Estimation of Exposure -Time Trends Using Nonlinear Segmented Models

Linda Levin, PhD
Associate Professor
University of Cincinnati

University of Cincinnati

Purpose of Analyses

- Reconstruct occupational exposure levels
- Estimate the impact of exposure on workers' health

Data

- IH samples- airborne fiber levels
- 1972-1994
- Samples identified by job and year of sampling
- Number of samples per year varied

Why Model Exposure Continuously?

- Insufficient data to calculate estimates of mean exposure each year
- Interpolation between data-rich years unreliable
- ‘Bumpy’ lines
- Exposures known to decrease with time

Preliminary Investigations of Exposure Trends -- LOESS Method

- A non-parametric method for estimating local regression
- Useful for exploring the parametric form of a regression curve which is unknown
- Assumes the regression curve can be *locally* approximated by values of a parametric function of the independent variable x
- Uses weighted least squares
- Fits linear or quadratic functions of x in neighborhoods of x
- **Linear** functions are default method

LOESS Method (Cont)

- Smoothing parameter
- Determines number of points used in local fitting
- Two types of fitting
- Direct: fitting done at each data point
- Computationally intensive
- KD trees (default): points selected for fitting.
- Results are then ‘blended’ linearly or quadratically for observed data points

LOESS Method (Cont)

- Strategies for choosing smoothing parameter
-
- Graphing: Residuals vs predictor variable
-

Look for lack of structure

or

Automatic method

Example: Minimization of Akaike Information Criteria

= log (residual SS) + f(smoothing parameter)

f decreases as smoothness increases

EXAMPLE of SAS CODE

- Estimate Changes in Sample Concentrations (Exposure) from 1972 to 1994
- PROC LOESS
- Fiber= Dependent variable (Exposure)   dt = Independent variable (Sample Date)
- N=170

```
proc loess;
model fiber=dt/details (modelsummary);
run;
```

Note:

Sample date was transformed using 1/1/1970 as an arbitrary frame of reference

Facilitated model convergence

dt = years (1/1/1970 to each sample date) / years (1/1/1970 to first sample date)

First sample data=5/30/1972 ; Last sample data=9/29/1994

Max value of dt =10.3 , Min=1.0

SAS OUTPUT

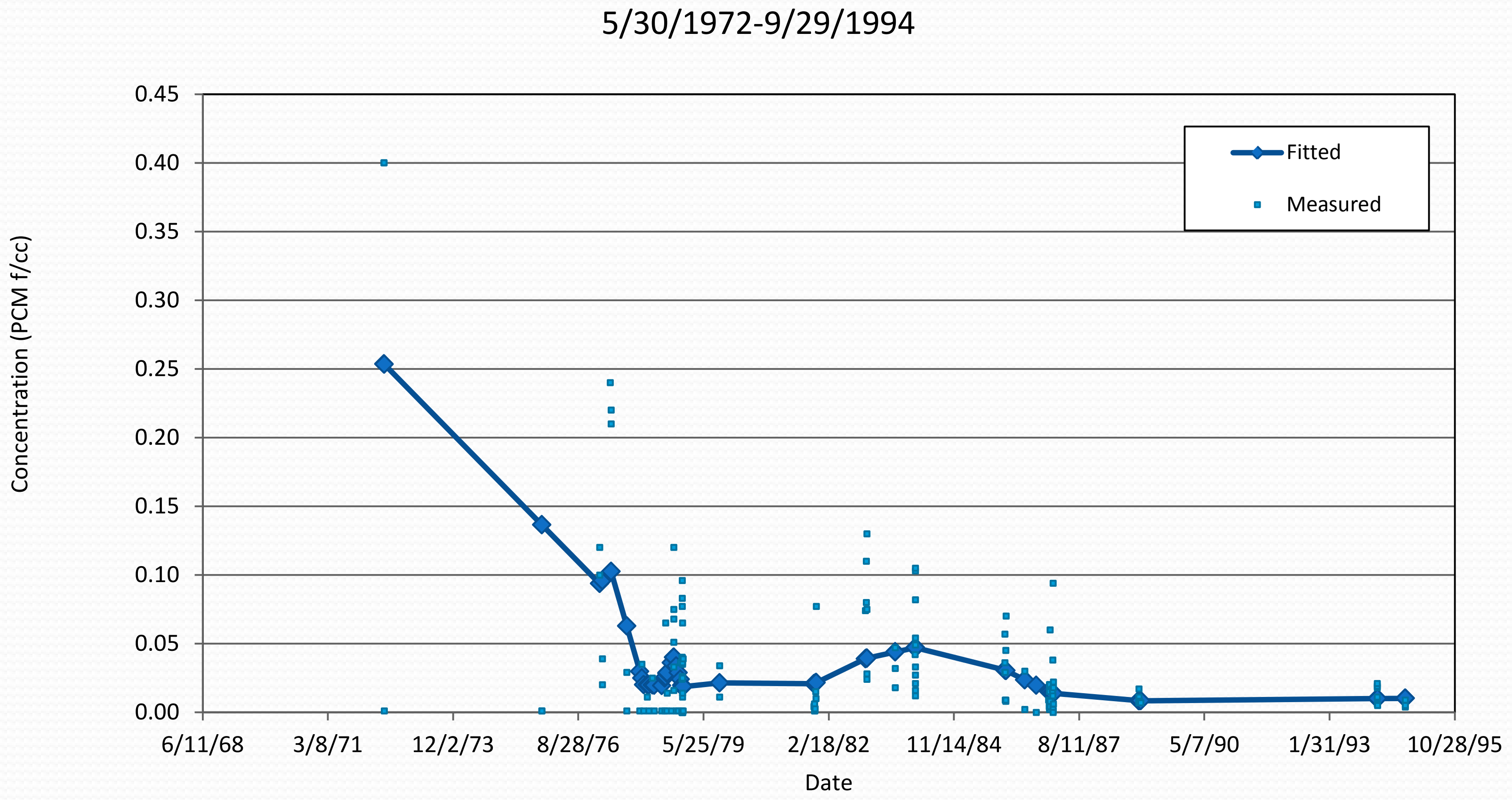
The LOESS Procedure
Dependent Variable: fiber

Smoothing Parameter	Local Points	Residual SS	AICC
0.39118	66	0.32735	-5.14977
0.63235	107	0.36350	-5.08624
0.25000	42	0.29679	-5.18578
0.16176	27	0.28728	-5.11985
0.30882	52	0.30105	-5.18782
0.33824	57	0.32212	-5.15355
0.27941	47	0.29770	-5.19214
0.26765	45	0.29660	-5.19154
0.29706	50	0.30035	-5.18828
0.27353	46	0.29653	-5.19344
0.26765	45	0.29660	-5.19154
0.27353	46	0.29653	-5.19344

Optimal Smoothing Criterion

AICC	Smoothing Parameter
-5.19344	0.27353

Figure1: LOESS Graph of Fiber Data for Grouped Jobs



Results of Job-Specific Exploratory Analyses

- Smoothness of curves varied by job
- Variations in exposure levels *and a*mount of data
- Between-sample variances increased as yearly exposure means increased

Results of Job-Specific Exploratory Analyses (Cont)

- Verified decrease in exposure over time
- Steeper in mid 1970s
- Less decline in later years

Conclusion

<u>Exponential models are a reasonable parametric form to model exposure trends over time</u>

Nonlinear Exponential Regression Model For Mean Exposure

- Dependent variable C(t) = fiber concentration at time t
- $C(t) = \mu(t) + e_t$
- $\mu(t)$ = mean of C(t) at time t
- $\mu(t)$ = two parameter exponential function of t
- e_t = normally distributed error term with mean 0
- Time t coded as number of years from 1/1/1970 to sample date

Two Parameter Exponential Model For Mean Exposure

- $C(t)$ = fiber concentration at time t
- $C(t) = \mu(t) + e_t$
- $\mu(t) = a \cdot \exp(-b \cdot t)$
- a>o intercept parameter; b>o slope parameter
- a and b expressed as exponential functions to guarantee positivity of $\mu(t)$
- $a = \exp(a_o)$ $b = \exp(b_o)$

How to Describe the **Variability** of Fiber Concentrations ?

- Define the relation between exposure variance and mean exposure at each year

by

'Power of the mean' variance function

Commonly used in nonlinear regression

$$Var\{C(t)\} = \sigma^2 \cdot \mu(t)^\theta$$

- θ = variance parameter determined from the data
- σ^2 = scale parameter describing precision of C(t) (Similar to σ^2 in ordinary regression)
- Consistently achieved model convergence

Implementation of Exponential Regression Analyses

- PROC NLIN
- SAS for Windows, Version 9.3

Estimation of Parameters of Mean Value Function

$$\mu(t) = a \cdot \exp(-b \cdot t)$$

IRWLS --- Iteratively reweighted least squares

a,b parameters estimated iteratively

Weighted least squares

Weights = inverse of variance function
(mean concentration to the power θ)

Variances updated from a,b estimates

… repeated until convergence

Estimation of Variance Parameter θ

- Initially set = 0
- If convergence not reached ,other values in range (0.1 to 2) manually selected
- Value at convergence identified
- Post hoc sensitivity analyses
- Other values for θ manually selected
- Confirmed convergence for θ~1 for each job

Assess Fit of Exponential Regression Models

- Mean Squared Error = σ^2
- Weighted sum of squared deviations/ df
- $\sum$ (Observed minus mean concentration)^2

  (n- number of parameters)

- Number of parameters= 2 for this model
- Weights = inverse of mean concentration to the power θ at each time.

Nonlinear Fitting Strategy A

- Individual jobs fitted 1972-1994
- Job-specific intercept and slope parameters
- Results unrealistic
- When jobs in the same work area
- Were allowed to have different slopes

Nonlinear Fitting Strategy B

- Area specific: JOINTLY modeled fiber data from all jobs in the same area
- Reasonable to believe similar rates of decline in fiber levels across jobs
- Single slope parameter estimated
- Data of all jobs
- 1972-1994

Nonlinear Fitting Strategy C

Segmented Modeling Approach

- **Area specific: JOINTLY modeled data from jobs in same area**
- **Assumed slopes differed at different time intervals**
- 1972- 1975 1976- 1980 1981-1994
- Job slopes **equal** on each interval
- Intervals determined by documented changes in work environment and worker information

Choosing a Strategy

- Consistency with the impact of engineering controls

- Statistical goodness of fit of the model (MSE)

  - Segmented approach C yielded lower MSE
  - Compared to the un-segmented approach B for all job areas

Note: A two- or three segmented modeling approach was optimum in all job areas

Examples

- Strategy A (program and results shown)
- Strategy B (results not shown)
- Strategy C (program and results shown)

EXAMPLE of SAS CODE- Strategy A

```
proc nlin method=gauss  nohalve;    * turn off step-halving  in  IRWLS;

      parms a =  5.24
            b =  -1;        * Initialize intercept and slope parameters;

       ea=exp(a);eb=exp(b);    *  Model exponential functions of parameters;

       θ= 0.7;     * Set θ at a value that was known to achieve convergence
                     for other jobs with similar variability patterns;

      model fiber= ea* exp(-eb* dt);   * dt is a transformation of sample date;

      fiber2= model.fiber ** θ;   * power of the mean variance function;

      _weight_= 1/fiber2;   * weights used in minimizing SS at each iteration;

       output out=outnlin p=pred  sse=sigma;    * used to graph curve;
run;
```

Output - Strategy A

The NLIN Procedure
Dependent Variable fiber
Method: Gauss-Newton

Iterative Phase

Iter	a	b	Weighted SS
0	5.2400	-1.0000	2043.8
1	4.2400	-1.0021	551.9
2	3.2399	-1.0078	147.1
3	2.2399	-1.0229	38.4241
4	1.2414	-1.0623	10.5923
5	0.2549	-1.1574	4.6150
6	-0.6633	-1.3493	3.7515
. . .			
12	-1.6342	-1.8213	3.3276
13	-1.6343	-1.8214	3.3276

NOTE: Convergence criterion met

Method	Gauss-Newton
Iterations	13
Objective	3.327635
Observations Read	170
Observations Used	170
Observations Missing	0

NOTE: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	Approx F Value	Pr > F
Model	2	2.0196	1.0098	50.98	<.0001
Error	168	3.3276	0.0198		
Uncorrected Total	170	5.3473			

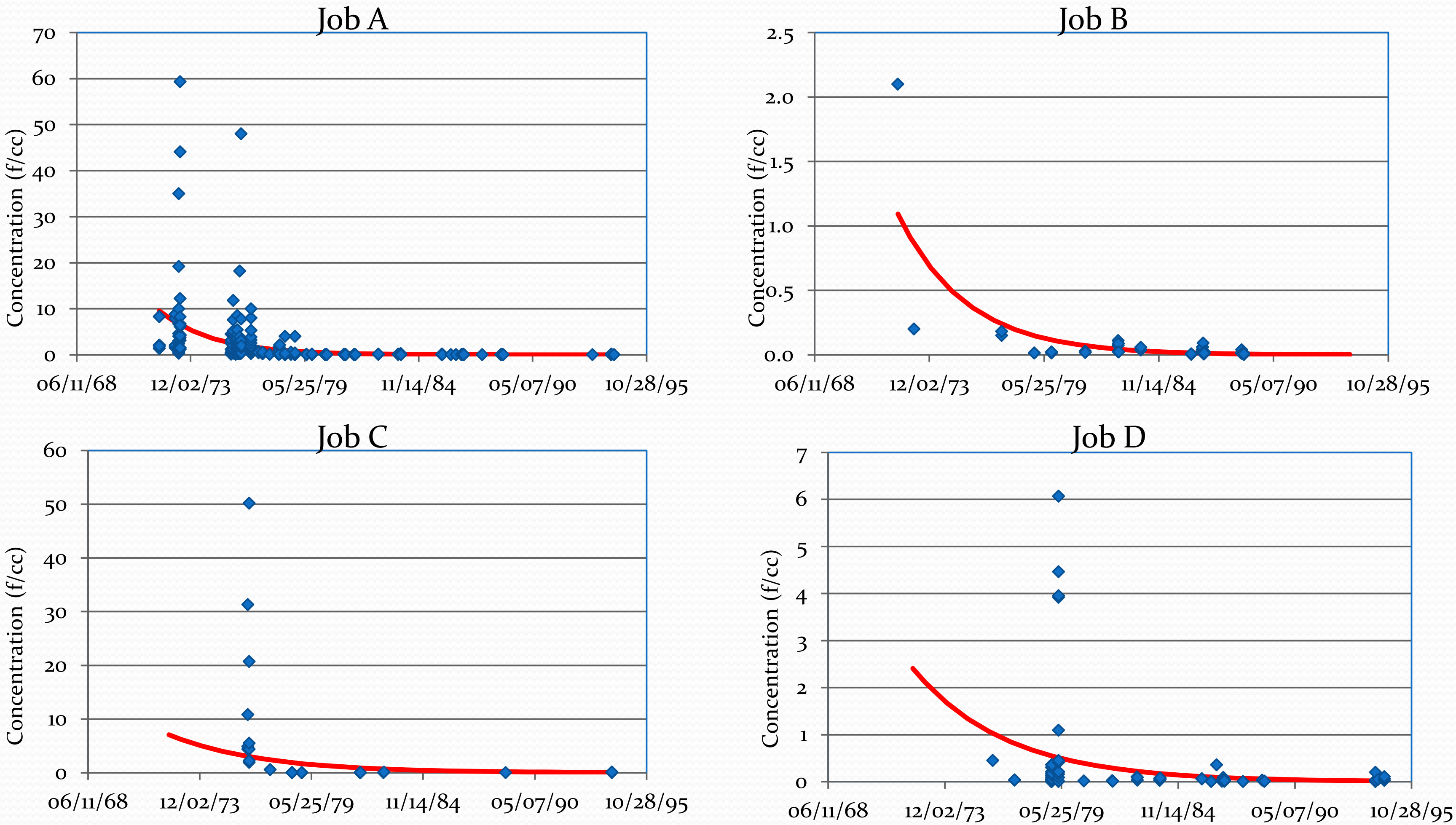
Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
a	-1.6343	0.2604	-2.1484	-1.1201
b	-1.8214	0.1625	-2.1423	-1.5006

From output

Intercept= 0.20 Slope= 0.16

Figure2. Strategy A Exponential Graphs of Jobs Data
A Curve Was Fitted for Each Job Separately

5/30/1972-9/29/1994



Terms for Strategy C (Three-Segment Model)

Example: Four Jobs

Time Intervals

1972- 1975   1976- 1980   1981-1994

Job	Segment 1	Segment 2	Segment 3
1	ea11*exp(eb1*dt)	ea21*exp(eb2*dt)	ea31*exp(eb3*dt)
2	ea12*exp(eb1*dt)	ea22*exp(eb2*dt)	ea32*exp(eb3*dt)
3	ea13*exp(eb1*dt)	ea23*exp(eb2*dt)	ea32*exp(eb3*dt)
4	ea14*exp(eb1*dt)	ea24*exp(eb2*dt)	ea34*exp(eb3*dt)

Estimated fiber values forced (by programming) to be connected at the endpoints of contiguous time segments

Example: Job 1

Intercepts are constrained to be equal

At dt1 = cutpoint between segments 1 and 2

ea21*exp(eb2*dt1) = ea11*exp(eb1*dt)

ea21= ea11 +exp ([eb2-eb1]*dt1)

At dt2= cutpoint between segments 2 and 3

ea31=ea11*exp(dt1*(eb2-eb1))*exp(dt2*(eb3-eb2));

EXAMPLE of SAS CODE- Strategy C

```
******Define indicator variables to estimate intercept and

  slope parameters for each job on each time segment;

data jobs4;set temp;
j=(job=1);  k=(job=2); m=(job=3); n=(job=4);
x=(dt le dt1); y=(dt1 lt dt le dt2); z=(dt gt dt2);
θ= 1;  *model converged for theta =1;
run;

proc nlin method=gauss nohalve;
      bounds a11-a14 <=4;  bounds b3 >=-12 ;

      parms a11=2.5 a12=0.9 a13=2.2 a14=1.7

            b1=-1.6 b2=-1.0 b3=-4;
```

- a11-a14 initial estimates of intercept parameters for each job on first segment;
- b1-b3 initial slopes for each segment;
- Exponentiate parameters;

EXAMPLE of SAS CODE- Strategy C (Cont)

- Intercepts of each job on segments 2 and 3 are interpolated from segment 1 and estimated slopes

```
eb1= exp(b1);eb2=exp(b2);eb3=exp(b3);

ea11=exp(a11);ea12=exp(a12);ea13=exp(a13);ea14=exp(a14);

ea21=ea11*exp(dt1*(eb2-eb1));ea31=ea11*exp(dt1*(eb2-eb1))*exp(dt2*(eb3-eb2));

ea22=ea12*exp(dt1*(eb2-eb1));ea32=ea12*exp(dt1*(eb2-eb1))*exp(dt2*(eb3-eb2));

ea23=ea13*exp(dt1*(eb2-eb1));ea33=ea13*exp(dt1*(eb2-eb1))*exp(dt2*(eb3-eb2));

ea24=ea14*exp(dt1*(eb2-eb1));ea34=ea14*exp(dt1*(eb2-eb1))*exp(dt2*(eb3-eb2));

model fiber= ea11*exp(-eb1*dt)*(j=1 and x=1) + ea21* exp(-eb2*dt)*(j=1
and y=1)+ ea31* exp(-eb3*dt)*(j=1 and z=1)+ ea12*exp(-eb1*dt) * (k=1
and x=1)+ ea22* exp(-eb2*dt)*(k=1 and y=1)+ ea32* exp(-eb3*dt)*(k=1 and
z=1)+ ea13*exp(-eb1*dt) * (m=1 and x=1)+ ea23* exp(-eb2*dt)*(m=1 and
y=1)+ ea33* exp(-eb3*dt)*(m=1 and z=1)+ ea14*exp(-eb1*dt) * (n=1 and
x=1)+ ea24* exp(-eb2*dt)*(n=1 and y=1)+ ea34* exp(-eb3*dt)*(n=1 and z=1);
_weight_= 1/fiber2;
output out=outnlin p=pred sse=sigma;run;
```

Output- Strategy C

The NLIN Procedure
Dependent Variable fiber
Method: Gauss-Newton

Iterative Phase

Iter	a11	a12	a13	a14	b1	b2	Weighted SS
0	2.5000	0.9000	2.2000	1.7000	-1.6000	-1.0000	2437.7
1	2.5560	0.9740	2.2353	1.7120	-1.5226	-0.6964	2432.5
2	2.5654	0.9942	2.2326	1.7009	-1.5154	-0.4729	2463.0
....							
38	2.5159	0.9412	2.1757	1.7393	-1.5977	-0.0781	2707.9

NOTE: Convergence criterion met.

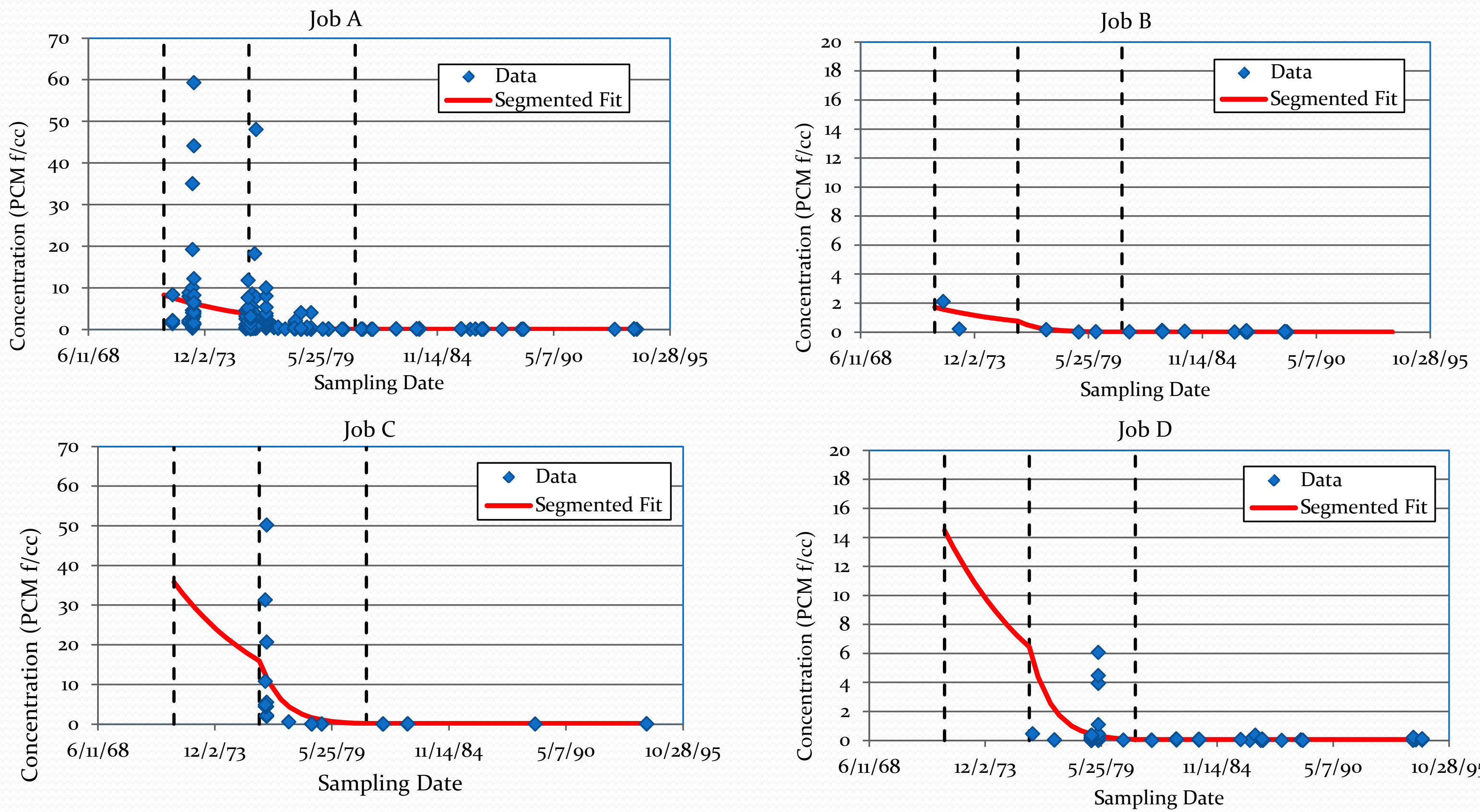
Method	Gauss-Newton
Iterations	38
Observations Read	542
Observations Used	542
Observations Missing	0

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	9	830.8	92.3075	18.17	<.0001
Error	33	2707.9	5.0804		
Uncorrected Total	542	3538.6			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
a11	2.5159	0.3101	1.9067	3.1251
a12	0.9412	1.1805	-1.3777	3.2602
a13	2.1757	0.5095	1.1748	3.1767
a14	1.7393	1.1554	-0.5304	4.0090
b1	-1.5977	0.3246	-2.2354	-0.9600
b2	-0.0781	0.1247	-0.3231	0.1669
Bound0	9.277E-7	0.000227	-0.00044	0.000445

Figure 3. Strategy C Three-Segment Exponential Graphs Curves Fitted Jointly for Each Job

Time segments: 1972-1975, 1976-1980, 1980-1994



Acknowledgement

- Grace LeMasters, PhD
- James Lockey, MD
- Tim Hilbert, MS
- Shu Zheng, MS

- Thanks for providing the data, assisting with graphs and PowerPoint presentation.