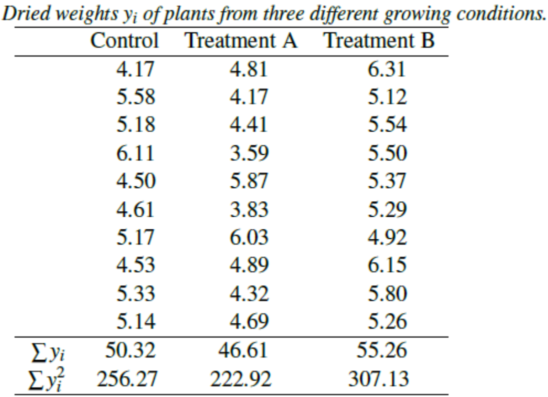
Lecture 7: Analysis of variance

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

ANOVA

Analysis of variance is the term used for statistical methods for comparing means of groups of continuous observations where the groups are defined by the levels of factors.

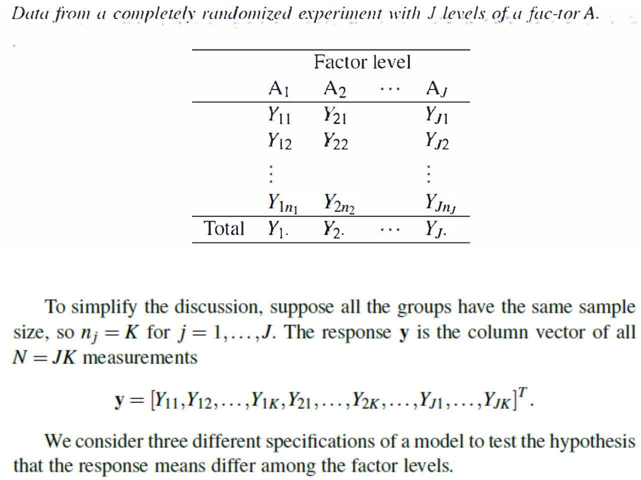


Dried weights y_i of plants from three different growing conditions.

	Control	Treatment A	Treatment B
	4.17	4.81	6.31
	5.58	4.17	5.12
	5.18	4.41	5.54
	6.11	3.59	5.50
	4.50	5.87	5.37
	4.61	3.83	5.29
	5.17	6.03	4.92
	4.53	4.89	6.15
	5.33	4.32	5.80
	5.14	4.69	5.26
$\sum y_i$	50.32	46.61	55.26
$\sum y_i^2$	256.27	222.92	307.13

One-factor ANOVA

If experimental units are randomly allocated to groups corresponding to J levels of a factor, this is called a **completely randomized experiment**.



Data from a completely randomized experiment with J levels of a fac-tor A.

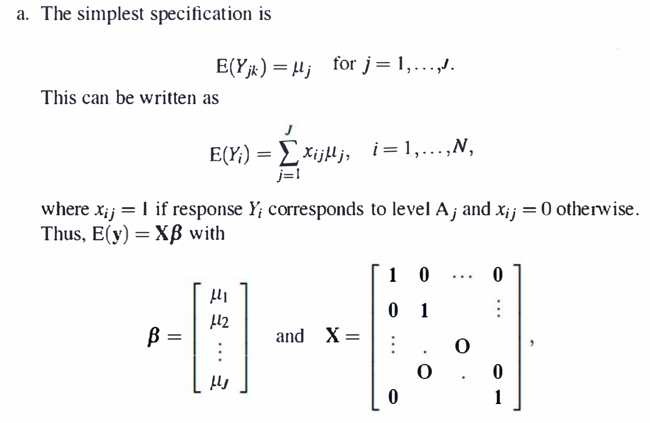
	Factor level			
	A$_1$	A$_2$	$\cdots$	A$_J$
	Y_{11}	Y_{21}		Y_{J1}
	Y_{12}	Y_{22}		Y_{J2}
	$\vdots$			$\vdots$
	Y_{1n_1}	Y_{2n_2}		Y_{Jn_J}
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{J.}$

To simplify the discussion, suppose all the groups have the same sample size, so $n_j = K$ for $j = 1, \ldots, J$. The response $\mathbf{y}$ is the column vector of all $N = JK$ measurements

$$\mathbf{y} = [Y_{11}, Y_{12}, \ldots, Y_{1K}, Y_{21}, \ldots, Y_{2K}, \ldots, Y_{J1}, \ldots, Y_{JK}]^T.$$

We consider three different specifications of a model to test the hypothesis that the response means differ among the factor levels.

Model 1



a. The simplest specification is

$$\mathrm{E}(Y_{jk}) = \mu_j \quad \text{for } j = 1,\ldots,J.$$

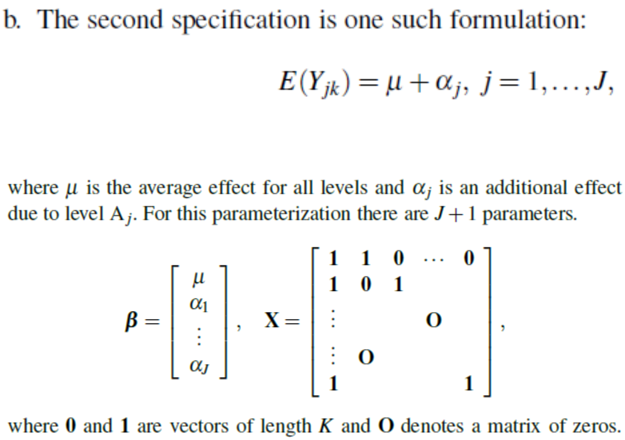
This can be written as

$$\mathrm{E}(Y_i) = \sum_{j=1}^{J} x_{ij}\mu_j, \quad i = 1,\ldots,N,$$

where $x_{ij} = 1$ if response Y_i corresponds to level A_j and $x_{ij} = 0$ otherwise. Thus, $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ with

$$\boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & . & O & \\ & O & . & 0 \\ 0 & & & 1 \end{bmatrix},$$

Model 2



b. The second specification is one such formulation:

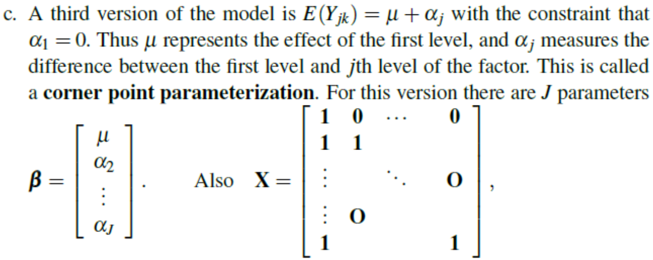
$$E(Y_{jk}) = \mu + \alpha_j, \ j = 1, \ldots, J,$$

where μ is the average effect for all levels and α_j is an additional effect due to level A_j. For this parameterization there are $J+1$ parameters.

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_J \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & & \\ \vdots & & & \mathbf{O} & \\ \vdots & \mathbf{O} & & & \\ \mathbf{1} & & & & \mathbf{1} \end{bmatrix},$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors of length K and $\mathbf{O}$ denotes a matrix of zeros.

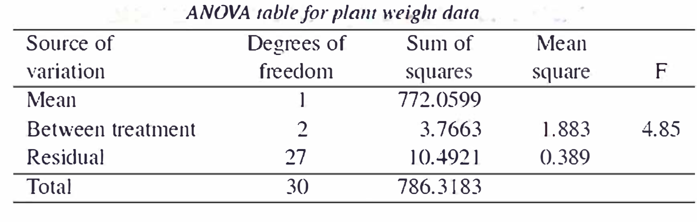
Model 3



c. A third version of the model is $E(Y_{jk}) = \mu + \alpha_j$ with the constraint that $\alpha_1 = 0$. Thus μ represents the effect of the first level, and α_j measures the difference between the first level and jth level of the factor. This is called a **corner point parameterization**. For this version there are J parameters

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_J \end{bmatrix}. \qquad \text{Also} \quad \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & & \\ \vdots & & \ddots & \mathbf{O} \\ \vdots & \mathbf{O} & & \\ \mathbf{1} & & & \mathbf{1} \end{bmatrix},$$

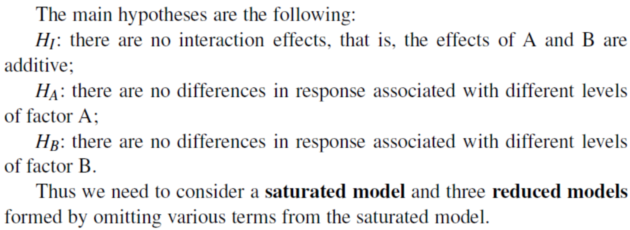
ANOVA Table



ANOVA table for plant weight data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Mean	1	772.0599		
Between treatment	2	3.7663	1.883	4.85
Residual	27	10.4921	0.389	
Total	30	786.3183		

Two-factor analysis of variance



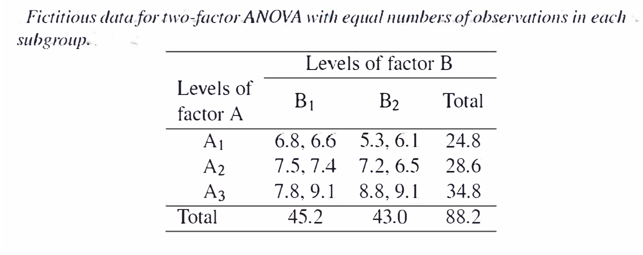
The main hypotheses are the following:

H_I: there are no interaction effects, that is, the effects of A and B are additive;

H_A: there are no differences in response associated with different levels of factor A;

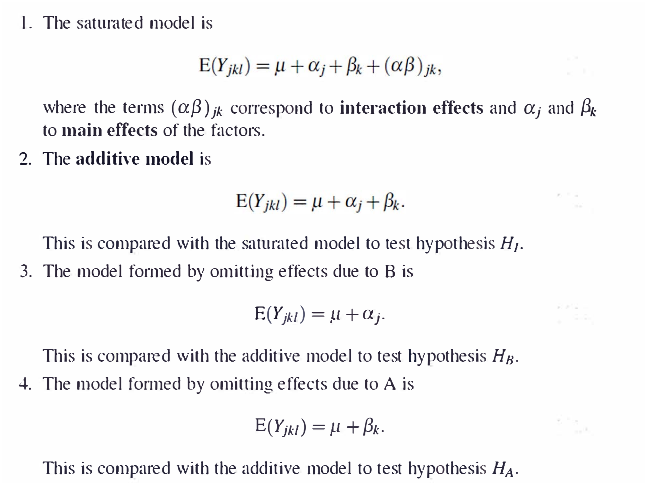
H_B: there are no differences in response associated with different levels of factor B.

Thus we need to consider a **saturated model** and three **reduced models** formed by omitting various terms from the saturated model.



Fictitious data for two-factor ANOVA with equal numbers of observations in each subgroup.

Levels of factor A	Levels of factor B		
	B_1	B_2	Total
A_1	6.8, 6.6	5.3, 6.1	24.8
A_2	7.5, 7.4	7.2, 6.5	28.6
A_3	7.8, 9.1	8.8, 9.1	34.8
Total	45.2	43.0	88.2



1. The saturated model is

$$\mathrm{E}(Y_{jkl}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

   where the terms $(\alpha\beta)_{jk}$ correspond to **interaction effects** and α_j and β_k to **main effects** of the factors.

2. The **additive model** is

$$\mathrm{E}(Y_{jkl}) = \mu + \alpha_j + \beta_k.$$

   This is compared with the saturated model to test hypothesis H_I.

3. The model formed by omitting effects due to B is

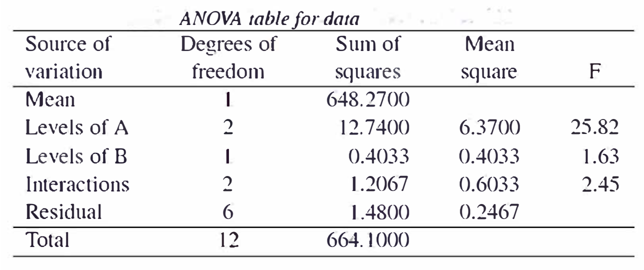
$$\mathrm{E}(Y_{jkl}) = \mu + \alpha_j.$$

   This is compared with the additive model to test hypothesis H_B.

4. The model formed by omitting effects due to A is

$$\mathrm{E}(Y_{jkl}) = \mu + \beta_k.$$

   This is compared with the additive model to test hypothesis H_A.

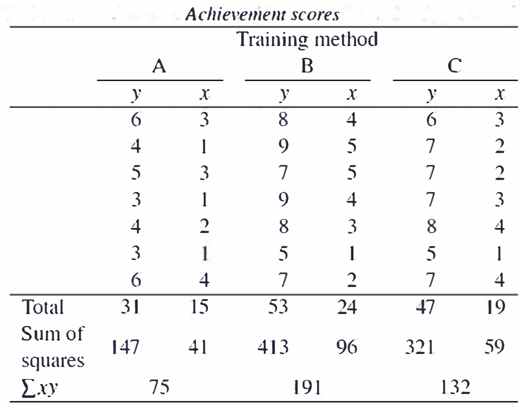


ANOVA table for data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Mean	1	648.2700		
Levels of A	2	12.7400	6.3700	25.82
Levels of B	1	0.4033	0.4033	1.63
Interactions	2	1.2067	0.6033	2.45
Residual	6	1.4800	0.2467	
Total	12	664.1000		

ANCOVA

Analysis of covariance is the term used for models in which some of the explanatory variables are dummy variables representing factor levels and others are continuous measurements called covariates. As with ANOVA, we are interested in comparing means of subgroups defined by factor levels, but recognizing that the covariates may also affect the responses, we compare the means after **adjustment for covariate effects.**



Achievement scores

	Training method					
	A		B		C	
	y	x	y	x	y	x
	6	3	8	4	6	3
	4	1	9	5	7	2
	5	3	7	5	7	2
	3	1	9	4	7	3
	4	2	8	3	8	4
	3	1	5	1	5	1
	6	4	7	2	7	4
Total	31	15	53	24	47	19
Sum of squares	147	41	413	96	321	59
$\sum xy$	75		191		132	

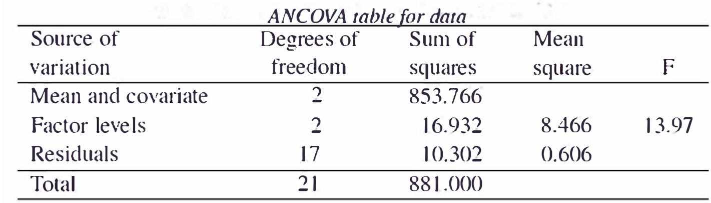
To test the hypothesis that there are no differences in mean achievement scores among the three training methods, after adjustment for initial aptitude, we compare the saturated model

$$E(Y_{jk}) = \mu_j + \gamma x_{jk},$$

with the reduced model

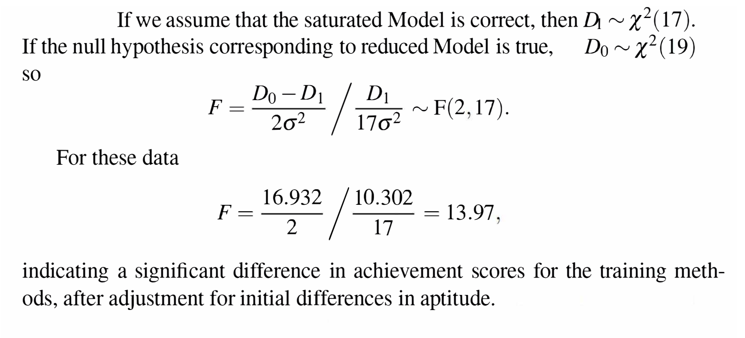
$$E(Y_{jk}) = \mu + \gamma x_{jk},$$

for $j = 1, 2, 3$, $k = 1, \ldots, 7$.



ANCOVA table for data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Mean and covariate	2	853.766		
Factor levels	2	16.932	8.466	13.97
Residuals	17	10.302	0.606	
Total	21	881.000		



If we assume that the saturated Model is correct, then $D_1 \sim \chi^2(17)$. If the null hypothesis corresponding to reduced Model is true, $D_0 \sim \chi^2(19)$ so

$$F = \frac{D_0 - D_1}{2\sigma^2} \bigg/ \frac{D_1}{17\sigma^2} \sim \mathrm{F}(2, 17).$$

For these data

$$F = \frac{16.932}{2} \bigg/ \frac{10.302}{17} = 13.97,$$

indicating a significant difference in achievement scores for the training methods, after adjustment for initial differences in aptitude.

Polynomial Regression

Regression models play an important role in many data analyses, providing prediction and classification rules, and data analytic tools for understanding the importance of different inputs. Although attractively simple, the traditional linear model often fails in these situations: in real life, effects are often not linear.

The standard linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

The polynomial regression

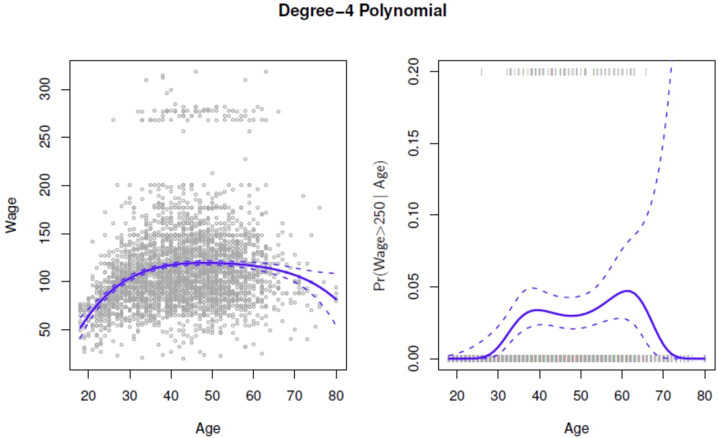
$$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i.$$

Generally speaking, **it is unusual to use d greater than 3 or 4** because for large values of d, the polynomial curve can become overly flexible and can take on some very strange shapes. This is especially true near the boundary of the X variable

- Create new variables $X_1 = X$, $X_2 = X^2$, etc and then treat as multiple linear regression.
- Not really interested in the coefficients; more interested in the fitted function values at any value x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4.$$

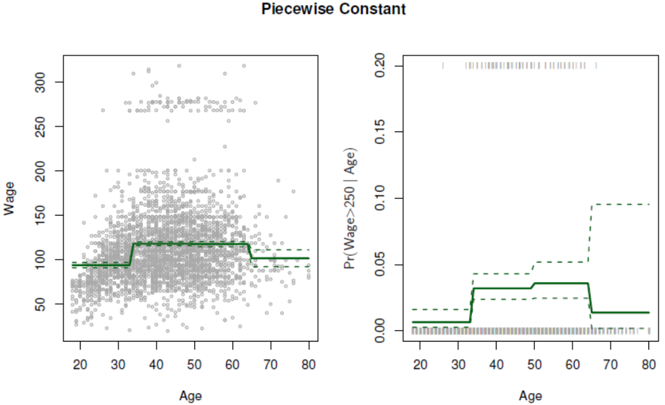
- Since $\hat{f}(x_0)$ is a linear function of the $\hat{\beta}_\ell$, can get a simple expression for *pointwise-variances* $\mathrm{Var}[\hat{f}(x_0)]$ at any value x_0. In the figure we have computed the fit and pointwise standard errors on a grid of values for x_0. We show $\hat{f}(x_0) \pm 2 \cdot \mathrm{se}[\hat{f}(x_0)]$.
- We either fix the degree d at some reasonably low value, else use cross-validation to choose d.


Degree−4 Polynomial
Wage
300
250
200
150
100
50
20
30
40
50
60
70
80
Age
Pr(Wage>250 | Age)
0.20
0.15
0.10
0.05
0.00
20
30
40
50
60
70
80
Age

Step functions

Another way of creating transformations of a variable — cut the variable into distinct regions.

$$C_1(X) = I(X < 35), \quad C_2(X) = I(35 \leq X < 50), \ldots, C_3(X) = I(X \geq 65)$$



Step functions

- Easy to work with. Creates a series of dummy variables representing each group.
- Useful way of creating interactions that are easy to interpret. For example, interaction effect of `Year` and `Age`:

$$I(\texttt{Year} < 2005) \cdot \texttt{Age}, \quad I(\texttt{Year} \geq 2005) \cdot \texttt{Age}$$

  would allow for different linear functions in each age category.
- In R: `I(year < 2005)` or `cut(age, c(18, 25, 40, 65, 90))`.
- Choice of cutpoints or *knots* can be problematic. For creating nonlinearities, smoother alternatives such as *splines* are available.

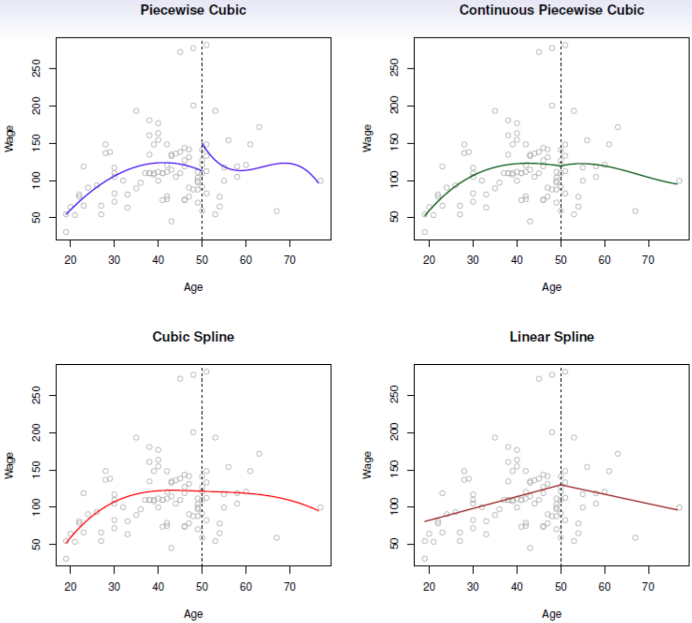
Piecewise Polynomials

- Instead of a single polynomial in X over its whole domain, we can rather use different polynomials in regions defined by knots. E.g. (see figure)

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

- Better to add constraints to the polynomials, e.g. continuity.
- *Splines* have the “maximum” amount of continuity.

Piecewise Polynomials



Linear Splines

A linear spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise linear polynomial continuous at each knot.

We can represent this model as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

where the b_k are *basis functions.*

$$\begin{aligned} b_1(x_i) &= x_i \\ b_{k+1}(x_i) &= (x_i - \xi_k)_+, \quad k = 1, \ldots, K \end{aligned}$$

Here the $()_+$ means *positive part*; i.e.

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

Cubic Splines

A cubic spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.

Again we can represent this model with truncated power basis functions

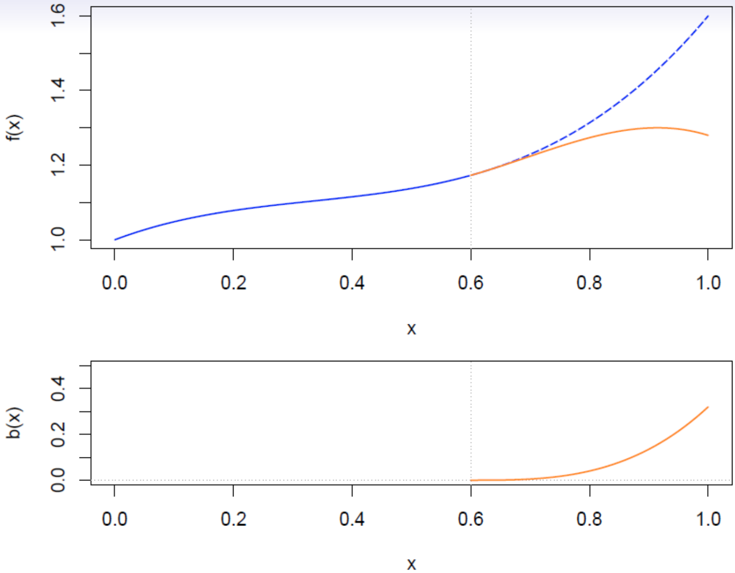
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, \quad k = 1, \ldots, K \end{aligned}$$

where

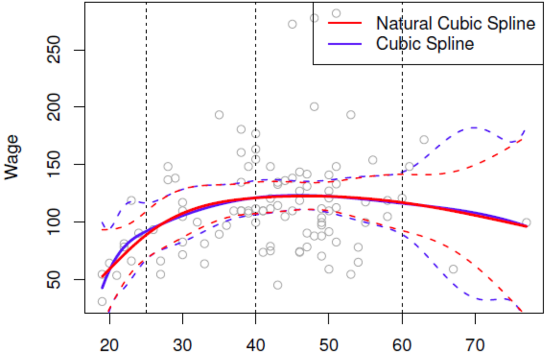
$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

Cubic Splines



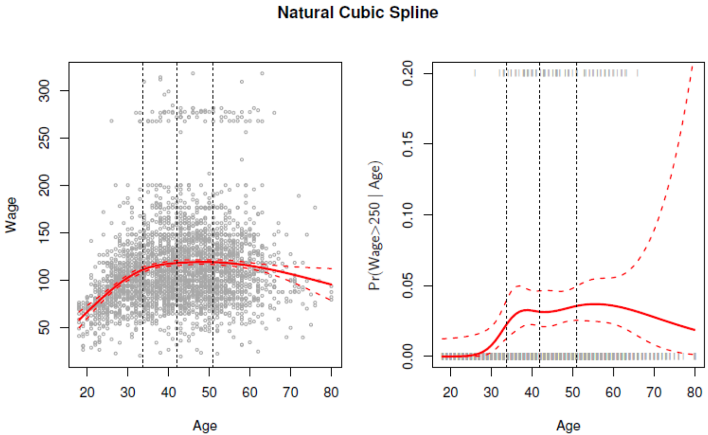
Natural Cubic Splines

A natural cubic spline extrapolates linearly beyond the boundary knots. This adds $4 = 2 \times 2$ extra constraints, and allows us to put more internal knots for the same degrees of freedom as a regular cubic spline.



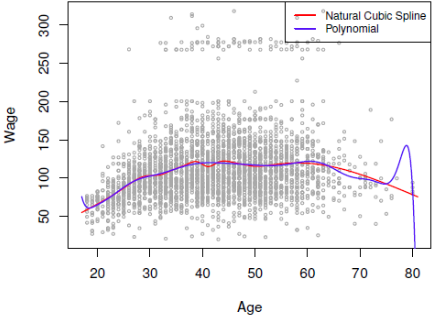
Natural Cubic Splines

Fitting splines in R is easy: `bs(x, ...)` for any degree splines, and `ns(x, ...)` for natural cubic splines, in package `splines`.



Knot placement

- One strategy is to decide K, the number of knots, and then place them at appropriate quantiles of the observed X.
- A cubic spline with K knots has $K + 4$ parameters or degrees of freedom.
- A natural spline with K knots has K degrees of freedom.



Comparison of a degree-14 polynomial and a natural cubic spline, each with 15df.

```
ns(age, df=14)
poly(age, deg=14)
```

Smoothing Splines

Consider this criterion for fitting a smooth function $g(x)$ to some data:

$$\underset{g \in \mathcal{S}}{\text{minimize}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- The first term is RSS, and tries to make $g(x)$ match the data at each x_i.
- The second term is a *roughness penalty* and controls how wiggly $g(x)$ is. It is modulated by the *tuning parameter* $\lambda \geq 0$.
  - The smaller λ, the more wiggly the function, eventually interpolating y_i when $\lambda = 0$.

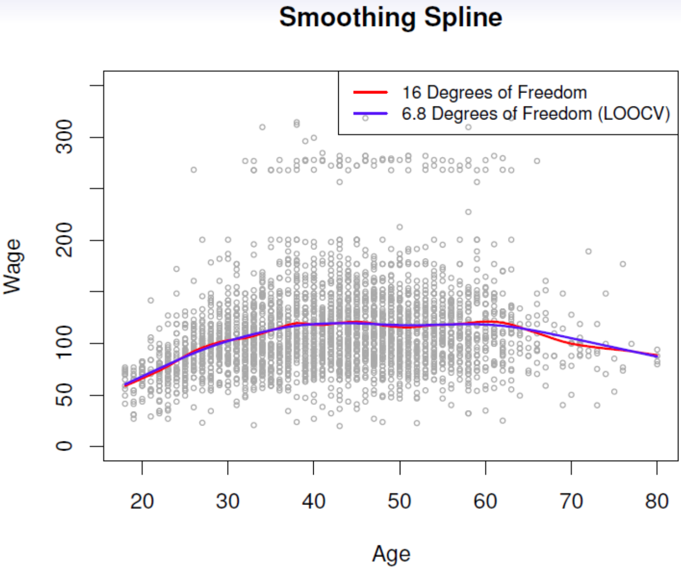
Smoothing Splines

The solution is a natural cubic spline, with a knot at every unique value of x_i. The roughness penalty still controls the roughness via λ.

Some details

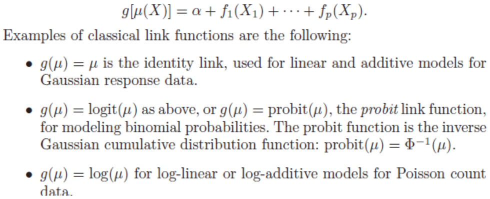
- Smoothing splines avoid the knot-selection issue, leaving a single λ to be chosen.
- The algorithmic details are too complex to describe here. In R, the function `smooth.spline()` will fit a smoothing spline.
- The vector of n fitted values can be written as $\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$, where $\mathbf{S}_\lambda$ is a $n \times n$ matrix (determined by the x_i and λ).
- The *effective degrees of freedom* are given by

$$df_\lambda = \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii}.$$


Smoothing Spline
16 Degrees of Freedom
6.8 Degrees of Freedom (LOOCV)
Wage
0
50
100
200
300
Age
20
30
40
50
60
70
80

Generalized additive model

In general, the conditional mean of a response Y is related to an additive function of the predictors via a link function g:

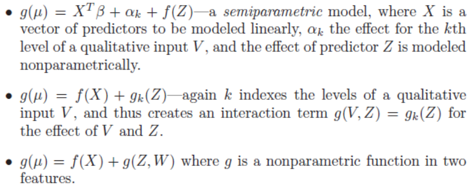


$$g[\mu(X)] = \alpha + f_1(X_1) + \cdots + f_p(X_p).$$

Examples of classical link functions are the following:

- $g(\mu) = \mu$ is the identity link, used for linear and additive models for Gaussian response data.
- $g(\mu) = \text{logit}(\mu)$ as above, or $g(\mu) = \text{probit}(\mu)$, the *probit* link function, for modeling binomial probabilities. The probit function is the inverse Gaussian cumulative distribution function: $\text{probit}(\mu) = \Phi^{-1}(\mu)$.
- $g(\mu) = \log(\mu)$ for log-linear or log-additive models for Poisson count data.

The functions $\hat{f}_j$ are estimated in a flexible manner, using an algorithm whose basic building block is a scatterplot smoother. Note that not all of the functions f_j need to be nonlinear.



- $g(\mu) = X^T\beta + \alpha_k + f(Z)$—a *semiparametric* model, where X is a vector of predictors to be modeled linearly, α_k the effect for the kth level of a qualitative input V, and the effect of predictor Z is modeled nonparametrically.
- $g(\mu) = f(X) + g_k(Z)$—again k indexes the levels of a qualitative input V, and thus creates an interaction term $g(V, Z) = g_k(Z)$ for the effect of V and Z.
- $g(\mu) = f(X) + g(Z, W)$ where g is a nonparametric function in two features.