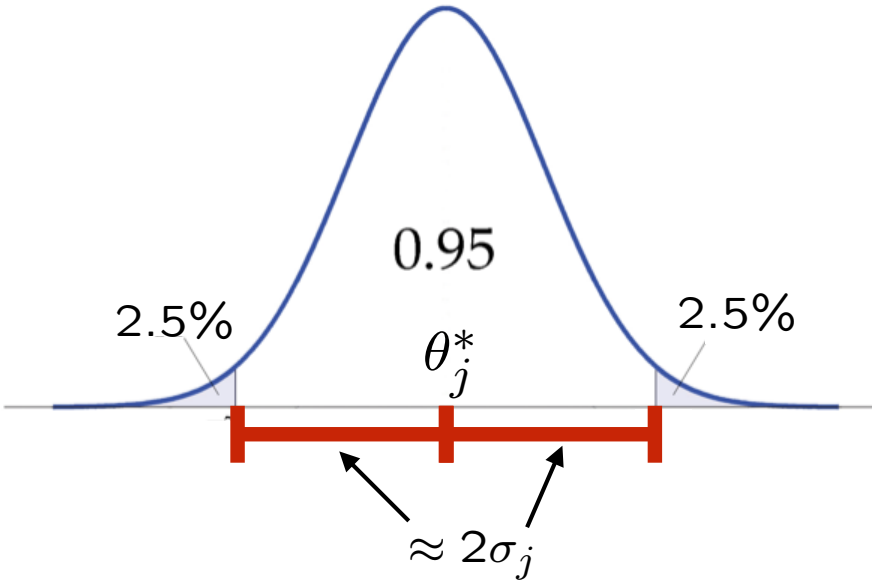
slight downwards bias

negligible bias if $m \ll n$

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

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

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: $\left[\widehat{\Theta}_j - 2\hat{\sigma}_j, \quad \widehat{\Theta}_j + 2\hat{\sigma}_j\right]$

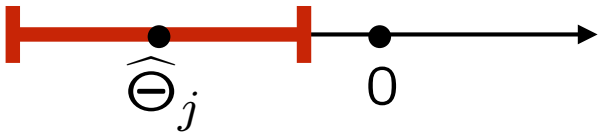
$$\mathbb{P}(\theta_j^* \in \text{CI}) \approx 0.95$$

- Needs careful ("frequentist") interpretation

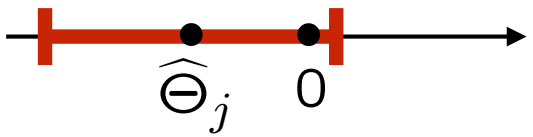
Testing the hypothesis $\theta_j^* = 0$

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

Wald test:



reject the null



do not reject the null
(data compatible with $\theta_j^* = 0$)

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

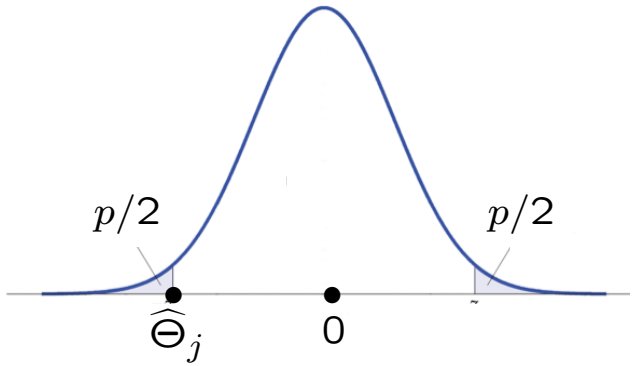
$$= \mathbb{P}(\text{the CI ``misses'' } 0 \mid \theta_j^* = 0)$$

$$\approx 5\%$$

---

p-value: probability of seeing something at least as extreme as the observed $\widehat{\Theta}_j$, under $\theta_j^* = 0$

- 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**

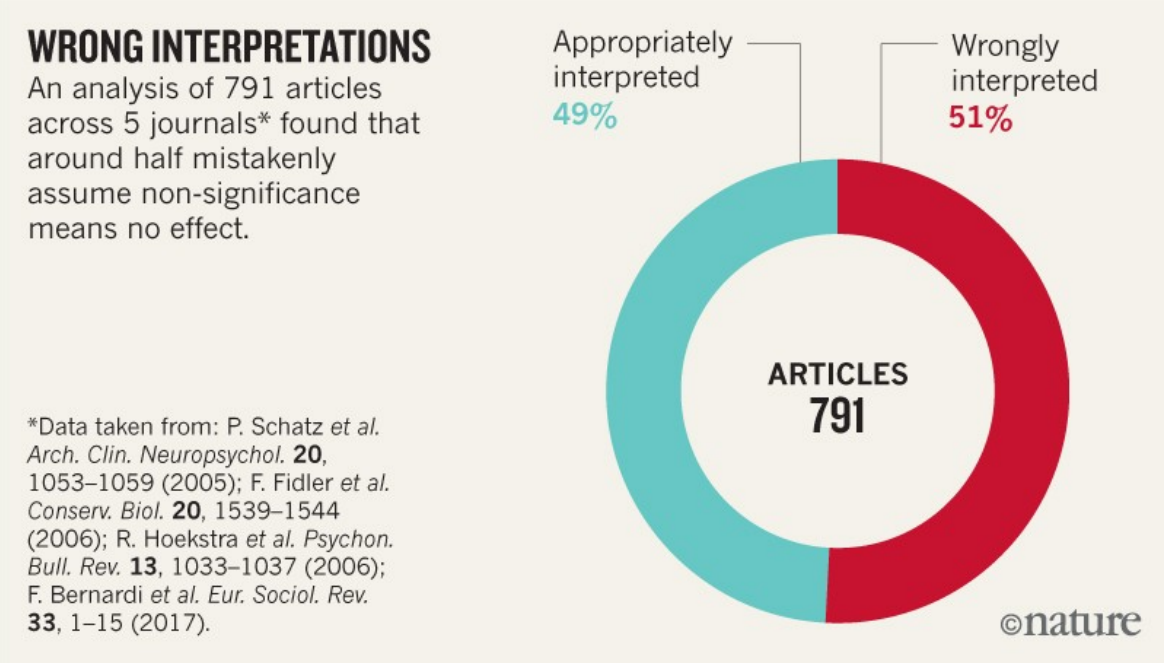
	coef	std err	Confidence intervals [0.025	0.975]
Intercept	2.9389	0.312	2.324	3.554
TV	0.0458	0.001	0.043	0.049
Radio	0.1885	0.009	0.172	0.206
Newspaper	-0.0010	0.006	-0.013	0.011

- 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)

Interpretation needs care

Scientists rise up against statistical significance, *Nature*, 20 March 2019



- Reject the null $\theta_j^* = 0$:
  what we see is unlikely to have been generated by a model with $\theta_j^* = 0$

- Do not reject the null
  data do not provide compelling evidence that $\theta_j^* \neq 0$
  - no effect: θ_j^* is zero
  - small effect: θ_j^* is so close to zero that data cannot detect it
  - too few data: θ_j^* may be nonzero, but need more data to "see it"

Making new predictions

- After running the regression

  given some new $\mathbf{X}$, predict Y $\qquad \widehat{Y} = \widehat{\boldsymbol{\theta}}^T \mathbf{X}$

$\mathbf{X}_1$	Y_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\mathbf{X}_n$	Y_n
$\mathbf{X}$	Y?

- Keep assuming structural model: $Y = (\boldsymbol{\theta}^*)^T \mathbf{X} + W$

- $\widehat{\boldsymbol{\theta}}$ is unbiased estimate of $\boldsymbol{\theta}^*$

  $\Rightarrow \widehat{\boldsymbol{\theta}}^T \mathbf{X}$ is unbiased estimate of $(\boldsymbol{\theta}^*)^T \mathbf{X}$ (and of Y)

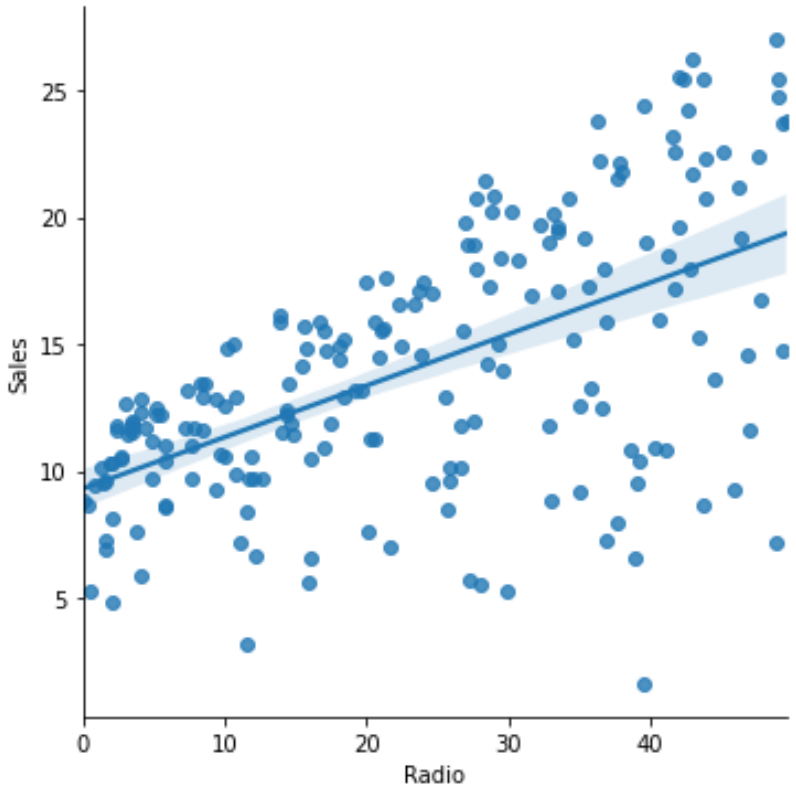
- Two sources of error:
  - unavoidable, from W; variance σ^2
  - variance of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{X}$ $\qquad \sigma^2 \mathbf{X}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{X}$
  - Total prediction error variance: $\qquad \sigma^2 + \sigma^2 \mathbf{X}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{X}$

Confidence bands

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

  $(\widehat{\boldsymbol{\theta}})^T\mathbf{X}$ plus or minus $\quad 2 \cdot \widehat{\sigma}\sqrt{\mathbf{X}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{X}}$

  - 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...

Appendix:
math details and extras

Plugin estimators: use data instead of $\mathbb{P}$

$\mu = \mathbb{E}[X]$ $\quad \widehat{M} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \xrightarrow[n \to \infty]{} \mu \quad$ (LLN)

$\gamma = \mathbb{E}[X^3]$ $\quad \widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} X_i^3$

$\phi = \mathbb{E}\big[g(X)\big]$ $\quad \widehat{\Phi} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$

$v = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big]$ $\quad \widehat{V} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{M})^2$

$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 \quad$ (if we know μ)

$a = \text{median}(\mathbb{P})$ $\quad \widehat{A} = \text{median(data)}$

Maximum likelihood estimators

- Assume X_i discrete, i.i.d.; $X = (X_1, \ldots, X_n)$

Likelihood function: $L^\theta(x_1, \ldots, x_n) = \mathbb{P}^\theta(X = x) = \prod_{i=1}^{n} \mathbb{P}^\theta(X_i = x_i)$

- Find θ under which the observed x is most likely to have been observed

having seen x: $\max_\theta L^\theta(x)$ $\max_\theta \sum_{i=1}^{n} \log\left(\mathbb{P}^\theta(X_i = x_i)\right)$

Sometimes analytically Usually numerically

- Example: $X \sim \text{Binomial}(n, \theta)$

$\max_\theta \binom{n}{x} \theta^x (1-\theta)^{1-x}$ algebra $\longrightarrow$ $\widehat{\theta} = \frac{x}{n}$

Maximum likelihood — continuous data

- X_i, independent, PDF $\sim f^\theta_{X_i}(\cdot)$ $\qquad X = (X_1, \ldots, X_n)$

Likelihood function: $L^\theta(x_1, \ldots, x_n) = f^\theta_X(x) = \prod_{i=1}^n f^\theta_{X_i}(x_i)$

having seen x: $\max_\theta L^\theta(x)$ $\qquad \max_\theta \sum_{i=1}^n \log\left(f^\theta_{X_i}(x_i)\right)$

- Example: $X_i \sim \text{Exp}(\theta)$, $\quad f^\theta(x) = \theta e^{-\theta x}$, $x \geq 0$

Bayesian estimators

- Bayesian philosophy
  - parameter treated as (unobserved) realized value of random variable Θ
  - prior distribution $\mathbb{P}_\Theta$: $p_\Theta(\theta)$ or $f_\Theta(\theta)$
    joint PMF/PDF if θ is a vector
    - Prior comes from: symmetry; known range; earlier studies; subjective beliefs
- Observation model $\mathbb{P}_{X|\Theta}$: $p_{X|\Theta}(x \mid \theta)$ or $f_{X|\Theta}(x \mid \theta)$
- Let $D = (X_1, \ldots, X_n)$

$$\mathbb{P}_\Theta(\theta) \cdot \mathbb{P}_{D|\Theta}(d \mid \theta) = \mathbb{P}_{\Theta,D}(\theta, d) = \mathbb{P}_D(d) \cdot \mathbb{P}_{\Theta|D}(\theta \mid d)$$

$$\mathbb{P}_{\Theta|X} = \frac{\mathbb{P}_\Theta \cdot \mathbb{P}_{X|\Theta}}{\mathbb{P}_X}$$