

Lecture 13

1. ODS select and ODS output
2. Back-transformations in SAS
3. Confounding
4. Mediation
5. Segmented regression (broken-stick models)

1

Presenting SAS output

Basic:

1. Save output window as file and open this file with MSWord; or copy all and paste into MSWord.
2. In MSWord, select all and change font to Consolas 10pt (or 9pt).
3. Run MSWord macro to fix special characters in tables and printer plots.

Often needs much editing: eliminate extra stuff, add comments, etc.

2

8th grade math scores example:

make table comparing small, medium, large districts.

	Number of Students in District			P-value
	< 500	500 – 1500	> 1500	
Number (<i>n</i>)				
Math Score				
% LEP				
% Free Lunch				
% Special Ed				
% Mobility				
% Drop-out				
8th grader students				

3

```
Proc GLM data=B;
  class district_category;
  model mathscore log_lep log_lunch Special_Ed_pct log_mobility
    = district_category;
  LSmeans district_category/ stderr pdiff CL ;
```

- multiple response variables with same model
- get SEs for variables on original scale, 95% CI for those on log scale
- pdiff *p*-values apply to back-transformed and original comparisons

4

Output:

8 pages of ANOVA tables (need only F -test result for p -value)

10 pages of LSmeans and pdiff

Alternative to editing or deleting parts in MSWord:

ODS SELECT which parts of output are produced

Need SAS Tablenames—names SAS uses for each part of output, given in Details section of SAS Help for each procedure.

5

Details: GLM Procedure

- [Statistical Assumptions for Using PROC GLM](#)
- [Specification of Effects](#)
- [Using PROC GLM Interactively](#)
- [Parameterization of PROC GLM Models](#)
- [Hypothesis Testing in PROC GLM](#)
- [Effect Size Measures for \$F\$ Tests in GLM](#)
- [Absorption](#)
- [Specification of ESTIMATE Expressions](#)
- [Comparing Groups](#)
- [Multivariate Analysis of Variance](#)
- [Repeated Measures Analysis of Variance](#)
- [Random-Effects Analysis](#)
- [Missing Values](#)
- [Computational Resources](#)
- [Computational Method](#)
- [Output Data Sets](#)
- [Displayed Output](#)
- [ODS Table Names](#)
- [ODS Graphics](#)

6

ODS Table Names

PROC GLM assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) create output data sets. These names are listed in [Table 39.8](#). For more information about ODS, see Chapter 20, [Using the Output Delivery System](#).

Table 39.8 ODS Tables Produced by PROC GLM

ODS Table Name	Description	Statement / Option
Aliasing	Type 1,2,3,4 aliasing structure	MODEL / (E1 E2 E3 or E4) and ALIASING
AltErrContrasts	ANOVA table for contrasts with alternative error	CONTRAST / E=
AltErrTests	ANOVA table for tests with alternative error	TEST / E=
Bartlett	Bartlett's homogeneity of variance test	MEANS / HOVTEST=BARTLETT
CLDiffs	Multiple comparisons of pairwise differences	MEANS / CLDIFF or DUNNETT or (Unequal cells and r
CLDiffsInfo	Information for multiple comparisons of pairwise differences	MEANS / CLDIFF or DUNNETT or (Unequal cells and r
CLMeans	Multiple comparisons of means with confidence/comparison interval	MEANS / CLM
CLMeansInfo	Information for multiple comparison of means with confidence/comparison interval	MEANS / CLM
CanAnalysis	Canonical analysis	(MANOVA or REPEATED) / CANONICAL
CanCoef	Canonical coefficients	(MANOVA or REPEATED) / CANONICAL
CanStructure	Canonical structure	(MANOVA or REPEATED) / CANONICAL
CharStruct	Characteristic roots and vectors	(MANOVA / not CANONICAL) or (REPEATED / PRINTF
ClassLevels	Classification variable levels	CLASS statement
ContrastCoef	L matrix for contrast	CONTRAST / EST
Contrasts	ANOVA table for contrasts	CONTRAST statement

Need to look farther down in this list: none of these are LSmeans, SEs, 95% CIs

7

Diff	PDiff matrix of least squares means	LSMEANS / PDIFF=ALL and more than two LS-means
Epsilons	Greenhouse-Geisser and Huynh-Feldt epsilons	REPEATED statement (MANOVA or REPEATED) / PRINTE
ErrorSSCP	Error SSCP matrix	MODEL / (E1 E2 E3 or E4)
EstFunc	Type 1,2,3,4 estimable functions	ESTIMATE statement
Estimates	Estimate statement results	RANDOM statement
ExpectedMeanSquares	Expected mean squares	default
FitStatistics	R-Square, Coeff Var, Root MSE, and dependent mean	
GAliasing	General form of aliasing structure	MODEL / E and ALIASING
GEstFunc	General form of estimable functions	MODEL / E
HOVFTest	Homogeneity of variance ANOVA	MEANS / HOVTEST (MANOVA or REPEATED) / PRINTH
HypothesisSSCP	Hypothesis SSCP matrix	MODEL / INVERSE
InvXPX	inv(X'X) matrix	LSMEANS / CL
LSMeanCL	Confidence interval for LS-means	LSMEANS / E
LSMeanCoef	Coefficients of least squares means	LSMEANS / PDIFF and CL
LSMeanDiffCL	Confidence interval for LS-mean differences	LSMEANS statement
LSMeans	Least squares means	

8

```

Proc GLM data=B;
  class district_category;
  model mathscore log_lep log_lunch Special_Ed_pct log_mobility
    drop_out_pct log_8th_grade operating_budget
    = district_category;
  lsmeans district_category/ stderr pdiff CL;
  ODS select LSMeans Diff LSMeansCL;

```

Only tablenames in ODS Select list are output.

Tablenames for LSmeans, SEs, pdiff output, 95% CIs

Omits 8 pages of ANOVA tables, lists of factor levels, etc.

9

Backtransforming Results in SAS

SAS can only work on variables in a dataset.

Need to get LSmeans and 95% CIs in a dataset to backtransform.

```

Proc GLM data=B;
  class district_category;
  model mathscore log_lep log_lunch Special_Ed_pct log_mobility
    drop_out_pct log_8th_grade operating_budget
    = district_category;
  lsmeans district_category/ stderr pdiff CL;
  ODS select LSMeans Diff;
  ODS output LSMeansCL = logCI;
  ODS output Tablename = dataset-name;

```

10

First step: identify variable names in output dataset:

```
Proc Print data = logCI (obs=6);
```

Obs	Miss Pattern	Effect	Dependent	district_ category
1	1	district_category	mathscore	large (>1500)
2	1	district_category	mathscore	med (500-1500)
3	1	district_category	mathscore	small (<500)
4	1	district_category	log_lep	large (>1500)
5	1	district_category	log_lep	med (500-1500)
6	1	district_category	log_lep	small (<500)

Obs	LowerCL	LSMean	UpperCL
1	51.214431	51.906542	52.598653
2	53.737589	54.388430	55.039271
3	53.675008	54.195767	54.716526
4	0.818231	1.012742	1.207253
5	-0.380983	-0.198070	-0.015157
6	-0.562633	-0.416279	-0.269925

11

Second step: SET logCI in a new data step, calculate back-transformed values

```
Data C;
```

```
set logCI;
```

```
where dependent IN ("log_lep", "log_lunch",  
"log_mobility", "log_8th_grade"); combines multiple OR conditions
```

```
geometric_mean = exp(lsmear); back transform LSmean, CI endpoints
```

```
left_CI = exp(LowerCL);
```

```
right_CI = exp(UpperCL);
```

```
Proc Print data=C;
```

```
var Dependent district_category geometric_mean left_CI right_CI;
```

12

Obs	Dependent	district_ category	geometric_ mean	left_CI	right_CI
1	log_lep	large (>1500)	2.753	2.266	3.344
2	log_lep	med (500-1500)	0.820	0.683	0.985
3	log_lep	small (<500)	0.659	0.570	0.763
4	log_lunch	large (>1500)	21.745	19.340	24.448
5	log_lunch	med (500-1500)	21.533	19.286	24.042
6	log_lunch	small (<500)	28.486	26.082	31.111
7	log_mobility	large (>1500)	11.330	10.231	12.548
8	log_mobility	med (500-1500)	9.474	8.607	10.429
9	log_mobility	small (<500)	8.277	7.665	8.938
10	log_8th_grade	large (>1500)	227.322	199.306	259.275
11	log_8th_grade	med (500-1500)	150.755	133.078	170.779
12	log_8th_grade	small (<500)	51.091	46.253	56.435

Advantage: automatically updates results when data is changed.

13

Confounding

We are interested in the effect of X on Y :

```
Proc GLM;
  model Y = X;
```

We could add another predictor W to get an adjusted effect of X on Y :

```
Proc GLM;
  model Y = X W;
```

What happens when we add W ?

14

$$Y = a_0 + a_1X + e_1 \quad (1)$$

$$Y = b_0 + b_1X + b_2W + e_2 \quad (2)$$

a_1 unadjusted (crude) effect of X from (1)

b_1 adjusted effect of X from (2)

$a_1 \neq b_1$!

adjusted $b_1 < unadjusted a_1$ is common,

adjusted $> unadjusted$ can happen.

15

Confounding

$$Y = b_0 + b_1X + b_2W + e$$

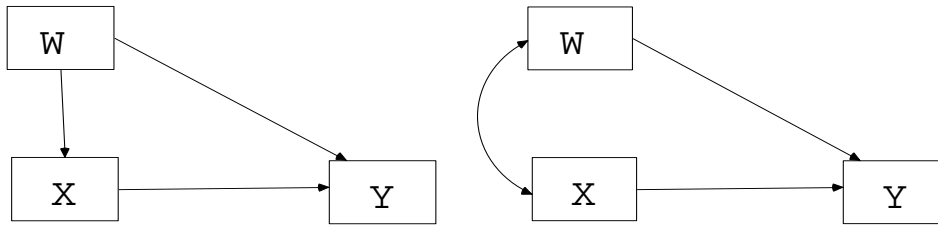
Definition 1. An adjustor W is *confounded* with X , the effect of interest, when W carries all or most of the same information in X about Y .

Definition 2. An adjustor W is *confounded* with X , the effect of interest, under any of these conditions:

- W and X are both causal determinants of Y
- W is a causal determinant of X
- another variable Z is a causal determinant of both X and W

(Def 2 is from Vittinghoff, Shiboski, Glidden, McCulloch *Regression Methods in Biostatistics*, 2005, Chapter 4)

16



$$Y = a_0 + a_1X + e_1 \quad (1)$$

$$Y = b_0 + b_1X + b_2W + e_2 \quad (2)$$

$$W = c_0 + c_1X + e_3 \quad (3)$$

17

Substitute (3) into (2):

$$\begin{aligned} Y &= b_0 + b_1X + b_2W + e_2 \\ &= b_0 + b_1X + b_2(c_0 + c_1X + e_3) + e_2 \\ &= (b_0 + b_2c_0) + (b_1 + b_2c_1)X + e_1 \end{aligned}$$

but this is the unadjusted regression

$$Y = a_0 + a_1X + e_1$$

18

so we have

$$a_1 = b_1 + b_2c_1$$

$$a_1 - b_1 = b_2c_1$$

$$\text{unadjusted effect} - \text{adjusted effect} = b_2c_1$$

$$Y = a_0 + a_1X + e_1 \tag{1}$$

$$Y = b_0 + b_1X + b_2W + e_2 \tag{2}$$

$$W = c_0 + c_1X + e_3 \tag{3}$$

19

Example 1. Buchanan vote in Florida, 2000

```
proc reg data=pred;
  title3 "Percent Bush alone";
  model log_buchanan = p_bush log_votes loghispanic
        income percent_65;
proc reg data=pred;
  title3 "Percent Bush + Percent Gore";
  model log_buchanan = p_bush p_gore log_votes loghispanic
        income percent_65;
proc reg data=pred;
  title3 "Percent Gore";
  model p_gore = p_bush ;
```

20

Percent Bush alone

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-3.70402	0.44582	-8.31	<.0001
p_bush	1	0.01680	0.00507	3.31	0.0016
log_votes	1	0.93020	0.04490	20.72	<.0001
logphhispanic	1	-0.28132	0.04898	-5.74	<.0001
income	1	-0.05496	0.01270	-4.33	<.0001
percent_65	1	-0.01477	0.00675	-2.19	0.0325

Percent Bush + Percent Gore

Variable	DF	Estimate	SE	t Value	Pr > t
Intercept	1	15.24473	7.29139	2.09	0.0409
p_bush	1	-0.17575	0.07412	-2.37	0.0210
p_gore	1	-0.19436	0.07466	-2.60	0.0117
log_votes	1	0.93612	0.04294	21.80	<.0001
logphhispanic	1	-0.25158	0.04815	-5.22	<.0001
income	1	-0.05565	0.01214	-4.59	<.0001
percent_65	1	-0.02031	0.00679	-2.99	0.0040

Percent Gore

Variable	DF	Estimate	SE	t Value	Pr > t
Intercept	1	97.46656	0.40825	238.74	<.0001
p_bush	1	-0.99093	0.00730	-135.70	<.0001

21

$$\text{unadjusted effect} - \text{adjusted effect} = b_2c_1$$

$$\text{unadjusted effect} - \text{adjusted effect} = 0.01680 - (-0.17575) = 0.19255$$

$$b_2c_1 = (-0.19436) * (-0.99093) = 0.1925972$$

22

Dealing with confounding:

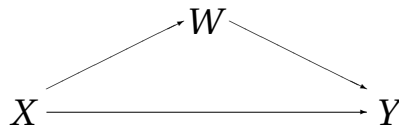
1. Make categories of the confounder, and stratify: look for effect of X when W is fixed
2. Check for interaction between X and W -categories: does effect of X change with category?

23

Mediation

$$Y = b_0 + b_1X + b_2W + e$$

Def (Vittinghoff, *et.al.* p 96) A *mediating variable* W is hypothesized to lie on the causal pathway between X and Y , and thus to mediate the effect of X on Y .



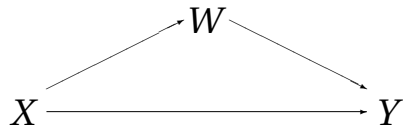
Example:

X = pregnant mother's smoking status; W = gestational age; Y = birth weight.

This definition appeals to external information about the causal relations.

Mediation cannot distinguished from confounding by statistical results only.

24



$$Y = a_0 + a_1X + e_1 \quad (1)$$

$$Y = b_0 + b_1X + b_2W + e_2 \quad (2)$$

$$W = c_0 + c_1X + e_3 \quad (3)$$

Checking whether a suspected mediator W may be acting as a mediator:

- X predicts W , so $c_1 \neq 0$
- W predicts Y , adjusting for X , so $b_2 \neq 0$
- adjusting for W attenuates the regression coefficient of X , so $b_1 < a_1$

25

Mediation terminology

$$Y = a_0 + a_1X + e_1 \quad (1)$$

$$Y = b_0 + b_1X + b_2W + e_2 \quad (2)$$

$$W = c_0 + c_1X + e_3 \quad (3)$$

a_1 **total effect** of X

b_1 **direct effect** of X

$b_2c_1 = a_1 - b_1$ **mediated effect** of X (equality holds for linear regression only)

$(a_1 - b_1)/a_1$ proportion of total effect mediated

26

(Baron-Kenny) Test for mediation

Null hypothesis of test: mediated effect of $X = b_2c_1$ is zero

Estimate of mediated effect: $\hat{b}_2\hat{c}_1$ (fitted regression coefficients)

Standard error of mediated effect:

$$SE(\hat{b}_2\hat{c}_1) = \sqrt{\hat{b}_2^2 SE(\hat{c}_1)^2 + \hat{c}_1^2 SE(\hat{b}_2)^2}$$

Test statistic: $Z = (\hat{b}_2\hat{c}_1)/SE(\hat{b}_2\hat{c}_1)$, compare to standard normal distribution.

27

This test follows the steps in Baron and Kenny (1986) The moderator-mediator variable distinction in social psychology research: conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

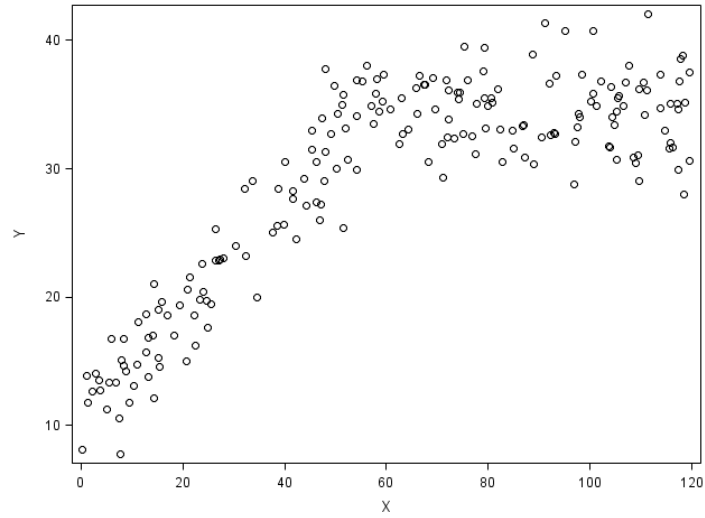
Extensive review:

MacKinnon DP, *Introduction to Statistical Mediation Analysis*, 2008, Psychology Press, Taylor & Francis.

Alternative test: get 95% confidence interval for proportion of total effect mediated $(a_1 - b_1)/a_1$ using bootstrap, see whether CI covers zero.

28

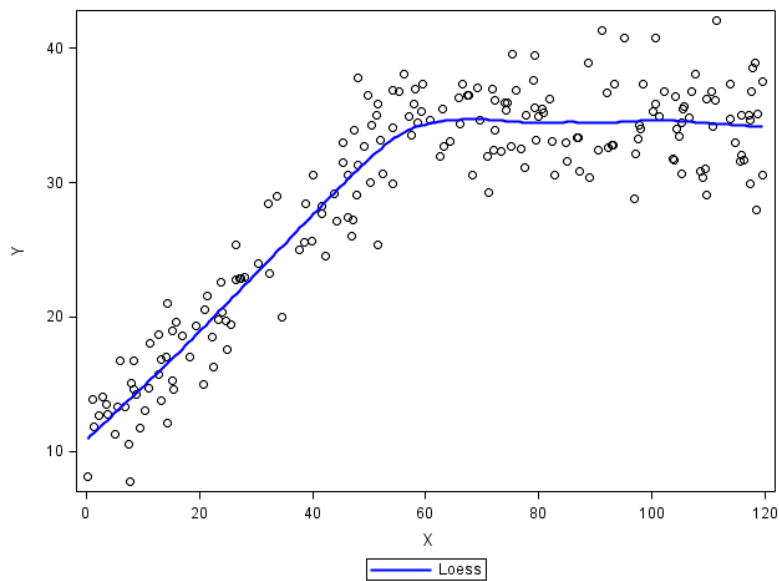
Nonlinear data



Examine shape of mean function by fitting a smoother:

```
Proc SGplot; loess x=X y=Y;
```

29



Fit with $y = \sqrt{x}$?

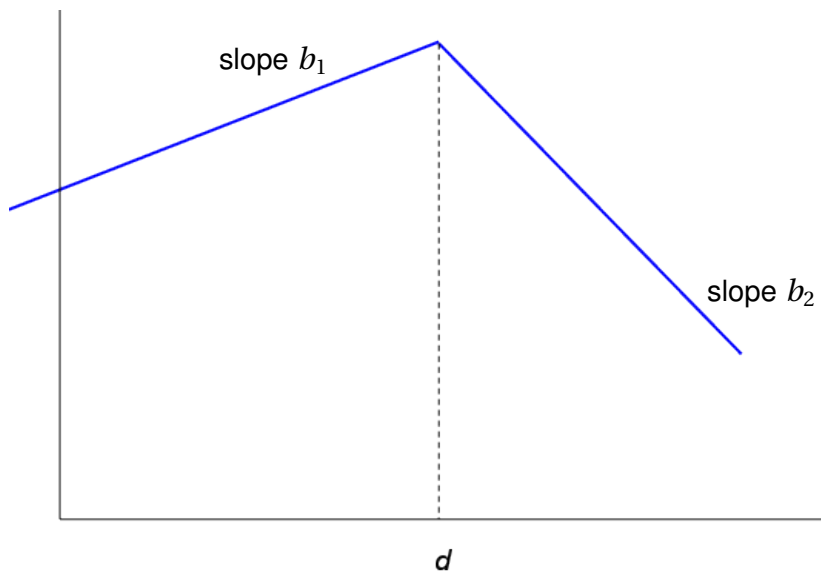
Fit two joined line segments?

30

Segmented regression, or broken-stick models

Fit two or more continuous line segments as a regression model

Advantages: simpler than polynomials, can estimate slopes and break-point(s)



31

Segmented regression model:

$$y = \begin{cases} b_1 + b_2x + \varepsilon & \text{if } x < d \\ b_1 + (b_2 - b_3)d + b_3x + \varepsilon & \text{if } x \geq d \end{cases}$$

with independent errors $\varepsilon \sim \text{Normal}(0, \sigma^2)$

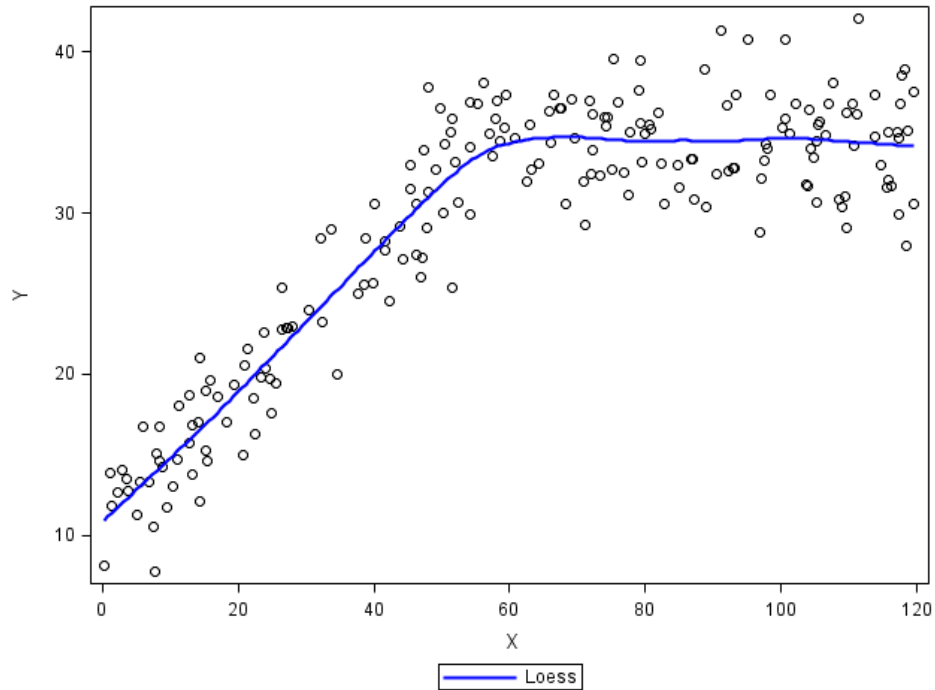
Middle term $(b_2 - b_3)d$ in line 2 joins the two regression lines at $x = d$

Model 1: use guess for breakpoint d

Model 2: estimate d from data

32

Model 1: guess the breakpoint



33

Proc NLin

```
Proc NLin data=ph6470.segmented_reg;  
  parms b1=10.0 b2=0.5 b3=0; starting values for parameters  
  if (X < 60.0) then mean = b1 + b2*X;  
  else mean = b1 + (b2 - b3)*60.0 + b3*X;  
  model Y = mean;  
  output out=B predicted=yhat ;
```

$$y = \begin{cases} b_1 + b_2x + \varepsilon & \text{if } x < d \\ b_1 + (b_2 - b_3)d + b_3x + \varepsilon & \text{if } x \geq d \end{cases}$$

34

The NLIN Procedure
 Dependent Variable Y
 Method: Gauss-Newton

Iterative Phase				
Iter	b1	b2	b3	Sum of Squares
0	10.0000	0.5000	0	5036.4
1	10.7947	0.4117	-0.0248	1525.9

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	12203.4	6101.7	787.78	<.0001
Error	197	1525.9	7.7454		
Corrected Total	199	13729.2			

35

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
b1	10.7947	0.5147	9.7797	11.8098
b2	0.4117	0.0125	0.3871	0.4363
b3	-0.0248	0.0120	-0.0484	-0.00122

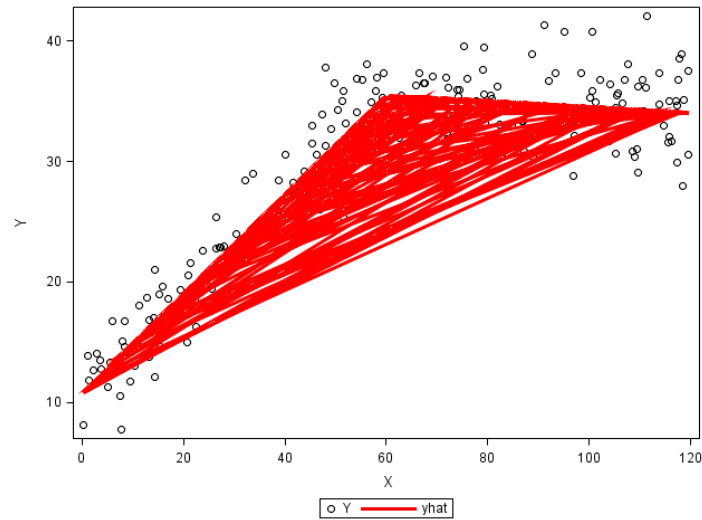
Approximate Correlation Matrix				
	b1	b2	b3	
b1	1.0000000	-0.8702565	0.2738237	
b2	-0.8702565	1.0000000	-0.6001420	
b3	0.2738237	-0.6001420	1.0000000	

Our initial estimates: `parms b1=10.0 b2=0.5 b3=0;`

Use predicted values in output dataset to plot fitted mean.

36

```
Proc SGplot data=A;
  scatter x=X y=Y;
  series x=X y=yhat; connects points with line
```



SAS connects obs1 to obs2 to obs3 ..., so must sort by x first

37

Proc SGPlot allows you to add elements to a scatterplot, as long as each is a column in the data

```
proc sort data=A; by X;
```

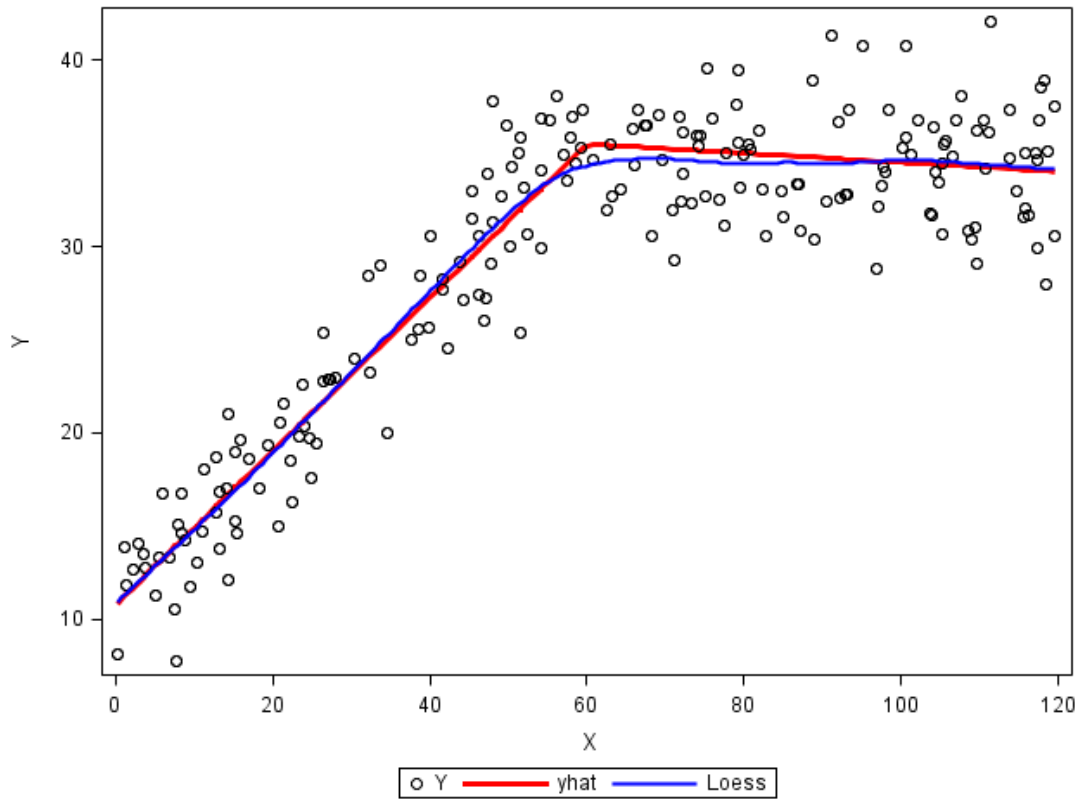
```
Proc SGplot data=A;
```

```
  scatter x=X y=Y; plots points
```

```
  series x=X y=yhat / LINEATTRS = ( color = red thickness=3);
  draws fitted mean by connecting dots at fitted values
```

```
  loess x=X y=Y / LINEATTRS = ( color = blue thickness=3);
  adds smoother
```

38



39

Model 2: estimate break-point

```
Proc NLin data=ph6470.segmented_reg;
  parms b1=10.0 b2=0.5 b3=0 d=60.0; starting values for parameters
  if (X < d) then mean = b1 + b2*X;
  else mean = b1 + (b2 - b3)*d + b3*X; d must appear in 'mean'
  model Y = mean;
  output out=B predicted=yhat ;
```

$$y = \begin{cases} b_1 + b_2x + \varepsilon & \text{if } x < d \\ b_1 + (b_2 - b_3)d + b_3x + \varepsilon & \text{if } x \geq d \end{cases}$$

40

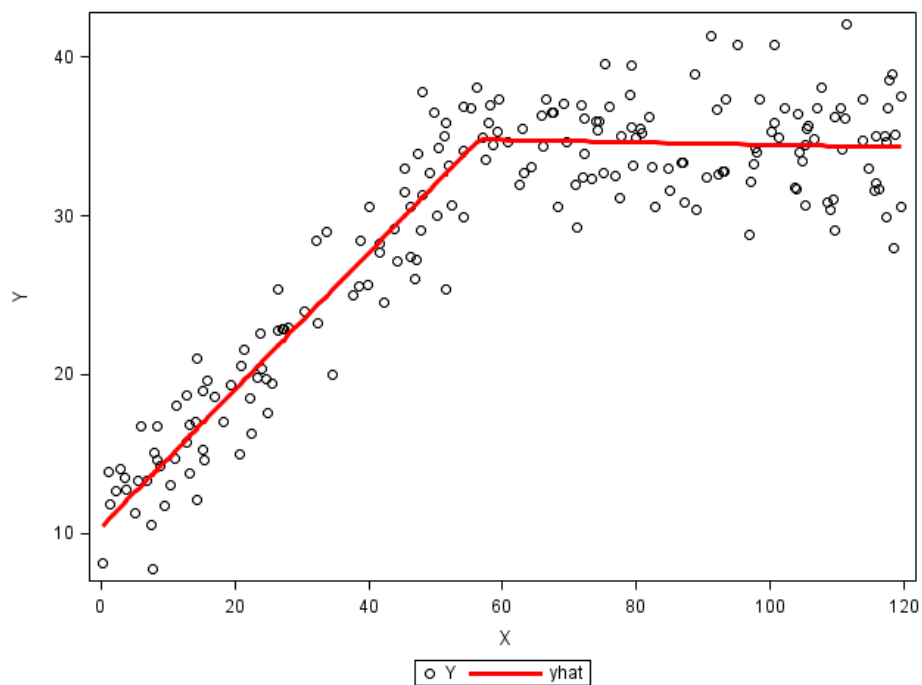
Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
b1	10.4094	0.5536	9.3176	11.5011
b2	0.4322	0.0169	0.3988	0.4656
b3	-0.00781	0.0138	-0.0351	0.0195
d	56.4480	1.7731	52.9511	59.9448

Breakpoint is estimated smaller than 60; final slope b_3 now not significant

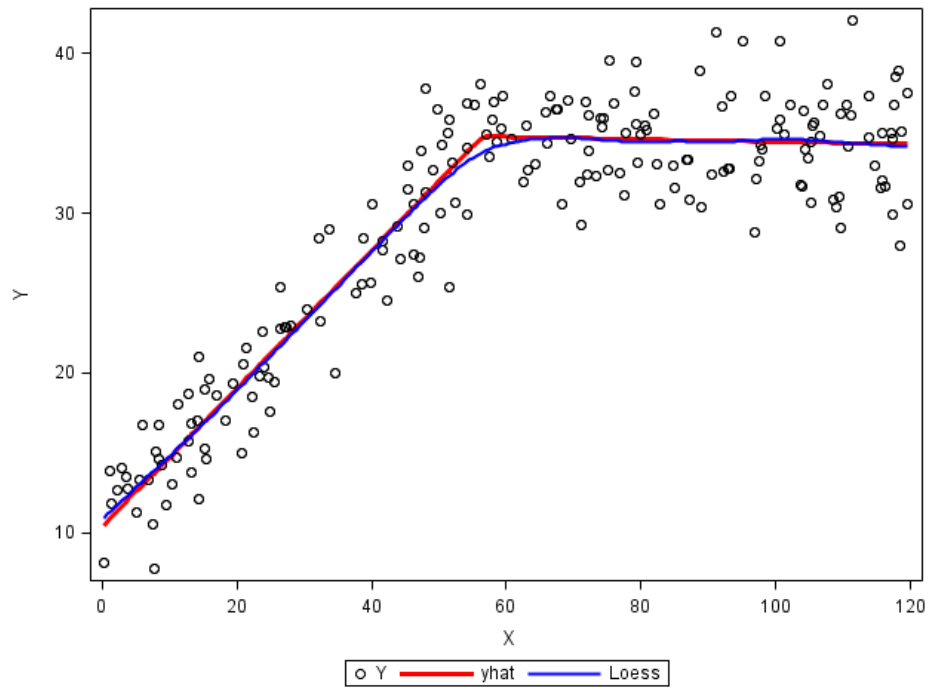
SEs are larger, because uncertainty about d is now included

More *honest* account for uncertainty: we didn't really know $d = 60$

41



42



More parameters, closer fit to data