Top 10 Chestii pe care orice cercetator ar trebui sa le stie despre regresie

Obiective

- Prezentare generala a regresiei statistice si aplicatii
  - Detalii tehnice
  - Context istoric: cum erau rezolvate probleme statistice inainte ca statistica sa existe? Cum a aparut si s-a dezvoltat regresia?

- Citeva conceptii gresite dar raspandite despre cum functioneaza regresia

Agenda

- **I.** $(X'X)^{-1}X'y$ e doar inceputul
- **II**. Vizualizare! Vizualizare! Vizualizare!
- **III**. Corelatie Nu e Cauzalitate
- **IV.** Regresie Catre Mediocritate
- **V.** Nu Cercetatorii Ti-au Creat Datele

Overview

- **VI.** Interpretarea Non-stiintifica a Rezultatelor
- **VII**. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie
- **VIII**. Sa Nu Incluzi Prea Putine Variabile
- **IX.** Sa Nu Incluzi Prea Multe Variabile
- **X.** Cash is King

I. $(X^TX)^{-1}X^Ty$ Este Doar Inceputul

- Sa consideram un sistem de ecuatii liniare (infinitate de solutii):

$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i, \ (i = 1, 2, \ldots, m),$$

- Deoarece nu exista un set unic de valori β solutii pt **X**β= y, putem incerca sa cautam un β care sa minimizeze distanta (sau, echivalent, patratul distantei) dintre **X**β si y :

$$\sum_{i=1}^{m} \left| \sum_{j=1}^{n} X_{ij}\beta_j - y_i \right|^2 \rightarrow \min.$$

- Solutia vine in forma ecuatiei normale de la regresii statistice:

$$(\mathbf{X^T X})\,\hat{\boldsymbol{\beta}} = \mathbf{X^T y}.$$

- *This concludes today's presentation …*

I. $(X^TX)^{-1}X^Ty$ Este Doar Inceputul

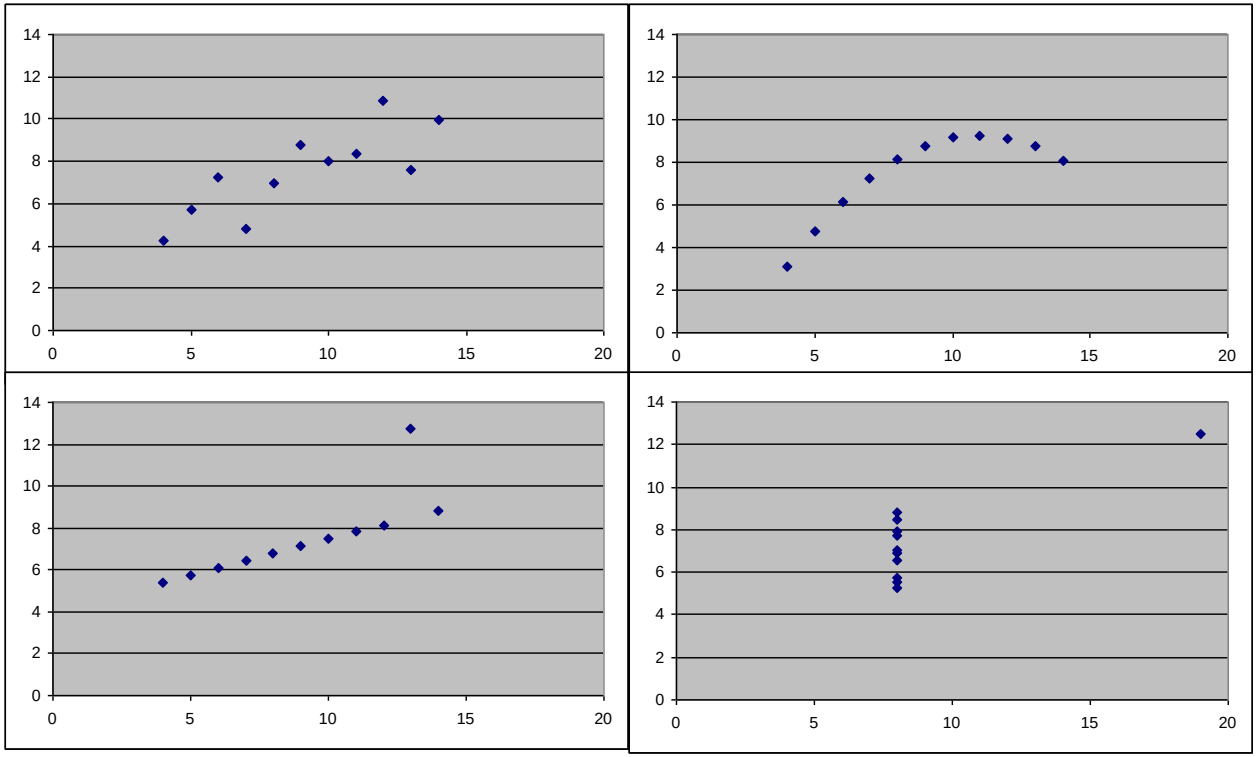
- **Quiz Time!**

- Urmatoarea pagina contine 4 seturi de date diferite reprezentate grafic

- Care grapfic are:

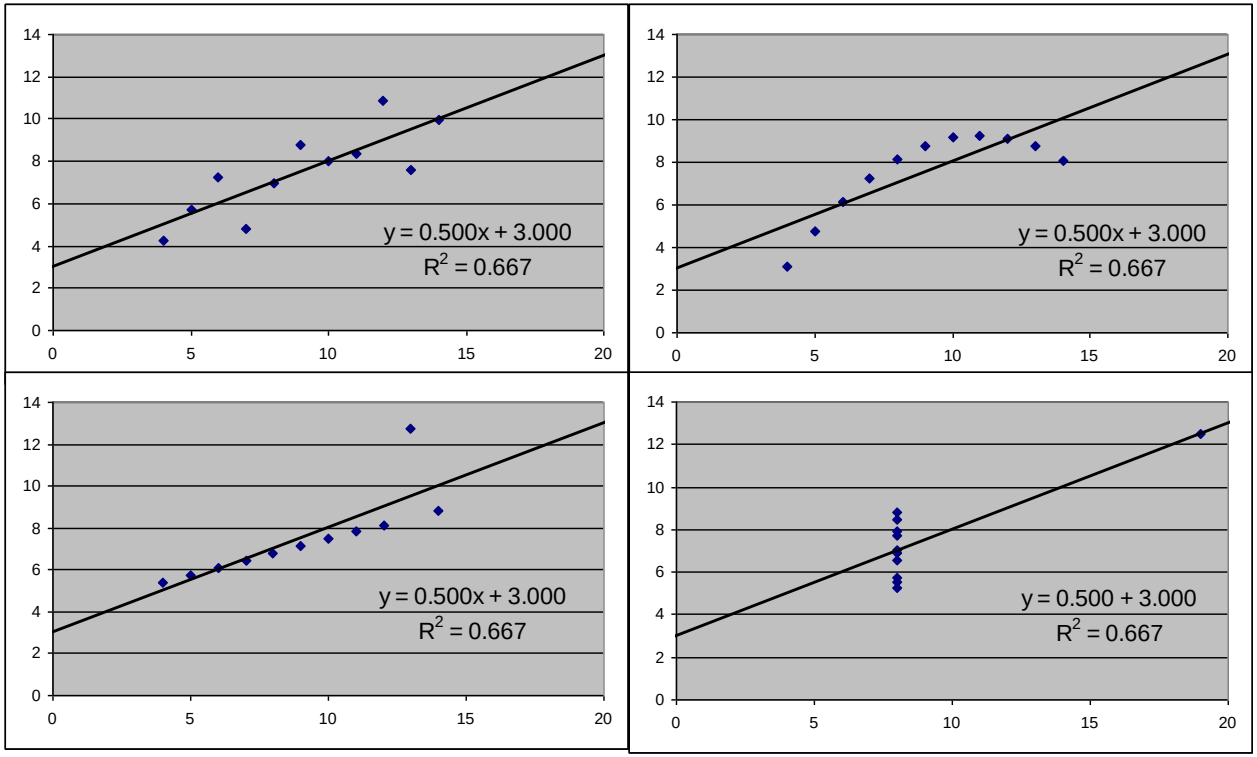
  - O dreapta de regresie $Y = 0.5X + 3$?
  - O valoare R^2 de 0.667?

I. $(X^TX)^{-1}X^Ty$ Este Doar Inceputul



- Dreapta de regresie: Y = 0.5*X+3
- $R^{2:}$ 0.667

I. $(X^TX)^{-1}X^Ty$ Este Doar Inceputul



- Felicitari, ai castigat ☺!

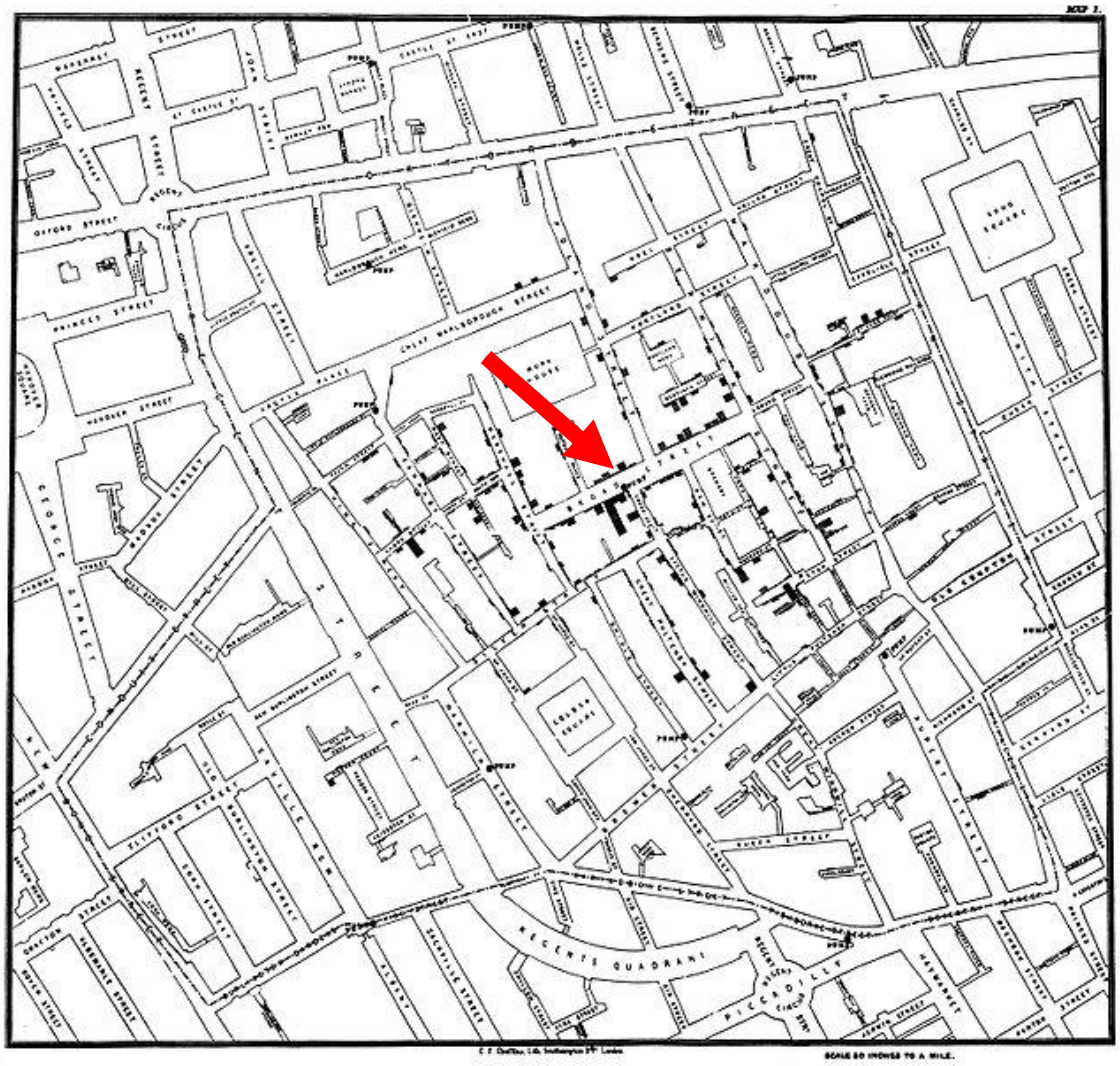
II. Vizualizare! Vizualizare! Vizualizare!

- O imagine buna valoreaza cat 1000 de drepte de regresie
  - Mai ales cand audienta nu e formata din statisticieni
  - ... sau, istoric vorbind, cand nu exista statistica

- Sa-l cunoastem pe John Snow, MD
  - Doctor si anestezist
  - Considerat adesea primul epidemiologist modern
  - A oprit raspandirea holerei in Londra prin colectarea datelor si analiza statistica descriptive
  - A murit cu 30 de ani inainte de desenarea primei drepte de regresie

John Snow, 1813-1858

II. Vizualizare! Vizualizare! Vizualizare!



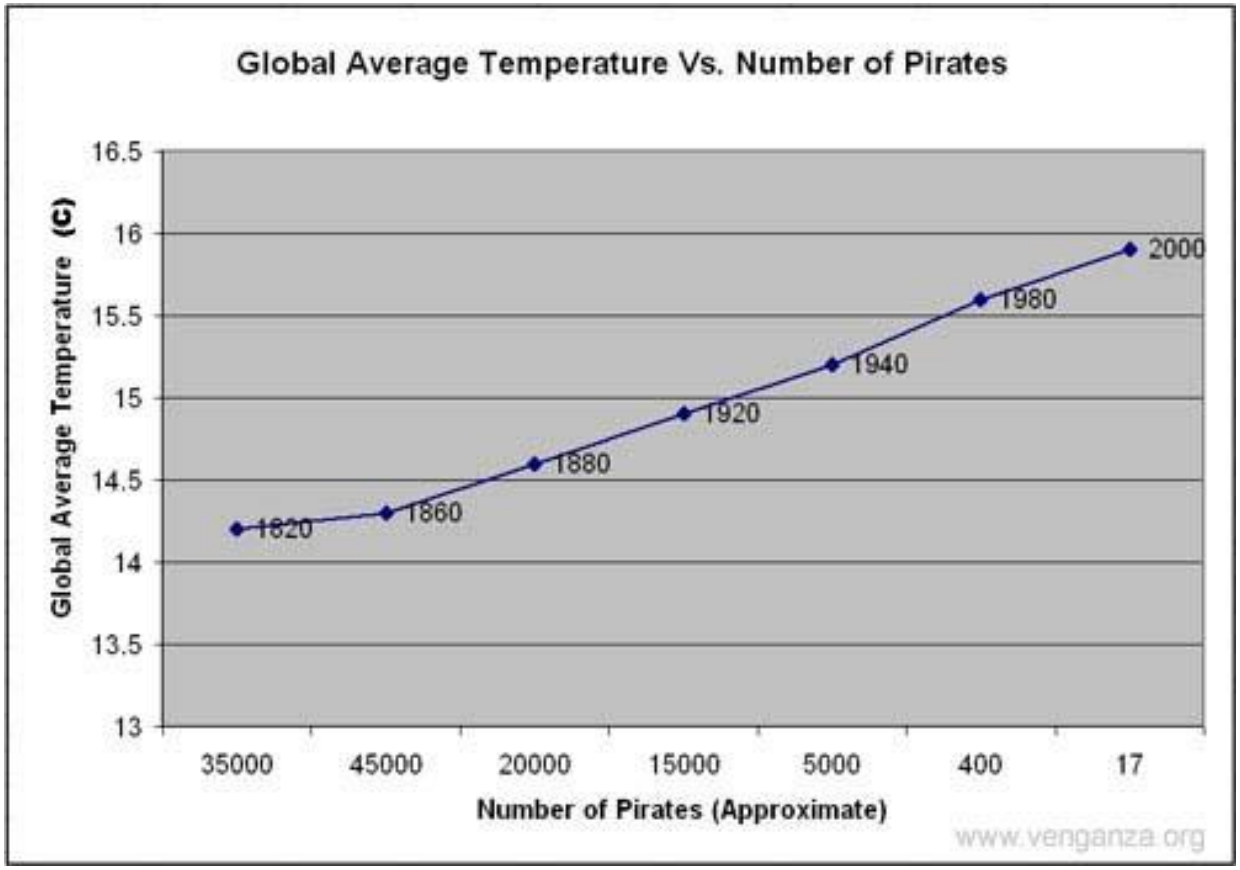
- 1854 London – izbucneste holera
- Harta regiunii afecate
- Punctele indica localizarea pacientilor cunoscuti
- Snow a colectat cu rabdare informatia si a desenat harta
- A dedus ca sursa holerei trebuie sa fie in epicentrul geografic

II. Vizualizare! Vizualizare! Vizualizare!

- Concluzie: ceva pe Broad Street cauza raspandire
  - Prin investigare amanuntita, a ajuns la concluzia ca pompa de apa era cauza
  - Credea in noua teorie a germenilor si a folosit harta ca suport
  - A examinat elementele extreme si a rezultat ca si acolo tot apa de la aceeasi pompa era folosita
  - A oprit pompa de apa de pe Broad Street si a oprit epidemia

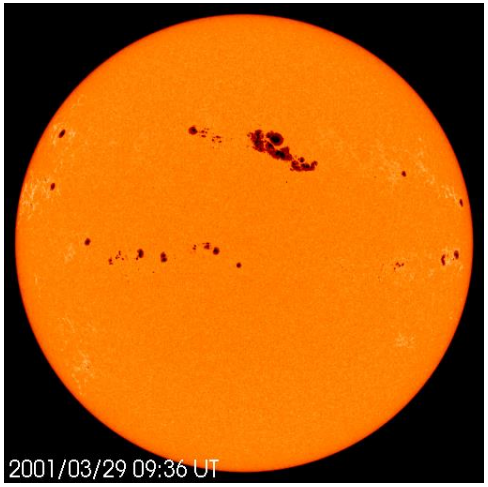
- Nota: fara ajutorul unei teorii stiintifice (germeni), rezultatele pot fi interpretate gresit si isi pierd din valoare

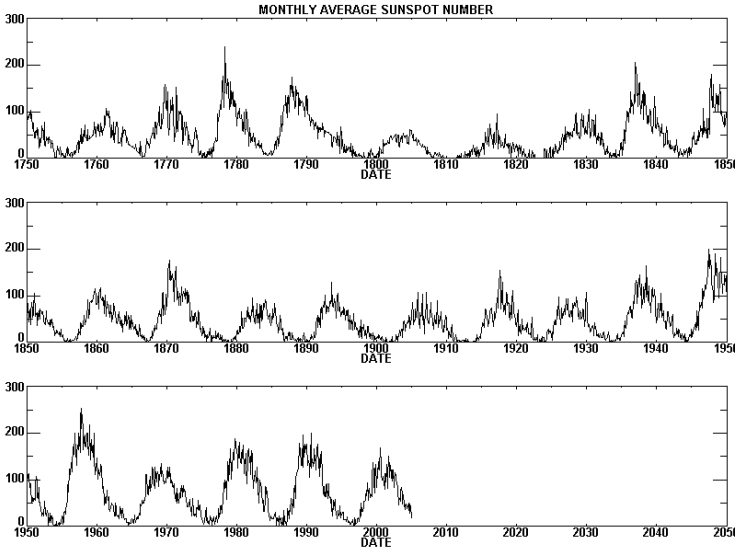
III. Correlatie nu e Cauzalitate



III. Corelatie Nu e Cauzalitate

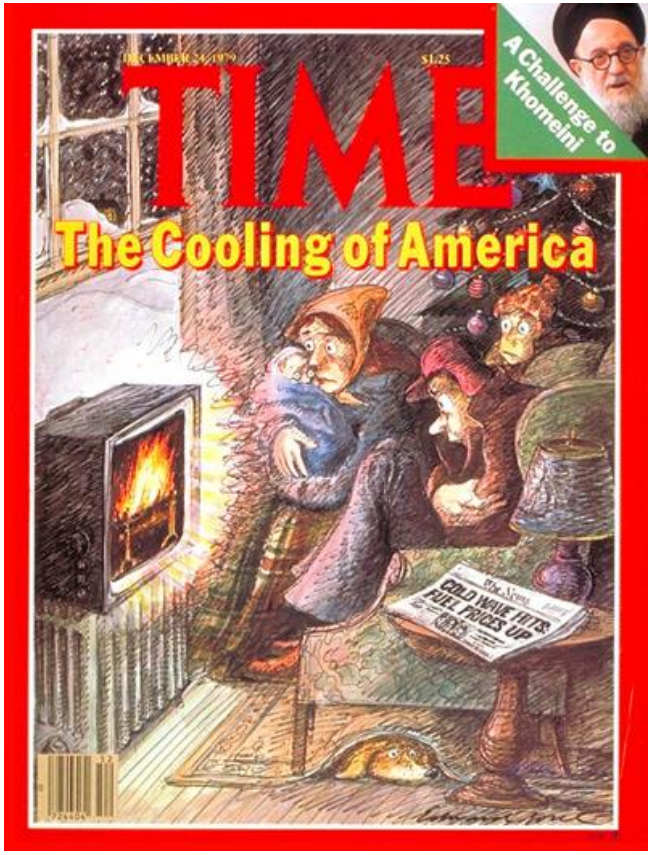
- Exemplu: pete solare
  - Indica furtuni pe suprafata soarelui
  - Approximativ odata la 11 ani (seasonality)
    - Maxime si minime Solare maxima and minima apar approximative odata la 11 ani, cu toate ca valorile extreme au o variatie foarte mare.
  - In cursul observarii s-au inregistrat si perioade mai lungi de 11 ani
  - Sunt legate de caldura degajata de soare
  - Regiunile cu pete sint sensibil mai reci
  - Petele acopera sufficient de multa suprafata ca sa produca reduceri semnificative in radiatia solara
  - Au potential in afectarea temperaturilor pe pamant





III. Corelatie Nu e Cauzalitate

- Coperte TIME Magazine publicate in anii cu cele mai multe (stanga) si cele mai putine (dreapta) pete solare raportate in istoria recenta:



III. Corelatie Nu e Cauzalitate

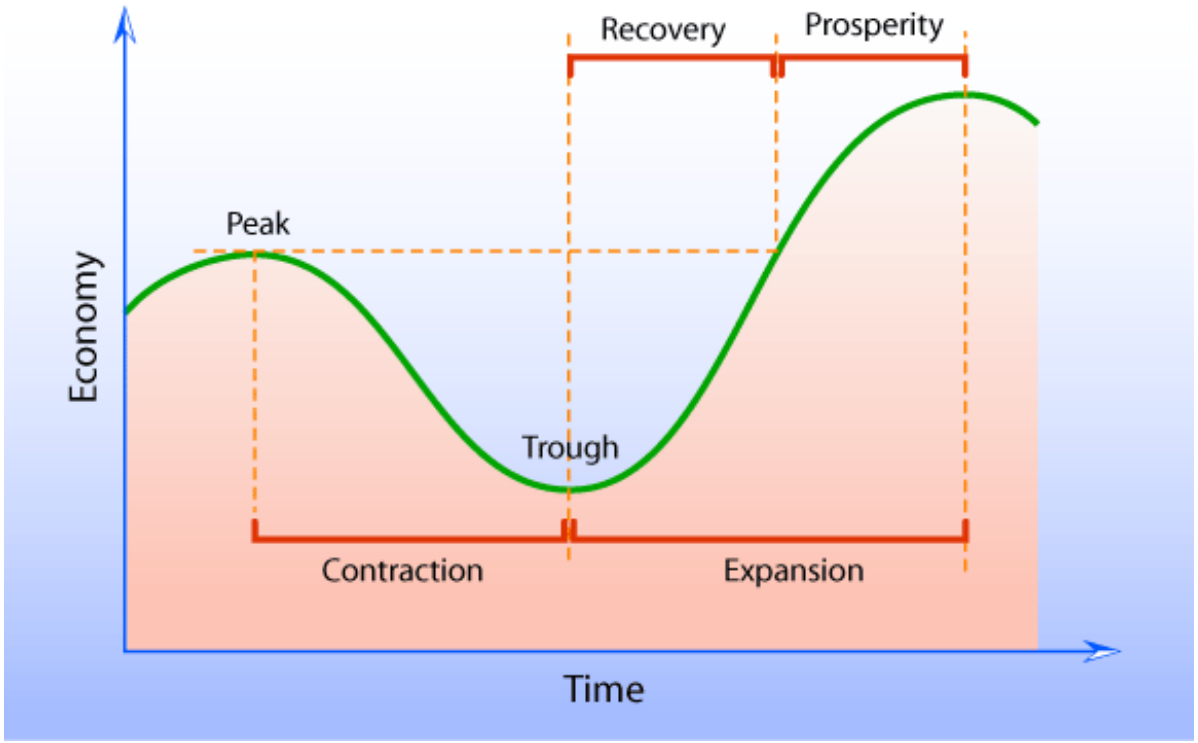
- Consens Stiintific atestand efectul puternic al petelor solare asupra vremii
  - Desi sunt disensiuni pe tema efectelor de alta natura decat temperature globala…
  - … si a importantei altor factori care afecteaza temperature globala
- In trecut au existat tentative de utilizare a acestor date in alte prognoze

- Sa facem cunostinta cu William Stanley Jevons
  - Economist si Logician
  - Contributii in Utilitate Marginala, Peak Oil Theory
  - A lucrat extensiv cu ciclurile economice
  - ‘celebru’ pentru ca a propovaduit o legatura intre economie si ciclurile petelor solare

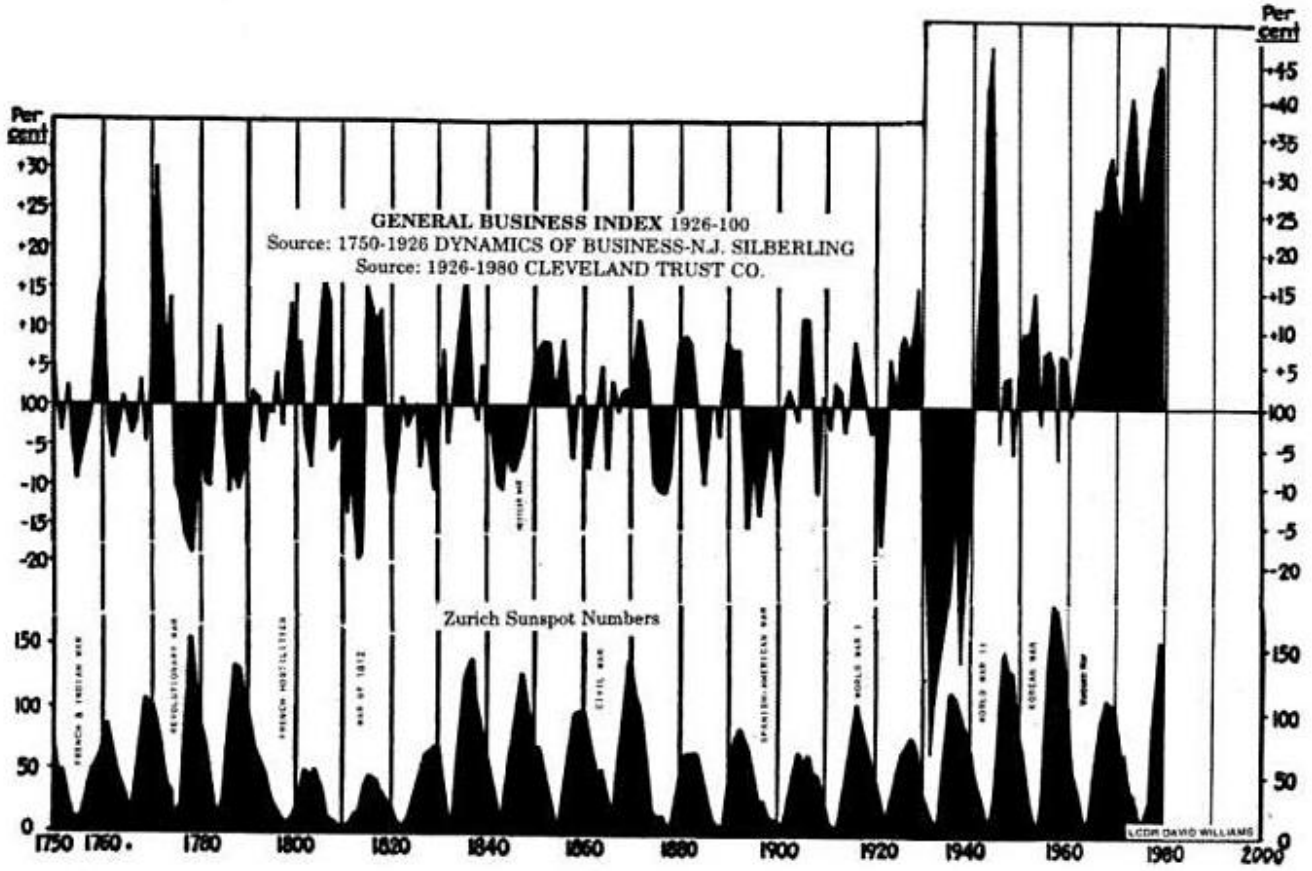
W.S. Jevons, 1835-1882

III. Corelatie Nu e Cauzalitate

- Ciclul economic
- Economia se dilata si se contracta relativ predictictibil
- Pe vremea lui Jevon, maximele economice erau cam odata la 11 ani …
- … si economia era majoritar bazata pe agricultura …



III. Corelatie Nu e Cauzalitate



- O relatie e sugestiva, insa nu a rezistat unei analize formale …

IV. Regresie Catre Mediocritate

- Motivatia care a dus la definite formala a regresie
- Sa presupunem ca luam 12 masteranzi si le dam fiecaruia o moneda de 0,5RON
- Fiecare este rugat sa ‘dea cu banul’ de 10 ori
- Fiecare is noteaza numarul de ‘steme’ rezultat
- Monedele sunt ok si 60 de steme

s-au inregistrat in total

IV. Regresie Catre Mediocritate

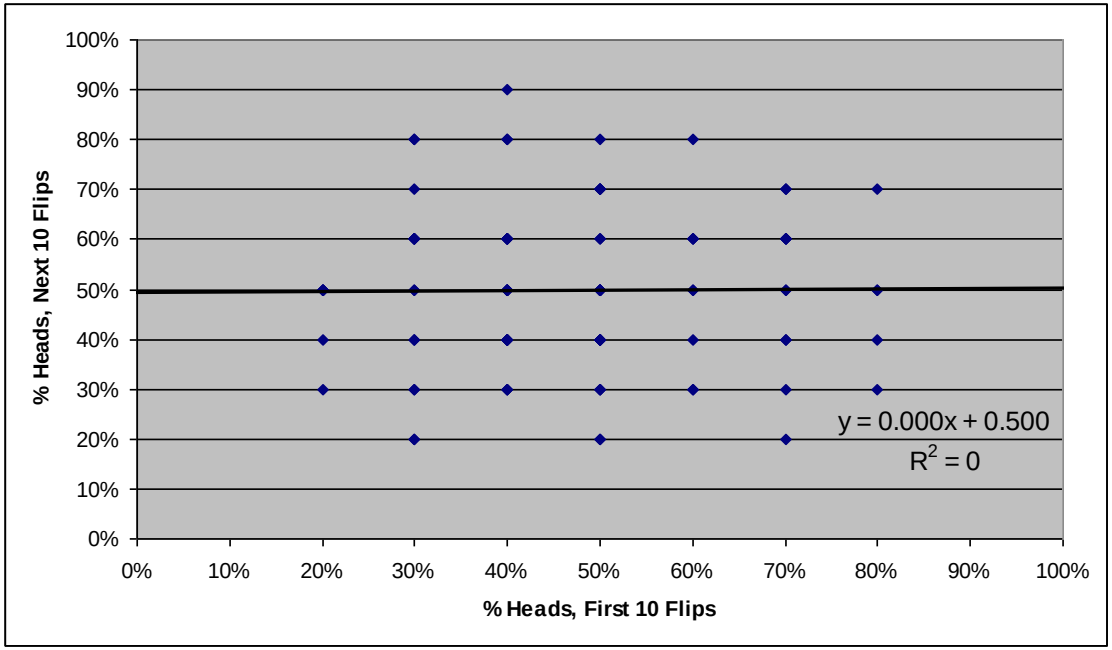
- Sa presupunem ca masterandul A a dat 8 steme din 10 incercari (80%)
- Si masterandul B 2 steme din 10 incercari (20%)
- Ce ar trebui sa ne asteptam din partea fiecaruia daca ii lasam sa mai ‘dea’ de 10 ori?
- Ne asteptam ca cel mai bun scor (A) sa scada? Scorul mai slab (B) sa creasca?

masterand A, 1992-

Hint: 50%, da, si da

masterand B, 1992-

IV. Regresie Catre Mediocritate



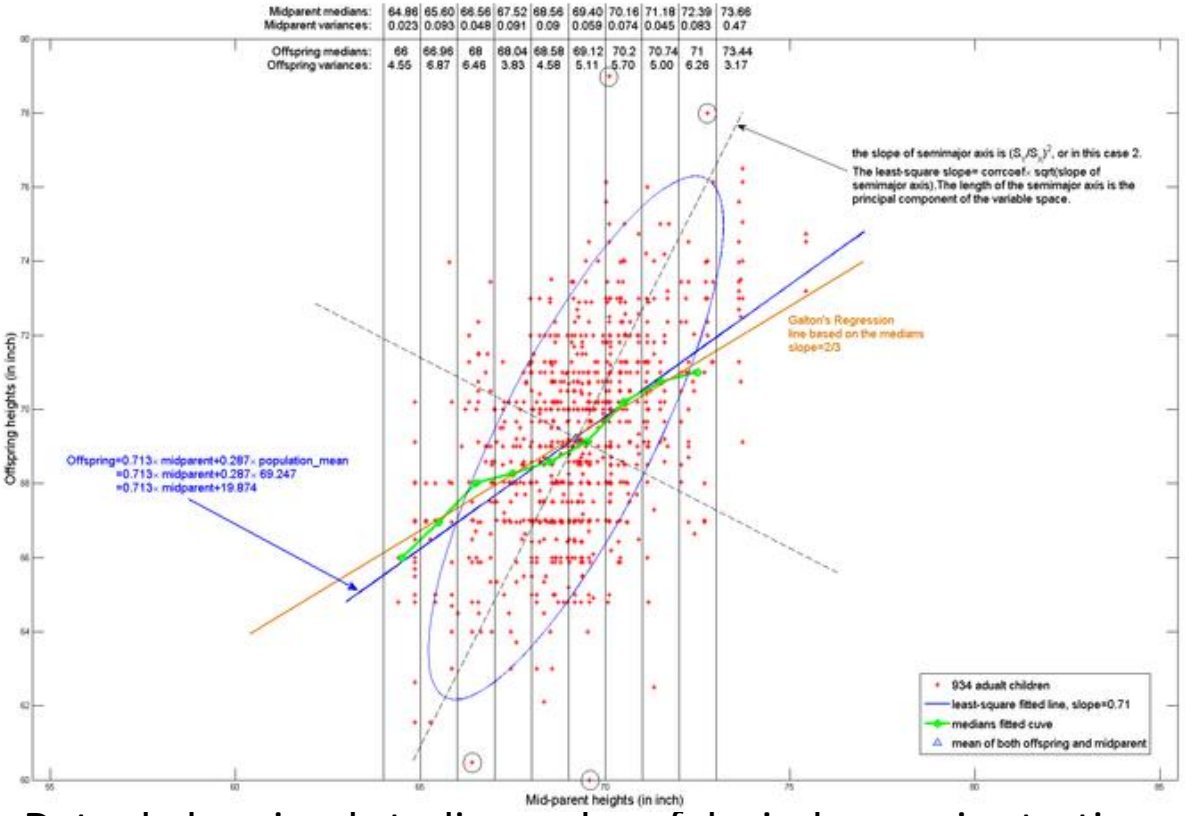
- Axa 0X: rezultatele primelor 10 observatii; axa 0Y: rezultatele urmatoarelor 10
- Regresie catre medie: un sir norocos urmat de un sir mai putin norocos (si vice versa)
- Dar daca procesul nu este aleator? Daca unii oameni se pricep mai bine la dat cu banul decat semenii lor? Ar fi runda a 2 –a mai asemanatoare cu prima?

IV. Regresie Catre Mediocritate

- Sa-l cunoastem pe Sir Francis Galton
  - Matematician, Antropolog, Psiholog
  - Responsabil pentru dezvolatrea notiunilor de regresie si corelatie
  - Initiator al cercetarii asupra ereditatii si geneticii
  - Pionier in apilcarea statisticii pe o populatie umana
- Studii Genetice
  - Ce attribute/trasaturi pot mosteni copiii de la parinti?
  - Se pot imbunatati caracteristici exceptionale in timp?
  - Sau toate trasaturile umane vor regresa spre mediocritate?

Francis Galton, 1822-1911

IV. Regresie Catre Mediocritate



- Date de la primul studiu modern folosind regresia stastica
- Inaltimea medie a parintilor pe axa X, cea a copiilor pe Y
- Cei mai inalti parinti au copii inalti care sunt mai mici decat ei
- Cei mai scunzi parinti au copii scunzi mai inalti decat ei

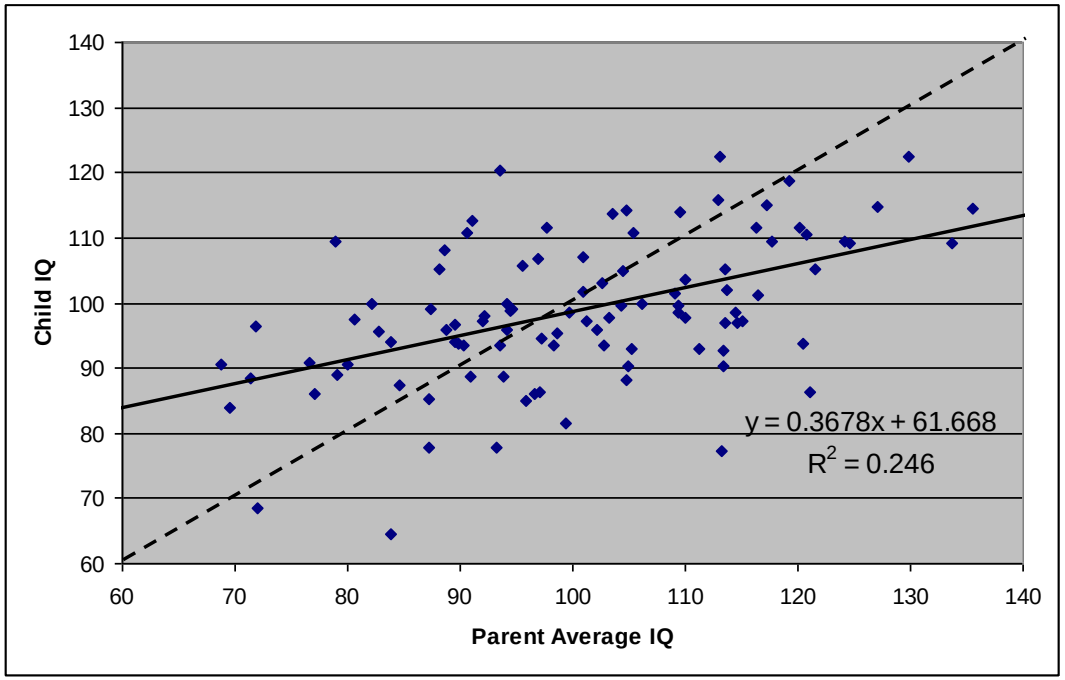
V. Nu Cercetatorii Ti-au Creat Datele

- Galton nu era prea interesat de inaltime
  - Dorea imbunatatirea genetica a tuturor caracteristicilor mostenite
  - Astfel a aparut o stiinta numita *Eugenics*

- Cel mai de seaman printre factorii Eugenici era *inteligenta*
  - Se presupunea ca se mosteneste – demonstrate in gluma chiar de Dalton referindu-se la arborele lui genetic (un verisor al lui era chiar Charles Darwin) si la cel al altor contemporani
  - A produs o demonstratie formala ulterior prin regresii elaborate
  - Problema: atunci, ca si acum, definitia exacta a inteligentei e eluziva

- A dorit sa culeaga date despre factorii care presupunea ca se mostenesc

V. Nu Cercetatorii Ti-au Creat Datele



- Inteligenta Mostenita: 38% (panta dreptei negru solid)
- Linia interupta reprezinta copii la fel de inteligenti ca parintii (sub ea, copii mai putin inteligenti; deasupra copii mai inteligenti ca parintii)
- Parinti mai inteligenti au mai multi copii sub interupta

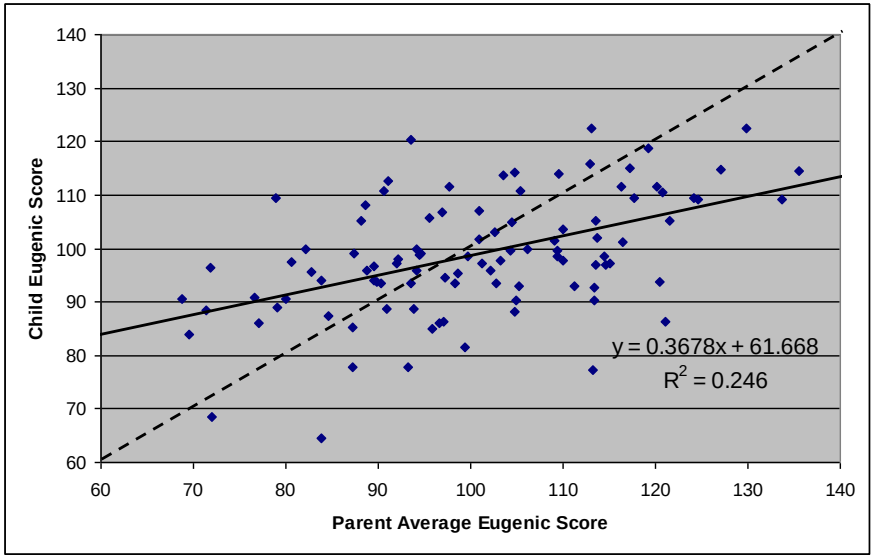
V. Nu Cercetatorii Ti-au Creat Datele

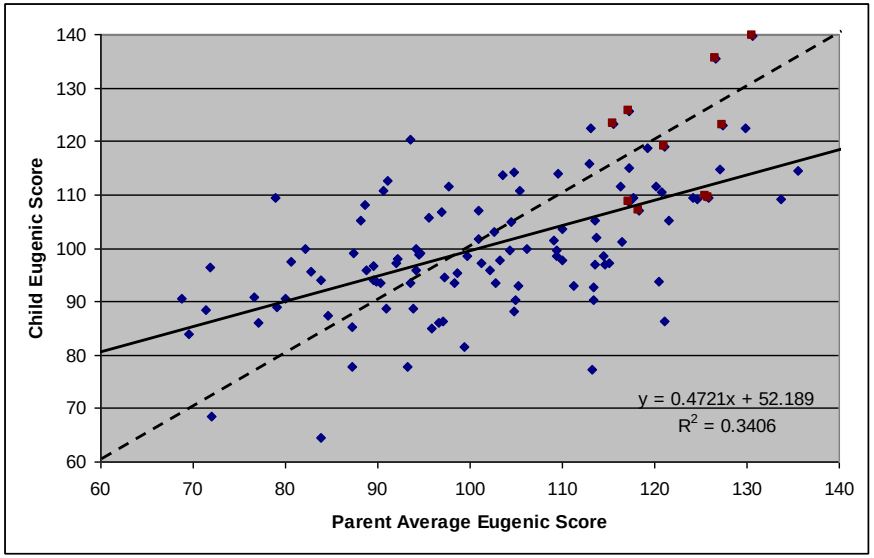
- Pentru fiecare cuplu studiat s-a calculat un scor Eugenic
  - Spera ca acel scor sa fie transmis urmasilor si tuturor generatiilor urmatoare
  - Astfel, peste cateva generatii, noua ‘mediocritate’ s-ar fi imbunatatit

- Componente importante ale scorului
  - Pentru tati: inteligenta (masurata relativ vag)
  - Pentru mame: frumusetea (masurata foarte vag)
  - Sanatate (lipsa in storic a bolilor)

- Desigur multe dintre caracteristici erau greu de masurat, si au recurs la surogate
  - Clasa sociala, Bogatie, Nationalitate, Religie, Rasa …

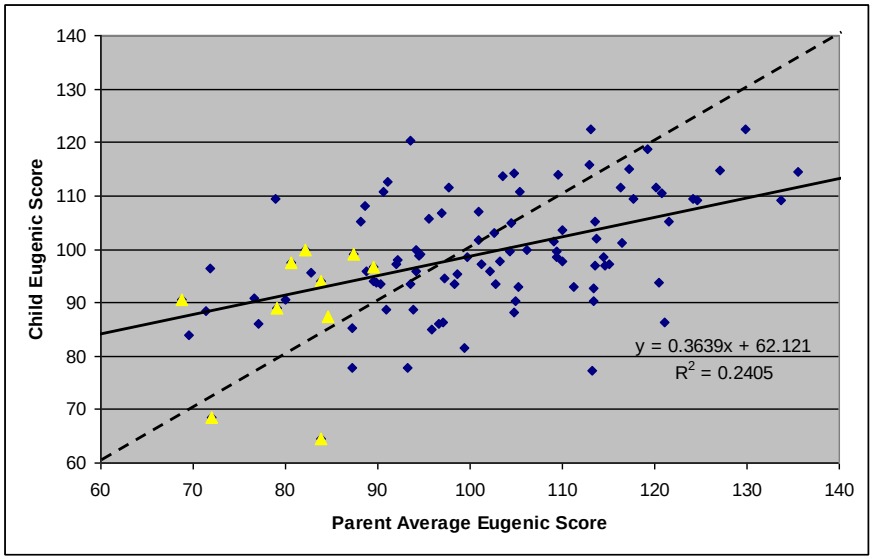
VI. Interpretarea Non-stiintifica a Rezultatelor

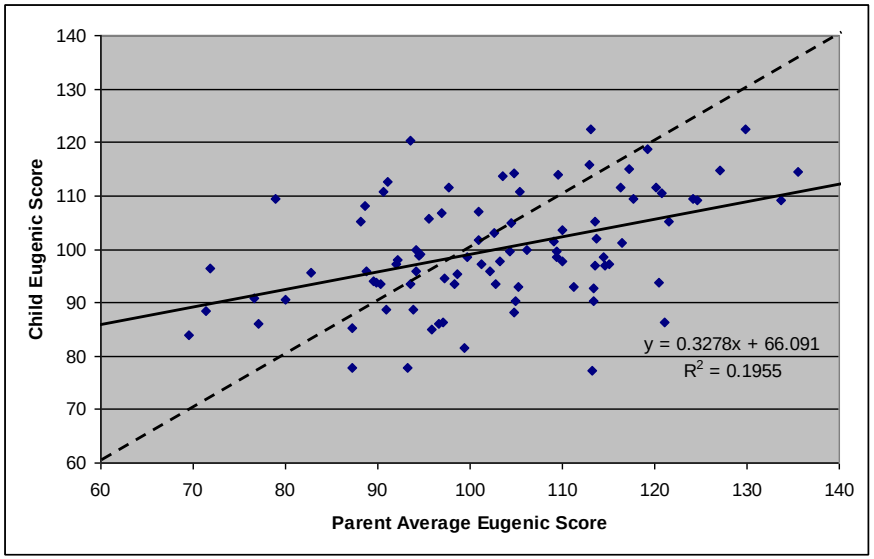




- Solutia lui Galton: Incurajarea imperecherii intre cei cu scor Eugenic mare (e.g. Oameni de stiinta/intelectuali cu femei frumoase)
- Al doilea graphic arata cum cu 10 cupluri cu scor mare s-ar imbunatati populatia genetica pentru generatiile urmatoare (in termeni de scor Eugenic)

VI. Interpretarea Non-stiintifica a Rezultatelor





- Slide-ul precedent este o aplicatie de *positive eugenics*.
- Succesorii lui Galton au realizat rapid beneficiile unei *negative eugenics* – reducerea numarului de copii in zona celor cu scoruri mici

VI. Interpretarea Non-stiintifica a Rezultatelor

- Nebunia a continuat: cum trebuie descurajata natalitatea celor cu scor mic? La 60 de ani de la moartea lui Galton, urmatoarele au fost incercate in SUA si Europa:
  - Diseminarea teoriei Eugenice catre audiente generale
  - Facilitati fiscale claselor superioare ale societatii pentru a incuraja natalitatea
  - Legalizarea contraceptiei, raspandirea mijloacelor ieftine de contraceptie
  - Sterilizarea disgenicilor contra unor plati
  - Legalizarea avortului
  - Reguli stricte de imigrare
  - Incriminarea casatoriilor inter-rasiale pana la interzicerea lor
  - Sterilizare fortata pentru bolnavii mental si infractori incarcerati (65,000)
  - Genocid (6,000,000)

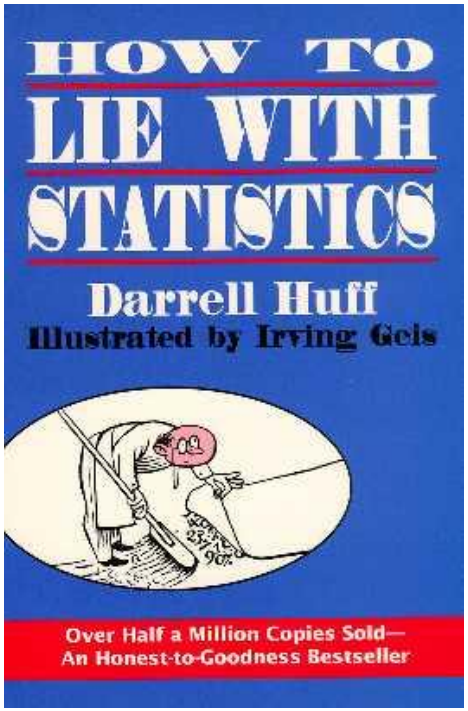
VI. Interpretarea Non-stiintifica a Rezultatelor

- Nu putem da vina doar pe cercetatori pentru actiunile politicienilor si afaceristilor
  - Asigurati-va ca ajutati la interpretarea datelor
  - Asigurati-va ca nu faceti presupuneri eronate sau erori fundamentale
  - Verificati modul de utilizare a rezultatelor si atrageti atentia unde este cazul
- Ca urmare: teribilul esec al Eugeniei s-a raspandit asupra tuturor aplicatiilor ale statisticii in genetica, medicina si biologie
  - Universitatile au inchis departamentele de Eugenics; a trecut o generatie pana cand programe academice legate de Genetica, Biostatistica au reaparut
  - Multe vieti ar fi putut fi salvate sau imbunatatite intre timp

VII. Corelatia Nu e Cauzalitate, dar Poate Fi o Indicatie

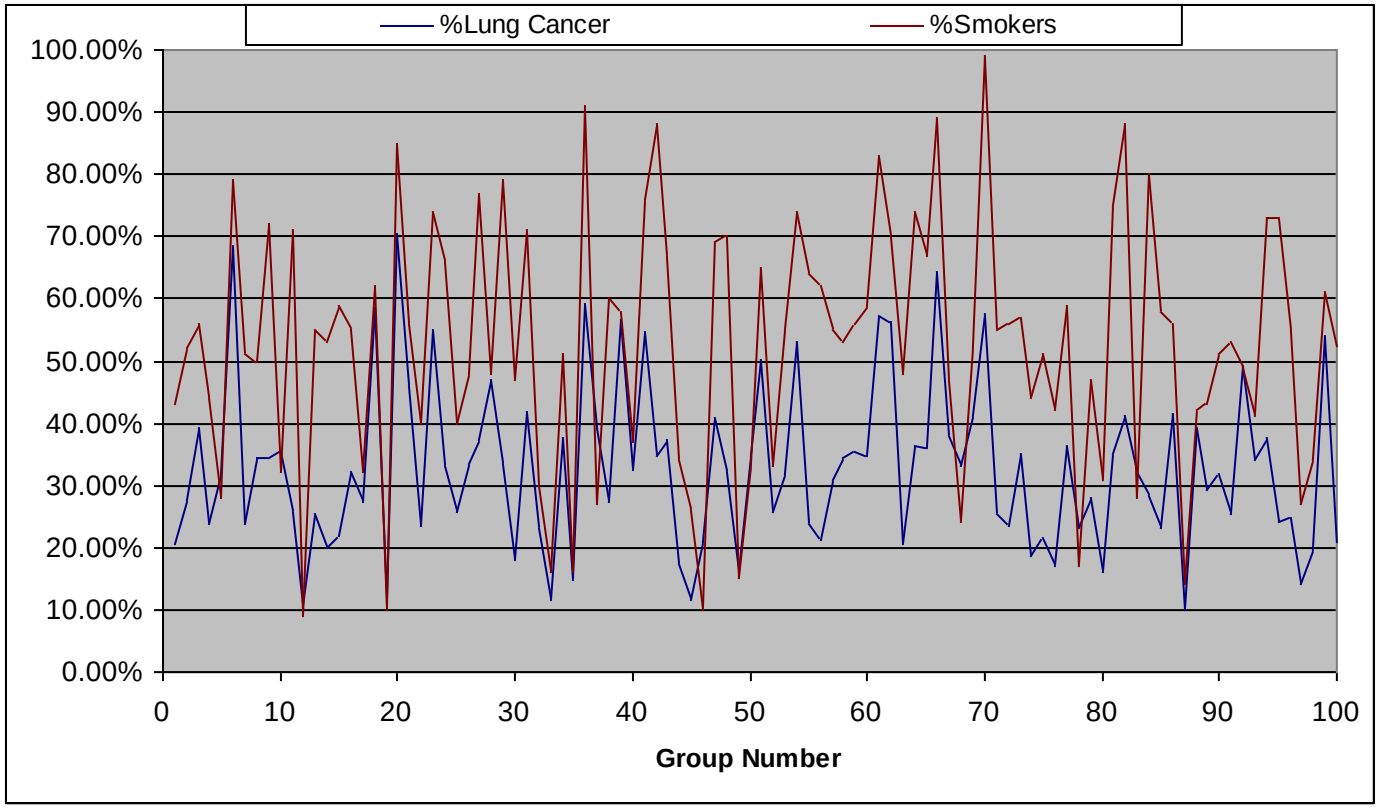
- Nu prea exista dubii ca fumatul are legatura cu cancerul la plamani
  - Si totus multi fumatori raman sanatosi …
  - … si nu putini nefumatori ajung sa aiba cancer la plamani
- Intrebare: cum demonstram o legatura cauzala intre tutun si cancer?

- Alternativ: cum putem folosi statistica sa demontam aceasta legatura sau chiar mai mult, sa promovam beneficii asupra sanatatii datorate tutunului?



1954

VII. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie



- Ca si mai devreme cu petele solare, un grafic poate ilustra o stare de fapt, dar nu putem trage nici o concluzie

VII. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie

- O solutie – o formula pentru corelatie

- Sa-l cunoastem pe Karl Pearson
  - Statistician, Matematician, Genetician
  - Student al lui Galton
  - Autor al multor concepte din statistica moderna
  - Pentru aceasta discutie ne rezumam la Corelatia Pearson (r)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

Karl Pearson, 1857-1936

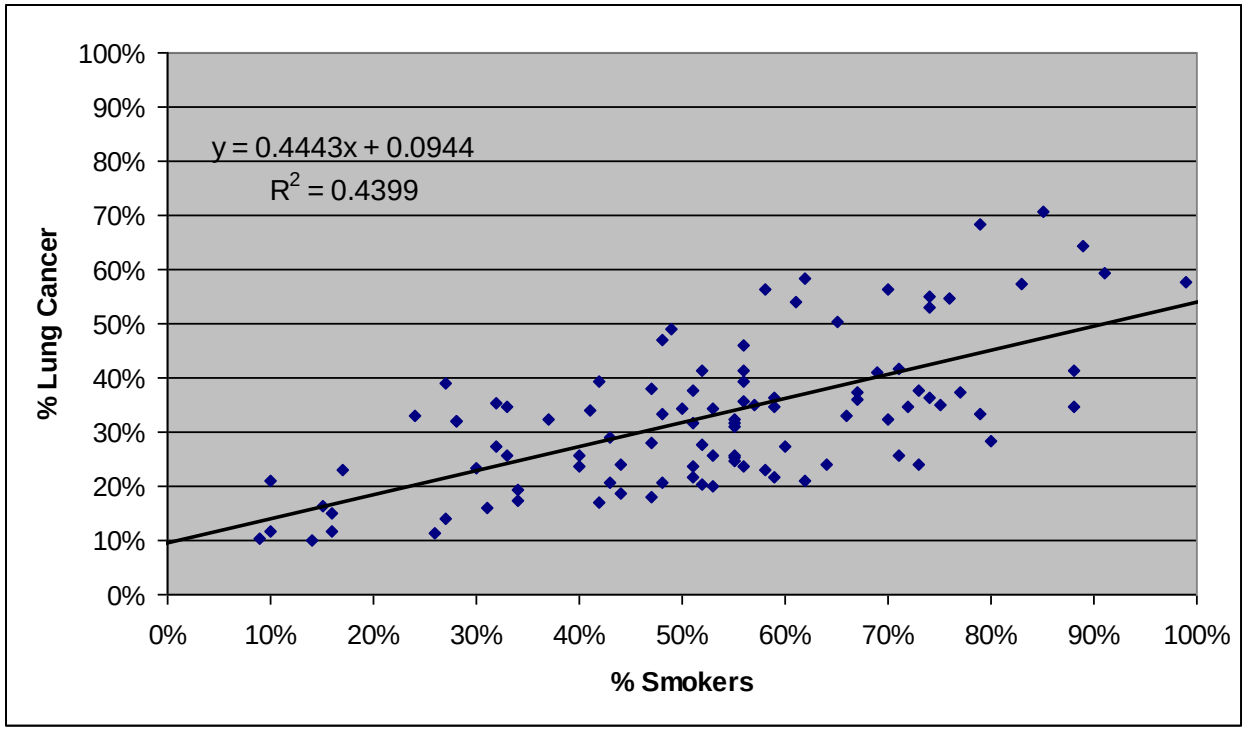
VII. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie

- Date de la 100 de spitale sunt examinate relative la mai multi factori:
  - % pacienti cu cancer de plamani
  - % fumatori
  - Varsta medie
  - % de barbati
  - Inaltime medie
  - Index de dieta
  - Index de exercitii fizice
  - Index de somn (sleep index)
  - Indicator al altor boli
  - Index de consum alcool

	Lung Cancer
Smoking	**66.32%**
Age	65.07%
Sex	20.67%
Weight	-37.66%
Diet	-29.26%
Disease	53.18%
Alcohol	53.80%
Sleep	26.42%

- Datele sugereaza un efect al fumatului dar si alte contributii

VII. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie



- Regresie a Cancerului pe Fumat (Smoking)
- Panta e mare, deci avem corelatie pozitiva. Dar este seminifcativa?
- Pe graf observam patratul corelatiei Pearson (43.99%)

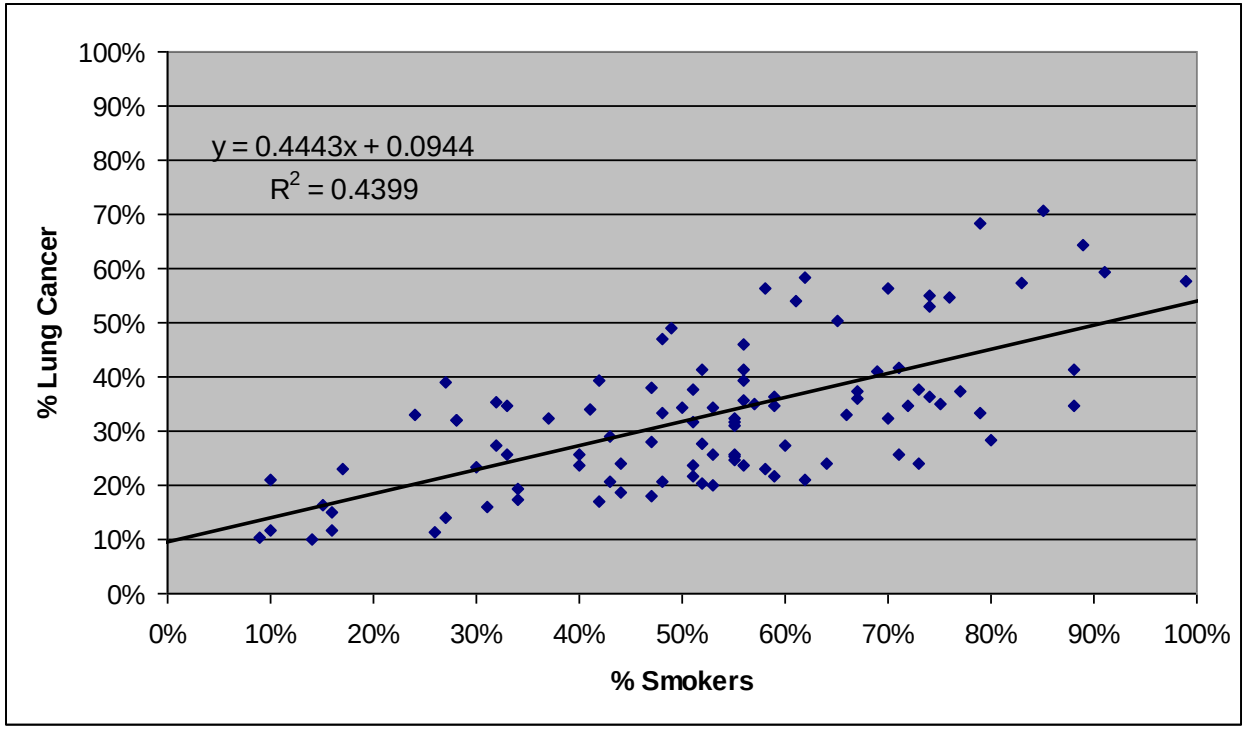
VII. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie

- Cum mai putem determina daca o panta pozitiva este semnificativa?

- Sa-l cunoastem pe Sir Ronald Fisher
  - Statistician, Matematician, Genetician
  - Rival inrait al lui Pearson
  - Analiza Variantei, Maximum Likelihood, etc
  - Definitia moderna a semnificatiei statistice
- Z-statistic – masura a semnificatiei
  - $|\beta/s.e.(\beta)|$ - comparatie cu distributia normala
  - <1 : nu e semnificativa ; >2 : semnificativa, >3 : foarte semnificativa
  - Se poate folosi si *t-statistic* (Gosset) care din motive tehnice e preferata uneori

R.A. Fisher, 1890-1962

VII. Corelatia Nu e Cauzalitate, dar Sigur e o Indicatie

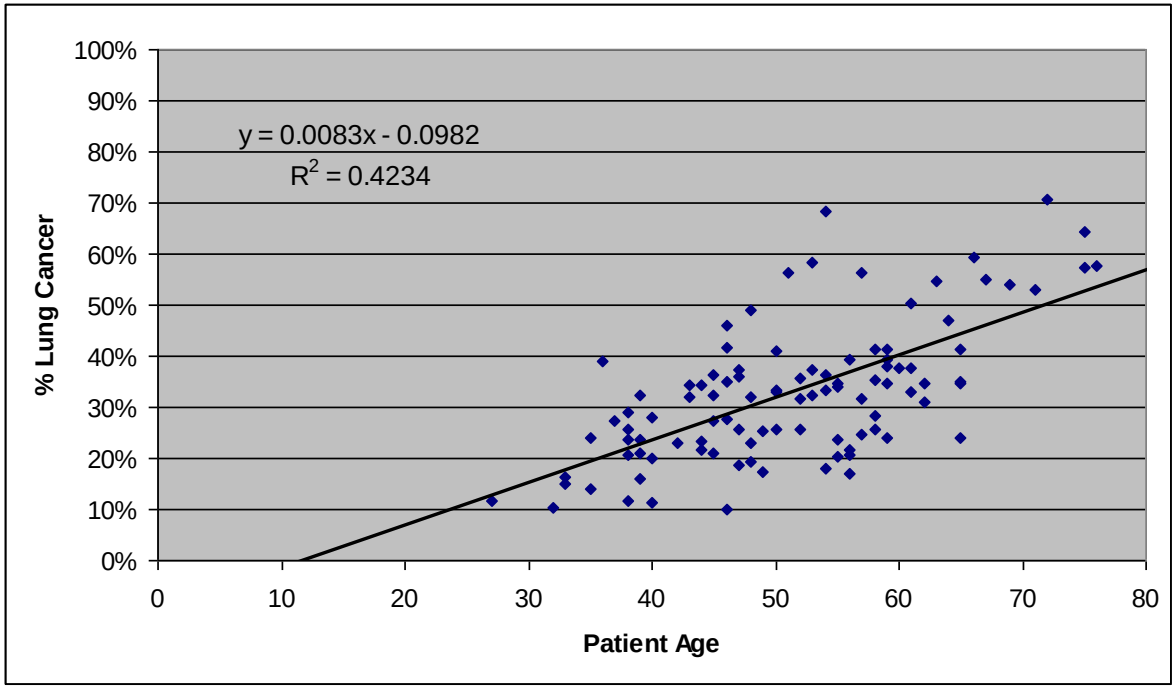


- Regresie: summary statistics
- Fumatul este foarte semnificativ

(considerat individual)

	Intercept	Smoking
Beta	0.094	0.444
s.e.(beta)	0.028	0.051
T-statistic	3.311	**8.773**
R-square	43.99%	

VIII. Sa Nu Incluzi Prea Putine Variabile



- Si alte variabile par importante luate individual- de exemplu varsta
- Oameni in varsta sunt mai frecvent bolnavi de cancer la plamani (t=7.85)
- Cum putem lua aceata informatie in considerare?

VIII. Sa Nu Incluzi Prea Putine Variabile

- Regresie multipla – acel $(X^TX)^{-1}X^Ty$ cu care am inceput

	Intercept	Smoking	Age
Beta	-0.058	0.274	0.0047
s.e.(Beta)	0.0483	0.0654	0.0012
T-statistic	-1.201	4.192	3.781
R-square		50.85%	

- Atat fumatul cat si varsta apar ca semnificative
  - Fumatul si varsta sunt amandoua correlate cu cancerul – mai grav daca esti fumator in varsta
  - Modelarea s-a imbunatatit dupa cum indica r^2
- Oare ce alte variabile/factori ar putea imbunatati modelul?

IX. Sa Nu Incluzi Prea Multe Variabile

- Regresie pe toate variabilele

	Intercept	Smoke	Age	Male	Weight	Diet	Disease	Alcohol	Sleep
Beta	-0.007	-0.0204	0.00473	0.1822	-0.0025	0.0012	0.0031	0.0774	-0.0201
s.e.(beta)	0.291	0.107	0.0011	0.112	0.0007	0.0028	0.0007	0.0967	0.015
T-statistic	-0.0263	-0.19	4.163	1.626	-3.215	0.4228	4.367	0.8015	-1.336
R-square					65.03%				

- De aici se vede ca varsta (Age), greutatea(Weight) si alte boli (Diseases) sunt semnificative
- Fumatul apare ca nesemnificativ in aceasta analiza:
  - Are chiar semn negativ, putand fi interpretat chiar un beneficiu
  - Multe din studiile anilor 50 (inclusive cele ale lui Fisher) au ajuns la concluzii similare

- Ce s-a intamplat?

IX. Sa Nu Incluzi Prea Multe Variabile

- Multe dintre variabile sunt intercorelate
- Prea multe variabile diminueaza efectele
  - In exemplul de mai sus: Prezenta Indicatorului de alte boli este puternic asociat cu Cancerul si intrucatva cu fumatul. Regresia surprinde aceste efecte dar nu le atribuie fumatului.

IX. Sa Nu Incluzi Prea Multe Variabile

- Cate variabile sunt prea multe?
- LASSO Regression
- Ridge Regression
  - Se cauta minimizarea

$$\sum_{i=1}^{m} \left| \sum_{j=1}^{n} X_{ij}\beta_j - y_i \right|^2 \to \min.$$

cu urmatoarele conditii: $\|\beta\|_1 = \sum_1^n |\beta_j| < s$ (LASSO)

Sau $\|\beta\|_2 = \sqrt{\sum_1^n {\beta_j}^2} < s$ (RIDGE)

LASSO si RIDGE

- RIDGE va minimiza importanta in model a variabilelor neinformative
- LASSO va elimina din model variabilele neinformative
  - Dintr-un grup de variabile corelate doar una va fi aleasa
  - Problema gasirii lui s este echivalenta cu problema de minimizare

$$\sum_{i=1}^{m} \left| \sum_{j=1}^{n} X_{ij} * \beta_j - y_i \right| + \lambda * \|\beta\|_1 \rightarrow min.$$

    - In problema de mai sus λ e fix, se afla β in functie de X si λ
    - Pt λ nu exista algoritm: se incearca multe valori si cea care da eroarea cea mai mica de predictive va castiga

Concluzii

- Regresia este o unealta in examinarea relatiilor din date
- Regresia poate fi o arta – o aplicatie automatizata bazata pe regresie poate da gres
- Vizualizarea datelor pot rezolva probleme de business
- Inainte de rularea unei regresii, e importanta o analiza formala sau informala a corelatiilor
- Atentie cand vi se cere sa discutati cauzalitate – regresia nu o poate demonstra
- Incercati o recenzie de la cineva avizat asupra analizei si intrepretarii rezultatelor
- Cand aveti variabile multiple (a.k.a. covariate), alegeti-le cu grija pe cele ce urmeaz a fi incluse in regresie
- Normalizati datele inainte de studiu

Q&A


constructive feedback
ur doin it wrong
08/10/2006
ICANHASCHEEZBURGER.COM