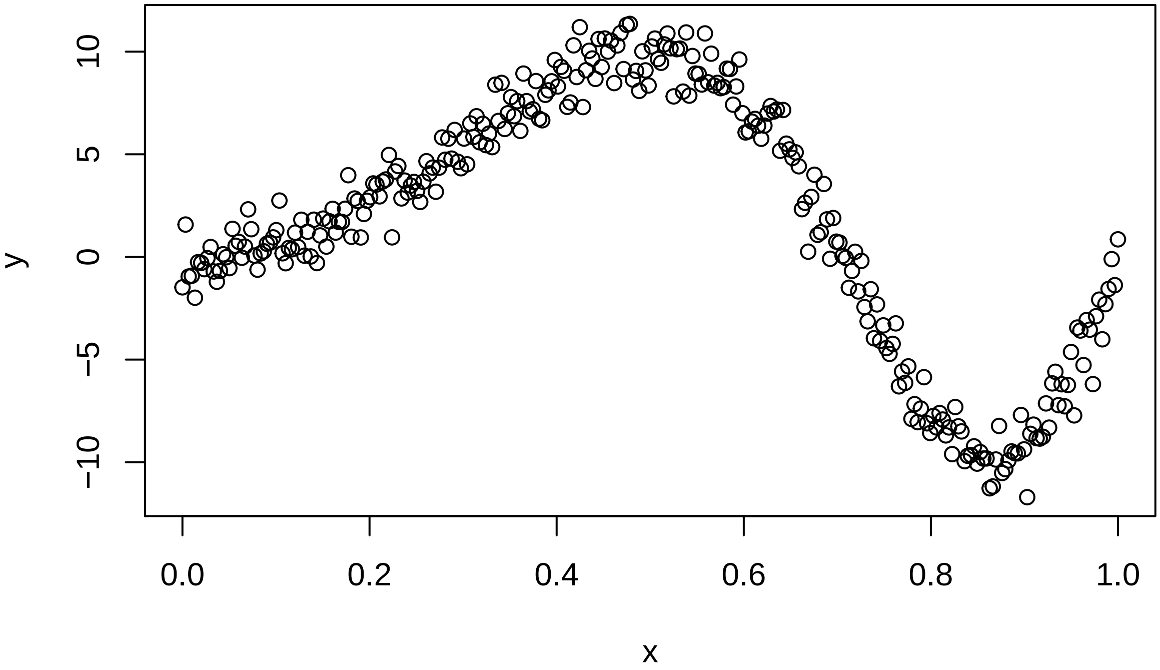
Introduction to Splines

Andy Shen, Devin Francom

Let's say you want to fit a model using some wiggly data. Maybe

```r
set.seed(12)
n<-300
x<-seq(0,1,length.out=n)
y<-sin(2*pi*x^2)*10+rnorm(n)
plot(x,y)
```



One way to fit a model to data like this is to come up with a linear basis and fit a linear model using the basis as the X matrix (which we will call B). People often use splines as a basis. The simplest set of spline basis functions would be to make the ith basis function (i.e., the ith column of B) look like

$$B_{ij} = [s_i(x_j - t_i)]_+$$

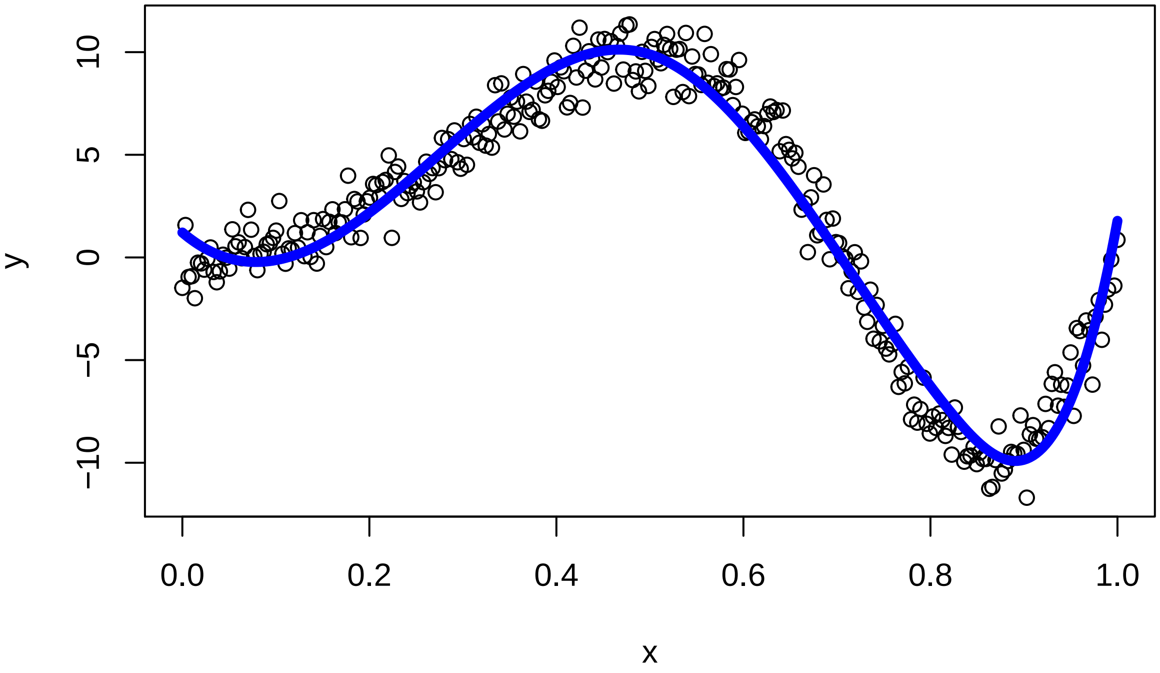
where $s \in \{-1, 1\}$, which we'll call the sign, and t is a value in the domain of x, which we will call a knot. Also, $[a]_+ = max(0, a)$.

Try some combinations of s and t to see what your basis functions look like, and what the corresponding linear model fit looks like (using the lm function or your Bayesian linear model code). Try with different numbers of basis functions, also.

Using the `bs()` Function

2 Knots (Expected)

```r
library(splines)
df <- data.frame(y, x)
m1 <- lm(y ~ bs(x, knots = c(0.5, 0.82)), data = df)
pred <- predict(m1)

plot(x,y)
lines(x, pred, lwd = 5, col = "blue")
```



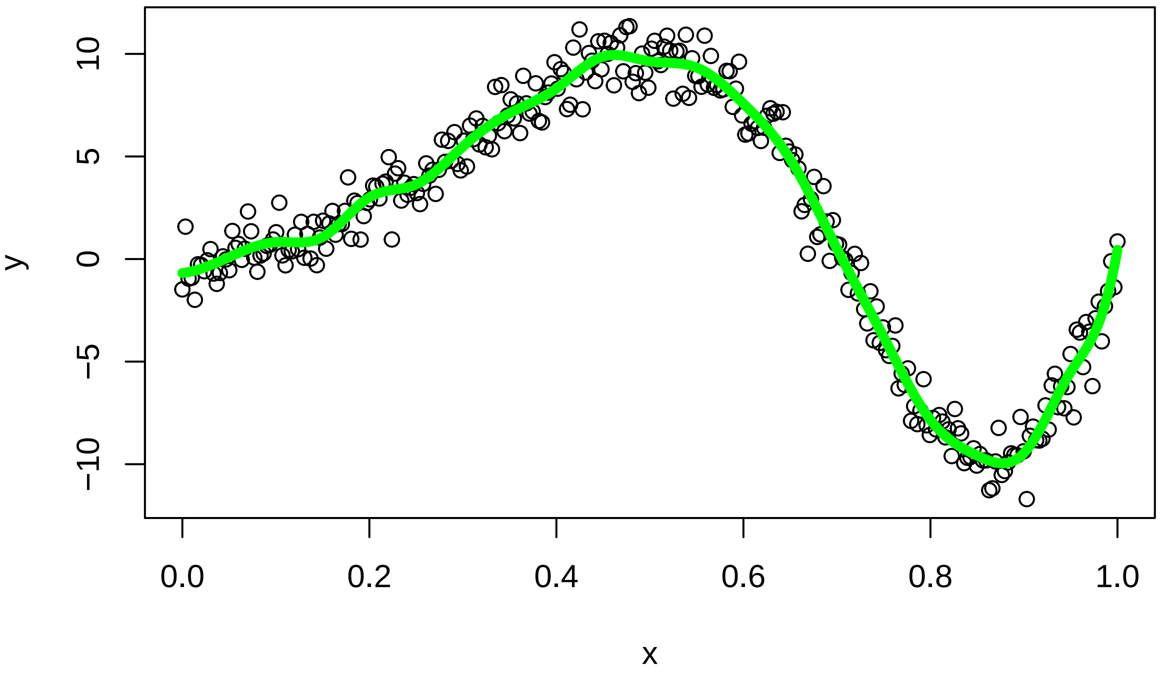
```r
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ bs(x, knots = c(0.5, 0.82)), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90576 -0.91688  0.09772  0.83283  3.14932
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.2153     0.3524   3.448 0.000647 ***
## bs(x, knots = c(0.5, 0.82))1  -6.4639     0.7560  -8.550 6.85e-16 ***
## bs(x, knots = c(0.5, 0.82))2  22.3566     0.5093  43.893  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))3  -8.1253     0.6274 -12.951  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))4 -14.2312     0.5277 -26.966  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))5   0.5685     0.6349   0.895 0.371271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.271 on 294 degrees of freedom
## Multiple R-squared:  0.9593, Adjusted R-squared:  0.9586
## F-statistic:  1385 on 5 and 294 DF,  p-value: < 2.2e-16
```

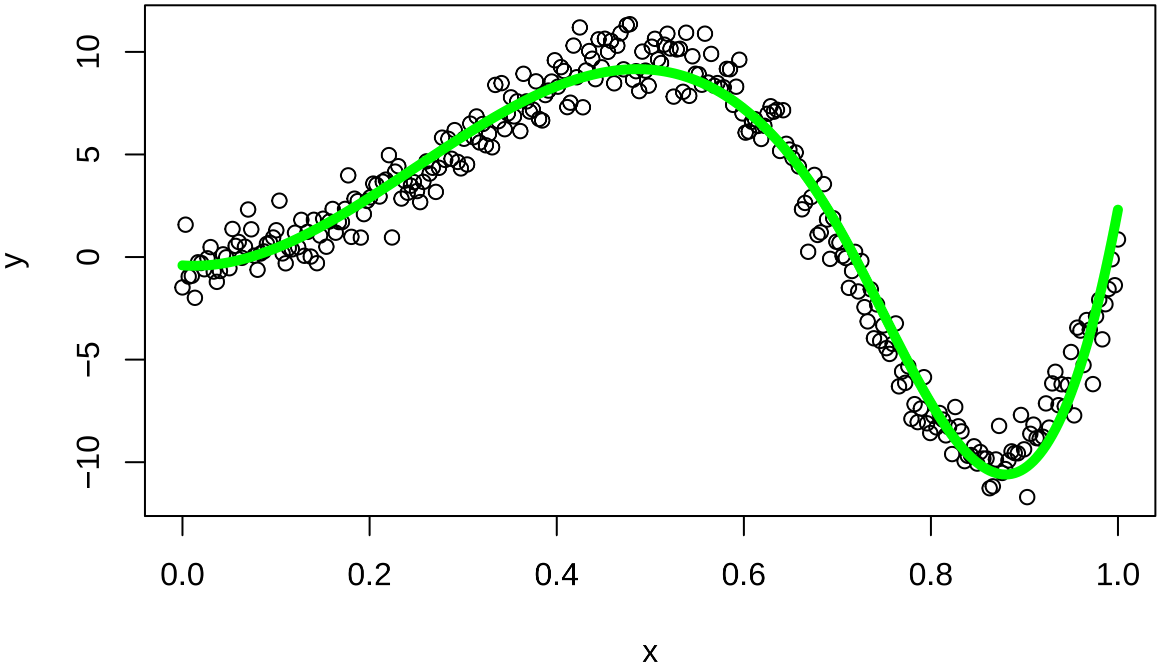
Too Many Knots

```
m2 <- lm(y ~ bs(x, knots = seq(0.1,1,by=0.05)), data = df)
pred <- predict(m2)

plot(x,y)
lines(x, pred, lwd = 5, col = "green")
```



1 Knot

```
m2 <- lm(y ~ bs(x, knots = 0.7), data = df)
pred <- predict(m2)

plot(x,y)
lines(x, pred, lwd = 5, col = "green")
```

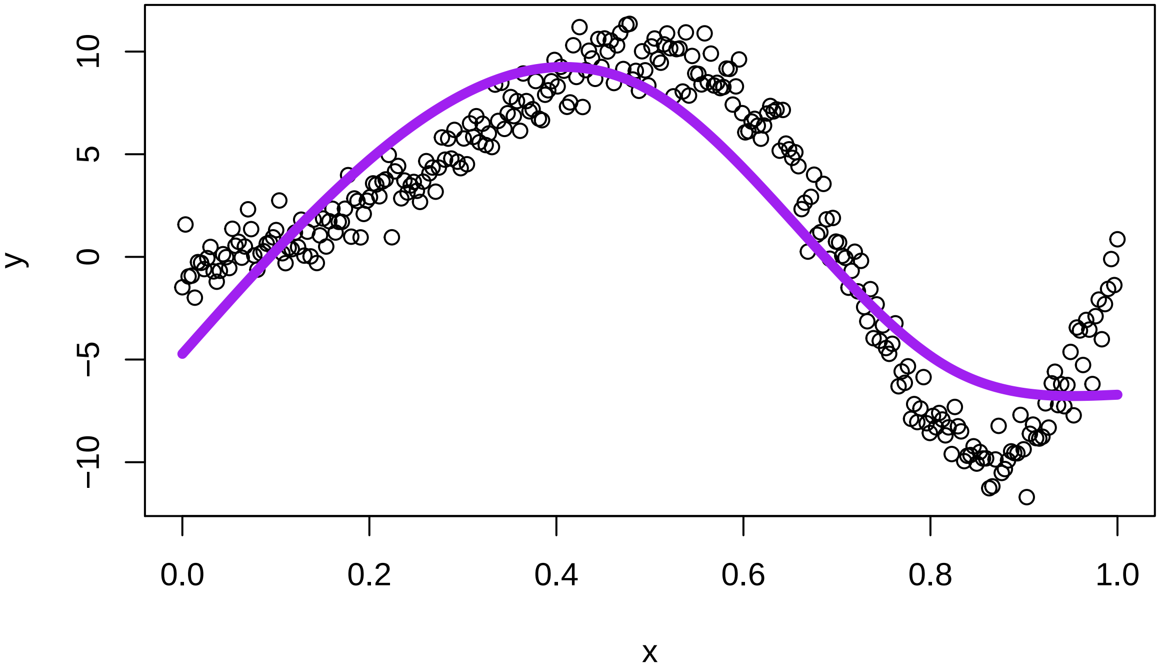


```
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ bs(x, knots = 0.7), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5298 -0.7746  0.0108  0.6896  2.7934
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.4070     0.2847  -1.430    0.154
## bs(x, knots = 0.7)1  -0.7528     0.6829  -1.102    0.271
## bs(x, knots = 0.7)2  29.5988     0.4900  60.401  < 2e-16 ***
## bs(x, knots = 0.7)3 -21.2465     0.5042 -42.138  < 2e-16 ***
## bs(x, knots = 0.7)4   2.7155     0.4379   6.201 1.89e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.148 on 295 degrees of freedom
## Multiple R-squared:  0.9667, Adjusted R-squared:  0.9662
## F-statistic:  2141 on 4 and 295 DF,  p-value: < 2.2e-16
```

Natural Splines

```
m3 <- lm(y ~ ns(x, knots = c(0.5, 0.82)), data = df)
pred <- predict(m3)

plot(x,y)
lines(x, pred, lwd = 5, col = "purple")
```



```
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ bs(x, knots = c(0.5, 0.82)), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90576 -0.91688  0.09772  0.83283  3.14932
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.2153     0.3524   3.448 0.000647 ***
## bs(x, knots = c(0.5, 0.82))1  -6.4639     0.7560  -8.550 6.85e-16 ***
## bs(x, knots = c(0.5, 0.82))2  22.3566     0.5093  43.893  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))3  -8.1253     0.6274 -12.951  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))4 -14.2312     0.5277 -26.966  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))5   0.5685     0.6349   0.895 0.371271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.271 on 294 degrees of freedom
## Multiple R-squared:  0.9593, Adjusted R-squared:  0.9586
## F-statistic:  1385 on 5 and 294 DF,  p-value: < 2.2e-16
```