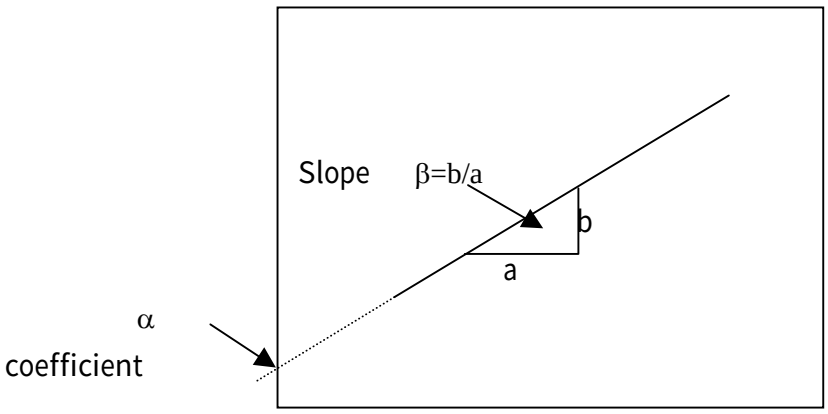
Simple Linear Regression

Correlation 解析のように２つの連続して変化する(yes/no ではない)変数を比較する際、simple linear regression が用いられます。Correlation analysis においては x と y を入れ替えても問題ありませんが、single linear regression model では x が explanatory variable として変化すると y が response として変化します。よって X を independent variable, Y を dependent variable とも表現します。そして１つの直線を引くことにより新たに与えられた値 x から y を算出することができます。例えば成長曲線は年齢を x、身長を y としたグラフで、年齢を当てはめればその年齢に相当する平均身長が得られるといった具合です。



$$\mu_{y/x} = \alpha + \beta x$$

であり直線となります。このμ を使って表される公式は不変の真理です。極端な話、世界全てのデータを集めたものをμとします。しかしこれを知ることは実際不可能です。そこで我々は手元にある sample からこの世界のデータを推論するわけです。しかし sample 数が十分でなければ、世界のデータから少しずれてしまうかもしれません。その分をε (error)として表します。Y は１部のサンプルを用いていることを示しています。

$$y = \alpha + \beta x + \varepsilon$$

Simple linear regression において、変数 X は観察された範囲において変数 Y と直線的関係にあることが期待されます。そうでなければ simple linear regression にはなりません。また X と Y はそれぞれ独立しながらも同じ分布を示します。そうでなければやはり

simple linear regression になりません。これは仮定です。もちろん実際は直線に近くなることはあっても直線にはならないことがほとんどです。。

The Method of Least Squares

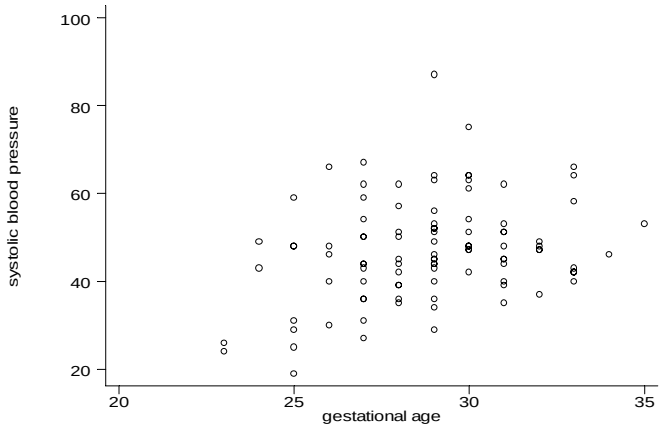
下の表はある病院で出生した低出生体重児 100 人の収縮期血圧(sbp), 性別(sex)、母親の中毒症の有無(tox)、妊娠中の出血（切迫早産の兆候）(grmhem)、在胎週数(gestage)、5 分後のアプガースコア(apgar5)を示しています。

. list

	sbp	sex	tox	grmhem	gestage	apgar5
1.	43	Male	No	No	29	7
2.	51	Male	No	No	31	8
3.	42	Female	No	No	33	0
4.	39	Female	No	No	31	8
5.	48	Female	Yes	No	30	7
6.	31	Male	No	Yes	25	0
7.	31	Male	Yes	No	27	7
8.	40	Female	No	No	29	9
9.	57	Female	No	No	28	6
10.	64	Female	Yes	No	29	9
11.	46	Female	No	No	26	7
12.	47	Female	No	Yes	30	6
13.	63	Female	No	No	29	8
14.	56	Female	No	No	29	1
15.	49	Male	No	No	29	8
16.	87	Male	No	No	29	7
17.	46	Female	No	No	29	8
18.	66	Female	No	No	33	8
19.	42	Female	Yes	No	33	8
20.	52	Female	No	No	29	7
21.	51	Male	No	No	28	7
22.	47	Female	No	No	30	9
23.	54	Male	No	No	27	4
24.	64	Male	No	No	33	9
25.	37	Female	No	No	32	7
26.	36	Male	Yes	No	28	3
27.	45	Female	No	Yes	29	7
28.	39	Male	No	No	28	7
29.	29	Female	No	No	29	4
30.	61	Female	No	No	30	3
31.	53	Male	No	No	31	7
32.	64	Female	No	No	30	7
33.	35	Female	No	No	31	6
34.	34	Male	No	No	29	9
35.	62	Female	No	No	27	7
36.	59	Female	No	No	27	8
37.	36	Male	No	No	27	9
38.	47	Female	No	No	32	8
39.	45	Male	No	Yes	31	2
40.	62	Female	No	Yes	28	5
41.	75	Male	Yes	No	30	7
42.	44	Male	No	No	29	0
43.	39	Male	No	No	28	8
44.	48	Female	No	Yes	31	7
45.	43	Female	Yes	No	27	6
46.	19	Female	No	Yes	25	4
47.	63	Male	No	No	30	7
48.	42	Male	No	No	28	6
49.	44	Female	No	No	28	9
50.	25	Female	No	No	25	8
51.	26	Female	No	No	23	8
52.	27	Male	No	No	27	9
53.	35	Male	No	No	28	8
54.	40	Male	No	No	27	7
55.	44	Female	No	No	27	6
56.	66	Male	No	No	26	8

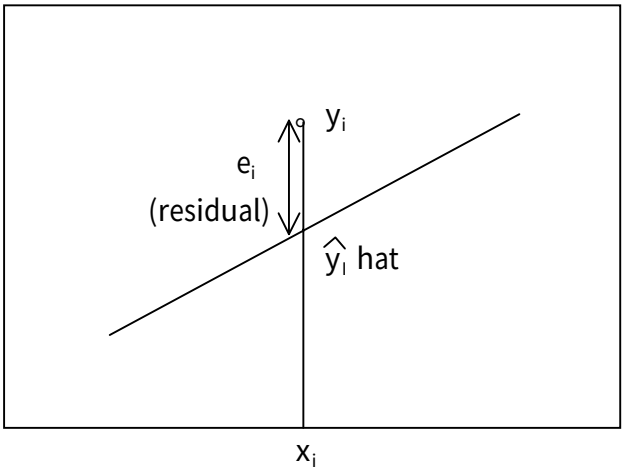
57.	59	Female	No	No	25	3
58.	24	Female	No	No	23	7
59.	40	Female	No	Yes	26	3
60.	49	Female	No	No	24	5
61.	53	Male	Yes	No	29	9
62.	45	Female	No	No	29	9
63.	50	Male	No	Yes	27	8
64.	64	Male	No	No	30	7
65.	48	Female	No	No	30	6
66.	48	Female	No	Yes	32	4
67.	58	Female	Yes	No	33	7
68.	67	Female	No	No	27	8
69.	40	Female	No	Yes	31	8
70.	48	Female	No	No	26	8
71.	36	Male	No	No	27	5
72.	44	Male	No	No	27	6
73.	53	Female	Yes	No	35	9
74.	45	Female	Yes	No	28	6
75.	54	Male	No	No	30	8
76.	44	Male	Yes	No	31	2
77.	42	Male	No	No	30	5
78.	50	Female	No	No	27	0
79.	48	Female	No	No	25	5
80.	29	Female	No	Yes	25	5
81.	30	Female	No	Yes	26	2
82.	36	Female	No	No	29	0
83.	44	Female	No	No	29	0
84.	46	Female	Yes	No	34	9
85.	51	Male	Yes	No	30	4
86.	51	Male	No	No	29	5
87.	43	Male	Yes	No	33	7
88.	48	Male	No	No	30	5
89.	52	Male	No	No	29	8
90.	43	Male	No	No	24	6
91.	42	Male	Yes	No	33	8
92.	48	Male	No	Yes	25	5
93.	49	Female	Yes	No	32	8
94.	62	Male	Yes	No	31	7
95.	45	Male	No	No	31	9
96.	51	Female	Yes	Yes	31	6
97.	52	Male	No	No	29	8
98.	47	Male	Yes	No	32	5
99.	40	Female	Yes	No	33	8
100.	50	Female	No	No	28	7

この表のデータを基に妊娠週数と収縮期血圧の関係をみてみましょう。



それぞれの点は広い幅を持って分布しています。しかし妊娠週数が進めば収縮期血圧も

上がる傾向にはあります。これらの点をよく表すように線を引くにはどうしらたよりでしょうか？2 人の人に書かせると微妙に異なった 2 種類の線になることでしょう。そこで皆の引く線が一致するように我々は method of least squares という概念を用います。



グラフ上のどの点(x_i, y_i)も描こうと思う直線からは一定の距離(e_i)をもって存在しています（中には直線上に乗るものもあるかもしれません）。y_i hat はx_i とy_iをつなぐ線と直線の交叉する点だとします。$y_i - y_i$ hat = e_iの関係にあります。e_i が 0 であると直線上に位置します。しかし通常どの点も直線より一定の距離をもって存在します。そこで e_i の総和を最小になるように直線を引けると理想的です。

Error sum of squares, or residual sum of squares

$$\Sigma e_i^2 = \Sigma (y_i - y_i \text{ hat})^2$$

よって least square line とは error sum of squares を最小にするように引いた直線のことであり、この方法を method of least square と呼びます。

$$y \text{ hat} = \alpha + \beta x$$

$$\Sigma e_i^2 = \Sigma (y_i - y_i \text{ hat})^2$$

$$= \Sigma (y_i \cdot a - bx_i)^2$$

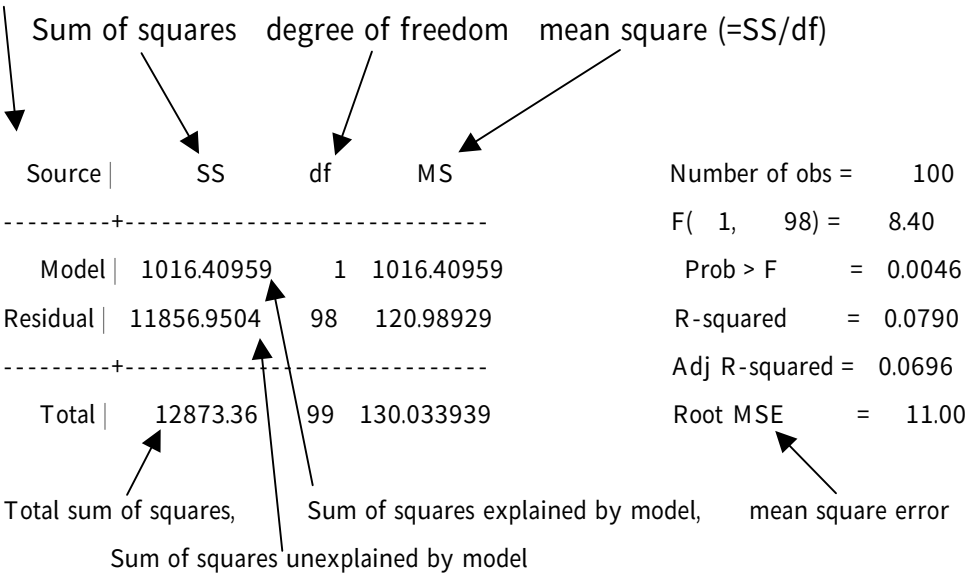
$$\beta = [\Sigma (x_i \cdot x_{mean})(y_i \cdot y_{mean})]/[\Sigma (x_i \cdot x_{mean})^2]$$

$$\alpha = y_{mean} - \beta x_{mean}$$

しかしこれを計算するのは至難の技で、コンピュータにやってもらいます。

. regress sbp gestage

analysis of variance (ANOVA)



sbp	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
gestage	1.26438	.4362311	2.898	0.005	.3986934	2.130066
_cons	10.55207	12.65063	0.834	0.406	-14.55269	35.65682

α β

yhat = 10.55 + 1.26 x, or systolic pressure = 10.55 + 1.26 gestational week

となります。上の一次方程式に妊娠週数を次々当てはめていけば収縮期血圧が予測できるというわけです。更にα, β に対してstandard error, t, p value (two sided), 95% CI が計算されます。t はstd err をcoef (α or β)で割ったものです。また上のR-squared はｒの二乗、すなわち相関係数の二乗なのでここで相関係数は$\sqrt{0.0790}$ = 0.28 となります。R-squred はcoefficient of determination であり、相関係数(Pearson's correlation coefficient)が －1 から＋１までの範囲であるのに対して、R^2は 0 から１までです。ここではR^2 は 0.0790 ですが、その意味は「出生時平均血圧の 7.9%は妊娠週数との直線関係で説明される」ことを意味しています。「たった 7.9%しか説明がつかない」というのは弱い相関を連想させますが、最初のグラフをみれば当然の話で、相当のばらつきを

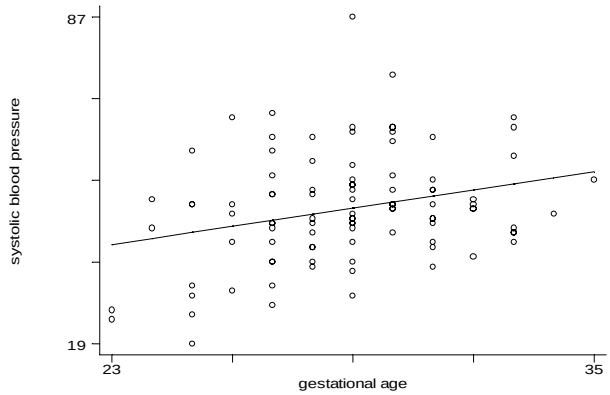
もって分布しています。妊娠週数から血圧を予想できても実際は当らないことが多いという見方もできます。またleast squares はoutliners により大きな影響を受けます。Outliner が測定ミスや記入上の間違いなど人為的ミスが明らかな場合は削除すればよいのですが、必ずしもそうはいかないときもあります。また２つの変数が必ずしも直線にならないこともあるでしょう。そのような場合にはtransformation といって、x2, log, √などにより変数を修飾します。

```
. generate    sbpsq= sbp^2

. regress    sbpsq    gestage

  Source |       SS       df       MS                  Number of obs =       100
---------+------------------------------                F(  1,    98) =      5.14
   Model |  6400141.93     1  6400141.93                Prob > F      =    0.0256
Residual |  122047633     98  1245384.01                R-squared     =    0.0498
---------+------------------------------                Adj R-squared =    0.0401
   Total |  128447775     99  1297452.28                Root MSE      =    1116.0

------------------------------------------------------------------------------
   sbpsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 gestage |    100.3317   44.25831     2.267   0.026      12.50248    188.1608
   _cons |   -553.3214   1283.483    -0.431   0.667     -3100.352    1993.709
```

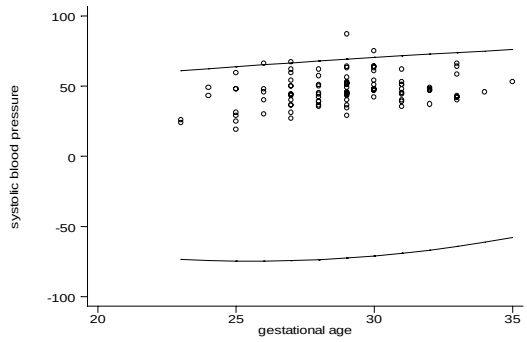
R-squared をみると（血圧）^{2}を変数とした場合かえって相関が下がってしまいました。よって無理に（血圧）^{2}を変数として用いる必要はなさそうです。

グラフを描かせてみましょう。

. predict hat
(option xb assumed; fitted values)
. graph　sbp hat gestage, c(.l) s(Oi) sort



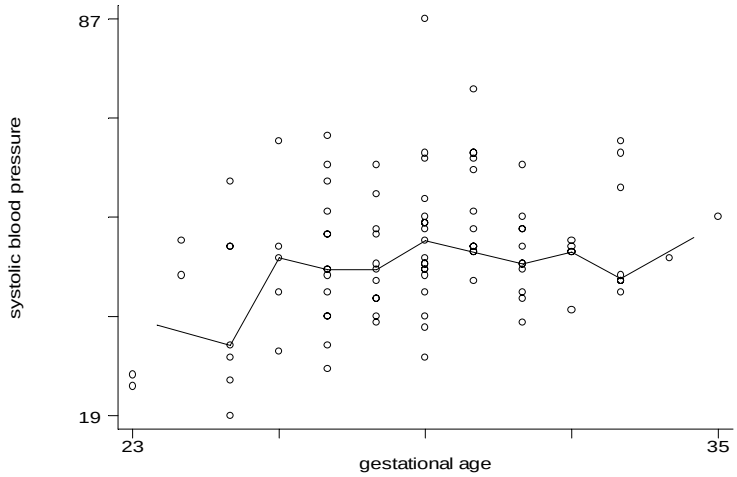
Standard error　２つ分の範囲を示してみます。
. predict s, stdr
. gen lo = hat-s*s
. gen hi = hat+2*s
. graph sbp hi lo gestage, c(.lll) s(Oiii) sort ylab xlab



Standard error が大き過ぎるためにはみだしてしまいました。

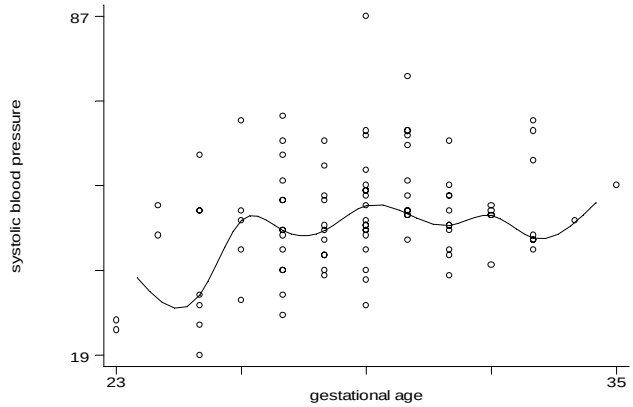
ｘ軸を例えば 10 等分して、その中での中央値を計算し結んでみます。この方が outliers による影響を少なくすることができます。

. graph　sbp gestage,c(m) bands(10)



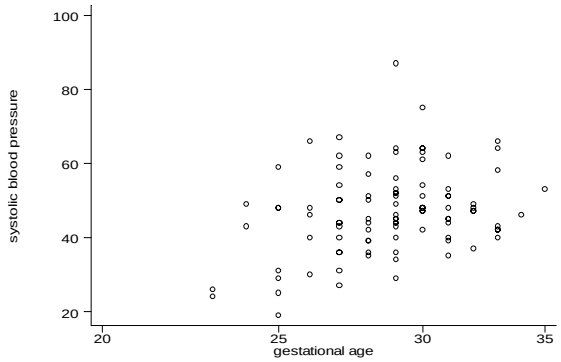
分割の数を増やすと折れ線グラフに近くなります。もう少しなめらかな(smooth)グラフにするとしたらどうしますか？

. graph　sbp gestage,c(s) bands(10)



グラフの上で log を用いる方法を紹介します。

. graph　sbp gestage, xlog ylab xlab



ｘ軸に注目してください。ちゃんとログになっています。

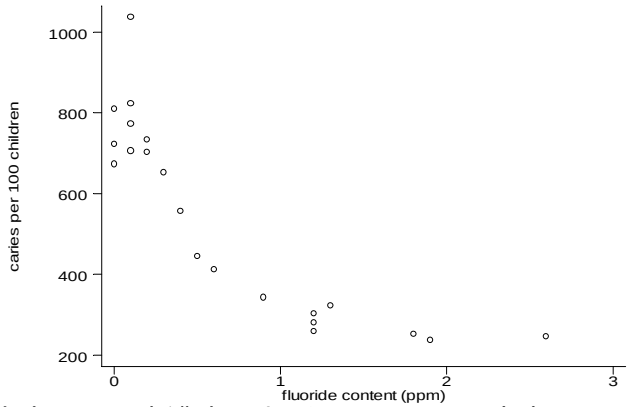
もう１つの例をとってグラフを描いてみましょう。アメリカでは「水道水にフッ素を含めることにより虫歯が減るかどうか」について検討がなされました。

. list

	fluoride	caries
1.	1.9	236
2.	2.6	246
3.	1.8	252
4.	1.2	258
5.	1.2	281
6.	1.2	303
7.	1.3	323
8.	.9	343
9.	.6	412
10.	.5	444
11.	.4	556
12.	.3	652
13.	0	673
14.	.2	703
15.	.1	706
16.	0	722
17.	.2	733
18.	.1	772
19.	0	810
20.	.1	823
21.	.1	1037

さてこれをグラフに描くと、

. graph caries　fluoride, ylab xlab



水道水中のフッ素濃度を上げると明らかに虫歯の Rate が減っています。日本でも検討した方が良いかもしれません。

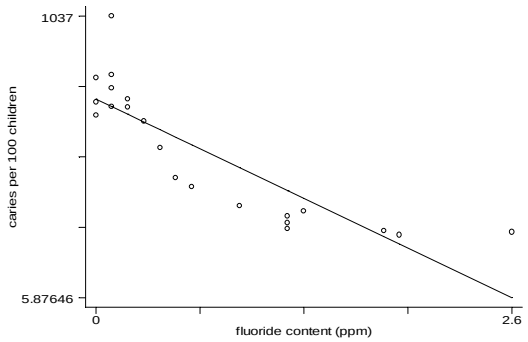
方程式はどうですか？

```
. regress  caries fluoride

  Source |       SS       df       MS              Number of obs =      21
---------+------------------------------           F(  1,    19) =   52.56
   Model |  870184.778     1  870184.778           Prob > F      =  0.0000
Residual |  314548.174    19  16555.1671           R-squared     =  0.7345
---------+------------------------------           Adj R-squared =  0.7205
   Total |  1184732.95    20  59236.6476           Root MSE      =  128.67

------------------------------------------------------------------------------
  caries |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
fluoride |  -279.7392   38.58464    -7.250  0.000    -360.4978   -198.9806
   _cons |   733.1984   38.95947    18.820  0.000     651.6553    814.7415
------------------------------------------------------------------------------
```

y = 733.2 – 279.7*フッ素濃度

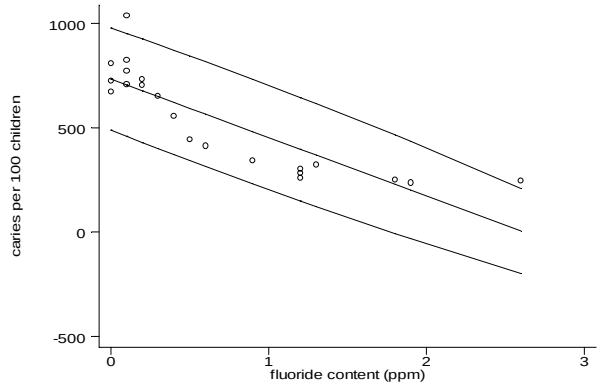
今度は高い相関が得られました。$\sqrt{0.7345}$ = 0.86 です。

```
. predict hat
(option xb assumed; fitted values)
. graph caries hat  fluoride, c(.l) s(Oi) sort
```

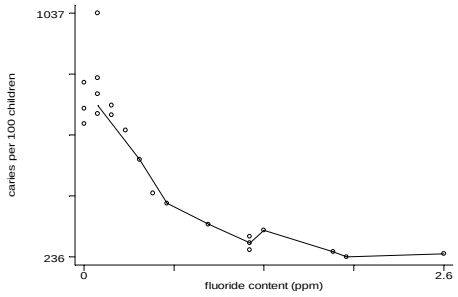


先と同じく２SE の範囲でグラフを描いてみましょう。

. predict s, stdr
. gen lo = hat-2*s
. gen hi = hat+2*s
. graph caries hat hi lo fluoride, c(.lll) s(Oiii) sort ylab xlab

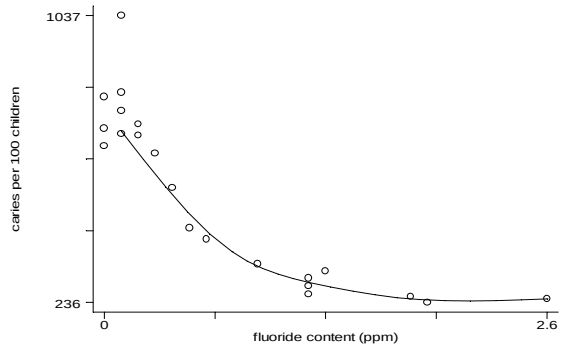


ｘ軸を例えば 10 等分して、その中での中央値を計算し結んでみます。この方が outliers による影響を少なくすることができます。
. graph caries　fluoride, c(m) bands(10)



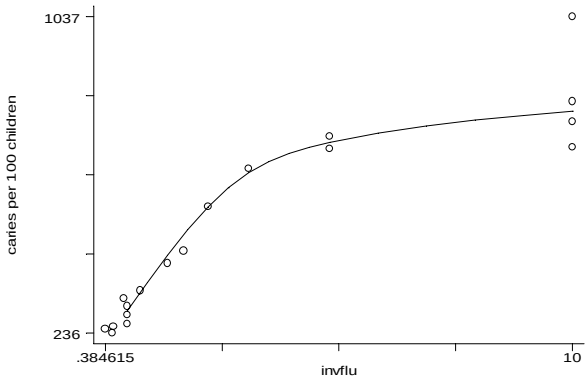
しかしかえって不自然な感じもします。

. graph caries fluoride, c(s) bands(5)



大分いいかんじです。この研究に関しては直線で示すよりはこの方が良いかもしれません。中学生のとき数学で習った y= 1/x に似ています。

. gen invflu = 1/fluoride
(3 missing values generated)
. graph caries invflu, c(s) bands(5)



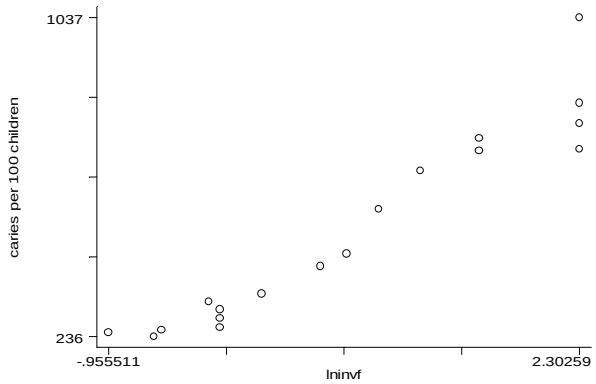
直線というよりは log の関数グラフに近くなってしまいました。

. gen lninvf = ln(invflu)
(3 missing values generated)

. graph caries lninvf

. graph caries lninvf, c(.l) s(Oi) sort

```
. regress caries lninvf

  Source |       SS       df       MS              Number of obs =      18
---------+------------------------------           F(  1,    16) =  175.81
   Model |  951786.782     1  951786.782            Prob > F      =  0.0000
Residual |  86621.6625    16  5413.85391            R-squared     =  0.9166
---------+------------------------------           Adj R-squared =  0.9114
   Total |  1038408.44    17  61082.8497            Root MSE      =  73.579

------------------------------------------------------------------------------
  caries |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  lninvf |   207.6445   15.66044    13.259   0.000      174.4459    240.8432
   _cons |    356.044   20.64066    17.250   0.000      312.2877    399.8002
------------------------------------------------------------------------------
```



大分直線に近くなったと思いませんか？transformation technique を用いて少し遊んでみました。