Inleiding programmeren

1e jaar wis-, natuur- en sterrenkunde

Universiteit van Amsterdam

november 2013

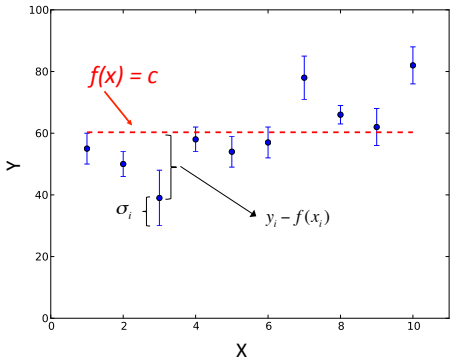
Opgaves bij college 3 (dagdeel 2)

fitten van data

1 Fitten van data en foutenbepaling

Om de onderliggende fenomenen van (natuurkundige) verschijnselen te achterhalen wordt data verzameld om afhankelijkheden te onderzoeken. Dat kan de massa van het Higgs boson zijn, de vervaltijd van uranium, maar ook het aantal kinderen in een gezin als functie van de gemiddelde lengte van de ouders. Je kan dan zoeken naar een (causaal) verband: lineair, exponentieel,etc. en daarbij ook de bijbehorende parameters bepalen met hun onzekerheid. Als je een goede beschrijving hebt gevonden kan je daarmee vervolgens ook voorspellingen doen.

Om de 'beste' waarde te vinden hebben we een maat nodig om de *'goedheid'* van de fit quantifiseert. We doen dat hier met de χ^2-maat: de som van de gemiddelde afwijking van de meetpunten tot het model gewogen met hun fout: 'hoeveel standaardafwijkingen ligt dit punt weg van mijn functie'.



$$\chi^2 = \sum_{i \text{ (datapunten)}} \left(\frac{y_i - f(x_i|\vec{\alpha})}{\sigma_i} \right)^2,$$

met $\vec{\alpha}$ de vector functie-parameters.

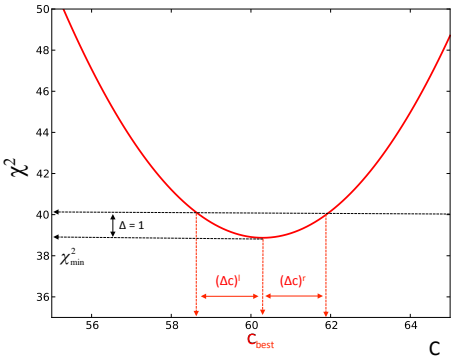
Voorbeeld: 1 dimensie/parameter (zie plot) $f(x|\vec{\alpha}) = c$.

a) de beste waarde van c: c_{best}

De waarde van c waarbij de χ^2 minimaal is.

b) de onzekerheid op $c_{\text{best}}(\Delta_c)$

De fout in de positieve richting $(\Delta_c)^r$ en negatieve richting $(\Delta c)^l$ zijn die waardes van c waarbij de χ^2 1 hoger is dan χ^2_{min}. In de meeste gevallen geldt $(\Delta c)^l = (\Delta c)^r = \Delta c$



Het eindresultaat van je meting is dan: $c = c_{\text{best}} \, {}^{+(\Delta_c)^r}_{-(\Delta_c)^l}$

opgavenset 4

opgave [1]: fitten van een model aan de data

De onderstaande data-set geeft voor een specifieke voetballer het percentage goede passes (y) weer als functie van het aantal gespeelde wedstrijden in oranje (x). De onzekerheid op het aantal goede passes is weergegeven als σ_y.

x	1	2	3	4	5	6	7	8	9	10
y	55	50	39	58	54	57	78	66	62	82
σ_y	5	4	9	4	5	5	7	3	6	6

a) maak een plot van deze data met fouten
Computing tip: gebruik de functie `plt.errorbar(x,y, yerr=yerror)`

b) bereken de beste waarde van c als $f(x)$ =c en de bijbehorende onzekerheid Δc.

c) Wat gebeurt er met Δc als de fout in elk meetpunt 2x kleiner wordt ?

Het lijkt erop of er een 'trend' zichtbaar is, dus we breiden ons model uit met een lineaire term: $f(x) = \mathrm{b}x + \mathrm{c}$.

d) bereken de beste waardes voor b en c als $f(x) = \mathrm{b}x + \mathrm{c}$.

Bij veel banken en verzekeraars werken er wis- en natuurkundigen in de zogenaamde risk-analysis departments. Laten we een opgave bekijken door de eindstand van de AEX te voorspellen in het jaar 2000 gebruikmakend van de data van de AEX in de jaren 1991-1999. Er zit geen 'fout' op de standen. Zet in de χ^2-formule de fout op de meetwaarde op 1.

jaar	1991	1992	1993	1994	1995	1996	1997	1998	1999
AEX	125.72	129.71	187.99	188.08	220.24	294.16	414.61	538.36	671.41

e) Maak een grafiek dan de eindstand van de AEX als functie van het jaar sinds 1991.

f) Fit de grafiek met een polynoom van graad 2: bereken dus de beste waardes voor a, b en c als $f(x) = \mathrm{a}x^2 + \mathrm{b}x + \mathrm{c}$.
Tip 1: gebruik x = 1,2,3 i.p.v. 1991, 1992, 1993 etc.
Tip 2: probeer de waarde van a, b en c te schatten voor je gaat fitten.

g) Wat is je voorspelling voor het jaar 2000 ? Zet beide waardes in de plot en vergelijk deze met de echte waarde in het jaar 2000. En ? Hoe kan dat ?