Reference growth charts for Saudi Arabian children and adolescents

Peter Foster and Tatjana Kecojević

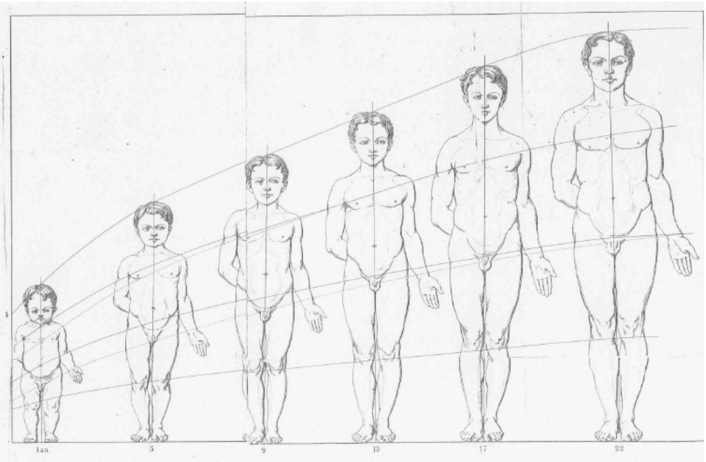
School of Mathematics, University of Manchester
Manchester, UK

Lancashire Business School, University of Central Lancashire
Preston, UK

Conference on Quantitative Social Science Research Using R

June 18, 2009

Quetelet's (1871) Growth Chart



Assumption about normality

The reference growth charts are based on normality assumptions for the data.

Age-specific mean $\mu(t)$ and standard deviation $\sigma(t)$ curves are estimated and chosen quantile curve for a $\alpha \in [0, 1]$ can then be constructed as:

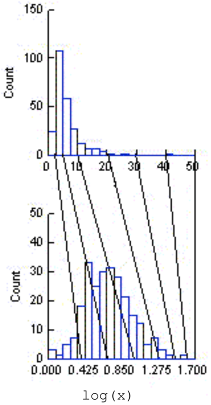
$$\hat{Q}\,(\alpha \mid t) = \hat{\mu}(t) + \hat{\sigma}(t)\Phi^{-1}(\alpha)$$

where $\Phi^{-1}(\alpha)$ denotes the inverse of the standard normal distribution function, in other words normal equivalent deviate of size α (corresponding to tail area).

Anthropometric data

Anthropometric data:

1. non-normally distributed,
2. tends to be right skewed rather than left, which is why a log transformation is often suggested to cope with it.



Penalised Maximum Likelihood Estimation

- Data: $\{Y\,(t_{i,j}) : j = 1, ..., J_i, \quad i = 1, ...n\}$
- Model: $Z(t) = \frac{[Y(t)/\mu(t)]^{\lambda(t)} - 1}{\lambda(t)\sigma(t)} \sim \mathcal{N}\,(0, 1)$
- Estimation:

$$\ell(\lambda, \mu, \sigma) = \sum_{i=0}^{n} [\lambda(t_i) \log \frac{Y(t_i)}{\mu(t_i)} - \log \sigma(t_i) - \frac{1}{2} Z^2(t_i)]$$

$$\max[\ell(\lambda, \mu, \sigma) - \nu_\lambda \int (\lambda''(t))^2 dt - \nu_\mu \int (\mu''(t))^2 dt - \nu_\sigma \int (\sigma''(t))^2 dt]$$

- Quantile: $Q\,(\alpha \mid t) = \mu(t)[1 + \lambda(t)\sigma(t)\Phi^{-1}(\alpha)]^{1/\lambda(t)}$

Data

SaudiR.csv

	A	B	C	D	E	F	G	H
1	REGION	IDNUMBER	SEX	WEIGHTOF	HEIGHTOF	HEADCIRC	AGEYEARS	AGEMON
2	1	1	1	.NA	.NA	.NA	.NA	.NA
3	1	1	2	.NA	.NA	.NA	.NA	.NA
4	1	1	2	64.5	148	59	17.9688	215.63
5	1	2	1	.NA	.NA	.NA	.NA	.NA
6	1	2	2	.NA	.NA	.NA	.NA	.NA
7	1	2	2	14.6	99	50	4.6525	55.83
8	1	3	1	.NA	.NA	.NA	.NA	.NA
9	1	3	2	.NA	.NA	.NA	.NA	.NA
10	1	3	1	8.5	74.5	49	0.914	10.97
11	1	3	2	9.5	82	48	1.3855	16.63
12	1	3	1	15.6	113	51	4.6197	55.44
13	1	3	2	29.5	129.5	55	7.6469	91.76
14	1	6	1	.NA	.NA	.NA	.NA	.NA
15	1	6	2	.NA	.NA	.NA	.NA	.NA
16	1	6	.NA	.NA	.NA	.NA	.NA	.NA
17	1	6	2	15.8	104	50	5.8009	69.61
18	1	6	1	35.2	130	51	9.6737	116.08
19	1	6	2	40	148	56	16.1846	194.21

SaudiR.csv
Ready
Sum=0

Codes and description of variables (health profile of Saudi children)

- Region: The ID number of the region. There are 13 regions in the Kingdom. All are covered in this survey.
- Id number: This is the id number of the household (family).
- Sex: 1=male, 2= female.
- Measure: The variables **weightof** (in Kg), **heightof** (in cm), **headcirc** (in cm), refer to the corresponding body measurements.
- Age: **Ageyears** and **agemon** refer to the date of measurement, recorded in Hijri calendar but subsequently converted to Gregorian.

Issues related to constructing the reference growth charts

- Detecting the **outliers**
- **Smoothing** the curves
- **Averaging** the overlapping period 2 to 3 years of age
- **Goodness-of-fit** of the centile curves
- **Comparison** between different geographical regions and between genders

Robust regression

R - a public domain language for data analysis

MASS package (contributed by W.N. Venables and B.D. Ripley)

```
> library(MASS)
> mp<-rlm(log(weight)~1+agey+I(agey^2)+I(agey^3), method="MM")
```

An object of class `rlm` inherited from `lm` is used to fit linear models and it can be used to carry out regression.

Using rlm fitting is done by iterated re-weighted least squares (IWLS).

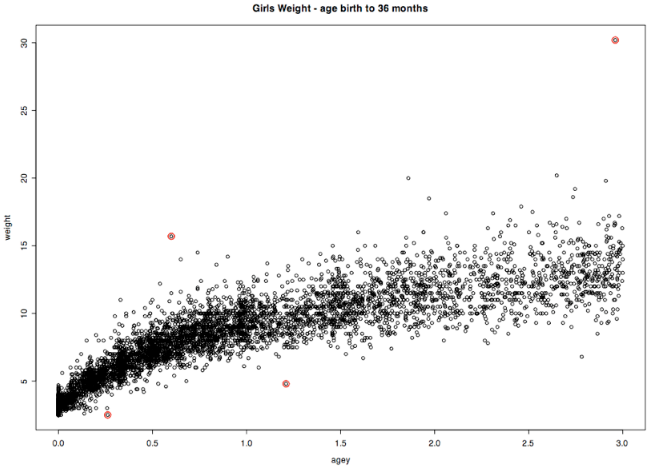
An additional component in `rlm` that is not in an `lm` object is:

`w` – the weights used in the IWLS process.

R Data Editor

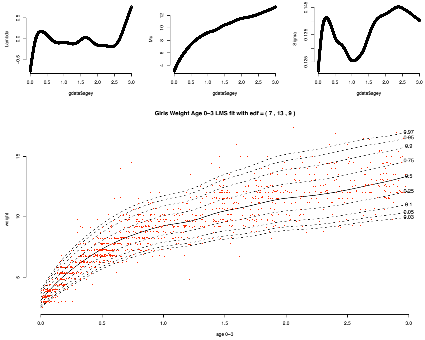
row.names	weight	agey	w
32	30.2	2.96	0
1676	4.8	1.21	0
2946	15.7	0.6	0
3979	2.5	0.2617	0
3776	11	0.33	0.002548249
186	6.8	2.7819	0.002738172
872	20	1.86	0.02485536
4361	8	0.15	0.02486028
3913	9.8	0.3	0.06277439
4141	8.4	0.2	0.06360341
2831	14	0.65	0.1121182
754	18.5	1.97	0.1242003
3985	3	0.259	0.1309745

Detecting the outliers

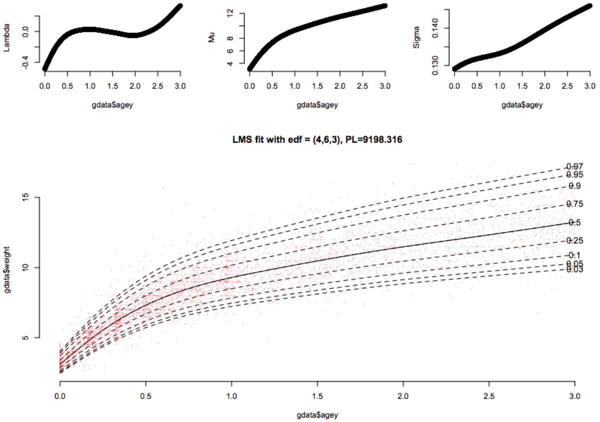


R package lmsqreg (contributed by V. J. Carry)

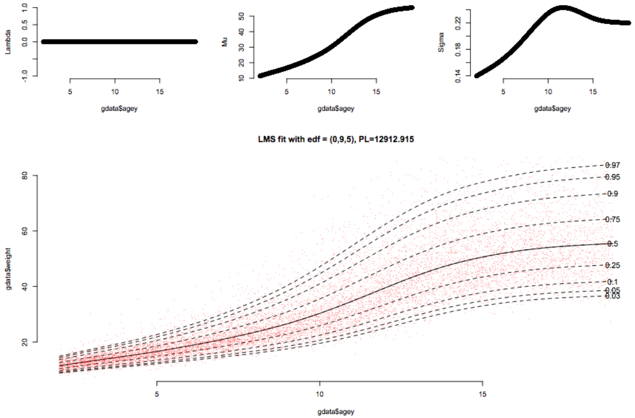
```
> library(lmsqreg)
> mw3<-lmsqreg.fit(gdata$weight, gdata$agey, edf=c(7, 13, 9),
pvec = c(0.03, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.97))
```



Smoothing with edf(4, 6, 3)



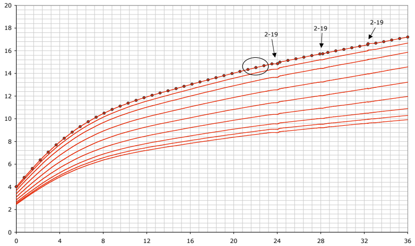
$\lambda = 0$

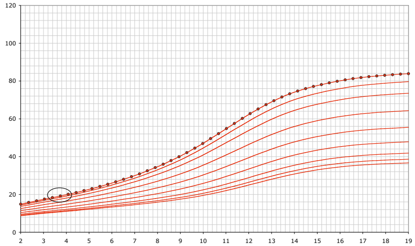

Lambda
gdata$agey
Mu
gdata$agey
Sigma
gdata$agey
LMS fit with edf = (0,9,5), PL=12912.915
gdata$weight
gdata$agey
0.97
0.95
0.9
0.75
0.5
0.25
0.1
0.05
0.03

Overlap for age 2 to 3 years

```
lmsqreg.fit(YY, TT, edf = c(3, 5, 3), targlen = 50, pvec = c(0.05, 0.1, 0.25, 0.5,
0.75, 0.9, 0.95))
```

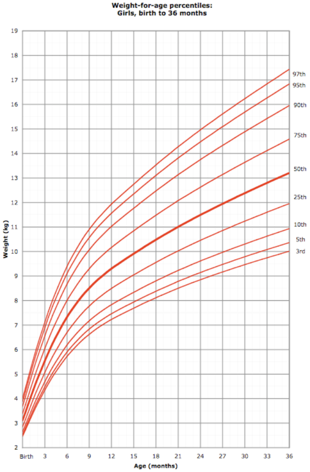
`targlen` - Number of points at which smooth estimates of L, M, S should be extracted for quantile plotting.





$$\hat{Y} = X\left[X'X\right]^{-1}X'Y$$

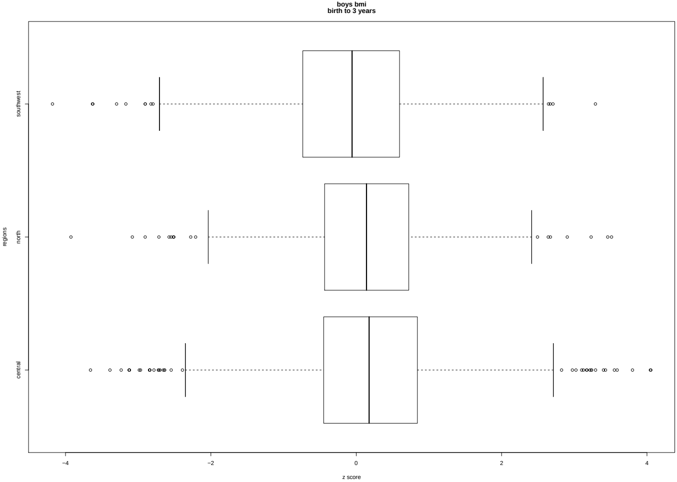
Averaged Chart



R package lmsqreg (mw3)

```
> mw3

lms quantile regression, version , fit date Thu Jun  4 14:37:52 2009

Dependent variable: gdata$weight , independent variable: gdata$agey
The fit converged with EDF=( 4,6,3 ), PL= 9198.316

nominal percentile   0.030 0.050 0.10 0.25 0.500 0.750 0.900 0.95 0.970
estimated percentile 0.025 0.052 0.09 0.24 0.506 0.755 0.905 0.95 0.972

KS tests: (intervals in gdata$agey //p-values)
   (-0.001,0]      (0,0.348] (0.348,0.802]  (0.802,1.54]       (1.54,3]        Overall
        0.000          0.000          0.271          0.324          0.676          0.001

t tests: (intervals in gdata$agey //p-values)
   (-0.001,0]      (0,0.348] (0.348,0.802]  (0.802,1.54]       (1.54,3]        Overall
        0.006          0.000          0.562          0.369          0.568          0.810

X2 tests (unit variance): (intervals in gdata$agey //p-values)
   (-0.001,0]      (0,0.348] (0.348,0.802]  (0.802,1.54]       (1.54,3]        Overall
        0.000          0.000          0.717          0.050          0.462          0.979
```

BoxPlots of mean SD scores of the three geographical regressions



R analysis of variance (ANOVA)

```
> summary(fm<-aov(z~group))
              Df Sum Sq Mean Sq F value    Pr(>F)
group          2   69.4    34.7  37.589 < 2.2e-16 ***
Residuals   3941 3640.0     0.9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> TukeyHSD(fm)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = z ~ group)

$group
                        diff        lwr         upr     p adj
north-central    -0.01495461 -0.1000653  0.07015605 0.9107089
southwest-central -0.33239976 -0.4255862 -0.23921328 0.0000000
southwest-north   -0.31744515 -0.4224195 -0.21247080 0.0000000
```

Procedure for using ANCOVA to compare the growth standards between the regions

1. STEP 1: Find the best fitting polynomials having the lowest possible common degree for each of the three regions.

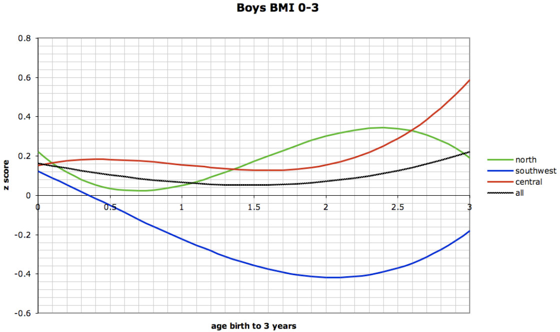
2. STEP 2: We want to answer the question "Is a common polynomial of the same degree as found in STEP 1 appropriate for all three regions or do the polynomials vary with region?" ie. for a particular measurement, sex and age group we want to test:

   $H_0 : E\left[z \mid age\right] = \beta_0 + ... + \beta_q age^q$ for each region, where $q \leq 3$ is the degree of the common best fitting polynomial.

   vs. H_1 : The polynomial for at least two regions differ.

3. STEP 3: After finding a significant result in STEP 2 carry out pairwise comparisons between the regions.

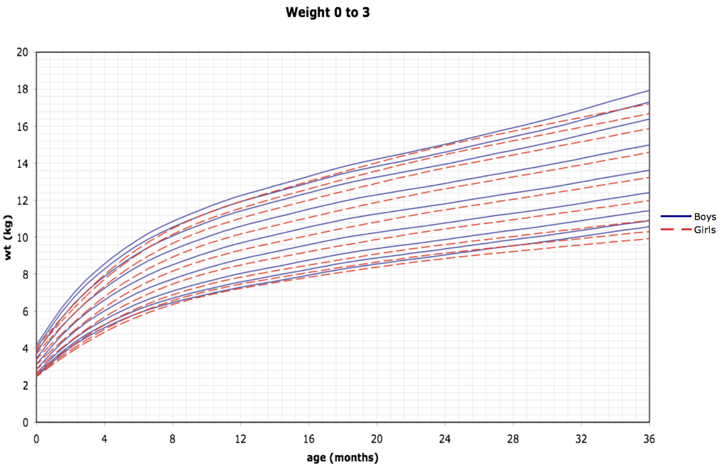
SD score regression models



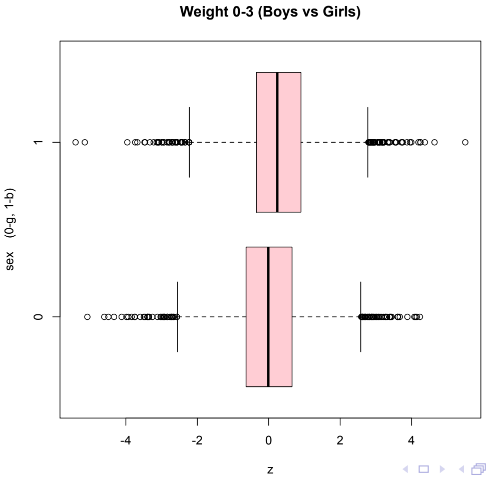
sex:male, age: birth to 36 moths

body mass index	
	p
north-central	0.061050
southwest-central	0.000000
southwest-north	0.000000

Weight, age birth to 36 months



Box Plots



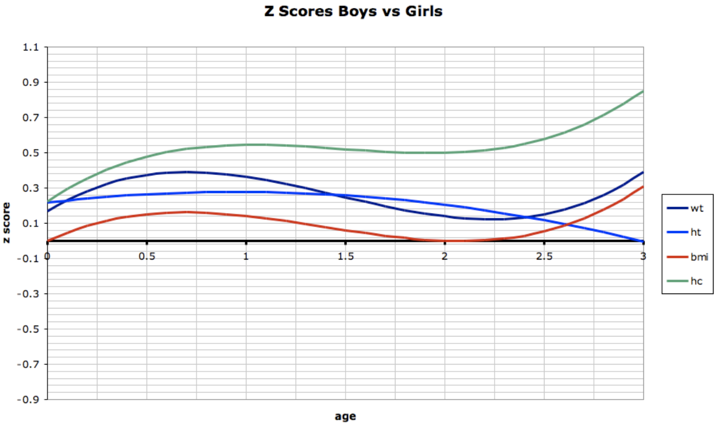
t-test

```
> t.test(z~sex)

Welch Two Sample t-test

data:  z by sex
t = -14.4148, df = 12473.99, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2958223 -0.2249998
sample estimates:
mean in group 0 mean in group 1
    0.003078598     0.263489639
```

lm function used to fit linear models

```
> m<-lm(z~sex+sex*x+sex*I(x^2)+sex*I(x^3))
> summary(m)

Call:
lm(formula = z ~ sex + sex * x + sex * I(x^2) + sex * I(x^3))

Residuals:
      Min        1Q    Median        3Q       Max
-5.727730 -0.642032 -0.002139  0.649591  5.140159

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0003369  0.0234113  -0.014 0.988518
sex1         0.1680739  0.0326333   5.150 2.64e-07 ***
x            0.1004580  0.1011812   0.993 0.320801
I(x^2)      -0.1245990  0.0993347  -1.254 0.209744
I(x^3)       0.0325240  0.0248738   1.308 0.191047
sex1:x       0.6006027  0.1418117   4.235 2.30e-05 ***
sex1:I(x^2) -0.5300762  0.1395383  -3.799 0.000146 ***
sex1:I(x^3)  0.1160172  0.0350394   3.311 0.000932 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.007 on 12473 degrees of freedom
Multiple R-Squared: 0.02064, Adjusted R-squared: 0.02009
F-statistic: 37.56 on 7 and 12473 DF,  p-value: < 2.2e-16
```

Boys vs Girls, age birth to 3 years



Things to do

1. Assessing the difference in fits of quantiles fitted by a parametric function and by a smooth non-parametric curve.
2. Test for a significant difference between the curves for the overlapping period with the original estimations.

- Box GEP, Cox DR. *An analysis of transformations.* Journal of the Royal Statistical Society, Series B, 1964; **26**:211–252.
- Carey VJ. *LMSqreg: An R package for Cole-Green reference centile curves*, 2002; http://www.biostat.harvard.edu/∼carey.
- Carey VJ, Youg FH, Frenkel LM, McKinney RM. *Growth velocity assessment in pediatric AIDS: smoothing, penalized quantile regression and the definition of growth failure.* Statistics in Medicine, 2004; **23**:509–526
- Cole TJ. *Fitting Smoothing Centile Curves to Reference Data.* Jurnal of the Royal Statistical Society, Series A–General 1988; **151**:385–418.
- Cole TJ, Green PJ. *Smoothing reference centile curves: the LMS method and penalized likelihood.* Statistics in Medicine, 1992; **11**:1305–1319.
- Huber PJ, 1981. *Robust Statistics.* New York: John Wiley and Sons
- Green PJ. *Penalized likelihood for general semi-parametric regression models.* International Statistical Review, 1987; **55**:245–259
- Koenker R, 2005 *Quantile Regression.* New York: Cambridge University Press
- Koenker R, Basset G. *Regression Quantiles.* Econometrica, 1978; **46**:33–50
- Venables WN, Ripley BD. *MASS: An R package in the standard library of Venables and Ripley*, 2007; http://cran.r-project.org/src/contrib/Descriptions/VR.html
- Venables WN, Ripley BD. (2002) *Modern applied statistics with S.* 4th Edition. New York: Spriger science+busines media, Inc.
- Wei Y, Pere A, Koenker R, He X. *Quantile regression methods for reference growth charts.* Statistics in Medicine, 2006; **25**:1396–1382