KJ-REGRESSION MODEL TO EVALUATE OPTIMAL MASSES OF EXTREME CASES
http://www.ngds-ku.org/Papers/J34.pdf

MATHEMATICS AND STATISTICS

Syed Arif Kamal[¶] *and Samira Sahar Jamil*

Additional File 1

Mathematical-Statistical Framework of KJ-Regression Model

Table of Contents

1. Regression Models

Height-versus-percentile data, at particular age, was considered, and a regression model was constructed out of it. This model was then tested for adequacy, through residual analysis, carried out using Minitab. Co-efficient of determination was calculated to determine the proportion of variation in the data explained by the models. This was a necessary step to assess the validity of extrapolated values, predicted through these models. Predictions were done for percentiles 0.01st, 0.1st, 1st, 99th, 99.9th and 99.99th. Same procedure was carried out on a mass-percentile data-set, for a particular age, to obtain a slightly different regression model. All of the above mentioned steps were carried out using Minitab. Still, it was difficult to manage a large amount of data, and to construct (74 + 74 =) 148 models using this software. A model was fitted to each row of the 4 growth tables. Note that each table has 37 rows representing ages 2.0-20.0 *years* in intervals of 0.5 *year*, making (37 + 37 =) 74 models for girls' tables and 74 models for boys' tables. A program was written on C#, to compute statistics for the regression models fitted to the data, at different ages. This program besides computing the results, also wrote them to Excel sheet. Source Code is given in Section 2. Summary of these results is given in this chapter. Extrapolated values for different ages, constituted extension of charts.

[¶] *Homepage*: http://www.ngds-ku.org/kamal • *project URL*: http://ngds-ku.org • *e-mail*: profdrakamal@gmail.com

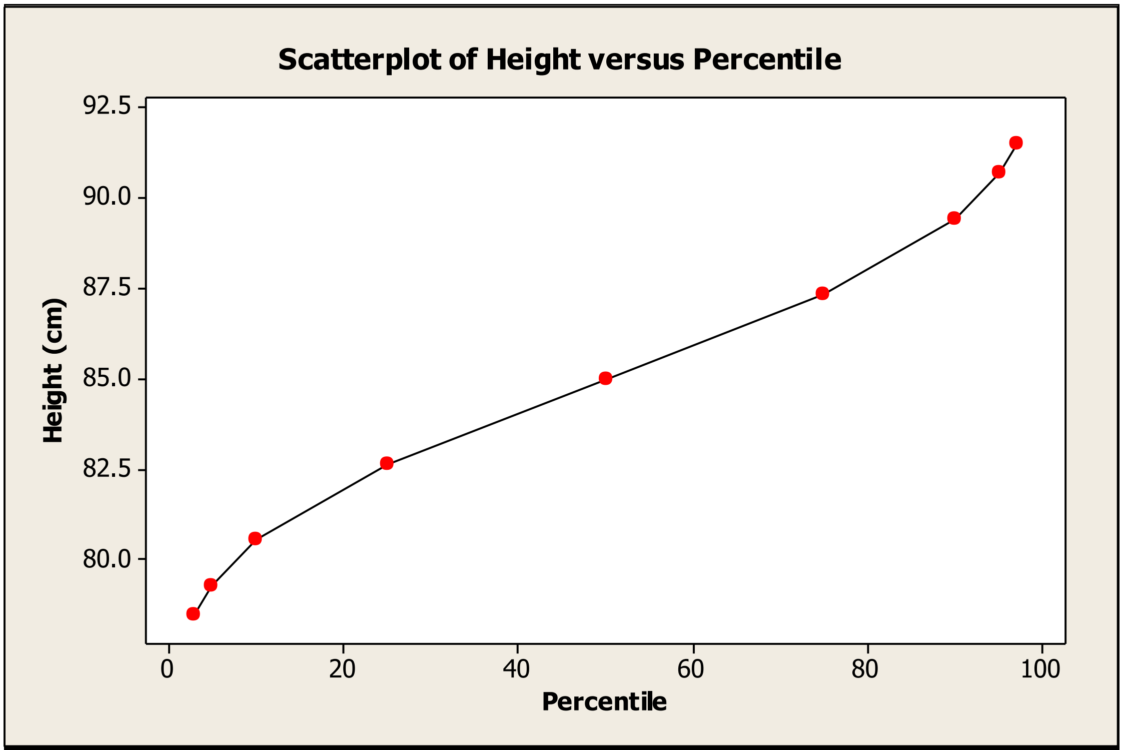

Fig. 1. Scatter plot of height versus percentile (girls, age: 2.0 *years*)

1.1 *Height-Percentile-Regression Models*

A few of the plots between heights and percentiles, obtained through CDC growth tables, were analyzed (one of them, from height-for-age growth table, for girls, age: 2.0 years, is shown in Figure 1). A curvilinear shape of these plots was seen, starting from a convex curvature, a linear trend and then a concave curvature.

However, when a transformation was applied to percentile data (equation 1), the transformed percentile values had, almost, a linear relation with height.

$$Logit(p) = ln\left(\frac{p}{100 - p}\right) \tag{1}$$

This transformation is denoted by $Logit(p)$, since it is a *Logit* transformation applied on $p/100$, *p* is representing percentile. Note that the argument of *ln* is never undefined,

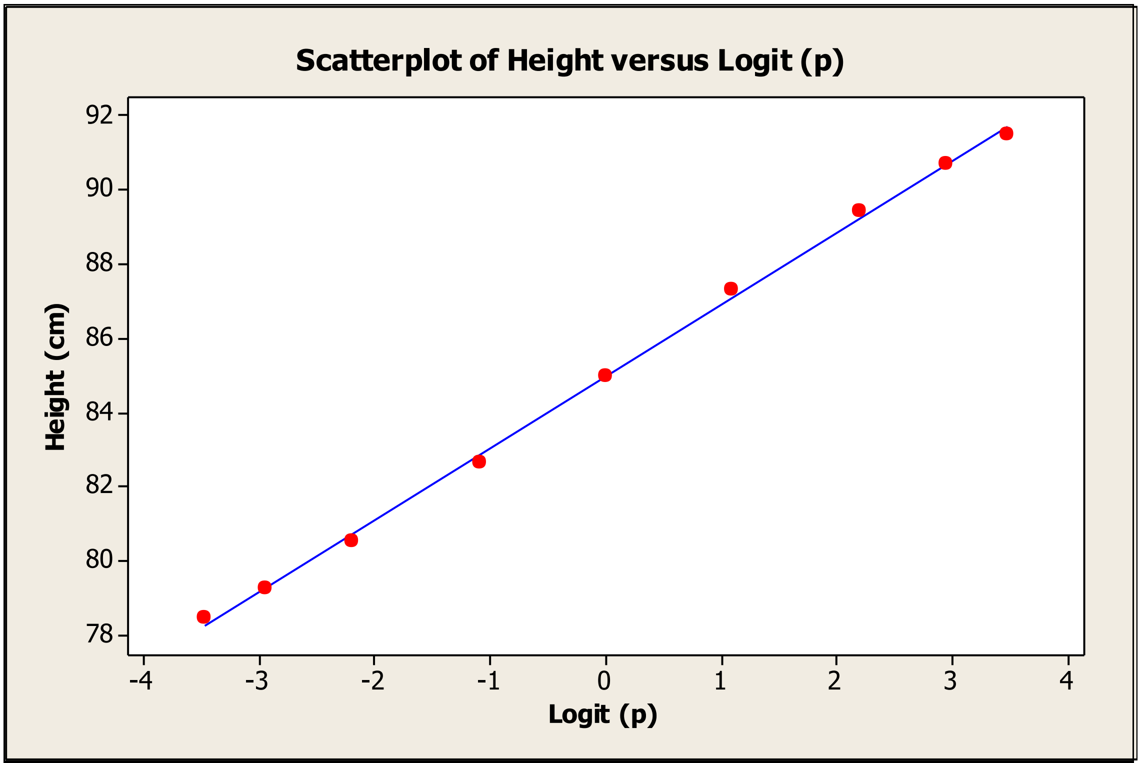

Fig. 2. Scatterplot of height versus *Logit(p)* with regression line fitted (girls, age: 2.0 *years*)

because percentile can never take an exact 100 value, because of case under consideration has to be excluded, when percentile is evaluated (Figure 2).

Although, a cubic polynomial had an even-better fit to height versus $Logit(p)$ data, this cubic model was not considered, since the addition of two more parameters served the purpose of regression, no better than the linear model. Thus techniques of linear regression were employed to model height versus transformed-percentile data.

The correlation between height and $Logit(p)$, at different ages, ranged from 0.99906 to 0.99938 for females, and from 0.99907 to 0.99938 for males, at different ages.

Transformed-linear regression model fitted to height-percentile data was**:**

(2) $$h = \alpha + \beta\ Logit(p)$$

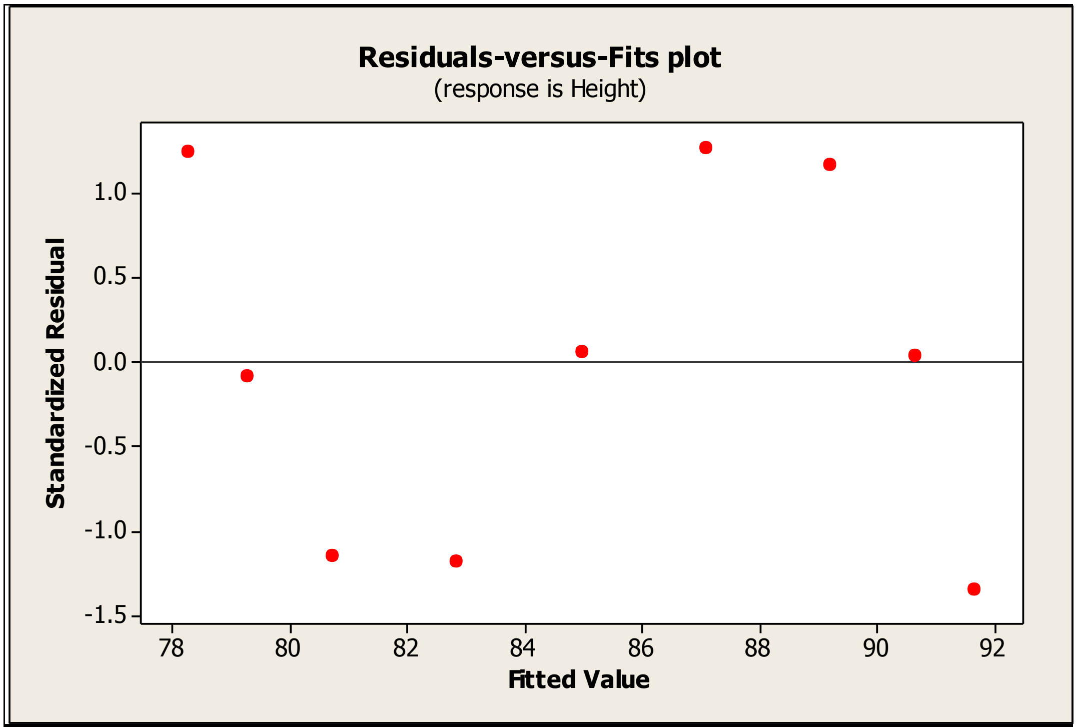

Fig. 3. Standardized residuals versus fits plot for height of girls (age: 2.0 years)

where h represents height (*centimeters*), at a particular age. The parameters α and β are intercept and slope of the linear model.

According to the results obtained through C# program, 99% (ranging from 99.81% to 99.87% for females and 99.81% to 99.87% for males, at different ages) of variation in heights was explained by $Logit(p)$.

The values of the slope parameter β ranged from 1.93 to 4.14 (SE 0.03 to 0.06) for females, and 1.94 to 4.49 (SE 0.03 to 0.07), for males, at different ages.

The values of intercept parameter α ranged from 84.97 to 163.31 for females, and 86.46 to 176.80 for males, at different ages.

Standardized residual plot obtained from Minitab is shown in Figure 3. This plot shows

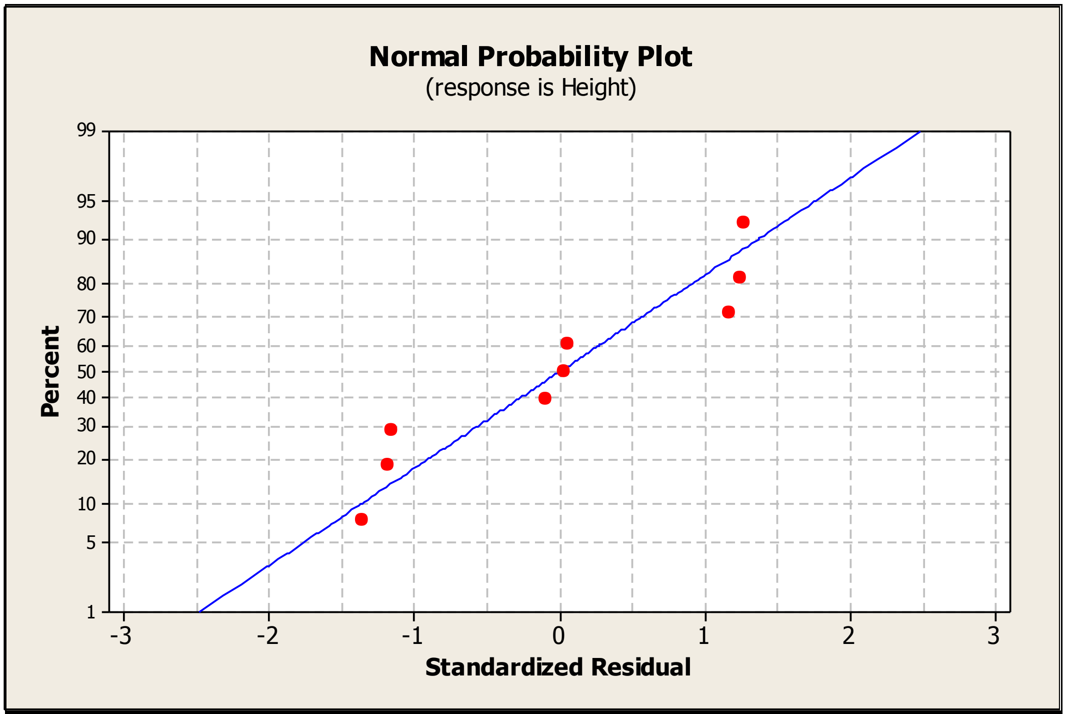

Fig. 4. Normal probability plot for standardized residuals for height versus *Logit(p)* percentile for girls (age: 2.0 years)

shows that the standardized residuals are randomly scattered around the line of 0 standardized residuals. Moreover, normal probability plot (Figure 4) shows that the residuals are somewhat normally distributed. These plots verify the assumptions of the linear model.

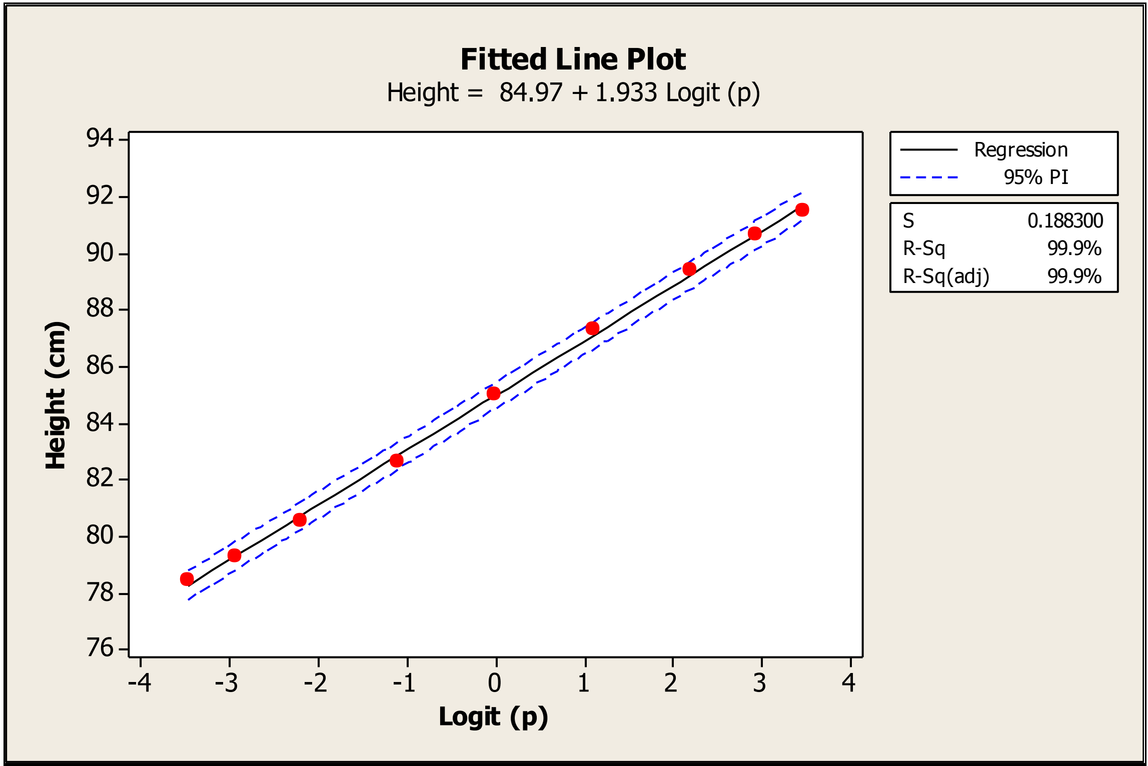

Fig. 5. Fitted line plot for height of girls (age: 2.0 years)

The 95% prediction interval, which is shown in the fitted-line plot (Figure 5) show that the prediction interval is quite narrow. More importantly, it doesn't widen up from the extreme sides, as it does usually in regression models. Thus extrapolation may be carried out, using the least-squares line (equation 2).

1.2 *Mass-Percentile-Regression Models*

Mass-versus-percentiles plots, for girls and boys, at different ages, were analyzed (Figure 6 shows one of these graphs, plotted from weight-for-age growth table, for girls, age: 2 years). For mass-percentile data, same transformation was employed on the percentiles as applied for height-percentile curve (equation 1). In addition to this transformation, the mass variable was *log* (base *e*) transformed. These transformations made the graph between the two variables, almost, linear, as depicted

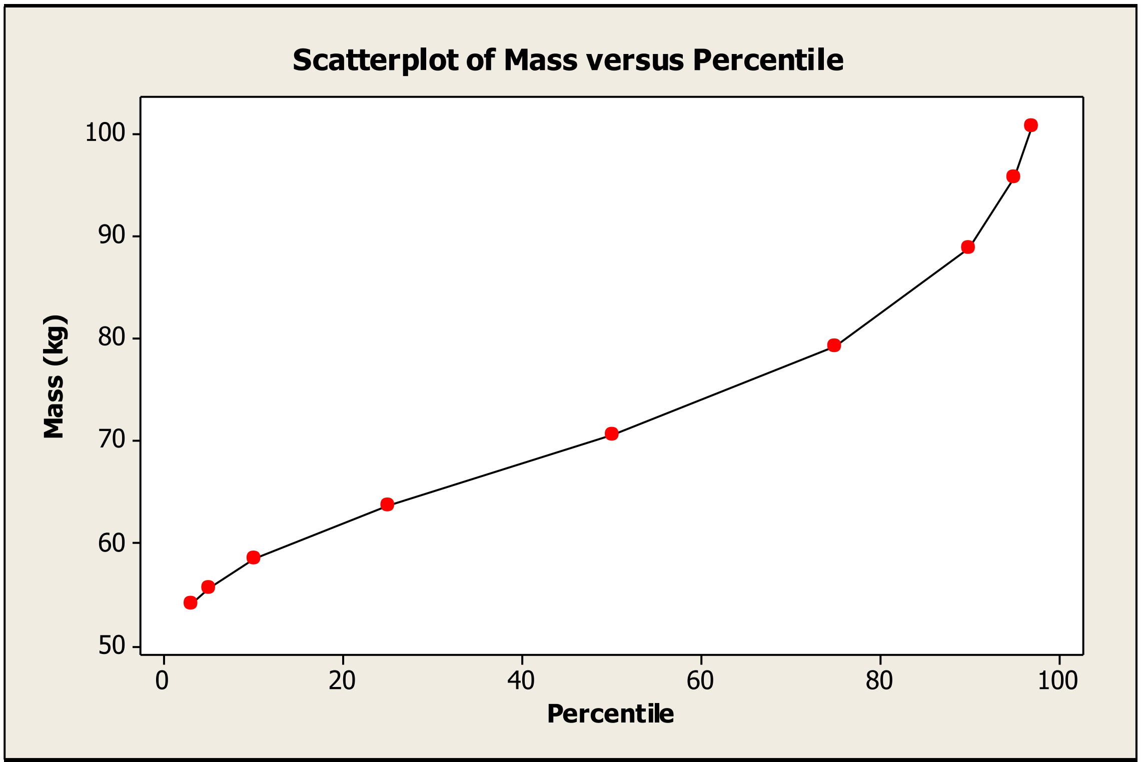

Fig. 6. Scatter plot of mass versus percentile of girls (age: 2.0 *years*)

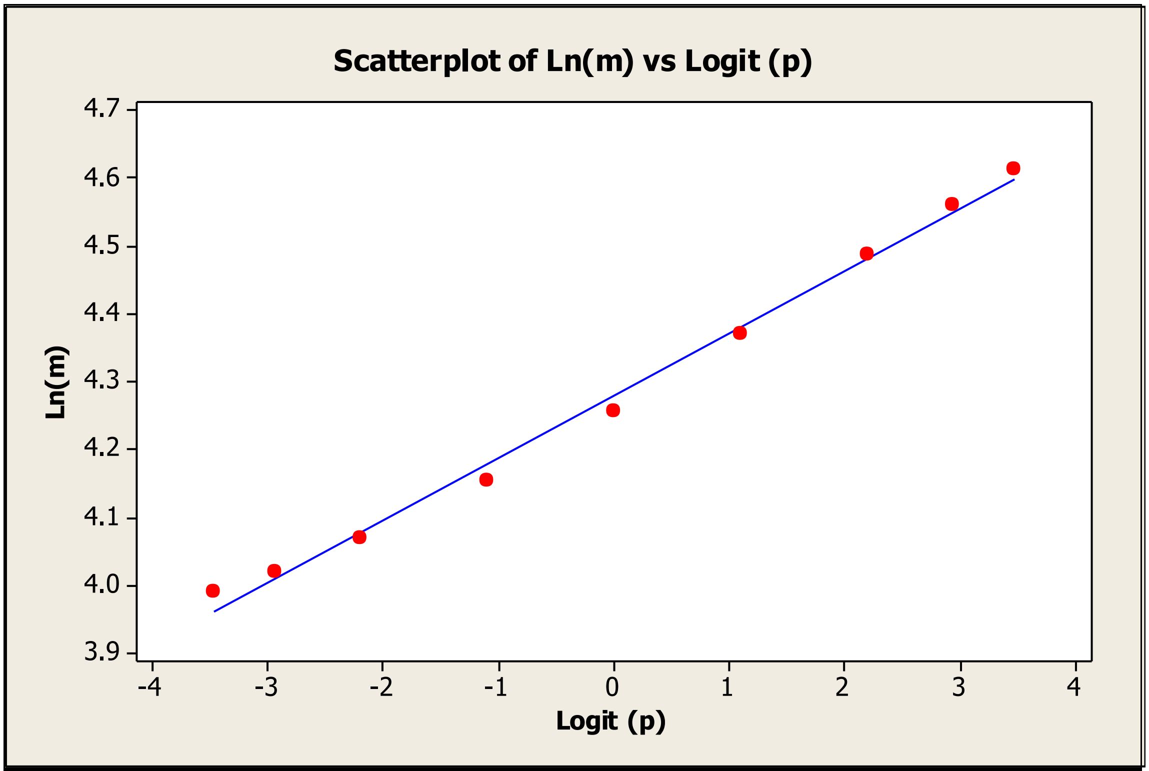

Fig. 7. Scatterplot of *ln(μ)* versus *Logit(p)* with regression line fitted

in Figure 7. According to the results of C# program, the correlation between $Ln(\mu)$ and $Logit(p)$, at different ages, ranged from 0.9883 to 0.9987 for females, and from 0.9938 to 0.9992 for males, at different ages.

Transformed linear-regression model fitted to mass-percentile data was**:**

(3) $$ln(\mu) = \gamma + \delta Logit(p)$$

where μ represents mass (*kilograms*), at a particular age. The parameters γ and δ are intercept and slope of the linear model.

Almost, 99% (ranging from 97.68% to 99.75% for females and 98.78% to 99.84% for males, at different ages) of variation in *ln(μ)* was explained by $Logit(p)$.

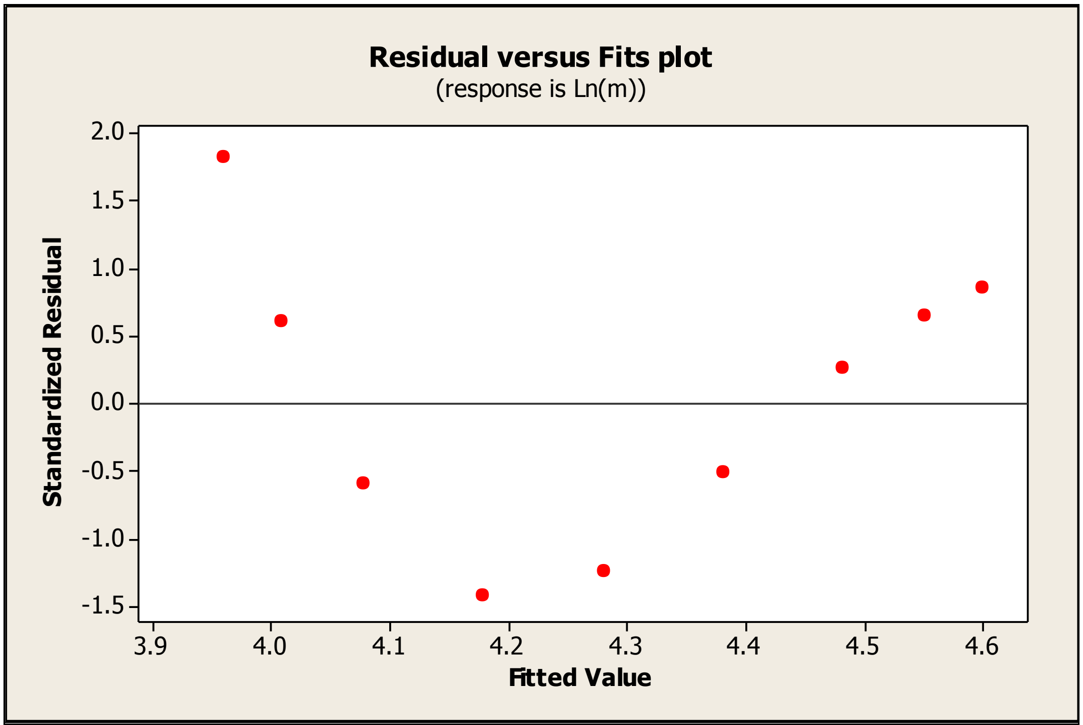

Fig. 8. Standardized residuals versus fits plot for height (girls, age: 2.0 years)

The values of the slope parameter δ ranged from 0.061 to 0.116 (SE 0.0083 to 0.0419) for females, and 0.061 to 0.112 (SE 0.0009 to 0.0042), for males, at different ages.

On the other hand, the values of intercept parameter γ ranged from 2.49 to 4.11 for females, and 2.56 to 4.28 for males, at different ages. Standardized residual (Figure 8)

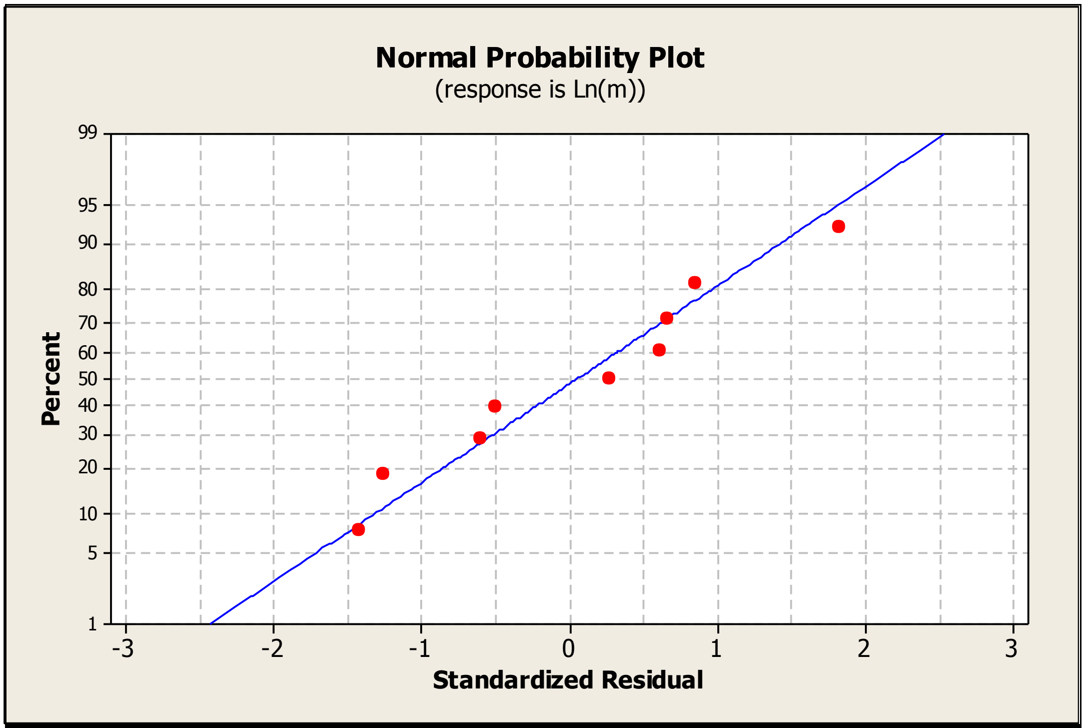

Fig. 9. Normal probability plot for standardized residuals for *ln(μ)* versus *Logit* percentile (girls, age: 2.0 years)

and normal probability plot (Figure 9) were drawn to verify the assumptions of the linear model.

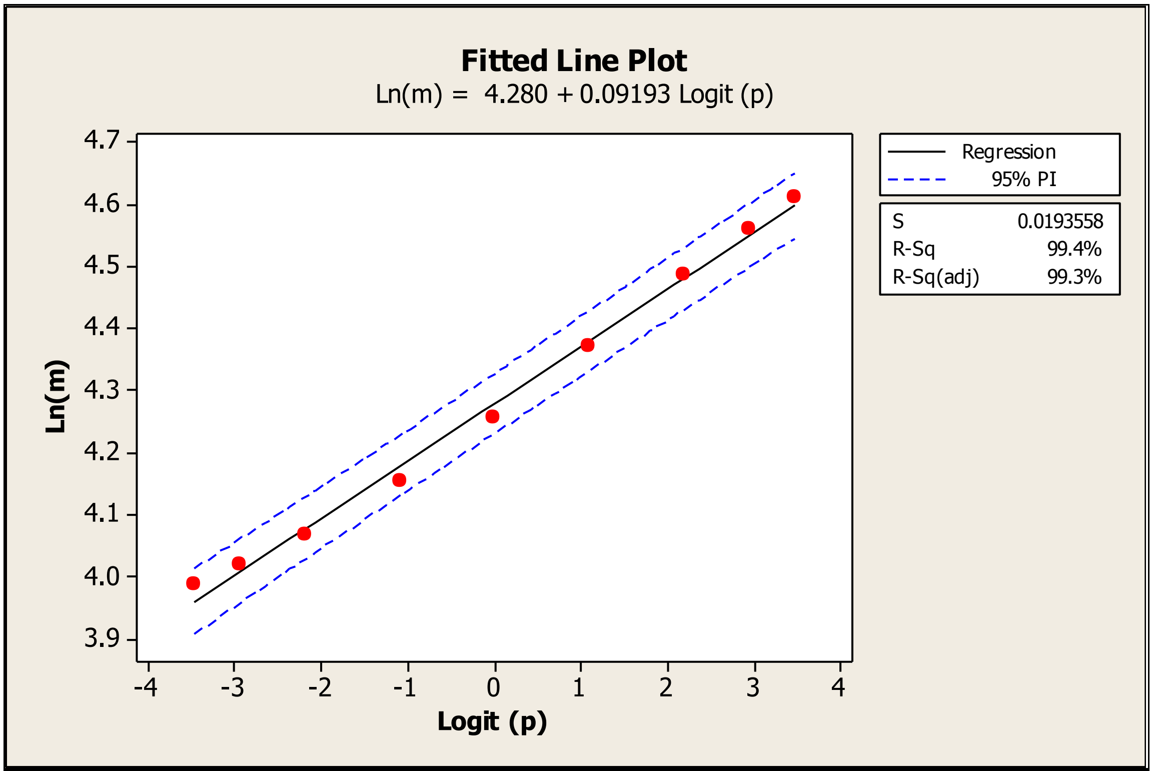

Fig. 10. Ten fitted line plots for *ln(μ)* (girls, age: 2.0 years)

The 95% prediction interval, which is shown in the fitted-line plot (Figure 10) show that the prediction interval is quite narrow, just like the prediction interval of height-percentile model (Figure 5). Moreover, it doesn't widen up from the extreme sides, as it, usually, does in regression models. Thus extrapolation may be carried out, using the least-squares line (equation 3).

3. Remarks

Although linear extrapolation failed outside the region defined by the interval $[3,97]$, it was observed that these two non-linear models didn't work better than linear interpolation, inside the region.

Besides these non-linear models, other models were also tried on the data, *e. g.*, step-wise regression (8 parameters), polynomial fitting (11 parameters), splining (6 parameters).

They were rejected on the basis of either being too complex (too many parameters) or not giving a good fit to the data.

Using extrapolation through these models, additional curves to height-for-age and weight-for-age charts representing 0.01st, 0.1st, 1st, 99th, 99.9th and 99.99th percentiles were incorporated in the already existing CDC charts. This was a significant value added to the research in this field.

Appendix: Source Code

```
using System;
using System.Collections.Generic;
using System.Text;
using System.Xml;
using System.Data;
using System.IO;
using Microsoft.Office.Core;
using Excel = Microsoft.Office.Interop.Excel;
using System.Web.UI.DataVisualization.Charting;


namespace ConsoleApplication1
{
  class Program
  {
    static void Main(string[] args)


{
    string path = @"C:\Users\Dell\Desktop";
    if (File.Exists(path + @"\CDC DataSheet.xlsx") == true)
    {
      Console.Write("Please wait");

      Excel.Application xlApp;
      Excel.Workbook xlWorkBook;
      Excel.Worksheet xlWorkSheet;
      Excel.Range chartRange;

      xlApp = new Excel.ApplicationClass();
      xlWorkBook = xlApp.Workbooks.Open(path + @"\CDC DataSheet.xlsx", 0,
false, 5, "", "", false, Microsoft.Office.Interop.Excel.XlPlatform.xlWindows,
"", true, false, 0, true, false, false);

      try
      {
        int l = 41;
        for (int k = 1; k <= 2; k++)
        {
        xlWorkSheet = (Excel.Worksheet)xlWorkBook.Worksheets.get_Item(k);
```

```
chartRange = xlWorkSheet.UsedRange;

if (k == 2)
  l = 40;
for (int j = 4; j <= l; j++)
{
  double[] h2 = new double[9];
  double sumh = 0;
  double sumh2 = 0;
  for (int i = 3; i <= 11; i++)
  {
  double h = (double)(chartRange.Cells[j, i] as Excel.Range).Value2;
    sumh = sumh + h;
    h2[i - 3] = h * h;
    sumh2 = sumh2 + h2[i - 3];
  }

  double avgh = sumh / 9;
  double[] p2 = new double[9];
  double[] p22 = new double[9];
  double[] hp2 = new double[9];
  double sump = 0;
  double sump2 = 0;
  double sump22 = 0;
  double sumhp2 = 0;
  for (int i = 3; i <= 11; i++)
  {
    double p = (double)(chartRange.Cells[3, i] as Excel.Range).Value2;
    sump = sump + p;
    double q = (p / 100) / (1 - (p / 100));
    p2[i - 3] = Math.Log(q, Math.E);
    sump2 = sump2 + p2[i - 3];
    p22[i - 3] = p2[i - 3] * p2[i - 3];
    sump22 = sump22 + p22[i - 3];
    double h = (double)(chartRange.Cells[j, i] as Excel.Range).Value2;
    hp2[i - 3] = p2[i - 3] * h;
    sumhp2 = sumhp2 + hp2[i - 3];
  }
 double avgp = sump / 9;
  double avgp2 = sump2 / 9;
  double Sxy = sumhp2 - sumh * sump2 / 9;
  double Sxx = sumh2 - (sumh * sumh) / 9;
  double Syy = sump22 - ((sump2 * sump2) / 9);
  double beta = Sxy / Syy;
  double alpha = avgh - beta * avgp2;
  double[] pEST = new double[9];
  double[] ppEST = new double[9];
  double[] ppEST2 = new double[9];
  double[] ppavg2 = new double[9];
  double sumppEST2 = 0;
  double sumppavg2 = 0;
  for (int i = 3; i <= 11; i++)
  {
    double h = (double)(chartRange.Cells[j, i] as Excel.Range).Value2;
    pEST[i - 3] = alpha + beta * p2[i - 3];
    double p = (double)(chartRange.Cells[3, i] as Excel.Range).Value2;
    ppEST[i - 3] = h - pEST[i - 3];
    ppEST2[i - 3] = ppEST[i - 3] * ppEST[i - 3];
    sumppEST2 = sumppEST2 + ppEST2[i - 3];
    ppavg2[i - 3] = (h - avgh) * (h - avgh);
    sumppavg2 = sumppavg2 + ppavg2[i - 3];
  }
  double SSE = sumppEST2;
  double MSE = SSE / 9;
  double SST = sumppavg2;
  double SSR = SST - SSE;
  double R2 = 1 - SSE / SST;
  double r = Sxy / (Math.Sqrt(Sxx) * Math.Sqrt(Syy));

  xlWorkSheet.Cells[j, 13] = Sxy;
  xlWorkSheet.Cells[j, 14] = Sxx;
  xlWorkSheet.Cells[j, 15] = Syy;
  xlWorkSheet.Cells[j, 16] = beta;
  xlWorkSheet.Cells[j, 17] = alpha;
  xlWorkSheet.Cells[j, 18] = SSE;
  xlWorkSheet.Cells[j, 19] = MSE;
  xlWorkSheet.Cells[j, 20] = SST;
  xlWorkSheet.Cells[j, 21] = SSR;
  xlWorkSheet.Cells[j, 22] = R2;
  xlWorkSheet.Cells[j, 23] = r;
}
}
for (int k = 3; k <= 4; k++)
{
xlWorkSheet = (Excel.Worksheet)xlWorkBook.Worksheets.get_Item(k);
chartRange = xlWorkSheet.UsedRange;

for (int j = 4; j <= 40; j++)
{
  double[] m1 = new double[9];
  double[] m12 = new double[9];
  double summ1 = 0;
  double summ = 0;
  double summ12 = 0;
  for (int i = 3; i <= 11; i++)
  {
    double m = (double)(chartRange.Cells[j, i] as Excel.Range).Value2;
    summ = summ + m;
    m1[i - 3] = Math.Log(m);
    summ1 = summ1 + m1[i - 3];
    m12[i - 3] = m1[i - 3] * m1[i - 3];
    summ12 = summ12 + m12[i - 3];
  }
  double avgm = summ / 9;
  double avgm1 = summ1 / 9;

  double[] p2 = new double[9];
  double[] p22 = new double[9];
  double[] m1p2 = new double[9];
  double sump = 0;
  double sump2 = 0;
  double sump22 = 0;
  double summ1p2 = 0;
  for (int i = 3; i <= 11; i++)
```

```
          {
            double p = (double)(chartRange.Cells[3, i] as Excel.Range).Value2;
            sump = sump + p;
            double q = (p / 100) / (1 - (p / 100));
            p2[i - 3] = Math.Log(q,Math.E);
            sump2 = sump2 + p2[i - 3];
            p22[i - 3] = p2[i - 3] * p2[i - 3];
            sump22 = sump22 + p22[i - 3];
            m1p2[i - 3] = p2[i - 3] * m1[i - 3];
            summ1p2 = summ1p2 + m1p2[i - 3];
          }
          double avgp = sump / 9;
          double avgp2 = sump2 / 9;
          double Sxy = summ1p2 - summ1 * sump2 / 9;
          double Sxx = summ12 - (summ1 * summ1) / 9;
          double Syy = sump22 - ((sump2 * sump2) / 9);
          double beta = Sxy / Syy;
          double alpha = avgm1 - beta * avgp2;

          double[] pEST = new double[9];
          double[] ppEST = new double[9];
          double[] ppEST2 = new double[9];
          double[] ppavg2 = new double[9];
          double sumppEST2 = 0;
          double sumppavg2 = 0;
          for (int i = 3; i <= 11; i++)
          {
            double m = (double)(chartRange.Cells[j, i] as Excel.Range).Value2;
            pEST[i - 3] = alpha + beta * p2[i - 3];
            double p = (double)(chartRange.Cells[3, i] as Excel.Range).Value2;
            ppEST[i - 3] = m1[i - 3] - pEST[i - 3];
            ppEST2[i - 3] = ppEST[i - 3] * ppEST[i - 3];
            sumppEST2 = sumppEST2 + ppEST2[i - 3];
            ppavg2[i - 3] = (m1[i - 3] - avgm1) * (m1[i - 3] - avgm1);
            sumppavg2 = sumppavg2 + ppavg2[i - 3];
}
          double SSE = sumppEST2;
          double MSE = SSE / 9;
          double SST = sumppavg2;
          double SSR = SST - SSE;
          double R2 = 1 - SSE / SST;
          double temp = Sxx * Syy;
          double r = Sxy / (Math.Sqrt(temp));
          double temp1 = SSE / 7;
          double Se = Math.Sqrt(temp1);
          double SEp = Se / (Math.Sqrt(Syy));
          double Tratio = beta / SEp;
          double Tstat = r / Math.Sqrt((1 - (r * r)) / 7);

          xlWorkSheet.Cells[j, 13] = Sxy;
          xlWorkSheet.Cells[j, 14] = Sxx;
          xlWorkSheet.Cells[j, 15] = Syy;
          xlWorkSheet.Cells[j, 16] = beta;
          xlWorkSheet.Cells[j, 17] = alpha;
          xlWorkSheet.Cells[j, 18] = SSE;
          xlWorkSheet.Cells[j, 19] = MSE;
          xlWorkSheet.Cells[j, 20] = SST;
          xlWorkSheet.Cells[j, 21] = SSR;
          xlWorkSheet.Cells[j, 22] = R2;
          xlWorkSheet.Cells[j, 23] = r;
          xlWorkSheet.Cells[j, 25] = Se;
          xlWorkSheet.Cells[j, 26] = SEp;
          xlWorkSheet.Cells[j, 27] = Tratio;
          xlWorkSheet.Cells[j, 30] = Tstat;
        }
        }
      }
      catch (Exception)
      {
Console.Clear();
        Console.WriteLine("Some data is missing."
                          "You might not get the appropriate results ");
      }
      xlWorkBook.Save();
      xlWorkBook.Close(true, null, null);
      xlApp.Quit();

      Console.Clear();
      Console.WriteLine("Success!");
}
    else
    {
      Console.WriteLine("File not found.");
    }
    }
  }
}
```

The above description is taken from MPhil Thesis of Samira Sahar Jamil
(figure and equation numbers have been adapted for this document)

Web address of the main document:
KJ-REGRESSION MODEL TO EVALUATE OPTIMAL MASSES OF EXTREME CASES
http://www.ngds-ku.org/Papers/J34.pdf

Web address of this document:
Additional File 1: MATHEMATICAL-STATISTICAL FRAMEWORK OF KJ-REGRESSION MODEL
http://www.ngds-ku.org/Papers/J34/Additional_File_1.pdf