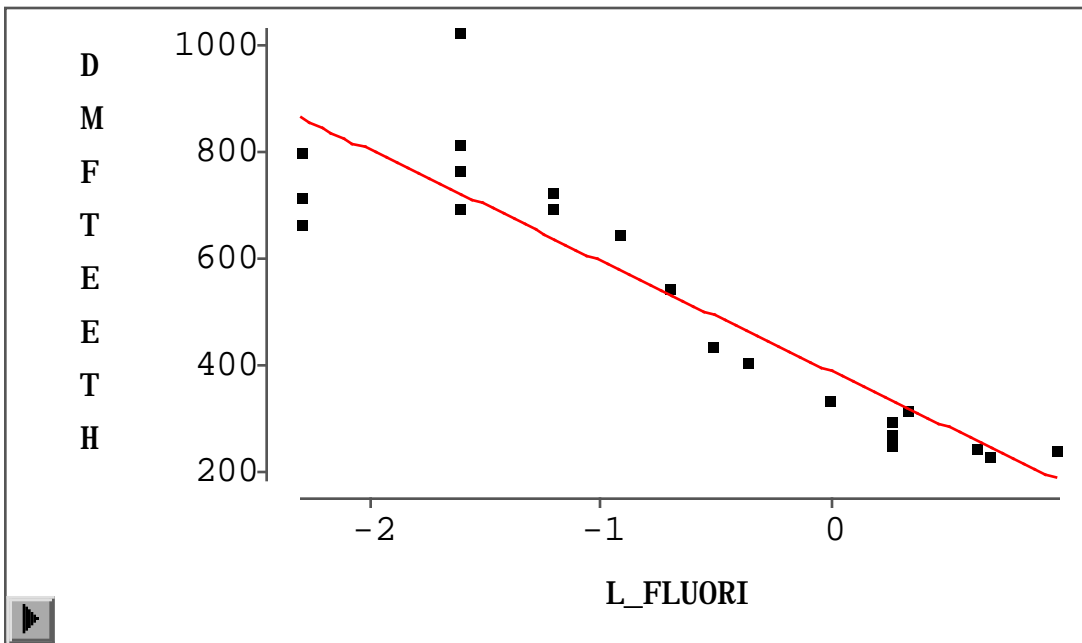


1 • Fit the relation $E[\text{DMF Teeth}] = \beta_0 + \beta_1 \log[\text{Fluoride}+0.1]$

▶	DMFTEETH	=	L_FLUORI
	Response Distribution:		Normal
	Link Function:		Identity

Notice the 3, 4, 2 and 3 "repeats", yielding $2+3+1+2 = 8\text{df}$ to estimate pure error (by pooling the squared deviations from the 4 corresponding means).




▶ Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	1	971816.253	971816.253	86.7218	0.0001
Error	19	212916.699	11206.1421		
C Total	20	1184732.95			

- Use the "pure error", estimable from the "repetitions at same fluoride level", to test the goodness of fit of this model.

Set up $\ln(\text{Fluoride}+0.1)$... called L_FLUORI by INSIGHT .. as a nominal variable with as many levels as there are unique values.

Fit model with these as the "X" (in fact the model now has 13 indicator variables, 1 of them redundant since the $X_0=1$ is used for the intercept)

 DMFTEETH	=	L_FLUORI
Response Distribution:		Normal
Link Function:		Identity

Nominal Variable Information	
Level	L_FLUORI
1	0.0000
2	0.2624
3	0.3365
4	0.6419
5	0.6931
6	0.9933
7	-0.3567
8	-0.5108
9	-0.6931
10	-0.9163
11	-1.2040
12	-1.6094
13	-2.3026

The 13 categories can be represented by an intercept term ($X_0=1$) and any other 12 indicators i.e., 1 indicator of the 13 will be redundant, and the category made redundant will be the "reference" category. The intercept will be the average Y level in the reference category and the other beta_hats will be the difference between the mean in those other categories and the average in the reference category. [If you had specified no intercept, then the 13 beta_hats will be directly interpretable.. as the mean Y level in each of the 13 categories]

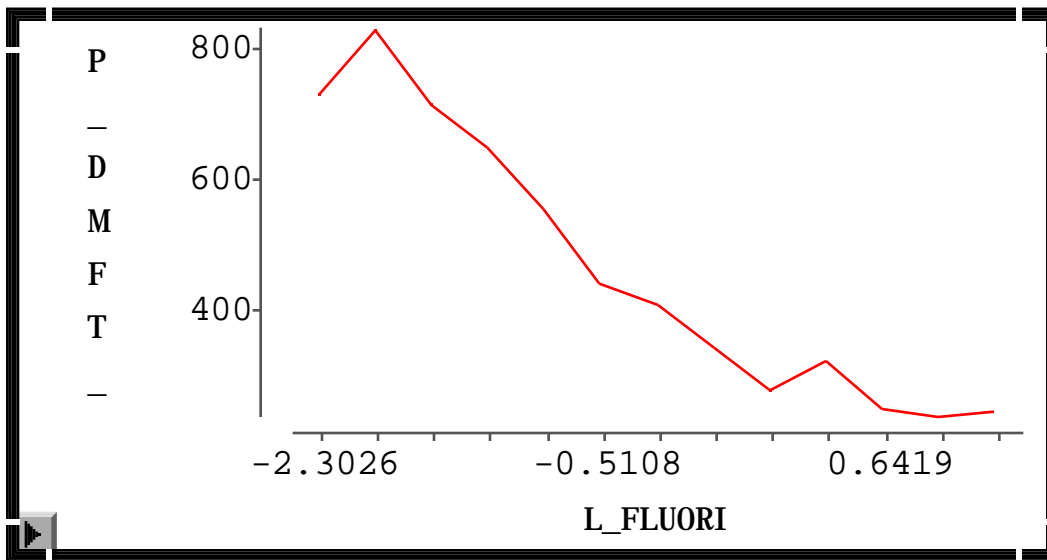
Notice below that INSIGHT sets the 1st fluoride category as the reference category (sets its beta to zero).

Parameter Information		
Parameter	Variable	L_FLUORI
1.0	INTERCEPT	
2.0	L_FLUORI	0.0000
3.0		0.2624
4.0		0.3365
5.0		0.6419
6.0		0.6931
7.0		0.9933
8.0		-0.3567
9.0		-0.5108
10.0		-0.6931
11.0		-0.9163
12.0		-1.2040
13.0		-1.6094
14.0		-2.3026

Model Equation											
DMFTEETH	=	735.000	-	392.000	P_2	-	454.333	P_3	-	412.000	
	-	483.000	P_5	-	499.000	P_6	-	489.000	P_7	-	323.000
	-	291.000	P_9	-	179.000	P_10	-	83.0000	P_11	-	17.0000
	+	99.5000	P_13								

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	12	1112075.29	92672.9405	10.2038	0.0014
Error	8	72657.6667	9082.2083		
C Total	20	1184732.95			

The model takes some 12 additional terms (beyond the intercept) to fit. You can plot the fitted model by plotting the fitted values against the fluoride values



All this is a "join the actual means" plot.. some of the means are based on 1 observation, four of them (as we remarked above, are based on 3, 4, 2 and 3 points. In a sense, the 13 means become the fitted model. The whole point now is whether the pattern of the 13 means could be described by a simpler (more economical) straight line model. Visually, the flexible model with 13 parameters looks almost like a straight line, so are we wasting parameters trying to accommodate all the twists and turns, which may be nothing more than noise?

Obviously, a model with 13 (12 + intercept) parameters will have a low SS_{residual} , but how much lower than the model with only 2 parameters?

	Model	SS_{residual}	
linear	2 parameters	213K	
flexible	<u>13 parameters</u>	<u>73K</u>	
difference	11 parameters	140K (lack of Fit)	ie. 12.7K per parameter

What is the Noise level per df?

8 df 73K i.e. 9K pure error

$$\text{Ratio } 12.7/9 = 1.4 < F_{11,8,95} = 3.3$$

This is a prototype for testing a "bigger" versus "smaller" model later in multiple regression.

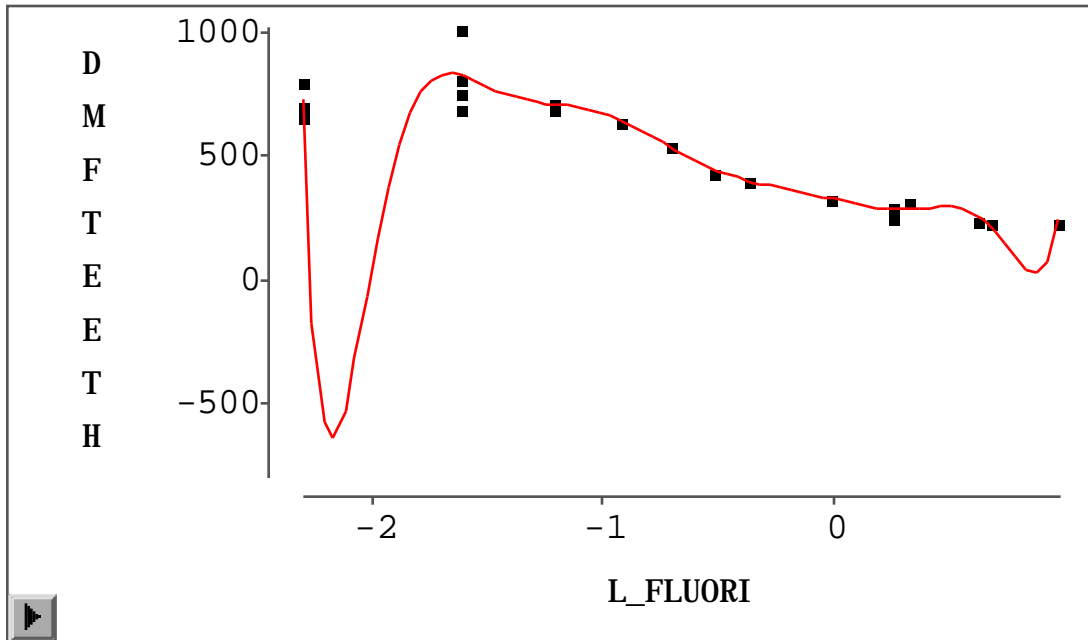
Note re Lack of Fit Test

The textbook uses "c" to denote the number of distinct values of X, but this definition is hidden in the middle of a paragraph. A few of you took the definition the wrong way.

*** Out of curiosity... what if we naively fitted the highest possible polynomial to these data...?
 Why do we get the dips"?

Because we are chasing noise..

ONE SHOULDN'T TRY TO PUT A CURVE EXACTLY THROUGH AS MANY POINTS AS POSSIBLE .. ESPECIALLY IF THE POINTS CONTAIN ERROR. AFTER ALL, THE DATAPOINTS HERE ARE BASED ON SAMPLES (FINITE NUMBERS) OF CHILDREN, AND SUBJECT TO SAMPLING VARIABILITY, OBSERVER VARIABILITY, OTHER UNMEASURED FACTORS ETC.. SO IT IS SILLY TO TRY TO HAVE THE CURVE "REPRODUCE" ALL THESE UNEXPLAINED VARIATIONS.



Parametric Regression Fit						
Curve	Degree(Polynomi al)	Model		Error		R-S
		DF	Mean Square	DF	Mean Square	
	12 <input type="text" value="12"/>	10	111005.471	10	7467.8237	0

2 Is caffeine "cleared" faster from smokers than non-smokers?

For each of your six subjects ...

- Obtain the estimated slope b_1 and its associated SE

$$\{ \text{model: } E[\log[\text{caffeine}]] = \beta_0 + \beta_1(\text{Time elapsed}) \}$$

DON'T FORGET TO REMOVE THE EARLY TIMEPOINTS!

- Rank the 6 b_1 's according to their precision (SE)
- Say why you think the estimates rank in this order

(base your explanation on visual inspection of 6 plots of $\log[\text{caffeine}]$ vs. Time)

The SE for a fitted slope is (when n and n-1 are close)

$$SE(b_1) = \frac{RMSE}{\sqrt{n} SD(X)}$$

so all 3 factors contribute...

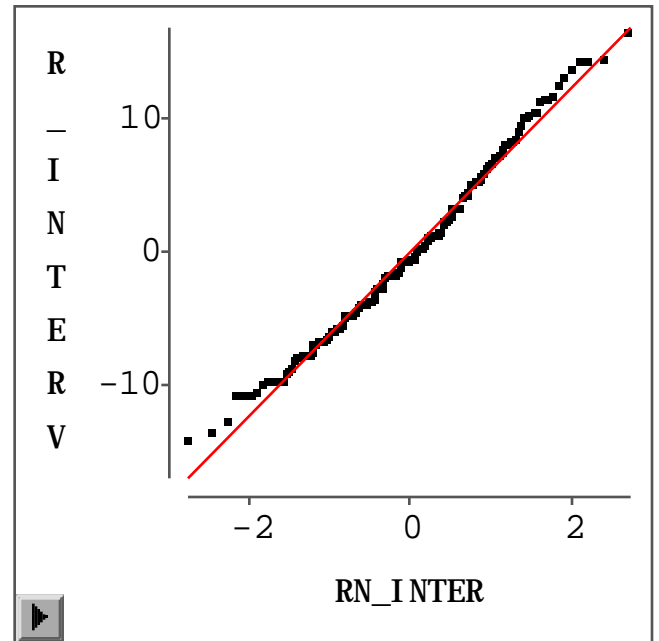
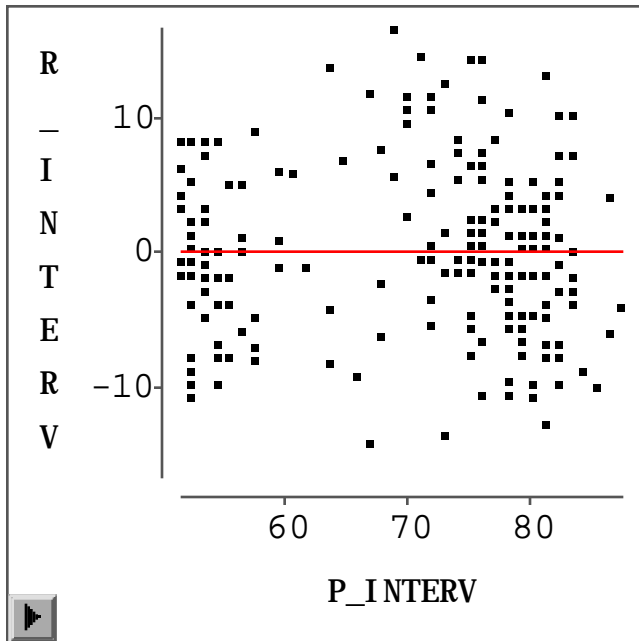
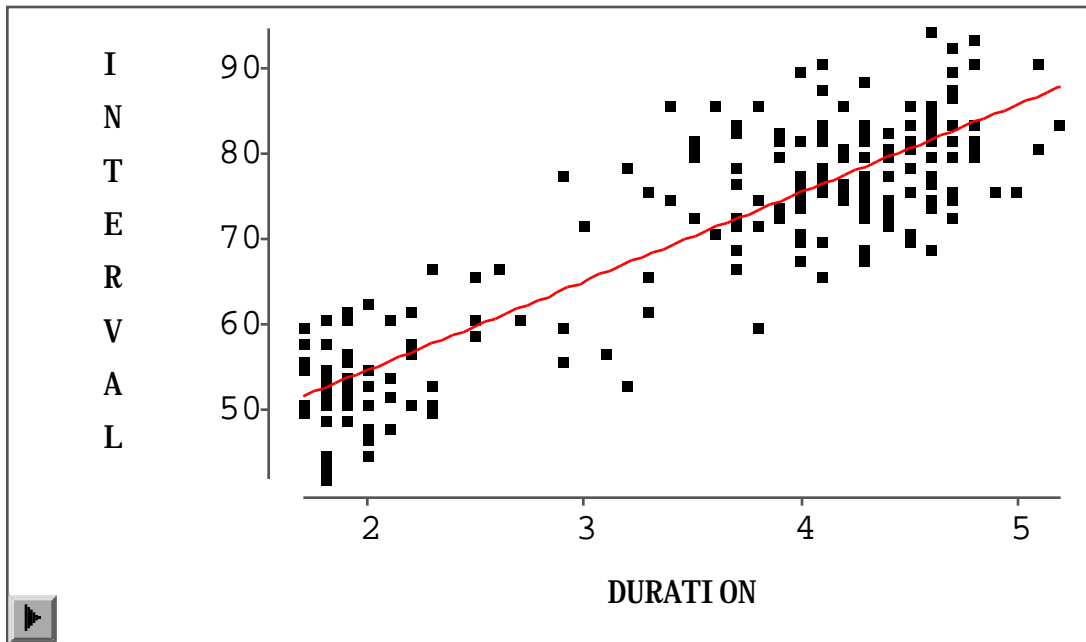
- **the more datapoints the better,**
- **the more spread out they are on the X axis the better, and**
- **the smaller the RMSE the better.**

3 Using Simple Linear Regression for Prediction: How Faithful was Old Faithful?

(Question 3 of homework due Friday June 9, 2000 in Course [678](#) this past June)

Some of you thought that the pattern violated the requirements for Simple Linear Regression, and in particular those for the Gaussian-Errors Model. In fact, apart from a slight tendency for the residuals to be slightly larger for the longer intervals associated with long durations (cf. non-homogeneity of variance in residual plot below), there isn't any serious issue (residual normal QQ plot (right, below) is quite "Gaussian")

Don't be worried by the "bi-modal distribution of "X" values. Remember that there is NO REQUIREMENT about the distribution of the X's. In this case, if one is going to use the data to make predictions about individual intervals, the (un)equal-variance Gaussian's would be relevant. One way to accommodate these (without a transformation) would be to model the variance as an increasing function of the mean-- thereby preserving the linearity of the regression function.



By the way, the textbook asked in an exercise in an earlier chapter: what if one had very little data near a certain X value, and a lot more data elsewhere, for inferences near this X would one rely on the few data or "borrow strength" from the data from elsewhere. this is a good example of the same question. near durations of 3 minutes, it might be better to go with interpolation than with the little "on-the-spot" data.

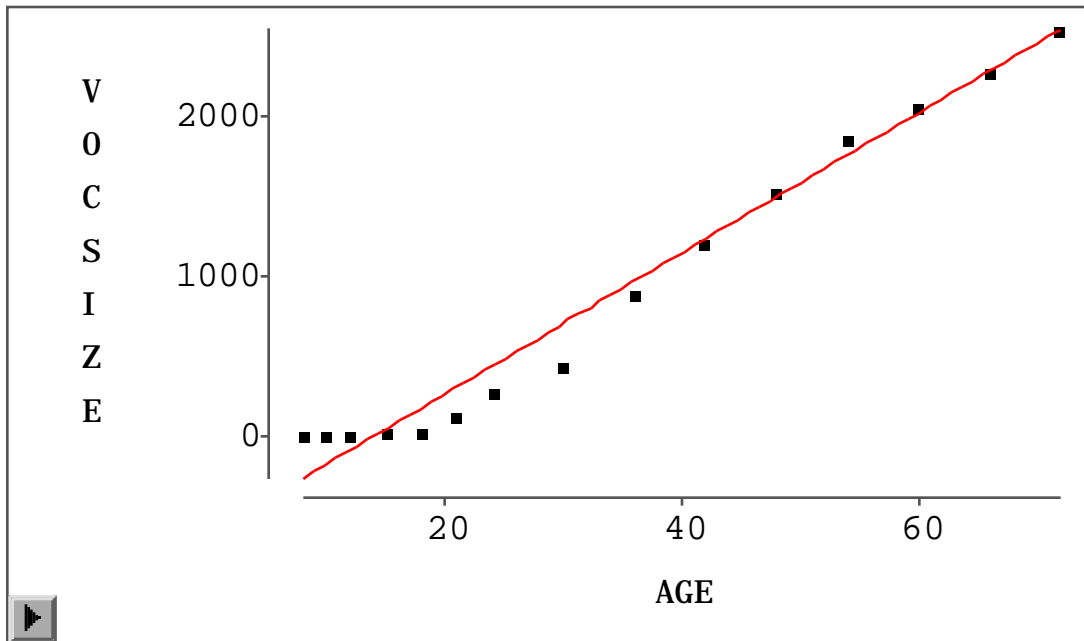
A propos the bimodal "X" distribution: how well would one do if one simplified the "X" to "X*" where $X^* = 1$ if $X >$ say 3.5 minutes, and $X^*=0$ otherwise. You might be surprised that the R^2 from this simplified X is not that much worse than the one based on the "full-grained" X.

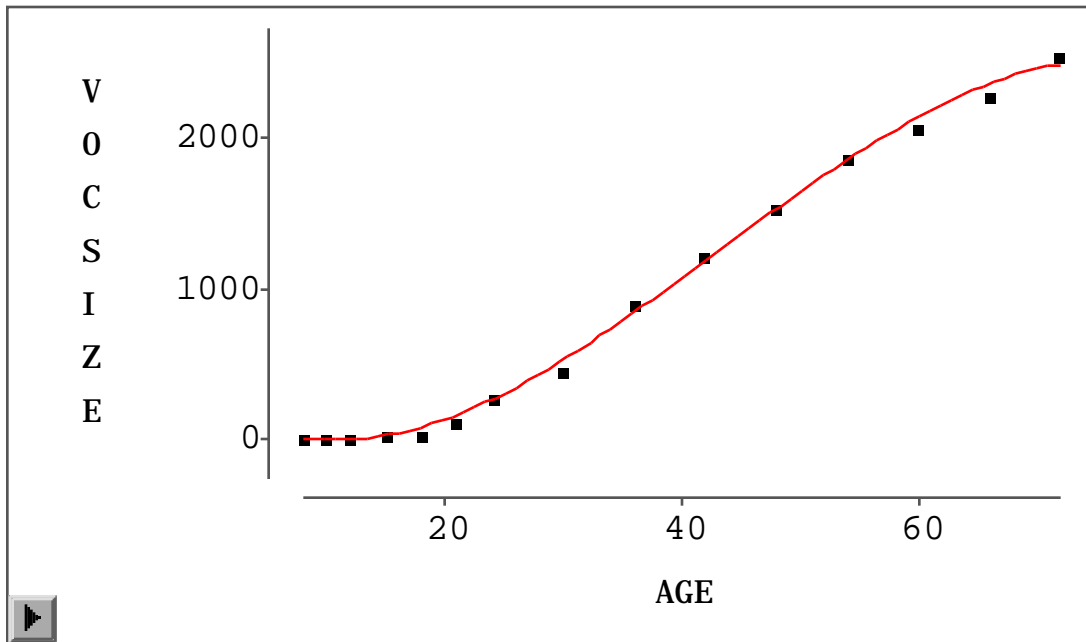
4 Using Non-linear relationships - vocabulary data for 1 child

(Question 6 of homework due Friday June 9, 2000 in Course 678 this past june)

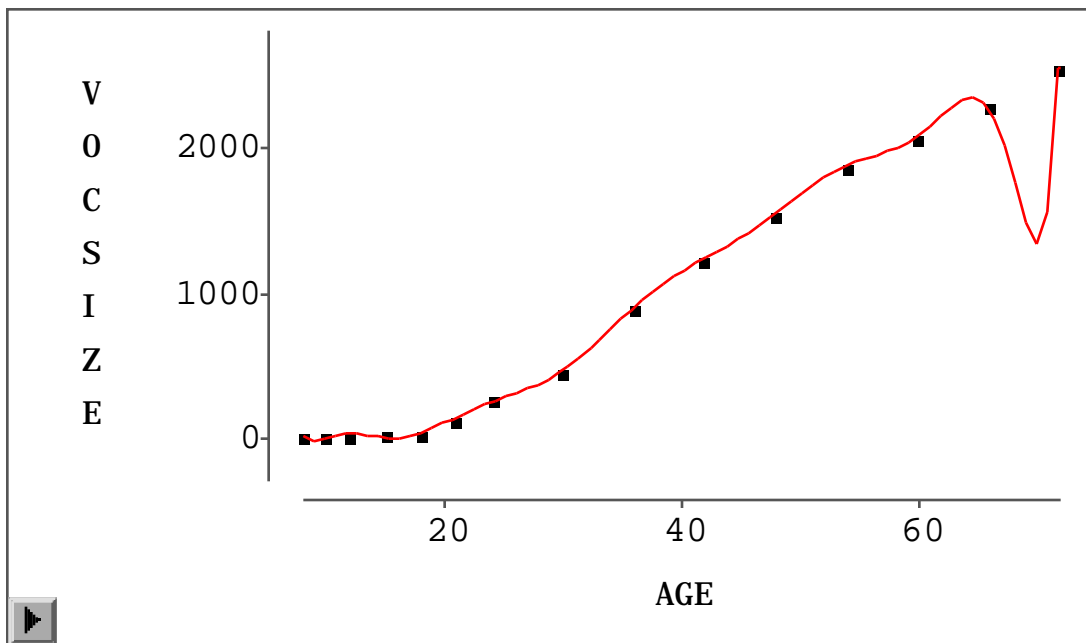
- a Does a linear growth model fit the data? [In their q's, KKMN ask if the point "0.0 years, 0 words" is on the line of vocabulary growth; they also ask that one add this point and re-fit!]
- b. Suggest and fit a better model.

It doesn't do very well early on. The possible parametric fit might be an S-shape, but even then it will be difficult to have it fit well everywhere. A polynomial of order 3 or 4 might have enough flexibility. I wouldn't go too much higher. And don't expect any such curve will extrapolate well (as the child starts school).





Don't be tempted to overfit, as here.



The vocabulary "dips" between ages 5 and 6 because of **overfitting**.. i.e. chasing error! None of the vocabulary counts is exact -- they are surely estimates, measured with some error. How can one say with any confidence that at age 5 the child has a vocabulary of EXACTLY 2072 words? It is thus a mistake to try to get the curve to go exactly through them all. Indeed, if each of the data points showed the margin of error in the measurement (and thus a CI for the true vocabulary at each age), one would

no longer wish to put the curve through all of the point estimates. One would simply hope that the curve went through most of the confidence intervals. As I have said on other occasions, the trouble with 300 dot per inch printers is that they give a false sense of precision to each datapoint. Understanding would be better served if one had put the data "points" on the graph with a paintbrush!

Another way to think of it: If one has no df remaining to assess error, how can one make a confidence band for the curve?

The investigator who fitted the highest possible degree polynomial to 10 daily WBC's (White Blood Counts)! and then wondered why the fitted curve for the patient's next value was "heading skywards" at a slope of near infinity, fell into the same trap. [this was back in the early 1980s when Lotus 123 first came out, and investigators were very impressed that it could fit very high degree polynomial curves]. This particular investigator was suspicious enough to trust his medical judgment and to ask what was the statistical reason for the "skywards pointing" curve. I asked him why all of the blood counts, 5600, 5900, 7200, 6600, etc. ended in 00. He answered that they are just estimates, since (if done manually) they are based on counting sufficient microscope fields to make a reasonable estimate. The ending 00's signify that the true WBC is *somewhere in this ballpark*. So again, it makes no sense to force the curve through them.

A few years ago, during an election, the Montreal Gazette reported that a study had shown that 1304 patients (I think that was the number) had gotten in medical trouble because of restrictions made in Québec's prescription insurance plan. When asked for reaction, the politician responded with a smart question: "I haven't seen the list of 1304 patients; can you get me the list?" The newspaper was then forced to concede that the estimate was based on multiplying another number in the study (itself a regression coefficient estimated from a sample, *but presented without a confidence interval*) by the total population size. The newspaper could easily have given the correct impression as to the (synthetic or model-based) nature of their figure of 1304 if it had instead reported "an estimated 1304 patients" or even "approximately 1000 patients".

d "The data appear to be from one child. If this is true, what assumption of the least-squares approach is most likely violated, and why?"

KKMN are probably getting at the INDEPENDENCE of the "error" components in the model.

If these were 15 observations from 15 DIFFERENT children, and one had a good model for the expected (MEAN) values, this clearly would not be a problem. But the only purpose here is to describe the progression of THIS ONE child, and to use the model for interpolation FOR THIS ONE CHILD.

Question: Do you think that the serial "errors" or residuals would be correlated? i.e., is it likely that observations on one side of the curve would tend to be immediately followed by observations on the same side of the curve (positive serial correlation) or by ones on the opposite side (negative serial correlation)? Why?

Given all the forces at play, I don't expect a reasonable low-order curve to be a good enough fit that the residuals are perfectly randomly distributed around it, with absolutely no ("serial") correlation between adjacent residuals. A child is likely to have small extra spurts and slight slowdowns because of increases in activity (say in daycare or because (s)he is no longer preoccupied with learning to walk, illnesses, holidays, local but corrected drifts in errors of measurement etc. If one looked at deviations from a trend of someone who put on weight gradually over time, one would find some "local" small cycles that tended to produce some runs of residuals on the same side of the trendline or trendcurve (could the same be the case with global warming trends?).