AN INTRODUCTION TO SPLINES

Trinity River Restoration Program
Workshop on Outmigration: Population Estimation

October 6–8, 2009
An Introduction to Splines

1. Linear Regression
   - Simple Regression and the Least Squares Method
   - Least Squares Fitting in \( \mathbb{R} \)
   - Polynomial Regression

2. Smoothing Splines
   - Simple Splines
   - B-splines
   - Overfitting and Smoothness
1 Linear Regression
- Simple Regression and the Least Squares Method
- Least Squares Fitting in R
- Polynomial Regression
1 Linear Regression

- Simple Regression and the Least Squares Method
- Least Squares Fitting in R
- Polynomial Regression
Simple Linear Regression

Daily temperatures in Montreal from April 1 (Day 81) to June 30 (Day 191), 1961.
Assumptions

Mean  On average, the change in the response is proportional to the change in the predictor.

Errors  1. The deviation in the response for any observation does not depend on any other observation.

2. The average magnitude of the deviation is the same for all values of the predictor.

Mathematically

For $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent with mean 0 and variance $\sigma^2$. 
The Least Squares Method

Example: The Montreal Data

Montreal Temp. --- April 1 to June 30, 1961

Introduction to Splines: Linear Regression, Simple Regression and the Least Squares Method
Definition
Given values for $\beta_0$ and $\beta_1$, the residual for the $i^{th}$ observation is the difference between the observed and the predicted response:

$$e_i = y_i - \hat{y}_i$$

where $\hat{y}_i = \beta_0 + \beta_1 x_i$. 
The least squares method defines the best values of $\beta_0$ and $\beta_1$ to be those that minimize the sum of the squared residuals:

$$SS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$
The Least Squares Method

Example: The Montreal Data

Montreal Temp. --- April 1 to June 30, 1961

\[ SS = 1549.37 \]

\[ SS = 1148.56 \]
1 Linear Regression
   - Simple Regression and the Least Squares Method
   - Least Squares Fitting in $\mathbb{R}$
   - Polynomial Regression
Suppose that the data is a data frame with elements:

- \( x \): the days from 90 to 181
- \( y \): the observed temperatures

```r
> data = read.table("MontrealTemp1.txt")
> summary(data)

     x        y
  Min.  :90.00  Min.  :-0.90
  1st Qu.:112.8  1st Qu.: 5.60
  Median :135.5  Median :11.55
  Mean   :135.5  Mean   :11.46
  3rd Qu.:158.2  3rd Qu.:16.70
  Max.   :181.0  Max.   :23.60
```

>
Fitting the model with \texttt{lm}:

\begin{verbatim}
> lm(y~x, data)

Call:
\texttt{lm(formula = y \sim x, \ data = data)}

Coefficients:
\begin{array}{cc}
\text{(Intercept)} & x \\
-15.3996 & 0.1982
\end{array}
\end{verbatim}
Fitting the model with `lm`:

```r
> lmfit = lm(y~x, data)
> attributes(lmfit)
$names
 [1] "coefficients"   "residuals"
 [3] "effects"        "rank"
 [5] "fitted.values"  "assign"
 [7] "qr"             "df.residual"
 [9] "xlevels"        "call"
[11] "terms"         "model"

$class
 [1] "lm"
> 
```
Plotting the fitted line over the raw data:

```r
# Plot the raw data
> plot(data$x, data$y, 
      main = "Montreal Temp. ...", 
      xlab = "Day of Year", ylab = "Temperature")

# Add the fitted line
> lines(data$x, lmfit$fit, col = "red", lwd = 3)
```
Least Squares Fitting in R

The Fitted Line

Montreal Temp. — April 1 to June 30, 1961

Day of Year

Temperature

Introduction to Splines: Linear Regression, Least Squares Fitting in R
Residual Diagnostics

The value of the residuals should not depend on $x$ or $y$ in any systematic way.

- Common indications of lack of fit:
  - trends with $x$ or $y$ (curves or clusters of high/low values)
  - constant increase/decrease (funnel shape)
  - increase followed by decrease (football shape)
  - very large ($+$ or $-$) values (outliers)

- Assessed by plotting $e$ versus $x$ and $y$. 
Plotting the residuals versus the predictor and response:

```r
## Plot the residuals versus day
> plot(data$x, lmfit$resid, 
    xlab="Day of Year", ylab="Residual")
> abline(h=0)

## Plot the residuals versus temperature
> plot(data$y, lmfit$resid, 
    xlab="Temperature", ylab="Residual")
> abline(h=0)
```
Residuals vs. Day

Residuals vs. Temperature
1. Montreal Temperature Data – April 1 to June 30, 1961
   File: Intro_to_splines\Exercises\montreal_temp_1.R
   Use the provide code to fit the simple linear regression model to the Montreal temperature data from the spring of 1961, plot the fitted line, and produce the residual plots.

   File: Intro_to_splines\Exercises\montreal_temp_2.R
   Repeat exercise 1 with the data from all of 1961.
1 Linear Regression

- Simple Regression and the Least Squares Method
- Least Squares Fitting in R
- Polynomial Regression
Polynomial Regression

Motivation
**Polynomial Regression**

**Motivation**

Residuals vs. Day

Residuals vs. Temperature
**Definition**

A polynomial of degree $D$ is a function formed by linear combinations of the powers of its argument up to $D$:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_D x^D$$

**Specific Polynomials**

- **Linear** $y = \beta_0 + \beta_1 x$
- **Quadratic** $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- **Cubic** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- **Quartic** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$
- **Quintic** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$
**Definition**

The design matrix for a regression model with \( n \) observations and \( p \) predictors is the matrix with \( n \) rows and \( p \) columns such that the value of the \( j^{th} \) predictor for the \( i^{th} \) observation is located in column \( j \) of row \( i \).

**Design matrix for a polynomial of degree \( D \)**

\[
\begin{bmatrix}
1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^D \\
1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^D \\
1 & x_3 & x_3^2 & x_3^3 & \cdots & x_3^D \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
n & x_n & x_n^2 & x_n^3 & \cdots & x_n^D
\end{bmatrix}
\]
The design matrix for polynomial regression can be generated with the function `outer()`:

```r
> D = 2
> X = outer(data$x, 1:D, "^")
> X[1:5,]
  [,1]  [,2]
[1,] 1 1
[2,] 2 4
[3,] 3 9
[4,] 4 16
[5,] 5 25
>
Note: we do not need to include the intercept column.
```
Polynomial Regression in R

Least Squares Fitting – Quadratic

> lmfit = lm(y~X, data)
> attributes(lmfit)
$names
 [1] "coefficients" "residuals" ...

$class
 [1] "lm"

> lmfit$coefficients
   (Intercept)       X1       X2
  -23.715358962  0.413901580 -0.001014625
Polynomial Regression in R

Fitted Model – Quadratic

Montreal Temp. — January 1 to December 31, 1961

Day of Year
Temperature
Residual

Temperature
Day of Year
Residual

Day of Year
File: Intro_to_splines\Exercises\montreal_temp_3.R
Use the provided code to fit polynomial regression models of varying degree to the data for all of 1961. Models of different degree are constructed by setting the variable D (e.g., D=2 produces a quadratic model). What is the minimal degree required for the model to fit well?

File: Intro_to_splines\Exercises\montreal_temp_4.R
Repeat this exercise using the data from both 1961 and 1962.
2. **Smoothing Splines**
   - Simple Splines
   - B-splines
   - Overfitting and Smoothness
2 Smoothing Splines
   - Simple Splines
   - B-splines
   - Overfitting and Smoothness
How is the temperature changing in the spring of 1962?

\[ y = -7.6 - 8.3x - 0.3x^2 - 5.2 \times 10^4 x^{-3} + 4.4 \times 10^{-6} x^4 \\
- 2.1 \times 10^{-8} x^5 + 6.0 \times 10^{-11} x^6 - 8.9 \times 10^{-14} x^7 + 5.5 \times 10^{-17} x^8 \]
How is the temperature changing in the spring of 1962?

\[ y = -144.5 + 0.3x \]
**Definition**

A *linear spline* is a continuous function formed by connecting linear segments. The points where the segments connect are called the *knots* of the spline.
**Definition**

A spline of degree $D$ is a function formed by connecting polynomial segments of degree $D$ so that:

- the function is continuous,
- the function has $D - 1$ continuous derivatives, and
- the $D^{th}$ derivative is constant between knots.
**Definition**

The truncated polynomial of degree $D$ associated with a knot $\xi_k$ is the function which is equal to 0 to the left of $\xi_k$ and equal to $(x - \xi_k)^D$ to the right of $\xi_k$.

\[
(x - \xi_k)^+ = \begin{cases} 
0 & x < \xi_k \\
(x - \xi_k)^D & x \geq \xi_k
\end{cases}
\]

The equation for a spline of degree $D$ with $K$ knots is:

\[
y = \beta_0 + \sum_{d=1}^{D} \beta_d x^d + \sum_{k=1}^{K} b_k (x - \xi_k)^D
\]
Simple Splines

The Design Matrix

The design matrix for a spline of degree $D$ with $K$ knots is the $n$ by $1 + D + K$ matrix with entries:

\[
\begin{bmatrix}
1 & x_1 & x_1^2 & \cdots & x_1^D & (x_1 - \xi_1)_+^D & \cdots & (x_1 - \xi_K)_+^D \\
1 & x_2 & x_2^2 & \cdots & x_2^D & (x_2 - \xi_1)_+^D & \cdots & (x_2 - \xi_K)_+^D \\
1 & x_3 & x_3^2 & \cdots & x_3^D & (x_3 - \xi_1)_+^D & \cdots & (x_3 - \xi_K)_+^D \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & x_n & x_n^2 & \cdots & x_n^D & (x_n - \xi_1)_+^D & \cdots & (x_n - \xi_K)_+^D 
\end{bmatrix}
\]
After defining the degree and the locations of the knots, the design matrix can be generated with the functions `outer` and `cbind`:

```
> D = 3
> K = 5
> knots = 730 * (1:K)/(K+1)
> X1 = outer(data$x,1:D,"^")
> X2 = outer(data$x,knots,">") * 
    outer(data$x,knots,"-")^D
> X = cbind(X1,X2)
> round(X[c(1,150,300),1:5],1)
```

```
[1,] 1  1  1  0.0  0.0
[2,] 150 22500 3375000 22745.4 0.0
[3,] 300 90000 27000000 5671495.4 181963
```

Fit the Spline Model

\texttt{lmfit = lm(y~X, data=data)}
Simple Splines in R
Fitted Cubic Spline

Montreal Temp. — January 1 to December 31, 1962

Temperature vs. Day of Year

Residual vs. Day of Year
File: Intro_to_splines\Exercises\montreal_temp_5.R
Use the code provided to fit splines of varying degree and with different numbers of knots to the data from 1961 and 1962.
2 Smoothing Splines
   - Simple Splines
   - B-splines
   - Overfitting and Smoothness
Splines computed from the truncated polynomials may be numerically unstable because:

- the values in the design matrix may be very large, and
- the columns of the design matrix may be highly correlated.
The B-spline Basis in R
Generating the Design Matrix and Fitting the Model

The B-spline design matrix can be constructed via the function `bs` provided by the `splines` library:

```r
> library(splines)
> D = 3
> K = 5
> knots = 730 * (1:K)/(K+1)
> X = bs(data$x, knots = knots, degree = D, intercept = TRUE)
> lmfit = lm(y~X-1, data = data)
> 
```
The B-spline Basis in R

Fitted Cubic B-spline Model


Introduction to Splines: Smoothing Splines, B-splines
   File: Intro_to_splines\Exercises\montreal_temp_6.R
   Fit B-splines to the data from 1961 and 1962 using the code in the file. Increase the number of knots to see how this affects the fit of the curve. What happens when the number of knots is very large, say $K = 50$?
2 Smoothing Splines
  - Simple Splines
  - B-splines
  - Overfitting and Smoothness
A cubic spline with 50 knots:
Concept
The shape of a spline can be controlled by carefully choosing the number of knots and their exact locations in order to:

1. allow flexibility where the trend changes quickly, and
2. avoid overfitting where the trend changes little.

Challenge
Choosing the number of knots and their location is a very difficult problem to solve.
Concept
We can also balance overfitting and smoothness by controlling the size of the spline coefficients.
Penalization for the Linear Spline

- Consider the equation for each segment of the spline:
  
  \[(0, \xi_1) : \ y = \beta_0 + \beta_1 x\]
  
  \[(\xi_1, \xi_2) : \ y = (\beta_0 - b_1 \xi_1) + (\beta_1 + b_1) x\]
  
  \[(\xi_2, \xi_3) : \ y = (\beta_0 - b_1 \xi_1 - b_2 \xi_2) + (\beta_1 + b_1 + b_2) x\]

- The spline is smooth if \(b_1, b_2, \ldots, b_K\) are all close to 0.

Penalized Least Squares

\[
PSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{K} b_k^2
\]
Penalization for the B-spline

The spline is smooth if $b_1, b_2, \ldots, b_K$ are all close to each other. (But not necessarily close to 0.)

Penalized Least Squares

\[
PSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=3}^{K} ((b_k - b_{k-1}) - (b_{k-1} - b_{k-2}))^2
\]
A penalized cubic B-spline with 50 knots and $\lambda = 5$:
File: Intro_to_splines\Exercises\montreal_temp_7.R
Fit penalized cubic B-splines to the Montreal temperature data for 1961 and 1962 using the provided code.