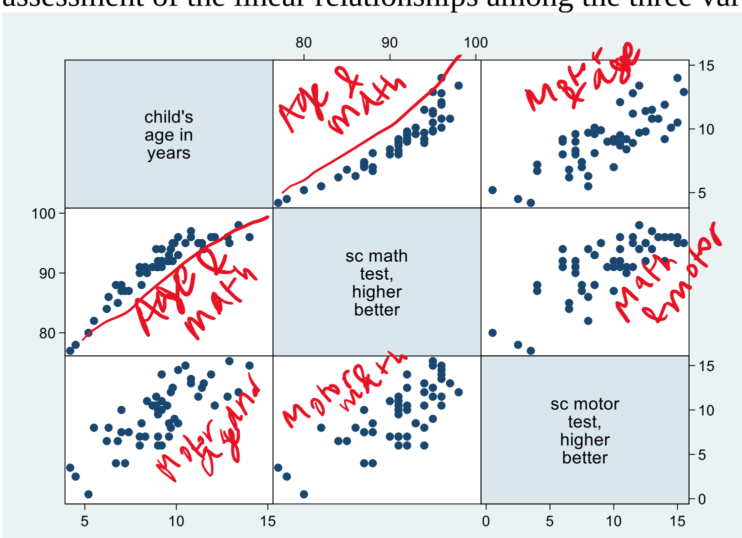
Lab 2 BIOSTAT100B – Lizette Romano

1) Printout with the summary statistics on all variables (i.e., results from the ‘sum’ command). Briefly describe the basic demographics of the children’s ages and their math and motor skills.

```
     Variable |        Obs        Mean
>    Std. dev.       Min         Max
-------------+------------------------
> ------------------------------
          oid |         50        25.5
>    14.57738          1          50
          age |         50       9.104
>    2.196933        4.2          14
         math |         50       91.08
>    4.831191         77          98
        motor |         50        9.48
>     3.46257         .5        15.5
```

The children were between 4.2 years old and 14 years old, with the mean being 9.10. The children’s motor skills ranged from 0.5 to 15.5, with the mean being 9.48. The children’s’ math skills ranged from 77 to 98, with the mean being 91.08.

2) Printout correlation and scatterplot matrices for the three variables age, math and motor. Briefly describe the magnitude of correlations among the three variables and give a visual assessment of the linear relationships among the three variables.



As children got older, they had higher math scores. There was a high correlation of 0.92 between age and math. There is no clear linear relationship between age and motor skills or between motor and math skills since the correlation lines are scattered.

3) Printout regression tables with summary statistics on each of the three regression models.

Regression Model i

```
      Source |       SS           df
>    MS      Number of obs   =        50
-------------+----------------------------------
> ---------  F(1, 48)        =     63.26
       Model |  650.273859         1  65
> 0.273859   Prob > F        =    0.0000
    Residual |  493.406141        48  10
> .2792946   R-squared       =    0.5686
-------------+----------------------------------
> ---------  Adj R-squared   =    0.5596
       Total |     1143.68        49  23
> .3404082   Root MSE        =    2.0261

------------------------------------------------
> ----------------------------------------------
        math | Coefficient  Std. err.
>    t
>       P>|t|
>                 [95% con
>                           f. interval]
-------------+----------------------------------
> ----------------------------------------------
       motor |   1.052087   .1322772
>  7.95
>       0.000
>                .7861257
>                               1.318048
       _cons |   81.10622   1.333444
> 60.82
>       0.000
>                78.42515
>                               83.78728
------------------------------------------------
```

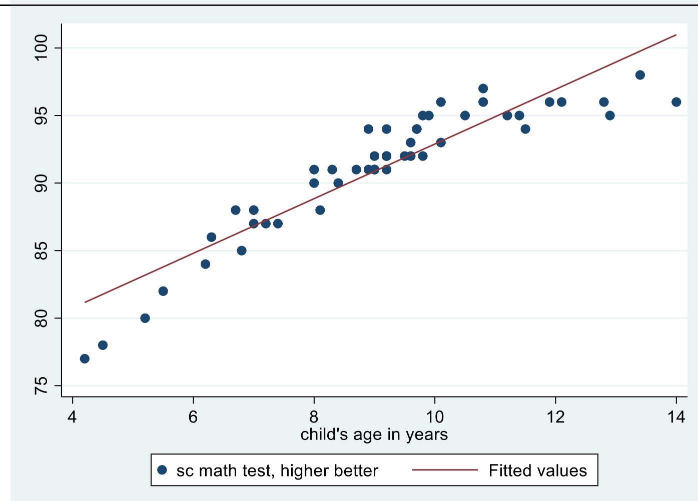
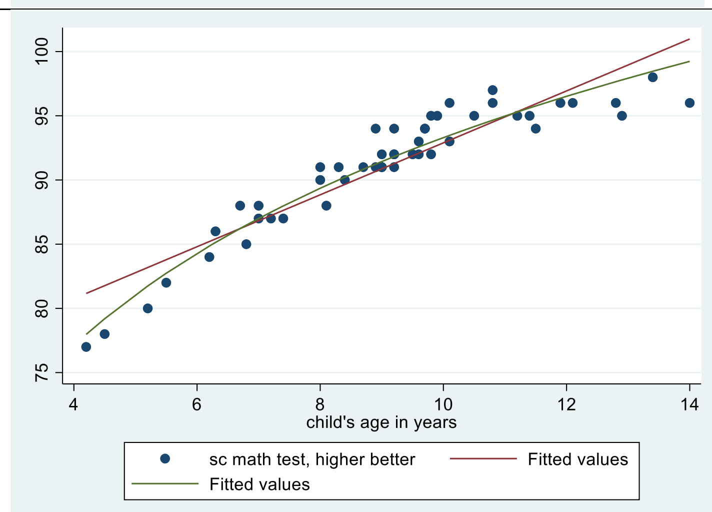
Regression Model ii

```
      Source |       SS           df
>    MS      Number of obs   =        50
-------------+----------------------------------
> ---------  F(1, 48)        =    263.17
       Model |  967.257335         1  96
> 7.257335   Prob > F        =    0.0000
    Residual |  176.422665        48  3.
> 67547218   R-squared       =    0.8457
-------------+----------------------------------
> ---------  Adj R-squared   =    0.8425
       Total |     1143.68        49  23
> .3404082   Root MSE        =    1.9172

------------------------------------------------
> ----------------------------------------------
        math | Coefficient  Std. err.
>    t
>       P>|t|
>                 [95% con
>                           f. interval]
-------------+----------------------------------
> ----------------------------------------------
         age |   2.022349   .1246642
> 16.22
>       0.000
>                1.771695
>                               2.273004
       _cons |   72.66853   1.166878
> 62.28
>       0.000
>                70.32237
>                               75.0147
------------------------------------------------
```

Regression Model iii:

```
      Source |       SS           df
>    MS      Number of obs   =        50
-------------+----------------------------------
> ---------  F(1, 48)        =    574.59
       Model |  1055.50584         1  10
> 55.50584   Prob > F        =    0.0000
    Residual |  88.1741632        48  1.
> 83696173   R-squared       =    0.9229
-------------+----------------------------------
> ---------  Adj R-squared   =    0.9213
       Total |     1143.68        49  23
> .3404082   Root MSE        =    1.3553

------------------------------------------------
> ----------------------------------------------
        math | Coefficient  Std. err.
>    t
>       P>|t|
>                 [95% con
>                           f. interval]
-------------+----------------------------------
> ----------------------------------------------
        lage |   17.66146   .7367945
> 23.97
>       0.000
>                16.18004
>                               19.14289
       _cons |   52.62924   1.615486
> 32.58
>       0.000
>                49.38109
>                               55.8774
------------------------------------------------
```

4) Printout scatterplots of variables with the fitted regression lines for each of the three models. Label regression lines with appropriate formulas showing point estimates of the intercept and slope as best you can by hand

Model i: **y = 80 + 1.4x**	
Model ii: **y = 80 + 2.5x** **Good model, close to line**	
Model iii: **y = 75 + 5x** **red line = second model linear regression, green line = 3rd model, Curve is quadratic because we are using scatter line of original scatterplot.**	

5) Regarding model(i), calculate individual 95% confidence intervals for the intercept and slope parameters, and E(math|motor=10). Based on the studentized residual plot, briefly assess model fit and assumptions. What would you advise the investigator about whether the data support the main hypothesis?

CI for slope/motor: (1.052087 .1322772)

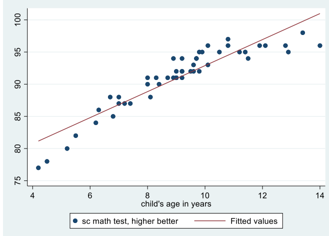
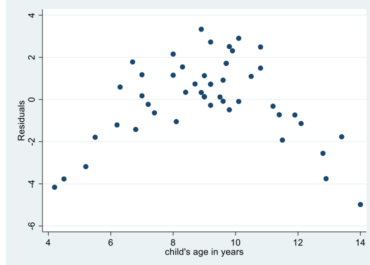
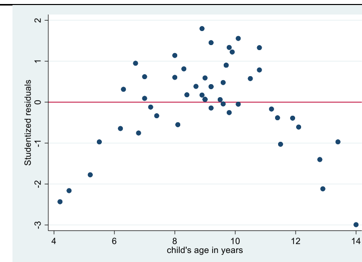
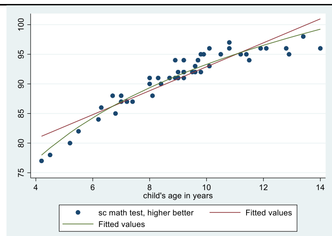
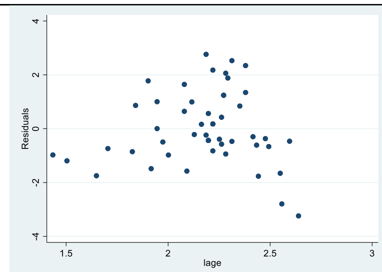
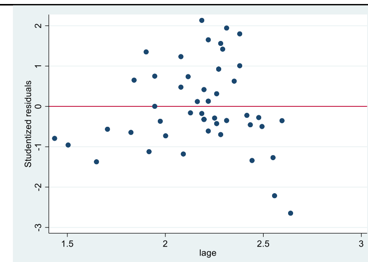
CI for intercept: (81.10622 1.333444)

CI for E: (2.022349 .1246642)

The studentized residual plot shows a random pattern spread evenly so I would assume the data is a random sample and has a normal distribution. The studentized residual values are centered at 0 so mean of residual is 0. The variance is constant since it is random. I would therefore advise that the data is a close representation to support the main hypothesis.

6) Regarding models ii and iii, discuss the two models and indicate which one is "best" in explaining the relationship between math and age. Explain with supporting scatter and diagnostic plots.

The best model would be model iii because the variance is constant compared to model ii. Variance is not constant in model ii since there is a pattern in the residual model. In model iii, values are centered around 0 and have no pattern. There are more points centered around 0 in the studentized residual plots of model making it a better fit. Both regression models are good but model iii would be better.

Model ii:	 **Good model, close to line**	 **Variance not constant since we can see pattern in residual model**	 **We would use log of x since it is same as previous outliers**
Model iii:	 **red line = second model linear regression, green line = 3rd model, Curve is quadratic because we are using scatter line of original scatterplot.**	 **Good residual analysis plot = values are centered by 0 and scattered randomly, no pattern = variance is constant**	 **Similar to previous, stud res values are within -3 and 3, no outliers, pattern is random, meets assumptions of linear regression models**

7) List one additional linear model to describe the relationship between math and age

y = 75 + 1.54x