

# Transformations

ST552 Lecture 20

Charlotte Wickham

Feb 22, 2016

# Transforming the response

**Motivation:** generally we are hunting for a transformation that makes the relationship simpler.

We might believe the relationship with the explanatory is linear only after a transformation of the response,

$$E(g(Y)) = X\beta$$

we might also hope on this transformed scale

$$\text{Var}(g(Y)) = \sigma^2 I$$

What is a good  $g$ ?

# Transforming the predictor

**Motivation:** we acknowledge that straight lines might not be appropriate and want to estimate something more flexible. For example, we believe the model is something like

$$Y = f_1(X_1) + f_2(X_2) + \dots + f_p(X_p) + \epsilon$$

or even,

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon$$

We are generally interested in estimating  $f$ .

# Transforming the response

In general, transformations make interpretation harder.

We usually want to make statements about the response (not the transformed) response.

Predicted values are easily back-transformed, as well as the endpoints of confidence intervals.

Parameters often **do not** have nice interpretations on the backtransformed scale.

## Special case: Log transformed response

Our fitted model on the transformed scale, predicts:

$$\log \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}$$

If we backtransform, by taking exponential of both sides,

$$\hat{y}_i = \exp \hat{\beta}_0 \exp (\hat{\beta}_1 x_{i1}) \dots \exp (\hat{\beta}_p x_{ip})$$

So, an increase in  $x_1$  of one unit, will result in the predicted response being multiplied by  $\exp(\beta_1)$ .

If we are willing to assume that on the transformed scale the distribution of the response is **symmetric**,

$$\text{Median}(\log(Y)) = E(\log(Y)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

and back-transforming gives,

$$\exp(\text{Median}(\log(Y))) = \text{Median}(Y) = \exp(\beta_0) \exp(\beta_1 x_1) \dots \exp(\beta_p x_p)$$

So, an increase in  $x_1$  of one unit, will result in the median response being multiplied by  $\exp(\beta_1)$ .

(Remember for monotone functions

$\text{Median}(f(Y)) = f(\text{Median}(Y))$ , but  $E(f(Y)) \neq f(E(Y))$  in general)

# Example

```
library(faraway)
data(case0301, package = "Sleuth3")
head(case0301, 2)
```

```
##   Rainfall Treatment
## 1   1202.6  Unseeded
## 2    830.1  Unseeded
```

```
summary(lm(log(Rainfall) ~ Treatment, data = case0301))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.13419    0.31787  16.1519 < 2e-16
## TreatmentUnseeded -1.14378    0.44953  -2.5444  0.01408
##
## n = 52, p = 2, Residual SE = 1.62082, R-Squared = 0.11
```

It is estimated the median rainfall for unseeded clouds is 0.32 times the median rainfall for seeded clouds.

(Assuming log rainfall is symmetric around its mean)

# Box-Cox transformations

Assume, the response is positive, and

$$g(Y) = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

and that  $g$  is of the form

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Estimate  $\lambda$  with maximum likelihood.

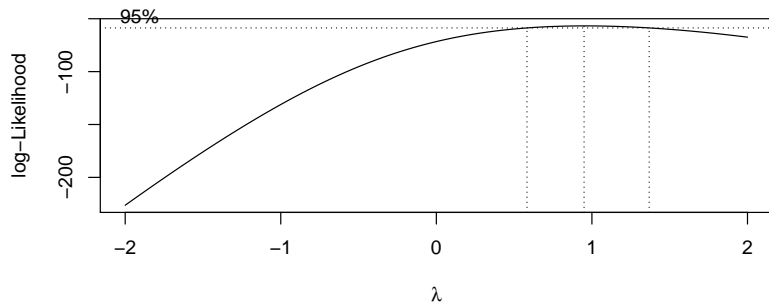
For prediction, pick  $\lambda$  as the MLE.

For explanation, pick “nice”  $\lambda$  within 95% CI.



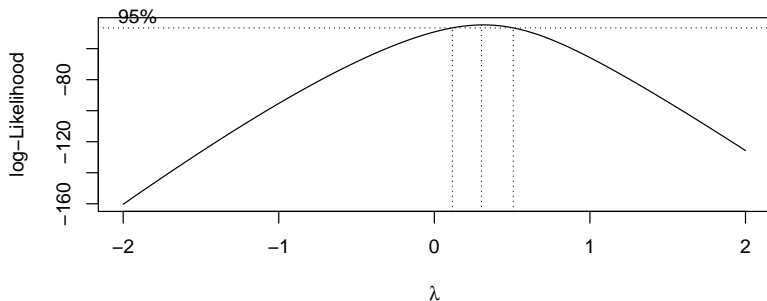
# Example:

```
library(MASS)
data(savings, package = "faraway")
lmod <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
boxcox(lmod, plotit = TRUE)
```



## Your turn:

```
data(gala, package = "faraway")  
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,  
  data = gala)  
boxcox(lmod, plotit = TRUE)
```



# Transforming the predictors

A very general approach is to let the function for each explanatory be represented by a finite set of basis functions. For example, for a single explanatory,  $X$ ,

$$f(X) = \sum_{k=1}^K \beta_k f_k(X)$$

where  $f_k$  are the known basis functions, and  $\beta_k$  the unknown basis coefficients.

Then

$$y_i = f(X_i) + \epsilon_i$$

$$y_i = \beta_1 f_1(X_i) + \dots + \beta_K f_K(X_i) + \epsilon_i$$

$$Y = X' \beta + \epsilon$$

where the columns of  $X'$  are  $f_1(X)$ ,  $f_2(X)$  and we can find the  $\beta$  with the usual least squares approach.

Your turn:

What are the columns in the design matrix for the model:

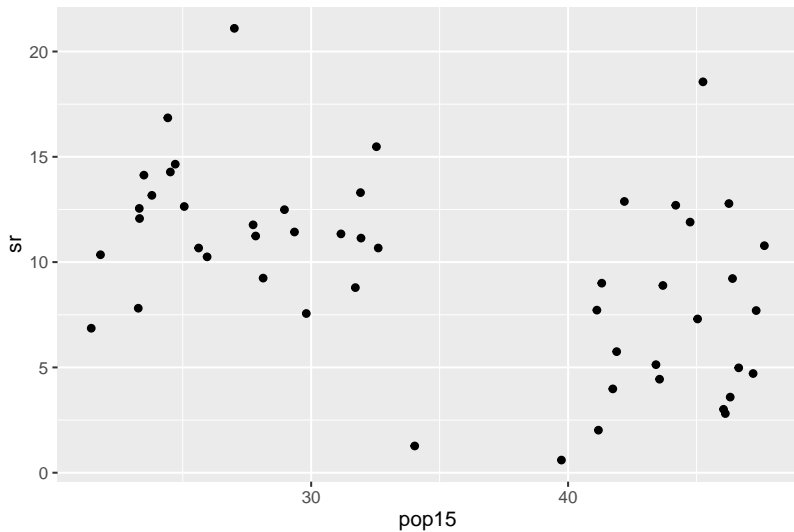
$$y_i = \beta_1 1_{\{X_i < 5\}} + \beta_2 1_{\{X_i \geq 5\}} + \beta_3 X_i 1_{\{X_i < 5\}} + \beta_4 X_i 1_{\{X_i \geq 5\}} + \epsilon_i$$

where  $X_i = i, i = 1, \dots, 10$ ?

What do the functions  $f_k(), k = 1, \dots, 4$  look like?

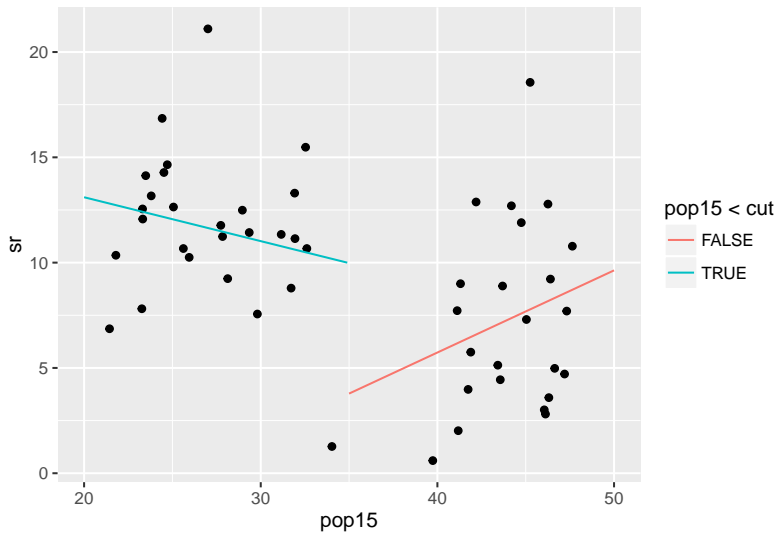


# Example: subset regression



```
cut <- 35
X <- with(savings, cbind(
  as.numeric(pop15 < cut),
  as.numeric(pop15 >= cut),
  pop15 * (pop15 < cut),
  pop15 * (pop15 >= cut)))

lmod <- lm(sr ~ X - 1,
  data = savings)
summary(lmod)
```



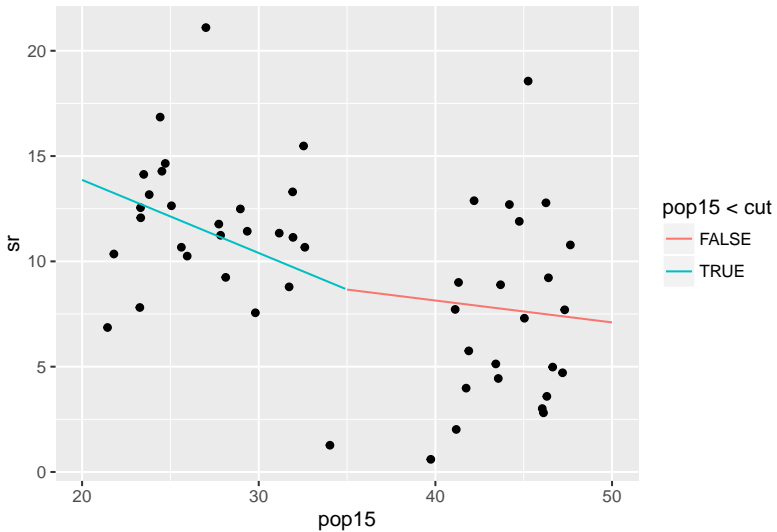


Broken stick:

$$f_1(x) = \begin{cases} c - x & \text{if } x < c \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(x) = \begin{cases} 0 & \text{if } x < c \\ x - c & \text{otherwise} \end{cases}$$

$$y_i = \beta_0 + \beta_1 f_1(X_i) + \beta_2 f_2(X_i) + \epsilon_i$$



##

# Polynomials

- ▶ **Polynomials:**

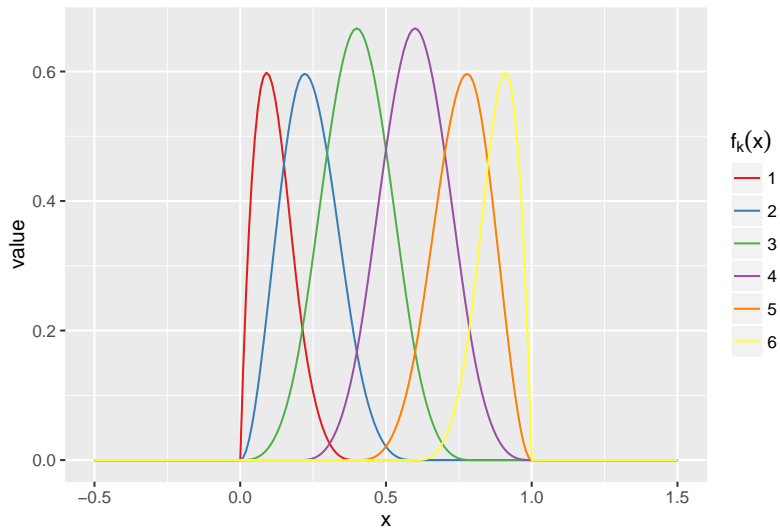
$$f_k(x) = x^k, \quad k = 1, \dots, K$$

- ▶ **Orthogonal polynomials**

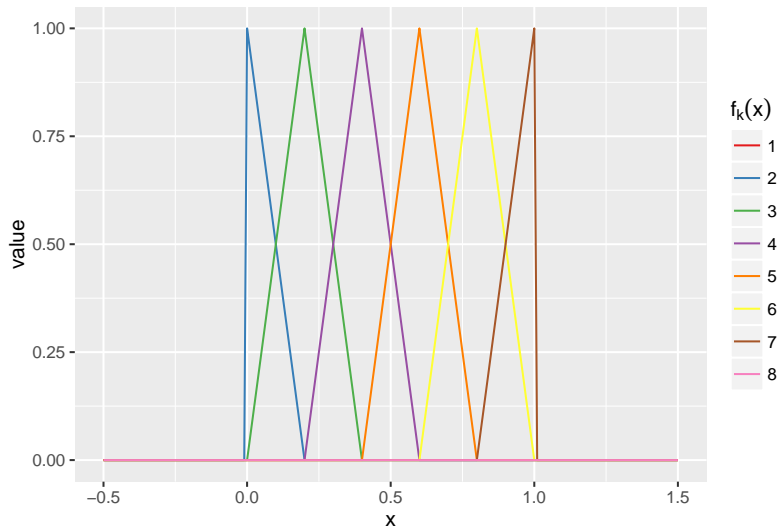
- ▶ **Response surface**, of degree  $d$

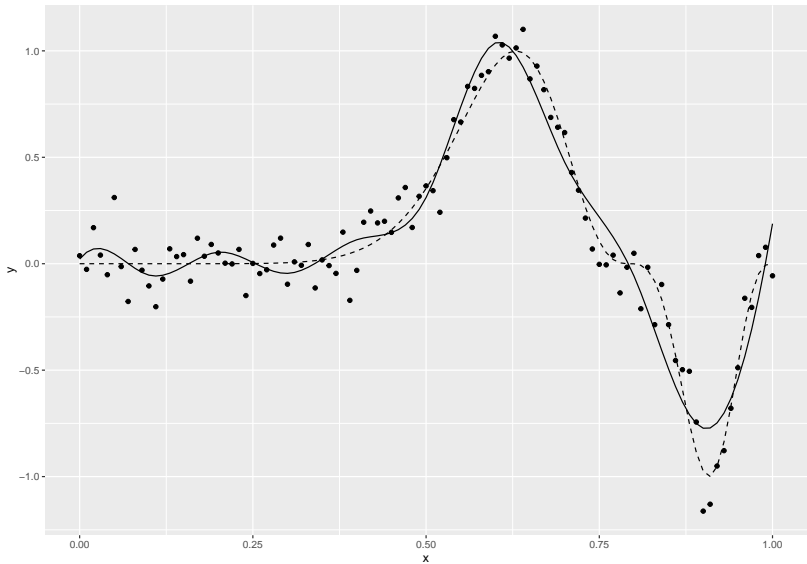
$$f_{kl}(x, z) = x^k z^l, \quad k, l \geq 0 \text{ s.t. } k + l = d$$

# Cubic Splines



# Linear splines





- ▶ **Smoothing splines:** have a large set of basis functions, but penalize against wiggleness
- ▶ **Generalized Additive Models:** simultaneously estimate

$$y_i = f(x_{i1}) + g(x_{i2}) + \dots + \epsilon_i$$

- ▶ The parameters in these regressions no longer have nice interpretations. The best way to present the results is a plot of the estimated function for each  $X$ , (or surfaces if variables interact),
- ▶ The significance of a variable can still be assessed with an Extra Sum of Squares F-test, comparing to a model without any of the terms relating to a particular variable.