# Handout 8

### STAT3530 Applied Linear Models

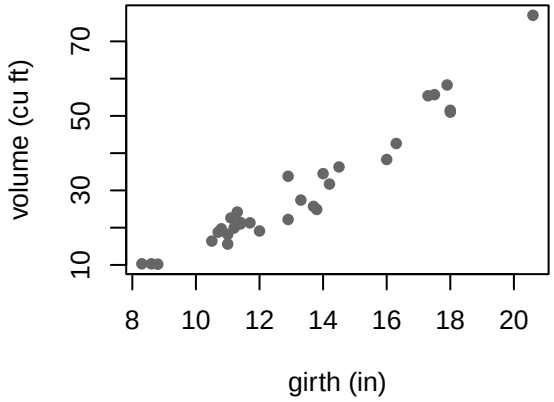
Least squares does not require the model to be linear in $x$ — only in $\beta$. We will consider a model of the form

$$Y_i = \beta_0 + \beta_1 f_1(x_i) + \cdots + \beta_k f_k(x_i) + \epsilon_i$$

to capture nonlinear relationships via the choice of functions $f_1, \ldots, f_k$.

## Polynomial models

Black cherry trees: volume against girth, for 31 trees.




1. Girth is a circumference, so geometry suggests volume $\propto$ _____.
   a. Write an appropriate model and label/identify the functions $f_j$.

$$\text{volume}_i =$$

The linear and nonlinear models are nested, so one can test for curvature.

```
m0 <- lm(Volume ~ Girth, data = trees)
m1 <- lm(Volume ~ Girth + I(Girth^2), data = trees)

anova(m0, m1)$`Pr(>F)`[2]
```

```
[1] 0.000152389
```

2. Using the output above, state $H_0$ in terms of model coefficients and interpret the result in context.

More broadly, a $k^{\text{th}}$ order polynomial is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$$

For example, here is cubic polynomial.

```
m2 <- lm(Volume ~ poly(Girth, degree = 3), data = trees)

anova(m1, m2)$`Pr(>F)`[2]
```

```
[1] 0.3725529
```

3. What hypothesis does this test, and what do you conclude?

### Orthogonal polynomials

Here is the same model fit two ways.

```
# raw powers
fit.r <- lm(Volume ~ poly(Girth, 4, raw = T),
            data = trees)
round(summary(fit.r)$coef[-1, 3:4], 3)
# orthogonalized
fit.o <- lm(Volume ~ poly(Girth, 4),
            data = trees)
round(summary(fit.o)$coef[-1, 3:4], 3)
```

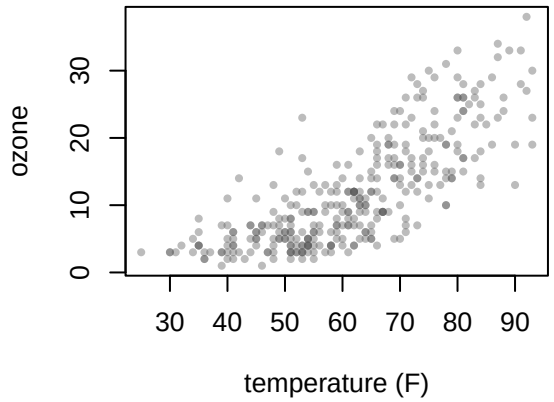
```
                       t value Pr(>|t|)                t value Pr(>|t|)
poly(Girth, 4, raw = T)1   0.769    0.449 poly(Girth, 4)1  25.761    0.000
poly(Girth, 4, raw = T)2  -0.722    0.477 poly(Girth, 4)2   4.317    0.000
poly(Girth, 4, raw = T)3   0.719    0.479 poly(Girth, 4)3   0.897    0.378
poly(Girth, 4, raw = T)4  -0.671    0.508 poly(Girth, 4)4  -0.671    0.508
```

4. The raw powers correlate at 0.994. Not one raw $t$ test is significant; explain why.

5. In the orthogonal version, each $t$ test is the test for adding that degree *last*. Interpret the partial $t$ tests in this model – what do they suggest is the correct polynomial specification?

## Basis functions

Ozone concentration against temperature, 330 days in the Los Angeles basin.




6. The relationship looks roughly linear below about 60°F and steeper above it. Sketch, on the plot, the shape you would want to fit.

7. Cut the range at a **knot** $t = 60$ using indicator functions.

$$I_1(x) = \begin{cases} 1, & x \le t \\ 0, & x > t \end{cases} \qquad I_2(x) = \begin{cases} 1, & x > t \\ 0, & x \le t \end{cases}$$

   a. Evaluate: $I_1(58) =$ _____, $I_2(58) =$ _____, $I_1(71) =$ _____, $I_2(71) =$ _____.

   b. Write the piecewise model using four derived predictors, and identify each one.

$$Y_i =$$

Here is the piecewise linear model fitted in R by manually computing indicators.

```
x <- ozone$temp
I1 <- function(x) as.numeric(x <= 60); I2 <- function(x) as.numeric(x > 60)
x1 <- I1(x); x2 <- x*I1(x); x3 <- I2(x); x4 <- x*I2(x)

free <- lm(O3 ~ x1 + x2 + x3 + x4 - 1, data = ozone)
round(coef(free), 3)
```

```
     x1      x2      x3      x4
 -1.688   0.160 -27.005   0.599
```

8. Using the output above:

   a. Write the equation of the line on each side of the knot.

   $x \leq 60$ : $\qquad\qquad$ $x > 60$ :

   b. Evaluate both at $x = 60$. Do they agree?

9. Require that the two pieces meet at the knot.

$$\underbrace{\beta_0 + \beta_1 t}_{\text{left piece at } t} = \underbrace{\alpha_0 + \alpha_1 t}_{\text{right piece at } t}$$

   Rearrange so that one side is zero, and write the result as a contrast.

$$c^T\beta = 0 \qquad \text{where} \qquad c^T = _____$$

10. Solving the constraint gives the **broken stick** model. Write the model.

$$Y_i =$$

    a. Write the model equation on either side of the knot.

    $x \leq t$ : $\qquad\qquad$ $x > t$ :

    b. Match the coefficients with the original model parametrization written in terms of $\alpha, \beta$.

$$\gamma_0 = _____ \qquad \gamma_1 = _____ \qquad \gamma_2 = _____$$

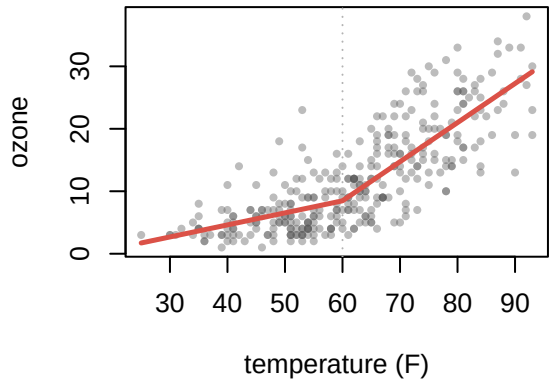
    c. Interpret $\gamma_2$ in the reparametrized model.

Here is the broken stick model fit in R

```
stick <- lm(O3 ~ temp + pmax(temp - 60, 0), data = ozone)

round(coef(stick), 3)
```

```
      (Intercept)              temp pmax(temp - 60, 0)
           -3.100             0.193             0.433
```




11. Using the output above:

    a. Report the slope below 60°F and the slope above it.

    b. Three columns, three dimensions. Reconcile that with the four columns in question 8.

Continuity was a contrast, so it is testable.

```
anova(stick, free)
```

```
Analysis of Variance Table

Model 1: O3 ~ temp + pmax(temp - 60, 0)
Model 2: O3 ~ x1 + x2 + x3 + x4 - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    327 7217.7
2    326 7189.5  1    28.257 1.2813 0.2585
```

12. Using the output above:

    a. State $H_0$ in words and give both degrees of freedom.

    b. Conclusion. Was imposing continuity a reasonable thing to do?

### More knots

With $K$ knots the basis simply gets longer:

$$1, \quad x, \quad (x-t_1)_+, \quad \ldots, \quad (x-t_K)_+$$

```
two <- lm(O3 ~ temp + pmax(temp - 50, 0) + pmax(temp - 60, 0), data = ozone)

round(coef(two), 3)
```

```
      (Intercept)              temp pmax(temp - 50, 0) pmax(temp - 60, 0)
           -0.556             0.132             0.136             0.345
```

13. Using the output above, give the slope on each of the three regions.

$x < 50$ : _____ $50 \le x < 60$ : _____ $x \ge 60$ : _____

### Local basis

Using the previous model, write the row of the design matrix as a function of $x$,

$$p(x)^T = \left[ \begin{array}{ccccc} 1 & x & (x-t_1)_+ & \cdots & (x-t_K)_+ \end{array} \right]$$

so the broken stick model is $Y = X\gamma + \epsilon$ with $X$ having rows $p(x_i)^T$. Now let $\eta_k = f(t_k)$ be the mean at the $k$th knot. Adding the boundary knots $t_0$ and $t_{K+1}$,

$$\eta = E\gamma \quad \text{where} \quad E = \left[ \begin{array}{cccc} p(t_0) & p(t_1) & \cdots & p(t_{K+1}) \end{array} \right]^T$$

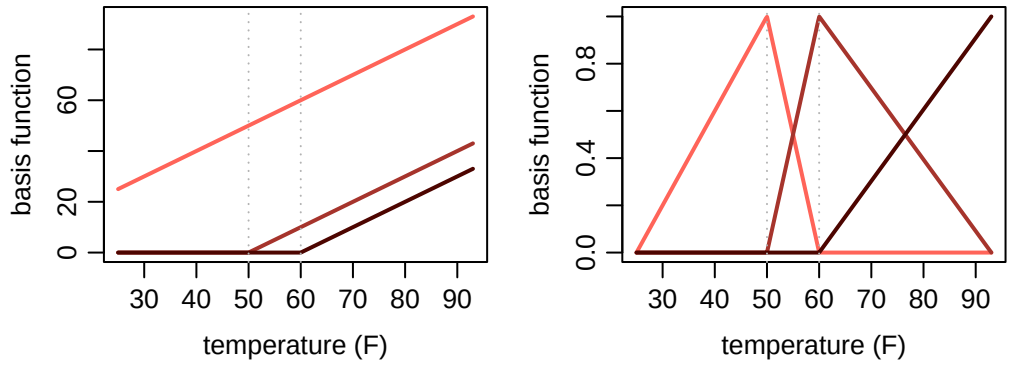
14. For the ozone model with knots at 50 and 60, the boundary knots are the smallest and largest temperatures, 25 and 93.

    a. Fill in $E$.

$$E = \left[ \begin{array}{cccc} 1 & ____ & ____ & ____ \\ 1 & ____ & ____ & ____ \\ 1 & ____ & ____ & ____ \\ 1 & ____ & ____ & ____ \end{array} \right]$$

    b. $E$ is square and invertible, so we can reparametrize the model. Write the change in coordinates and the new model in terms of the new coordinates.

The functions $h_k$ are **basis functions**: each is nonzero over two intervals only.




Any piecewise linear $f$ with knots $t_0, \ldots, t_{K+1}$ can be written $f(x) = \sum_{k=0}^{K+1} c_k h_k(x)$, and by construction $c_k = f(t_k)$.

```
hat <- lm(O3 ~ bs(temp, knots = c(50, 60), degree = 1, intercept = TRUE) - 1,
          data = ozone)

round(coef(hat), 3) |> setNames(c('h0', 'h1', 'h2', 'h3'))
```

```
    h0     h1     h2     h3
 2.740  6.036  8.712 28.942
```

15. Using the output above:

    a. There are two knots but four coefficients. What are the other two, and what does each coefficient estimate?

    b. The slope on $[t_{k-1}, t_k]$ is $(c_k - c_{k-1})/(t_k - t_{k-1})$. Give all three slopes.

25 to 50 : _____ 50 to 60 : _____ 60 to 93 : _____

### Level shifts via indicator bases

Consider indicator basis functions:

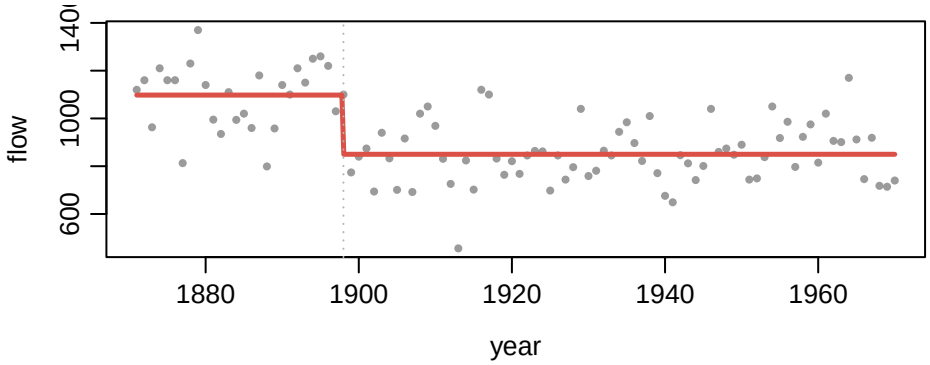
$$h_k^{(0)}(x) = I(t_k < x \leq t_{k+1})$$

and the same machinery gives a **level shift** model. Annual flow of the Nile, with one level change at 1898:

```
era   <- cut(nile$year, c(1871, 1898, 1970), include.lowest = TRUE, dig.lab = 4)
level <- lm(flow ~ era - 1, data = nile)

round(coef(level), 1)
```

```
era[1871,1898] era(1898,1970]
        1097.7          850.0
```




16. Using the output above, interpret each coefficient estimate.

### B-splines

The piecewise linear model can be recovered by blending two neighboring indicators with linear weights:

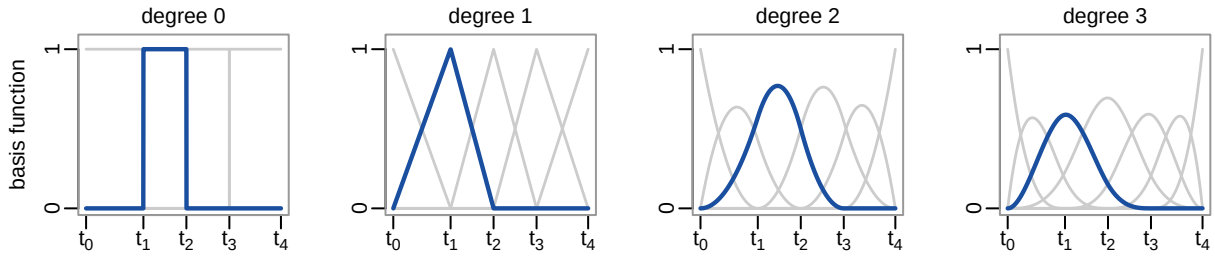
$$h_k^{(1)}(x) =$$

Iterating gives increasingly smooth bases, the **b-splines** of degree $d$:

$$h_k^{(d)}(x) =$$

Repeating the boundary knots to clamp the basis gives $k = 0, \ldots, K + d$.

Here are b-spline bases for degrees $d = 0, 1, 2, 3$.




17. Each panel uses the same three interior knots, with one basis function picked out in blue.
    a. Give the number of basis functions for $K$ interior knots at degree $d$.

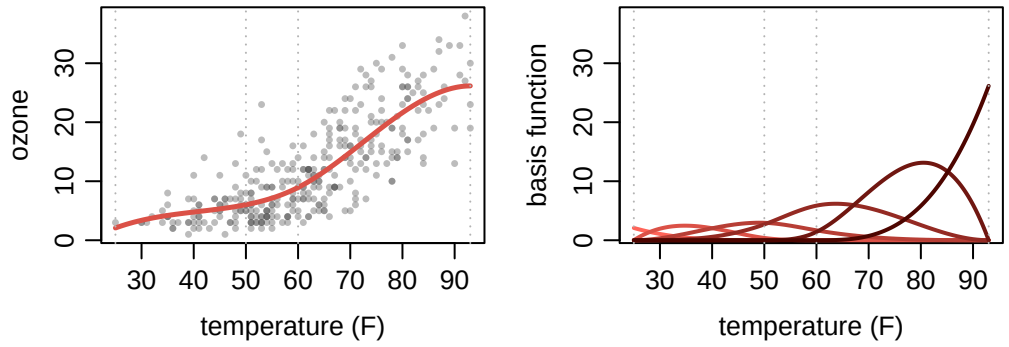
A nonlinear relationship is estimated by a linear model in the basis expansion,

$$Y_i = \beta_0 h_0^{(d)}(x_i) + \beta_1 h_1^{(d)}(x_i) + \cdots + \beta_{K+d} h_{K+d}^{(d)}(x_i) + \epsilon_i$$

```
H   <- bs(ozone$temp, knots = c(50, 60), degree = 3, intercept = TRUE)
cub <- lm(O3 ~ H - 1, data = ozone)

round(coef(cub), 2) |> setNames(paste0('h', 0:5))
```

```
   h0    h1    h2    h3    h4    h5
 2.08  4.75  4.50 10.22 26.08 26.16
```



18. Consider two knots at degree 3. How many coefficients should there be, and does the output agree?

Rather than naming knots, `df` asks for a number of basis functions and lets `bs()` place the knots at quantiles of the predictor.

```
bs(mcycle$times, df = 8, intercept = TRUE) |> attr('knots')
```

```
[1] 14.68 18.44 26.52 36.20
```

```
quantile(mcycle$times, c(0.2, 0.4, 0.6, 0.8))
```

```
  20%   40%   60%   80%
14.68 18.44 26.52 36.20
```

The models are nested, so we can compare via an $F$ test to select the number of knots.

```
anova(lm(accel ~ bs(times, df = 5),  data = mcycle),
      lm(accel ~ bs(times, df = 8),  data = mcycle),
      lm(accel ~ bs(times, df = 12), data = mcycle))
```

```
Analysis of Variance Table

Model 1: accel ~ bs(times, df = 5)
Model 2: accel ~ bs(times, df = 8)
Model 3: accel ~ bs(times, df = 12)
  Res.Df    RSS Df Sum of Sq       F  Pr(>F)
1    127 134906
2    124  65319  3     69587 45.7593 < 2e-16 ***
3    120  60829  4      4490  2.2142 0.07145 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

19. Interpret the results of the test.