Regression Splines

- A basis expansion approach:

$$y = \beta_1 h_1(x) + \beta_2 h_2(x) + \cdots + \beta_p h_p(x) + \text{err},$$

  where $p = m + 4$ for regression with cubic splines and $p = m$ for NCS.

- Represent the model on the observed n data points using matrix notation,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{F}\boldsymbol{\beta}\|^2,$$

where

$$\underset{n\times 1}{\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}} = \underset{n\times p}{\begin{pmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_p(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_p(x_2) \\ & & & \\ h_1(x_n) & h_2(x_n) & \cdots & h_p(x_n) \end{pmatrix}} \underset{p\times 1}{\begin{pmatrix} \beta_1 \\ \cdots \\ \beta_p \end{pmatrix}} + \text{err}$$

- We can obtain the design matrix F by commands `bs` or `ns` in `R`, and then call the regression function `lm`.

Understand how R counts the degree-of-feedom.

- To generate a cubic spline basis for a given set of x_i's, you can use the command `bs`.
- You can tell `R` the location of knots.
- Or you can tell `R` the df. Recall that a cubic spline with m knots has $m + 4$ df, so we need $m = \text{df} - 4$ knots. By default, `R` puts knots at the $1/(m+1), \ldots, m/(m+1)$ quantiles of $x_{1:n}$.

How R counts the df is a little confusing. The df in command bs actually means the number of columns of the design matrix returned by bs . So if the intercept is not included in the design matrix (which is the default), then the df in command bs is equal to the real df minus 1.

So the following three design matrices (the first two are of $n \times 5$ and the last one is of $n \times 6$) correspond to the same regression model with cubic splines of df 6.

```
> bs(x, knots=quantile(x, c(1/3, 2/3)));
> bs(x, df=5);
> bs(x, df=6, intercept=TRUE);
```

- To generate a NCS basis for a given set of x_i's, use the command `ns`.
- Recall that the linear functions in the two extreme intervals are totally determined by the other cubic splines, even if no data points are in the two extreme intervals (i.e., data points are inside the two boundary knots). By default, `R` puts the two boundary knots as the min and max of x_i's.
- You can tell `R` the location of knots, which are the interior knots. Recall that a NCS with m knots has m df. So the df is equal to the number of (interior) knots plus 2, where 2 means the two boundary knots.

- Or you can tell `R` the df. If `intercept = TRUE`, then we need $m = \text{df} - 2$ knots, otherwise we need $m = \text{df} - 1$ knots. Again, by default, `R` puts knots at the $1/(m+1), \ldots, m/(m+1)$ quantiles of $x_{1:n}$.

- The following three design matrices (the first two are of $n \times 3$ and the last one is of $n \times 4$) correspond to the same regression model with NCS of df 4.

```
> ns(x, knots=quantile(x, c(1/3, 2/3)));
> ns(x, df=3);
> ns(x, df=4, intercept=TRUE);
```