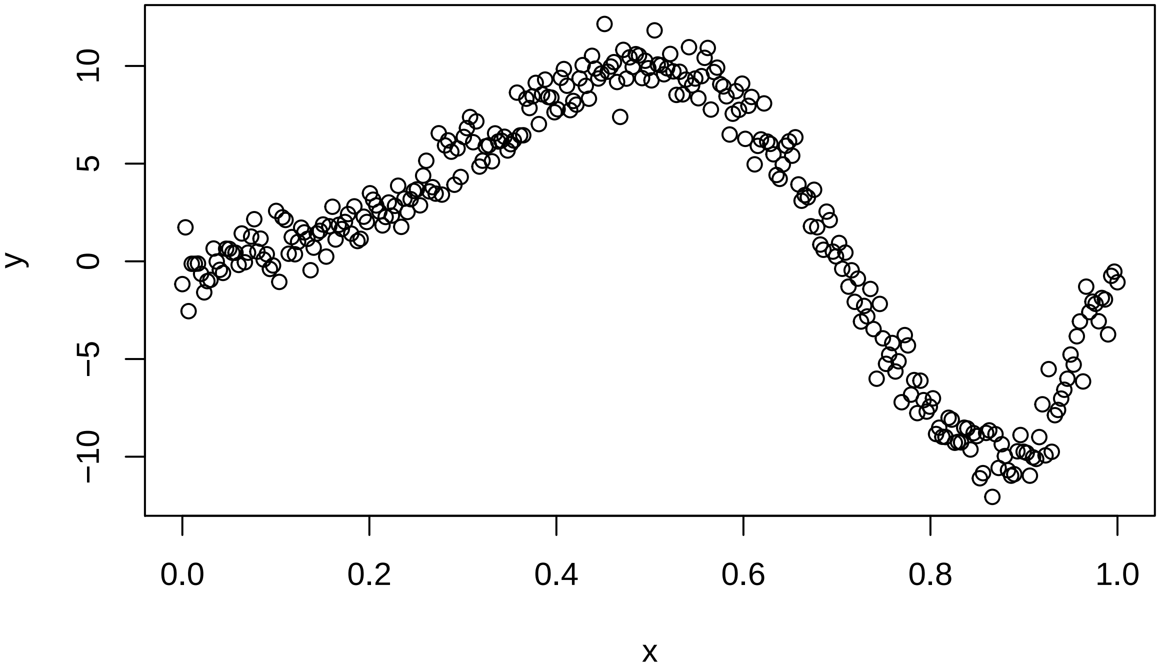
Introduction to Splines

Andy Shen, Devin Francom

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

```
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

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

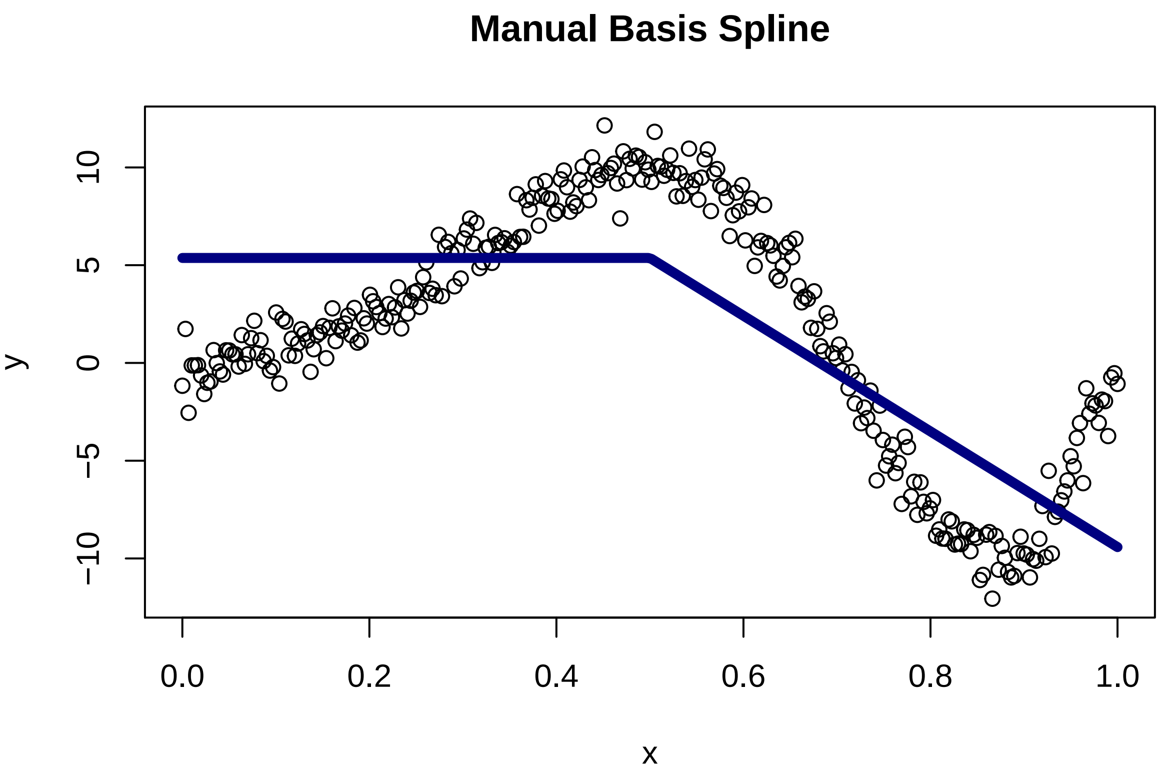
where $s \in \{-1, 1\}$, which we'll call the sign, and t is a value in the domain of x, which we will call a knot. Also, $[a]_+ = max(0, a)$.

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.

```r
t1 <- 0.5 #knot at 0.5
s <- 1
B1 <- rep(NA, length(x))

for(i in 1:length(x)) {
  B1[i] <- max(s * (x[i] - t1), 0)
}
mod <- lm(y ~ B1)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ B1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9201 -3.6652 -0.2204  3.9339  8.7934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.370      0.299   17.96   <2e-16 ***
## B1           -29.582      1.460  -20.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.095 on 298 degrees of freedom
## Multiple R-squared:  0.5794, Adjusted R-squared:  0.578
## F-statistic: 410.5 on 1 and 298 DF,  p-value: < 2.2e-16
```

```r
cf <- mod$coefficients
sq <- x
hs <- (sq - t1)
hs[sq < t1] <- 0
yfit <- cf[1] + cf[2]*hs

plot(x,y, main = "Manual Basis Spline")
lines(x, yfit, type = "l", lwd = 5, col="navy")
```

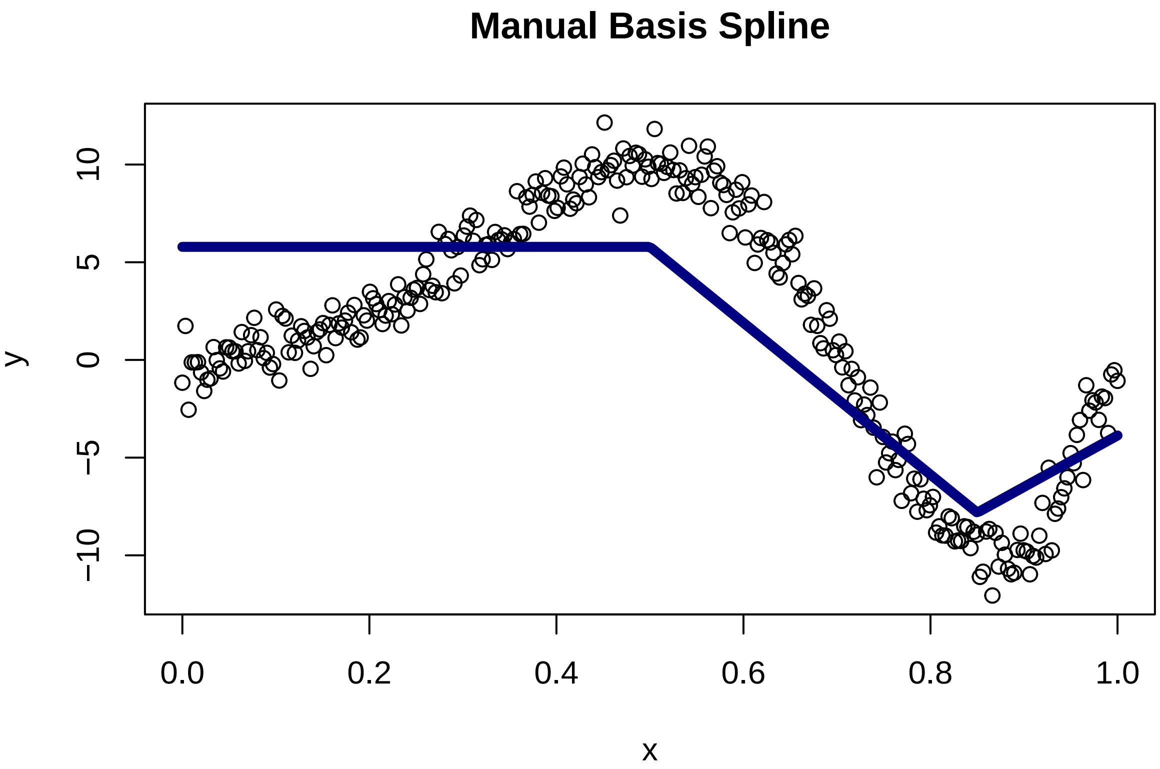


Add another knot

```
t1 <- 0.5 #knot at 0.5
t2 <- 0.85 #another knot at 0.85
s <- 1
B1 <- rep(NA, length(x))
B2 <- B1

for(i in 1:length(x)) {
  B1[i] <- max(s * (x[i] - t1), 0)
  B2[i] <- max(s * (x[i] - t2), 0)
}
mod <- lm(y ~ B1 + B2)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ B1 + B2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3372 -3.2381 -0.0274  3.5185  7.5381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.7874     0.2871   20.16  < 2e-16 ***
## B1          -38.8659     1.9675  -19.75  < 2e-16 ***
## B2           65.1927     9.9377    6.56 2.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 297 degrees of freedom
## Multiple R-squared:  0.6326, Adjusted R-squared:  0.6301
## F-statistic: 255.7 on 2 and 297 DF,  p-value: < 2.2e-16
```

```
cf <- mod$coefficients
sq <- x

hs1 <- (sq - t1)
hs1[sq < t1] <- 0

hs2 <- (sq - t2)
hs2[sq < t2] <- 0

yfit <- cf[1] + cf[2]*hs1 + cf[3]*hs2

plot(x,y, main = "Manual Basis Spline")
lines(x, yfit, type = "l", lwd = 5, col="navy")
```

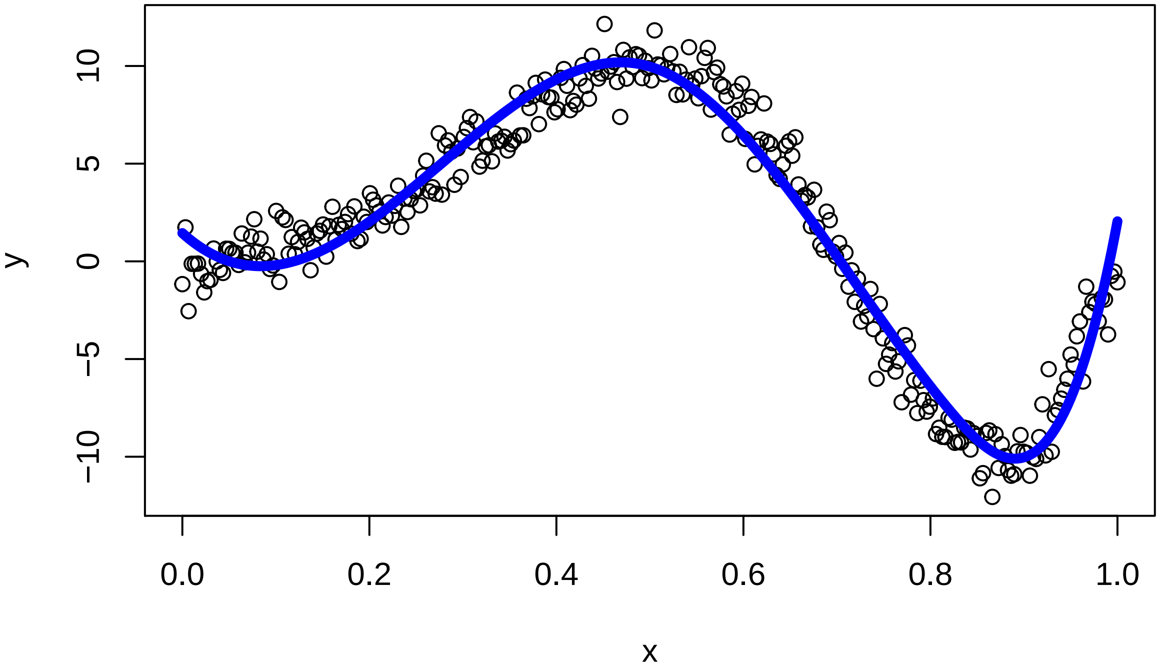


Using the `bs()` Function

2 Knots (Expected)

```
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")
```

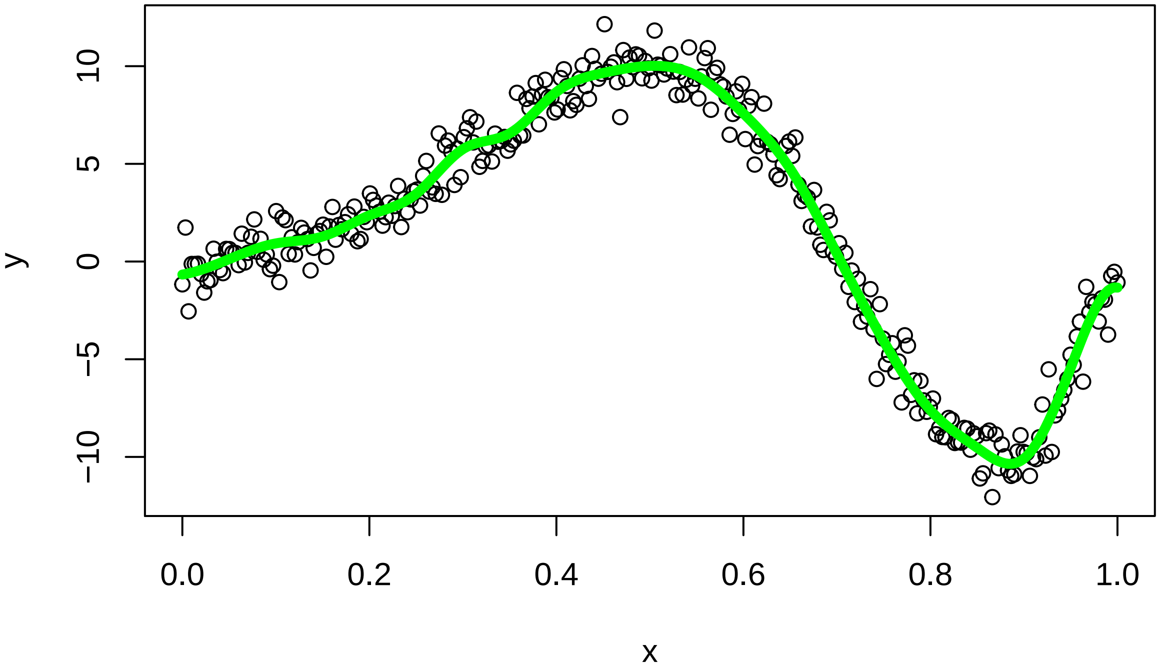


```
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ bs(x, knots = c(0.5, 0.82)), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7223 -0.8534  0.0052  0.8180  3.5369
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.4470     0.3549   4.077 5.87e-05 ***
## bs(x, knots = c(0.5, 0.82))1  -7.1743     0.7613  -9.424  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))2  22.4771     0.5129  43.825  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))3  -8.4905     0.6317 -13.440  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))4 -14.7232     0.5314 -27.706  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))5   0.6037     0.6393   0.944    0.346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 294 degrees of freedom
## Multiple R-squared:  0.9594, Adjusted R-squared:  0.9587
## F-statistic:  1391 on 5 and 294 DF,  p-value: < 2.2e-16
```

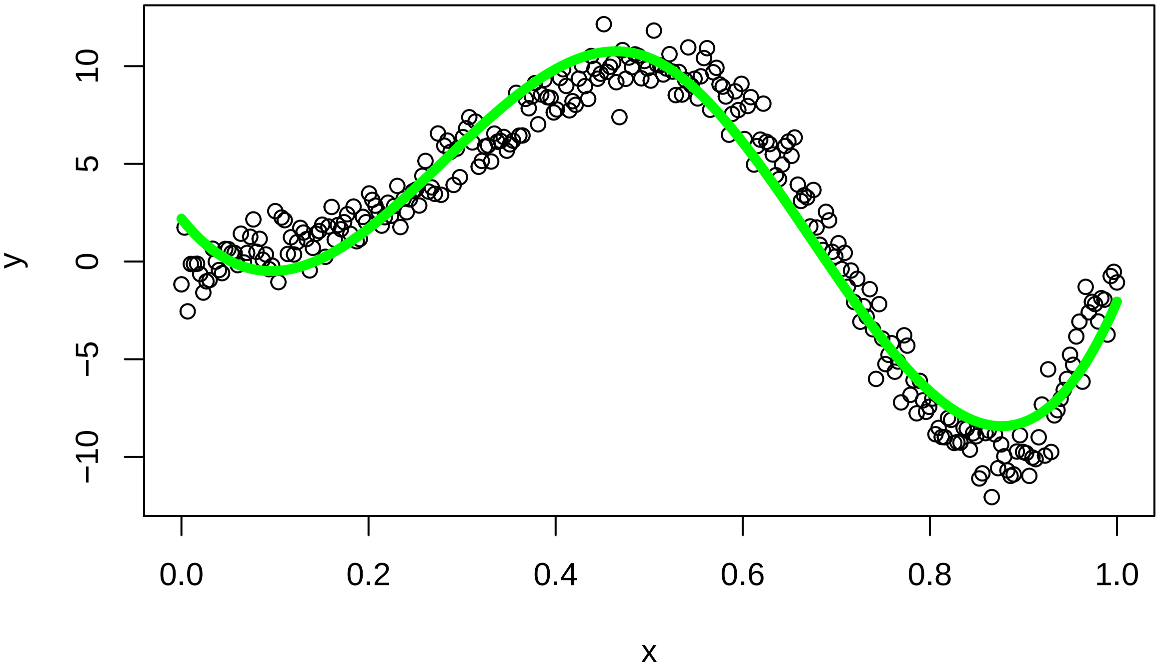
Too Many Knots

```
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 = 0.5), data = df)
pred <- predict(m2)

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

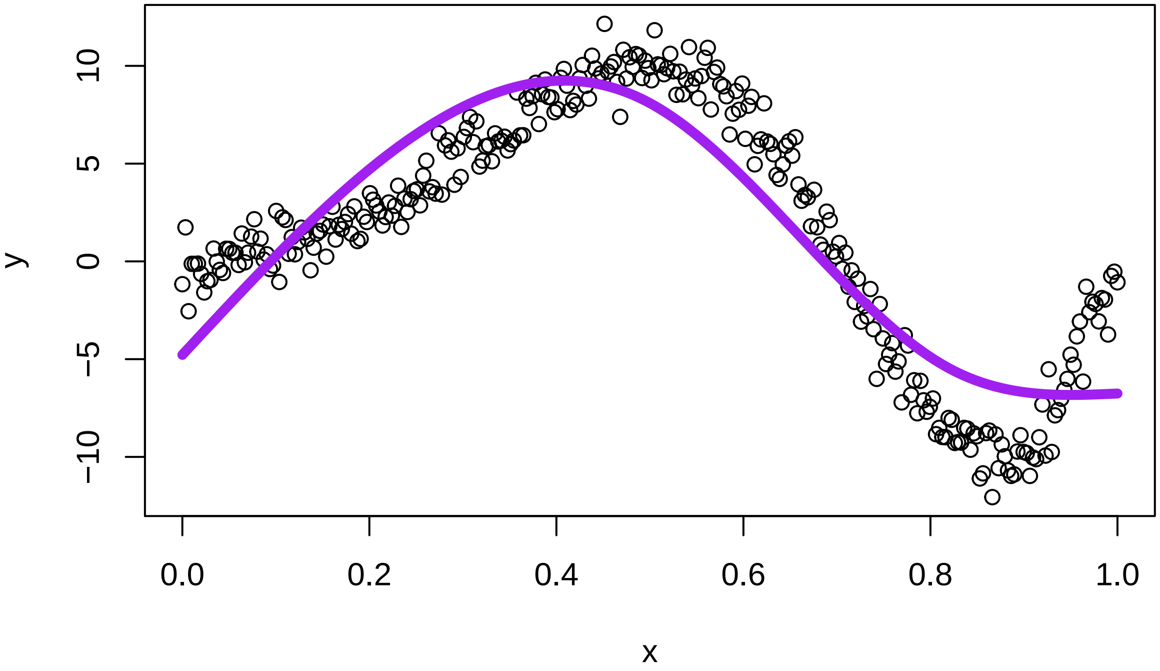


```
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ bs(x, knots = 0.5), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3590 -1.1207 -0.0527  1.1752  3.9509
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.1944     0.4123   5.322 2.03e-07 ***
## bs(x, knots = 0.5)1 -10.0163     0.8471 -11.824  < 2e-16 ***
## bs(x, knots = 0.5)2  32.9247     0.6654  49.484  < 2e-16 ***
## bs(x, knots = 0.5)3 -22.9028     0.7543 -30.364  < 2e-16 ***
## bs(x, knots = 0.5)4  -4.2431     0.5452  -7.782 1.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.516 on 295 degrees of freedom
## Multiple R-squared:  0.9429, Adjusted R-squared:  0.9422
## F-statistic:  1219 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 
## -3.7223 -0.8534  0.0052  0.8180  3.5369 
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    1.4470     0.3549   4.077 5.87e-05 ***
## bs(x, knots = c(0.5, 0.82))1  -7.1743     0.7613  -9.424  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))2  22.4771     0.5129  43.825  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))3  -8.4905     0.6317 -13.440  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))4 -14.7232     0.5314 -27.706  < 2e-16 ***
## bs(x, knots = c(0.5, 0.82))5   0.6037     0.6393   0.944    0.346    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 294 degrees of freedom
## Multiple R-squared:  0.9594, Adjusted R-squared:  0.9587 
## F-statistic:  1391 on 5 and 294 DF,  p-value: < 2.2e-16
```