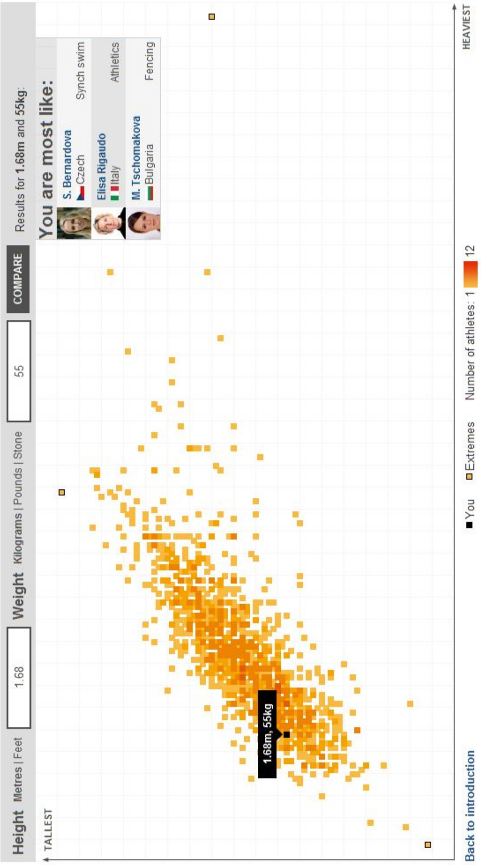
Olympic Outliers

Maria Bekker-Nielsen Dunbar

During the 2012 Summer Olympics, the BBC had a rather snazzy plot of a sample of all the Olympic athletes, where you could enter your height and weight to see whose body you possess:



Height Metres | Feet 1.68 Weight Kilograms | Pounds | Stone 55 COMPARE

Results for 1.68m and 55kg:

You are most like:

S. Bernardova Czech	Synch swim
Elisa Rigaudo Italy	Athletics
M. Tschomakova Bulgaria	Fencing

It also informs you of the tallest, shortest, lightest and heaviest athletes (the extremes). They are:

Tallest	219cm	Zhaoxu Zhang (China)
Shortest	136cm	Asuka Teramoto (Japan)
Lightest	30kg	Asuka Teramoto (Japan)
Heaviest	218kg	Ricardo Blas Jr (Guam)

Obviously, I cannot look at something like this without wanting to play around with the data, so I grabbed the .txt-file of data (which was created by the Press Association) and converted it to a .csv-file. Now it can be loaded into SAS, thusly:

```
proc import datafile= "C:\Users\Maria Dunbar\Desktop
                         \famos\olymp.csv"
     out=olymp2
     dbms=dlm
     replace;
     delimiter= ",";
     getnames=yes;
run;
```

Averages

In order to see what height and weight can be expected of these athletes, we can generate the means of these variables:

```
proc means data=olymp2 maxdec=2 mean var std n;
var Height_cm_ Weight_kg_;
run;
```

Variable	Mean	Variance	Std Dev	N
Height_cm_	177.45	127.41	11.29	1587
Weight_kg_	72.65	256.31	16.01	1587

The average athlete has a height of 177.45cm with a standard deviation of 11.29cm and a weight of 72.65kg with a standard deviation of 16.01kg.

Hence, practically everyone at our department can have 'the body of an athlete' (understood as being the same height and weight - muscle mass, fat levels and sporting ability may vary). However not everyone can be a *mathlete* since this year, for the first time (I think), all types of mathematical studies at UCPH have specific grade requirements!

Linear regression

Recall the picture from the first page of this article - obviously no sane person can look at it without wanting to draw a line through the points. This means we are looking at a model of the type

$$H_i = \alpha + \beta W_i + \varepsilon_i \tag{1}$$

where H_i and W_i are height and weight respectively, while ε_i are errors.

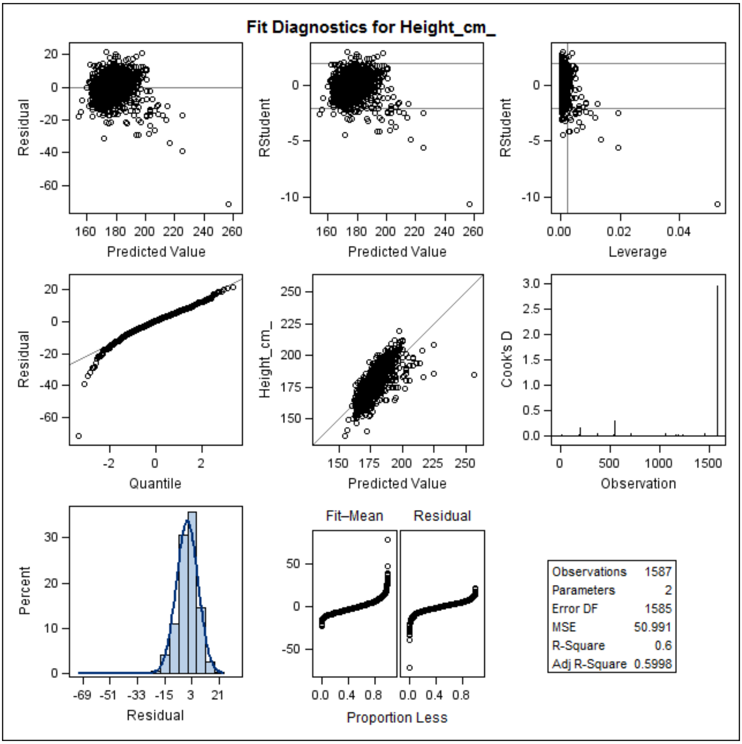
Is this a good model?

The assumptions of this error, which we have to check, before we estimate the parameters in our regression, are:

- The errors have a mean of 0, $E[\varepsilon_i] = 0$
- They have variance, $Var[\varepsilon_i] = \sigma^2, \sigma^2 > 0$

- They are independent, $\varepsilon_i \perp\!\!\!\perp \varepsilon_j, \forall i, j$
- They are distributed normally, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

These assumptions are checked by viewing the diagnostics plots of our model:



Observations	1587
Parameters	2
Error DF	1585
MSE	50.991
R-Square	0.6
Adj R-Square	0.5998

The assumptions on the error are fine (cf. the three leftmost plots) but the model has the potential to become *better* as the R^2-value is 0.6 (preferably it would be closer to 1).

The estimated model

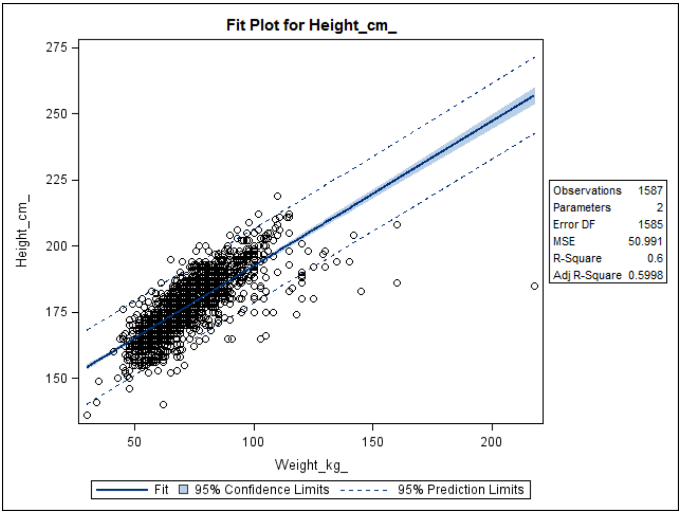
The parameters of this model (the intercept and the slope, α and β) can be estimated to[15]

$$H_i = 137.77 + 0.546W_i \tag{2}$$

using the command:

```
proc reg data=olymp2;
model Height_cm_=Weight_kg_;
run;
```

In order to see whether this line is good (i.e. fits well), we look at the following fit plot:



[15] α has a standard error of 0.83 and β has a standard error of 0.01

Note that the point on the far right seems to be ill-fitting. We will proceed to check whether it is an outlier, and if it is, identify which athlete it is.

Outliers

Recall that an outlier is a point which is radically different than the other points, so the assumption that the error on this observation follows the same distribution as the errors observed on other observations, may be incorrect.

The externally standardised residuals, called t-values, are calculated (these follow a t-distribution). Cutting off t-values larger than 4 leaves the following athletes:[16]

```
proc reg data=olymp2;
model Height_cm_=Weight_kg_/partial influence all;
id+1;
output out=olymp3
Rstudent=t
covratio=c
h=h;
proc print data=olymp3;
where abs(t)>4;
run;
```

[16] Cutoffs at 3 are also seen. SAS' standard setting is to cut off at 2 (which it uses when producing the Rstudent/Leverage-plot)

Obs.	*t*	Height[17]	Weight[18]	Athlete
187	-4.0907	165	103	Mami Shimamoto
202	-4.8221	183	145	Kazuomi Ota
343	-4.4582	140	62	Tuau Lapua Lapua
554	-5.5899	186	160	Artem Udachyn
1196	-4.1045	166	105	Adysangela Moniz
1586	-10.6983	185	218	Richardo Blas Jr

Using the highest *t*-value, -10.6983, and the 1587 variables our model uses, with its 2 degrees of freedom, the following code gives us a value of *t* which can be regarded as a critical limit/cutoff point, and as a result observations with a *t*-value higher than this are seen as outliers (for an explanation of what the code does see [3])

```
data a;
p1=probt(10.6983,1587-1-2);
p2=2*1-p1); p3=1-(1-p2)**1587;
tgraense=tinv(1-(1-(1-0.05)**(1/ 1587))/2, 1587-1-2)-
tbonf=tinv(1-0.05/2/ 1587, 1587-1-2);
proc print data=a;
var tgraense tbonf;
run;
```

tgraense	tbonf
4.16849	4.17437

Using 4.17 as our new cutoff point the following four athletes are outliers:

[18] The height is measured in cm.

[18] The weight is measured in kg.

```
proc print data=olymp3;
where abs(t)>4.17;
run;
```

Obs.	t	Height[19]	Weight[20]	Athlete
202	-4.8221	183	145	Kazuomi Ota
343	-4.4582	140	62	Tuau Lapua Lapua
554	-5.5899	186	160	Artem Udachyn
1586	-10.6983	185	218	Richardo Blas Jr

The athlete with the biggest t-value is thereby Richardo Blas Jr, who is the aforementioned point on the far right of the fit plot (he can also be spotted in the top three diagnostic plots and the one in the middle).

So the question one should ask oneself is - is he an atypical observation? In some ways yes, as he seems to be the only athlete with a higher value of weight (measured in kg) than value of height (measured in cm).

Removing observation 1586 might give the model a better fit. However, an outlier is not necessarily the same as ’a bad observation’ so he will be left in for the time being (whether or not he stayed in the Games is another story).[21]

[20] The height is measured in cm.

[20] The weight is measured in kg.

[21] If he was removed, however, the critical limit for t would have to be changed accordingly as we have one less observation. This could lead to a higher/lower number of outliers

Plot-command

All plots used in this article were created by running the following command:

```
ods graphics on;
proc reg data=olymp2 plots=all;
model Height_cm_=Weight_kg_;
run;
ods graphics off;
```

References

[1] Your Olympic body match (30 July) http://www.bbc.co.uk/news/uk-19050139

[2] Nils Kousgaard og Anders Milhøj: *Anvendt Regression for Samfundsvidenskaberne*; Akademisk forlag

[3] Anders Milhøj: *Statistik med SAS - Sommerskolen i Videregående Statistik 2011*

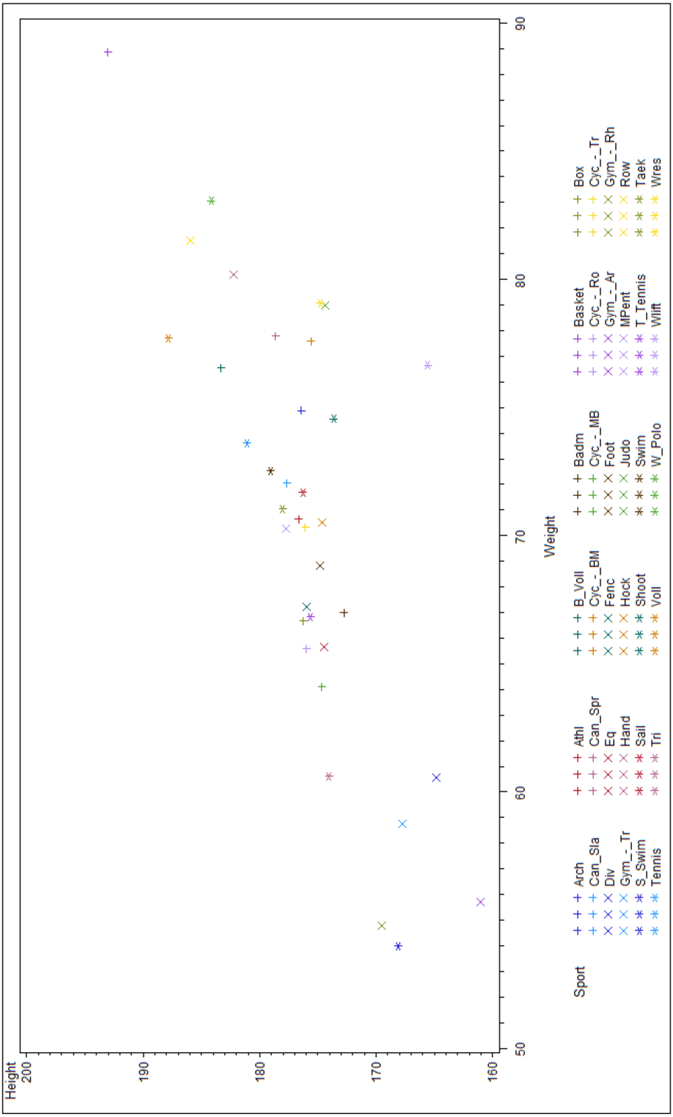
For good measure, on page 50 is a plot of the means of each sport individually.[22]

The code for which is:

```
data points;
   input Height Weight Sport $;
   datalines;
176.47 74.88 Arch
...
```

[22] I wanted to create oval-shaped bubbles with the standard errors as radius but alas, my SAS abilities do not yet reach that far

```
174.83 79.07 Wres
;
proc gplot data=points;
plot Height*Weight=Sport;
bubble Height*Weight=Sport $ Sport/;
run;
```


Height
200
190
180
170
160
Weight
50
60
70
80
90
Sport
Arch
Can_Sla
Div
Gym_-_Tr
S_Swim
Tennis
Athl
Can_Spr
Eq
Hand
Sail
Tri
B_Voll
Cyc_-_BM
Fenc
Hock
Shoot
Voll
Badm
Cyc_-_MB
Foot
Judo
Swim
W_Polo
Basket
Cyc_-_Ro
Gym_-_Ar
MPent
T_Tennis
Wlift
Box
Cyc_-_Tr
Gym_-_Rh
Row
Taek
Wres