Topics for this lecture:

- *Spline modeling*

Associated reading: Sections 1 and 2 of the 'Nonparametric and flexible longitudinal regression' notes.

1 Parametric, semiparametric and nonparametric regression: introduction and terminology

- In modeling a mean function over time (or more generally for predictor x), a researcher may need more flexibility than what standard polynomials or transformations can offer. In this chapter we consider methods to accomplish such flexible fits.

- Three classes of regression are parametric, semiparametric and nonparametric. These are discussed in some detail in this chapter (see course notes). Here, we focus on piecewise polynomial regression (parametric) and spline modeling (usually semiparametric or nonparametric).

- See the course notes for more detail on nonparametric and semiparametric regression methods.

2 Piecewise polynomial regression and splines

- Piecewise polynomial regression offers a researcher a more flexible way to model the mean function over time (or more generally over some predictor, x), where the pieces are usually joined together so that the function is continuous but not necessary differentiable.

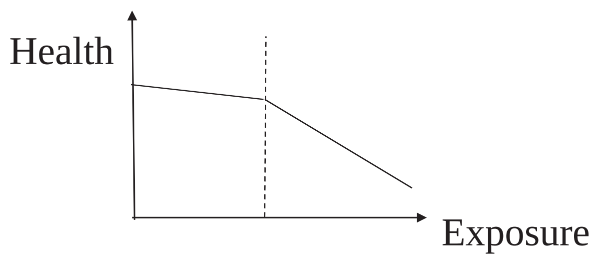
- Spline models further require differentiability so that the entire function is smooth. Cubic terms are commonly used in spline models since they yield a flexible and smooth fit. Quadratic splines can also be used but are less common.

- Although smoothness is intuitive in many cases, in certain cases it may be reasonable to allow the function to be continuous but not differentiable at one or more points, such as for a threshold model or when a treatment is applied during an experiment, resulting in a sharp change in the mean function. Such situations are discussed next.

2.1 Piecewise linear regression

- Consider a health outcome that is modeled as a function of exposure to an environmental risk factor. There may be a negligible or slight dose-response relationship until the level of the risk factor reaches a certain point. Beyond that point, there may be a strong dose-response relationship between this risk factor and the health outcome. Such a model (sometimes called a threshold model) can be fit by joining polynomial functions together into one function. The simplest such function joins two simple linear functions together; the *knot* is where the two linear pieces join together.

- As an introductory example, consider the threshold model described above. Let's assume there is some level of an environmental exposure variable that has the following relationship with a health outcome.



- Say that exposure/health data are collected across subjects and the data is 'cross-sectional' in nature.

- In a GLM regression model, if we know where the knot occurs (say k), we can use the following regression function to fit the linear spline.

$$Y = \beta_0 + \beta_1 \mathrm{x} + \beta_2 max(x-k,0) + \varepsilon$$

- Note that the extra linear piece only 'kicks in' for $x \geq k$.

For $x<k$: $Y = \beta_0 + \beta_1 x + \varepsilon$
For $x \geq k$: $Y = \beta_0 + \beta_1 x + \beta_2 (x-k) + \varepsilon$

$$= (\beta_0 - \beta_2 k) + (\beta_1 + \beta_2) x + \varepsilon$$

$$= \beta_0' + (\beta_1 + \beta_2) x + \varepsilon \text{ .}$$

Thus, the slope of x is β_1 for $x<k$, and $\beta_1+\beta_2$ for $x \geq k$. Often our data will be longitudinal or clustered in nature, but we can fit splines in a linear mixed model in the same way.

- Illustration: Here is a simplified example of a real data set that I have worked with. Subjects that work in Beryllium metal plants have an increased risk of developing Beryllium sensitization (BeS), which can progress into Chronic Beryllium disease (CBD). We are interested in modeling changes in health over time, and specifically we want to see if there is a pronounced change when they progress from BeS to CBD. The health outcome measure here is y = AADO2R (Alveolar-arterial O_2 tension difference at rest); a higher value indicates worse health.

<u>Description of variables</u>:
time can be thought of with units of years
CBDX is the time when subjects progressed from BeS to CBD
prog_group = 0/1/2 for those that progress before/during/after the observation period
stage = stage of illness, 0 for BeS, 1 for CBD
Y=AADO2R, as described above

Here is one approach to modeling the data using linear splines:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 max(x_{ij} - cbdx_i, 0) + pg_h + b_{0i} + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$,
$b_{i0} \sim N(0, \sigma_{b_0}^2)$,
h=1,…,3 (progression group);
i=1,…,n;
j=1,…,r_i.

Here, r_i=r=4 for all i; j=0,…,4.

Above, I'm using x for TIME, pg for PROG_GROUP. Note: since $cbdx_i$ depends on i, subjects can have knots at different times.
In the code below, *time_star* denotes the 'max' term.

```
options ps=60 ls=80;
data new;
input id time cbdx prog_group stage y @@;
time_star=max(time-cbdx,0);
datalines;
1 0 1 1 0 8  1 1 1 1 0 5  1 2 1 1 1 7  1 3 1 1 1 9  1 4 1 1 1 13
. . .
5 0 -2 0 1 5 5 1 -2 0 1 6 5 2 -2 0 1 7 5 3 -2 0 1 8 5 4 -2 0 1 16
;
*spline method;
proc mixed data=new; class prog_group;
model y=time time_star prog_group / outp=pred s;
random intercept / subject=id; run;
proc gplot data=pred; plot pred*time=id;
symbol1 c=red r=2 i=join;
symbol2 c=blue r=2 i=join;
symbol3 c=black r=1 i=join; run;
```


Output:

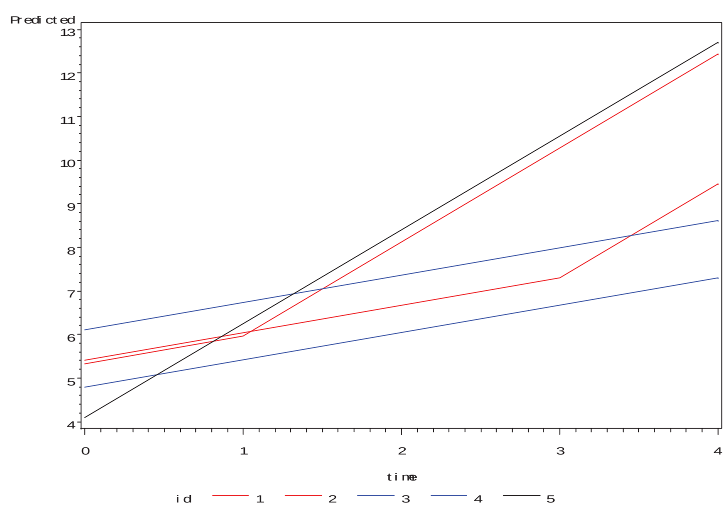
```
Model Information

Dependent Variable          y
Covariance Structure        Variance
Components
Subject Effect              id
Estimation Method           REML
Residual Variance Method    Profile
Fixed Effects SE Method     Model-Based
Degrees of Freedom Method   Containment

Class Level Information

Class         Levels    Values
prog_group         3    0 1 2
Dimensions

Covariance Parameters     2
Columns in X              6
Columns in Z Per Subject  1
Subjects                  5
Max Obs Per Subject       5
```

```
Type 3 Tests of Fixed Effects

              Num   Den
Effect        DF    DF   F Value   Pr>F
time           1    18     4.37   0.0511
time_star      1    18     9.60   0.0062
prog_group 2        18     2.07   0.1546

Covariance Parameter Estimates

Cov Parm      Subject     Estimate
Intercept     id            0.7586
Residual                    2.5811
Fit Statistics
-2 Res Log Likelihood         90.0
AIC (smaller is better)       94.0
AICC (smaller is better)      94.7
BIC (smaller is better)       93.2
```

```
Solution for Fixed Effects

            prog_
Effect      group    Estimate      SE    DF   t Value   Pr>|t|
Intercept              5.4425  0.9998     2      5.44   0.0321
time                   0.6287  0.3009    18      2.09   0.0511
time_star              1.5282  0.4932    18      3.10   0.0062
prog_group    0       -4.4126  2.4091    18     -1.83   0.0836
prog_group    1       -0.0697  1.1807    18     -0.06   0.9536
prog_group    2        0        .         .       .       .
```

- The test for *time_star* indicates that the progression from BeS to CBD causes significant changes to the health-time relationship (p=0.0062). With more data, we can try adding a few more parameters to the model to see if they help describe other patterns in the data.

- In the graph, BeS subjects are forced to have the same linear trend and those with CBD are forced to have the same linear trend, but subjects can progress from one stage to the next at different times. Subject 5 progressed before the observation period, so they have the CBD trend; subjects 3 and 4 have the BeS trend since they progress after the observation period; subjects 1 and 2 progress during the observation period, one at time 1 and the other at time 3.



2.2 Piecewise quadratic and cubic regression

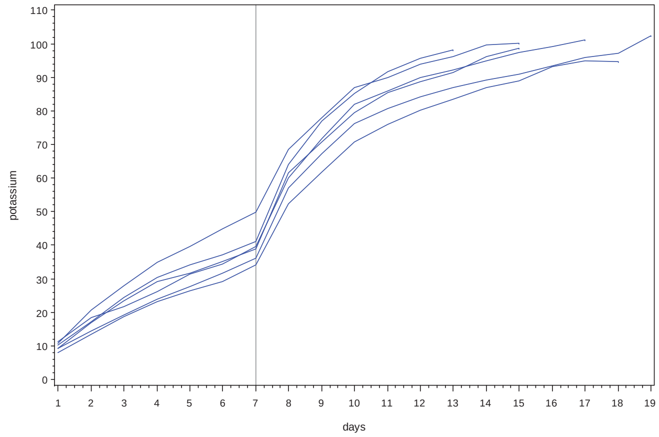
Quadratic and cubic piecewise polynomial functions can also be fit to data.

Example 1: Potassium data.

- These data were obtained via Ed Hess (a former graduate student) based on a consulting project he performed here at the university:
    - Units of blood were sampled daily over the course of several weeks and assayed for Potassium level (exterior to the cells).

-
    - The units were divided into four groups (see following page for plots of each group, given in the order listed as follows:
        - (1) control units that were not irradiated;
        - (2) units irradiated at study initiation;
        - (3) units irradiated at 7 days;
        - (4) units irradiated at 14 days.
    - The motivation for this study was the idea that irradiation of bags can cause a release of free potassium which could result in cardiac arrest (such events had been observed during transfusions).
    - The investigators wanted to characterize the rate of change in potassium level after irradiation for units of blood of different ages (i.e. that had been stored after donation for different lengths of time) to see if this had an impact on potassium release after irradiation.

Sample data



- Here, we consider units irradiated at 7 days. The data illustrate that there was an immediate effect of treatment on potassium levels. In the graph, potassium levels for 6 blood samples were each measured daily for up to 19 days. Responses within samples were joined to yield a spaghetti plot. In terms of spline modeling, it is clear that we want a knot at 7 days. Although the pattern appears to be that of two joined quadratic functions, we actually get a better model fit (lower AIC) including cubic terms in the spline model.

- Here is a possible model for the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + \beta_4 s_{ij1}^1 + \beta_5 s_{ij2}^2 + \beta_6 s_{ij3}^3 + b_{1i} x_{ij} + \varepsilon_{ij}$$

i indexes subject, j indexes observation, i=1,…,n; j=1,…,r_i
Y_{ij} = j^{th} weight observation for mouse i.
x_{ij} = day that j^{th} observation was taken on mouse i.
$s_{ijk} = max(x_{ij} - 7,0)$
$\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$, $b_{i1} \sim N(0,\sigma_{b_1}^2)$

```
data k;
 set long.potassium;
 if day<16;
 s1 = (max(0,day-7))**1;
 s2 = (max(0,day-7))**2;
 s3 = (max(0,day-7))**3;
run;
```

```
proc mixed data=k;
class sample;
model potassium= day day*day day*day*day
s1 s2 s3
  / solution outp=outer;
random day / solution subject=sample;
repeated / type=ar(1) subject=sample;
run;
```

Abbreviated output:

The Mixed Procedure

Dimensions

Covariance Parameters	3
Columns in X	7
Columns in Z Per Subject	1
Subjects	6
Max Obs Per Subject	15
Number of Obs. Used	88

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
day	sample	0.1368
AR(1)	sample	0.8678
Residual		8.5462

Fit Statistics

-2 Res Log Likelihood	344.8
AIC (smaller is better)	350.8
AICC (smaller is better)	351.1
BIC (smaller is better)	350.2

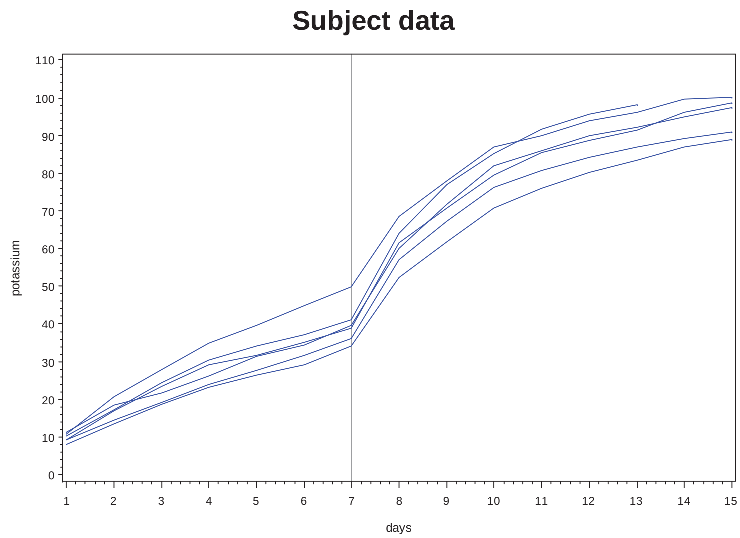
Solution for Fixed Effects

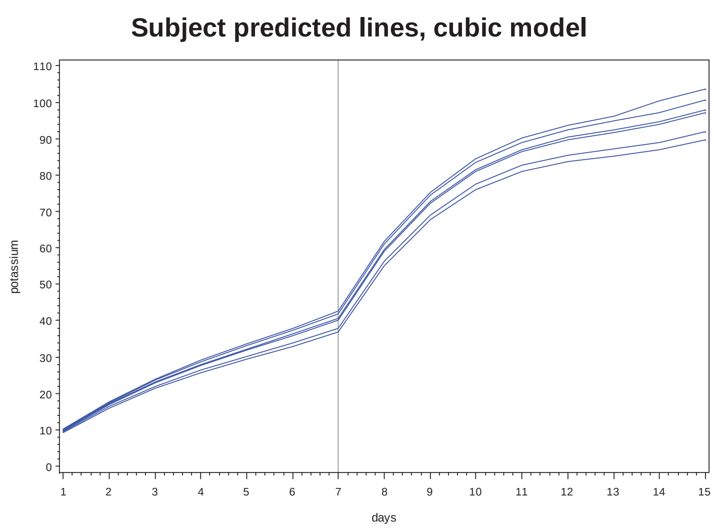
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	0.6236	1.7619	76	0.35	0.7243
day	10.2237	1.4932	5	6.85	0.0010
day*day	-1.1590	0.4193	5	-2.76	0.0397
day*day*day	0.07170	0.03440	5	2.08	0.0915
s1	17.3723	1.0657	76	16.30	<.0001
s2	-3.6920	0.3766	76	-9.80	<.0001
s3	0.1147	0.03818	76	3.00	0.0036

Solution for Random Effects

Effect	sample	Estimate	Std Err Pred	DF	t Value	Pr > |t|
day	1	0.09050	0.2155	76	0.42	0.6756
day	2	-0.4602	0.2155	76	-2.14	0.0359
day	3	0.3793	0.2283	76	1.66	0.1007
day	4	-0.3115	0.2155	76	-1.45	0.1523
day	5	0.02884	0.2155	76	0.13	0.8939
day	6	0.2731	0.2155	76	1.27	0.2088

- The raw data (up through day 15 only) is shown below, followed by a graph of predicted values. For both, responses are joined by lines within subjects.

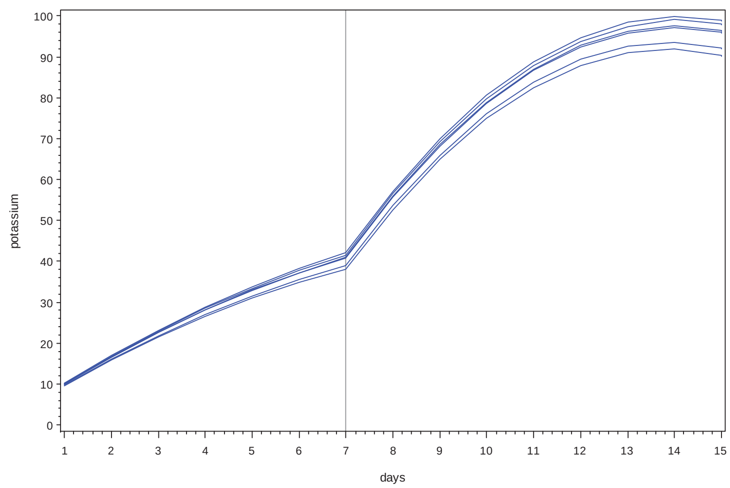




- The predicted values exhibit ‘shrinkage toward the mean’ that we previously discussed.

- If we drop the cubic terms (day and s3), we yield a much higher AIC of 416.6 for the quadratic model (shown below). Note that the predicted values start to bend back down at higher days, a pattern not evident in the data. Thus, the cubic model is superior both visually and quantitatively.

Subject predicted lines, quadratic model



- For the cubic model, we can test for significance of at least one of the spline terms at day 7, H_0: $\beta_4 = \beta_5 = \beta_6 = 0$, using an *F*-test. This is accomplished by adding the following contrast statement in the PROC MIXED code:

```
contrast 'test for spline terms' s1 1, s2 1, s3 1;
```

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
test for spline terms	3	76	177.31	<.0001

- It is not surprising that the test is very significant, given the previous output. This test is just confirming what we have already observed, that irradiation gives a strong boost to potassium levels.

- We can compare the slope just before vs. just after irradiation by taking the derivatives of the fitted function at fixed days. Specifically, let $f(x) = E(Y \mid x)$ for the mixed model, where *x*=days; let $f'(x)$ denote the derivative of $f(x)$. Note that

$$f'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x_{ij}^2 \qquad \text{for } x<7$$
$$f'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x_{ij}^2 + \beta_4 + 2\beta_5(x-7) + 3\beta_6(x-7)^2 \qquad \text{for } x\geq 7$$

- Using the fitted equations, we find that

$$\hat{f}'(6) =$$
$$\hat{f}'(8) =$$

Thus, potassium is increasing an average of ____ units per day one day before irradiation and is increasing an average of ____ units per day one day after irradiation.

2.3 Cubic spline model

Cubic splines have a natural appeal due to their flexible fit, and although they are considered in the class of nonparametric regression modeling, the model can still often be expressed easily in parametric form. So far we have considered piecewise polynomial functions that may have a hard change point (i.e., continuous but not differentiable), but now we consider piecewise polynomial functions that are smooth. To obtain smoothness, lower-order terms are not included at the change points. Specifically, a piecewise polynomial cubic spline model has the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{p} \beta_{k+3} s_k^3$$

where $s_k = \max(0, x - c_k)$ and c_k is the location of knot *k* with respect to the x-axis, *k*=1,…,*p*.

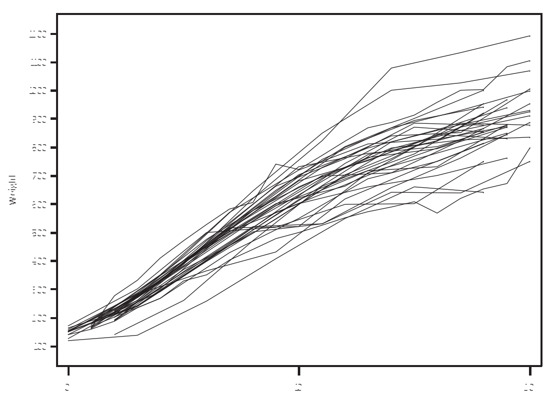
Unlike the previous examples, we only include the cubic terms (s_k^3), but not the lower-order terms (s_k, s_k^2), which forces differentiability across the entire function.

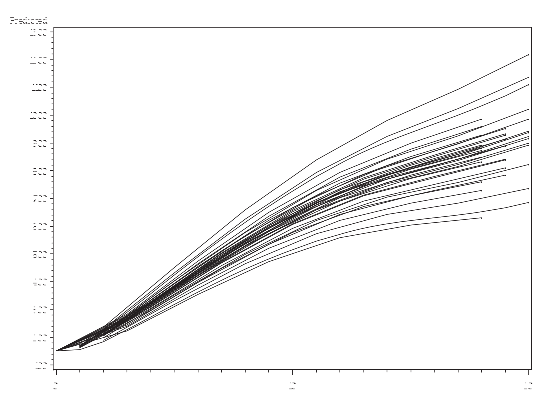
Example 2: Mouse growth data.

- In some cases, we may want to include multiple knots in the spline model, and it may not be so clear where the knots should be. These are true particularly when we are more concerned about getting a flexible fit for the data – in the direction of nonparametric regression. To illustrate, consider the mouse growth data graphed below.

- These data were obtained from Rob Weiss's (Dept. of Biostatistics, UCLA) web site: http://rem.ph.ucla.edu/rob/rm/examples/mice.html. In the graph to the lower left, the weights of mice are measured over their first

days of life; to the right are the predicted values based on the mixed model fit of the model described below.





- You may notice with the data that the quickest growth occurs around days 3 to 8, while the growth is not so steep shortly after birth, and then after day 10 or so. This suggests some type of cubic function may work for these data. Also, we may try modeling a random slope for time across subjects in order to account for the expanding variability between mice over time.

- Using knots at days 3, 8 and 13 (where change points seem to be occurring), here is a possible model for the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + \beta_4 s_{ij1}^3 + \beta_5 s_{ij2}^3 + \beta_6 s_{ij3}^3 + b_{1i} x_{ij} + \varepsilon_{ij}$$

i indexes subject, j indexes observation, i=1,…,n; j=1,…,r_i
Y_{ij} = j^{th} weight observation for mouse i
x_{ij} = day that j^{th} observation was taken on mouse i
$s_{ijk} = max(X_{ij} - v_k, 0)$ where k denotes knot,
knots were fixed at v_1=3.3, v_2=8.3, v_3=13.3 days
$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$
$b_{i1} \sim N(0, \sigma_{b_1}^2)$

- In this case, the lower order spline terms were not included in the model, which is often not done in spline modeling with multiple knots.

- For both of the examples in this subsection, note that we included a random term for time, but no random intercept. This worked since all experimental units had the same value, 0, at the start time.

- But generally, I would caution against such an approach unless it makes sense.

- Generally, I would warn against excluding the random intercept simply based on p-value, just as I would warn against dropping the fixed intercept term based on p-value.

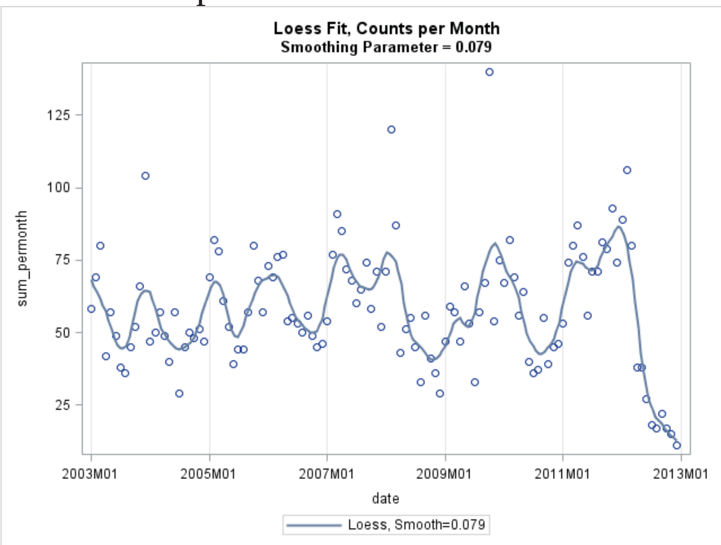
2.4 Case study: Alamosa asthma and pollution study

- The study took place in the San Luis Valley; hospital admission counts (for a medical facility in Alamosa) was compared with daily PM_{10} data (i.e., coarse particulate matter in the air) between 2003 and 2013.

- Here, we consider larger number of knots to be able to get a ‘nonparametric’ fit to the data. I use nonparametric in quotes since really the spline data can still technically be expressed parametrically. However, most consider it a class of nonparametric regression.

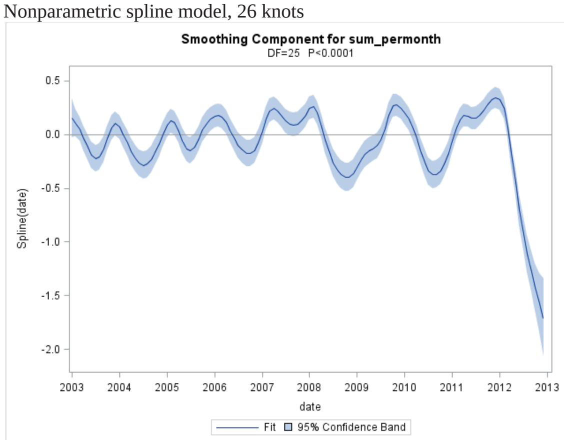
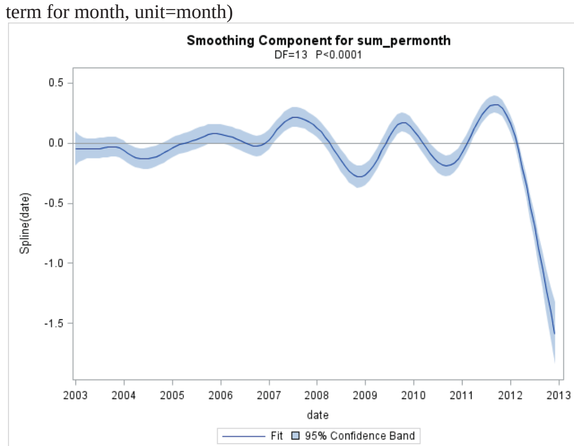
- When such spline variables are combined in models with predictors that are used in the standard way, then we typically call this a semi-parametric regression model. In the models discussed below, we use splines for time (so treat it ‘nonparametrically’), and use standard variables for the pollutant, meteorological variables, month and day of week, and thus have a semi-parametric model.

San Luis hospital admission counts and PM10.



- Circles show monthly hospital counts, and a LOESS (kernel-type) nonparametric regression was used to get the fitted function. This was used for descriptive purposes only. LOESS regression is discussed in more detail in the next section of notes.

- Models below are only initial models that examine hospital counts as a function of time (left) and time and month (right). Here, canned procedures were used to obtain fits.

Nonparametric spline model, 26 knots	Initial semiparametric model (splines for time using 14 knots, parametric term for month, unit=month)
	

- The final model needs to include a flexible fit for time, account for serial correlation, and allow for testing for effects of interest (primarily the pollutant variable).

- Once we define the variables associated with the splines, we actually have a parametric representation of the spline data and can include the variables in a standard parametric longitudinal model, like an LMM or GzLM with GEE. Since we have count data, we will use the latter to do all of this.

- Note that with these data, there is only one ‘subject’, the hospital at which we’re measuring the daily admission counts. We will be able to fit the model as we have ample longitudinal data, although inference is limited to the population that uses this facility.

- With a piecewise smooth cubic spline function, we include the (initial) intercept, linear, quadratic and cubic terms, and then *k* knots, where each knot has a related cubic 'spline' variable that kicks for *x* greater than the knot. By including only the cubic terms associated with the knots, we keep the function smooth. (Also see the mouse data described previously.)

- The initial analyses suggested placing knots at roughly yearly intervals. We have about 10 years of data, and we can place 9 equally spaced knots in the interior. This means there are 13 degrees of freedom including the initial intercept, linear, quadratic and cubic terms, and the spline terms associated with the 9 knots.

- A 'b-spline' approach is essentially a transformation of the X matrix (for the spline variables) so that rows add up to 1. In this case, x variables act more like weights, and variables will have 0's for some elements, indicating that certain spline parameters are not used in predicting values if they are far away from point of interest. (See the SAS Appendix for a comparison of piecewise splines (or 'psplines') that we're familiar with, and basis-splines (or 'bsplines').

- One advantage of b-splines is that the covariance between spline terms can be reduced, compared with p-splines. Another spline approach is to use natural b splines, which force the 2nd derivative of the function to be 0 at the beginning and ending knots.

- For practical purposes, I do not see much difference in models that use the pspline, bspline and nbspline approaches. While estimates and SE's of the spline terms may differ (including the intercept), those for the other terms in the model are either exactly the same or close to the same (they are exactly the same for pspline and bpline approaches, and close to the same for the natural bspline approach).

- Spline matrices can be obtained with software and code as follows.
  - SAS: PROC TRANSREG
    - PSPLINE for piecewise spline
    - BSPLINE for basis spline
  - R: SPLINE package
    - bs for basis splines
    - ns for natural splines

- There are many different types of spline approaches, only some of which are discussed here. Also, be careful with the terminology, it is not always consistent.

- The SAS code demonstrates the program for total hospital count , using 3-day moving average for the pollutant, total hospital count. One run for each of bspline and pspline approaches is shown.

```
%macro
june(var,dist,splinetype,svar_begin,svar_end,polvar1,polvar2,
polvar3);

*use SAS to get variables;

proc transreg data=alldata;
  model identity(&var)= &splinetype(cday
    / knots=0.350 to 3.150 by 0.350);
 output out=sas_splines iapproximations predicted; run;
data alldata2; merge alldata nbspline bspline sas_splines;
run;
```

```
proc genmod data=alldata2 /*descending*/;  *where muni02<300;
class dayofweek month subject;
model &var = &polvar1 &polvar2 &polvar3
/* if pspline is used, vars will be cday_1-cday_12*/
/* if bspline is used, vars will be cday_0-cday_12*/
/* if R version of b-spline is used, vars are bs0-bs12*/
/* if R version of natural b-spline is used, vars are nbs0-
nbs12*/
&svar_begin - &svar_end
dayofweek month temp pressure precip
/ dist=&dist corrb; output out=modfit predicted=p;
ods output GeeEmpPest=est1;
repeated subject=subject / type=/*ar(1)*/mdep(4) modelse;
*estimate 'line after knot' &polvar1 1 &polvar2 1; run;
%mend june;

%june(n_tot,poisson,bspline,cday_0,cday_12,logmuni02,,);
%june(n_tot,poisson,pspline,cday_1,cday_12,logmuni02,,);
```


BSPLINE approach	PSPLINE approach
The GENMOD Procedure Model Information Data Set WORK.ALLDATA2 Distribution Poisson Link Function Log Dependent Variable n_tot	The GENMOD Procedure Model Information Data Set WORK.ALLDATA2 Distribution Poisson Link Function Log Dependent Variable n_tot
Number of Observations Read 3469 Number of Observations Used 3274 Missing Values 195	Number of Observations Read 3469 Number of Observations Used 3274 Missing Values 195
Class Level Information Class Levels Values dayofweek 7 1 2 3 4 5 6 7 month 12 1 2 3 4 5 6 7 8 9 10 11 12 subject 1 1	Class Level Information Class Levels Values dayofweek 7 1 2 3 4 5 6 7 month 12 1 2 3 4 5 6 7 8 9 10 11 12 subject 1 1
GEE Model Information Correlation Structure 4-Dependent Subject Effect subject (1 levels) Number of Clusters 1 Clusters With Missing Values 1 Correlation Matrix Dimension 3469 Maximum Cluster Size 3274 Minimum Cluster Size 3274	GEE Model Information Correlation Structure 4-Dependent Subject Effect subject (1 levels) Number of Clusters 1 Clusters With Missing Values 1 Correlation Matrix Dimension 3469 Maximum Cluster Size 3274 Minimum Cluster Size 3274
Algorithm converged.	Algorithm converged.

GEE Fit Criteria
QIC 3406.1776
QICu 3474.1776

Analysis Of GEE Parameter Estimates using Model-Based SE Estimates (SE's were all 0 when using Empirical SE estimates)

Parameter	Estimate	SE	Z	Pr>|Z|
Intercept	6.3184	2.2385	2.82	0.0048
logmuni02	0.0202	0.0333	0.61	0.5437
cday_0	1.1343	0.3309	3.43	0.0006
cday_1	1.1469	0.3355	3.42	0.0006
cday_2	1.2261	0.3300	3.72	0.0002
cday_3	0.9632	0.3109	3.10	0.0019
cday_4	1.5810	0.3069	5.15	<.0001
cday_5	0.8336	0.3101	2.69	0.0072
cday_6	2.0013	0.3038	6.59	<.0001
cday_7	0.3656	0.3201	1.14	0.2533
cday_8	1.9698	0.2952	6.67	<.0001
cday_9	0.6050	0.3434	1.76	0.0781
cday_10	1.7775	0.2710	6.56	<.0001
cday_11	1.7970	0.4189	4.29	<.0001
cday_12	0.0000	0.0000	.	.
dayofweek 1	0.2668	0.0444	6.01	<.0001
dayofweek 2	0.1326	0.0468	2.83	0.0046
dayofweek 3	0.0487	0.0461	1.06	0.2911
dayofweek 4	0.0435	0.0459	0.95	0.3441
dayofweek 5	-0.0264	0.0482	-0.55	0.5838

GEE Fit Criteria
QIC 3406.1776
QICu 3474.1776

Analysis Of GEE Parameter Estimates using Model-Based SE Estimates (SE's were all 0 when using Empirical SE estimates)

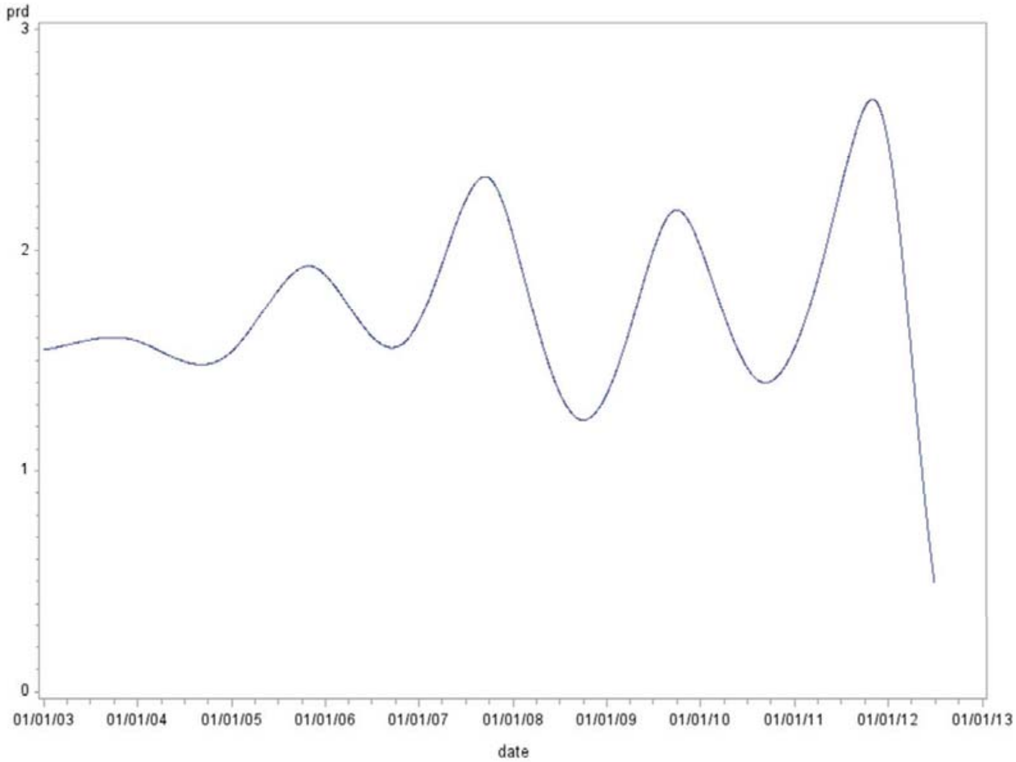
Parameter	Estimate	SE	Z	Pr>|Z|
Intercept	7.4526	2.2297	3.34	0.0008
logmuni02	0.0202	0.0333	0.61	0.5437
cday_1	0.1072	2.7133	0.04	0.9685
cday_2	0.6692	11.1875	0.06	0.9523
cday_3	-2.1248	13.0953	-0.16	0.8711
cday_4	7.0367	16.8737	0.42	0.6767
cday_5	-13.6431	7.2355	-1.89	0.0594
cday_6	21.4826	5.9278	3.62	0.0003
cday_7	-31.0938	5.7758	-5.38	<.0001
cday_8	41.8347	5.9230	7.06	<.0001
cday_9	-47.6282	6.1466	-7.75	<.0001
cday_10	45.6791	6.5321	6.99	<.0001
cday_11	-36.3350	8.1614	-4.45	<.0001
cday_12	-35.5942	23.6427	-1.51	0.1322
dayofweek 1	0.2668	0.0444	6.01	<.0001
dayofweek 2	0.1326	0.0468	2.83	0.0046
dayofweek 3	0.0487	0.0461	1.06	0.2911
dayofweek 4	0.0435	0.0459	0.95	0.3441
dayofweek 5	-0.0264	0.0482	-0.55	0.5838

Parameter	Estimate	SE	Z	Pr>|Z|
dayofweek 6	0.0452	0.0467	0.97	0.3331
dayofweek 7	0.0000	0.0000	.	.
month 1	0.0922	0.0746	1.24	0.2166
month 2	0.3782	0.0721	5.24	<.0001
month 3	0.2771	0.0784	3.54	0.0004
month 4	0.0604	0.0897	0.67	0.5005
month 5	-0.0048	0.0997	-0.05	0.9614
month 6	-0.1366	0.1130	-1.21	0.2270
month 7	-0.2933	0.1288	-2.28	0.0228
month 8	-0.2159	0.1222	-1.77	0.0773
month 9	0.0376	0.1071	0.35	0.7255
month 10	0.0599	0.0921	0.65	0.5150
month 11	-0.0044	0.0811	-0.05	0.9567
month 12	0.0000	0.0000	.	.
temp	0.0014	0.0019	0.75	0.4514
pressure	-0.3134	0.0977	-3.21	0.0013
precip	0.0044	0.1818	0.02	0.9808
Scale	1.0276	.	.	.

Note: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi-square.

Parameter	Estimate	SE	Z	Pr>|Z|
dayofweek 6	0.0452	0.0467	0.97	0.3331
dayofweek 7	0.0000	0.0000	.	.
month 1	0.0922	0.0746	1.24	0.2166
month 2	0.3782	0.0721	5.24	<.0001
month 3	0.2771	0.0784	3.54	0.0004
month 4	0.0604	0.0897	0.67	0.5005
month 5	-0.0048	0.0997	-0.05	0.9614
month 6	-0.1366	0.1130	-1.21	0.2270
month 7	-0.2933	0.1288	-2.28	0.0228
month 8	-0.2159	0.1222	-1.77	0.0773
month 9	0.0376	0.1071	0.35	0.7255
month 10	0.0599	0.0921	0.65	0.5150
month 11	-0.0044	0.0811	-0.05	0.9567
month 12	0.0000	0.0000	.	.
temp	0.0014	0.0019	0.75	0.4514
pressure	-0.3134	0.0977	-3.21	0.0013
precip	0.0044	0.1818	0.02	0.9808
Scale	1.0276	.	.	.

Note: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi-square.



- The graph above shows predicted counts based on the GzLM/GEE model fit.

- The fit represents month and day of week at reference values (December and Saturday, respectively).

- Otherwise, other covariates in the model besides those involving date (i.e., the spline terms) were set to their mean values.

- Predicted values are exactly the same, whether the PSPLINE or BSPLINE approaches are used.

- The pollutant effect is not significant, but is going in the expected direction (positive). Some other models yielded $p<0.05$ for the pollutant variable, e.g., model with a binary pollutant variable based on a particular cut point.

- Pspline correlation between spline parameter estimates (intercept not included)

	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10	Prm11	Prm12	Prm13	Prm14
Prm3	1.0000	-0.9815	0.9568	-0.9115	0.5505	-0.2803	0.1317	-0.0521	0.0071	0.0240	-0.0482	0.0525
Prm4	-0.9815	1.0000	-0.9943	0.9698	-0.6620	0.3627	-0.1803	0.0809	-0.0249	-0.0125	0.0412	-0.0517
Prm5	0.9568	-0.9943	1.0000	-0.9898	0.7277	-0.4210	0.2174	-0.1029	0.0384	0.0038	-0.0355	0.0498
Prm6	-0.9115	0.9698	-0.9898	1.0000	-0.8146	0.5159	-0.2852	0.1447	-0.0636	0.0121	0.0251	-0.0453
Prm7	0.5505	-0.6620	0.7277	-0.8146	1.0000	-0.8668	0.6021	-0.3652	0.2043	-0.1038	0.0382	0.0098
Prm8	-0.2803	0.3627	-0.4210	0.5159	-0.8668	1.0000	-0.8869	0.6460	-0.4097	0.2371	-0.1223	0.0370
Prm9	0.1317	-0.1803	0.2174	-0.2852	0.6021	-0.8869	1.0000	-0.8980	0.6688	-0.4305	0.2418	-0.0965
Prm10	-0.0521	0.0809	-0.1029	0.1447	-0.3652	0.6460	-0.8980	1.0000	-0.9044	0.6743	-0.4144	0.1828
Prm11	0.0071	-0.0249	0.0384	-0.0636	0.2043	-0.4097	0.6688	-0.9044	1.0000	-0.8984	0.6423	-0.3185
Prm12	0.0240	-0.0125	0.0038	0.0121	-0.1038	0.2371	-0.4305	0.6743	-0.8984	1.0000	-0.8843	0.5263
Prm13	-0.0482	0.0412	-0.0355	0.0251	0.0382	-0.1223	0.2418	-0.4144	0.6423	-0.8843	1.0000	-0.7845
Prm14	0.0525	-0.0517	0.0498	-0.0453	0.0098	0.0370	-0.0965	0.1828	-0.3185	0.5263	-0.7845	1.0000

- Bspline correlation between spline parameter estimates (intercept not included)

	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10	Prm11	Prm12	Prm13	Prm14
Prm3	1.0000	0.5602	0.8492	0.7575	0.8282	0.7982	0.8079	0.7990	0.7924	0.8063	0.7081	0.8261
Prm4	0.5602	1.0000	0.5545	0.8544	0.7372	0.7980	0.7680	0.7727	0.7719	0.7658	0.7094	0.7794
Prm5	0.8492	0.5545	1.0000	0.6725	0.8875	0.7757	0.8375	0.7933	0.8112	0.8057	0.7289	0.8318
Prm6	0.7575	0.8544	0.6725	1.0000	0.7596	0.9029	0.8205	0.8586	0.8339	0.8474	0.7585	0.8603
Prm7	0.8282	0.7372	0.8875	0.7596	1.0000	0.7798	0.9158	0.8247	0.8733	0.8468	0.7852	0.8765
Prm8	0.7982	0.7980	0.7757	0.9029	0.7798	1.0000	0.7694	0.9078	0.8183	0.8724	0.7509	0.8782
Prm9	0.8079	0.7680	0.8375	0.8205	0.9158	0.7694	1.0000	0.7568	0.9119	0.8198	0.8085	0.8656
Prm10	0.7990	0.7727	0.7933	0.8586	0.8247	0.9078	0.7568	1.0000	0.7409	0.9053	0.7050	0.8842
Prm11	0.7924	0.7719	0.8112	0.8339	0.8733	0.8183	0.9119	0.7409	1.0000	0.7381	0.8668	0.8283
Prm12	0.8063	0.7658	0.8057	0.8474	0.8468	0.8724	0.8198	0.9053	0.7381	1.0000	0.6073	0.9245
Prm13	0.7081	0.7094	0.7289	0.7585	0.7852	0.7509	0.8085	0.7050	0.8668	0.6073	1.0000	0.6537
Prm14	0.8261	0.7794	0.8318	0.8603	0.8765	0.8782	0.8656	0.8842	0.8283	0.9245	0.6537	1.0000

- Nbspline correlation between spline parameter estimates (intercept not included)

	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10	Prm11	Prm12	Prm13	Prm14	Prm15
Prm3	1.0000	0.0072	0.5104	0.2004	0.3780	0.3045	0.3048	0.3100	0.3254	0.3143	0.3351	0.2743	0.1938
Prm4	0.0072	1.0000	0.1650	0.6456	0.3883	0.5609	0.4226	0.5077	0.4855	0.4936	0.2832	0.7355	-0.0659
Prm5	0.5104	0.1650	1.0000	0.0310	0.6551	0.3238	0.4491	0.3930	0.4416	0.4188	0.3332	0.5449	0.0499
Prm6	0.2004	0.6456	0.0310	1.0000	0.0370	0.6485	0.2825	0.4803	0.3978	0.4365	0.2638	0.6012	-0.0015
Prm7	0.3780	0.3883	0.6551	0.0370	1.0000	0.1223	0.5815	0.3434	0.4904	0.4154	0.3301	0.5868	0.0110
Prm8	0.3045	0.5609	0.3238	0.6485	0.1223	1.0000	0.0824	0.6476	0.3592	0.5071	0.2960	0.6410	0.0238
Prm9	0.3048	0.4226	0.4491	0.2825	0.5815	0.0824	1.0000	0.0187	0.6256	0.2983	0.3355	0.5409	-0.0195
Prm10	0.3100	0.5077	0.3930	0.4803	0.3434	0.6476	0.0187	1.0000	0.0726	0.6191	0.1980	0.6123	0.0664
Prm11	0.3254	0.4855	0.4416	0.3978	0.4904	0.3592	0.6256	0.0726	1.0000	0.1295	0.5074	0.5880	-0.1101
Prm12	0.3143	0.4936	0.4188	0.4365	0.4154	0.5071	0.2983	0.6191	0.1295	1.0000	-0.1133	0.6392	0.2778
Prm13	0.3351	0.2832	0.3332	0.2638	0.3301	0.2960	0.3355	0.1980	0.5074	-0.1133	1.0000	0.2684	-0.4521
Prm14	0.2743	0.7355	0.5449	0.6012	0.5868	0.6410	0.5409	0.6123	0.5880	0.6392	0.2684	1.0000	0.1505
Prm15	0.1938	-0.0659	0.0499	-0.0015	0.0110	0.0238	-0.0195	0.0664	-0.1101	0.2778	-0.4521	0.1505	1.0000

2.5 Comparing piecewise polynomial and b-splines: bases and properties

Note: this section is taken from SAS Help Documentation, with some minor editing. An algorithm for generating the B-spline basis is given in de Boor (1978, pp. 134–135). B-splines are both a computationally accurate and efficient way of constructing a basis for piecewise polynomials; however, they are not the most natural method of describing splines. Consider an initial scaling vector $\mathbf{x} = (1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9)^t$ and a degree-three spline with interior knots at 3.5 and 6.5. The natural piecewise polynomial spline basis (X matrix for associated variables) is the left matrix, and the B-spline basis for the transformation is the right matrix.

Piecewise Polynomial Splines

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 4 & 8 & 0 & 0 \\ 1 & 3 & 9 & 27 & 0 & 0 \\ 1 & 4 & 16 & 64 & 0.125 & 0 \\ 1 & 5 & 25 & 125 & 3.375 & 0 \\ 1 & 6 & 36 & 216 & 15.625 & 0 \\ 1 & 7 & 49 & 343 & 42.875 & 0.125 \\ 1 & 8 & 64 & 512 & 91.125 & 3.375 \\ 1 & 9 & 81 & 729 & 166.375 & 15.625 \end{pmatrix}$$

B-Spline Basis

$$\begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 & 0 & 0 \\ 0.216 & 0.608 & 0.167 & 0.009 & 0 & 0 \\ 0.008 & 0.458 & 0.461 & 0.073 & 0 & 0 \\ 0 & 0.172 & 0.585 & 0.241 & 0.001 & 0 \\ 0 & 0.037 & 0.463 & 0.463 & 0.037 & 0 \\ 0 & 0.001 & 0.241 & 0.585 & 0.172 & 0 \\ 0 & 0 & 0.073 & 0.461 & 0.458 & 0.0008 \\ 0 & 0 & 0.009 & 0.167 & 0.608 & 0.216 \\ 0 & 0 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

The two matrices span the same column space. The numbers in the B-spline basis do not have a simple interpretation like the numbers in the natural piecewise polynomial basis. The B-spline basis has a diagonally banded structure and the band shifts one column to the right after every knot. The number of entries in each row that can potentially be nonzero is one greater than the degree. The elements within a row always sum to one. The B-spline basis is accurate because of the smallness of the numbers and the lack of extreme collinearity inherent in the piecewise polynomials.