Nonparametric Regression in R

- $y = m(x) + \epsilon$

- $m(x)$: in nonparametric regression, smooth, continuous function

https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Nonparametric-Regression.pdf

Nonparametric regression

- Simple regression: "scatterplot smoothing"

- Multiple regression: assume additive regression model

  – $y = m(x_1) + m(x_2) + \cdots m(x_p) + \epsilon$
  where partial-regression functions $m(x_j)$ are assumed to be smooth,

- semiparametric models
  – Example: $y = \beta_1 x_1 + m_{12}(x_1, x_2) + \cdots m(x_p) + \epsilon$

R packages

- Simple-regression smoothing-spline estimation
  - smooth.spline()

- Local polynomial regression
  - lowess()
  - loess()

- ˆGeneralized nonparametric regression
  - locfit

- Generalized additive models
  - gam

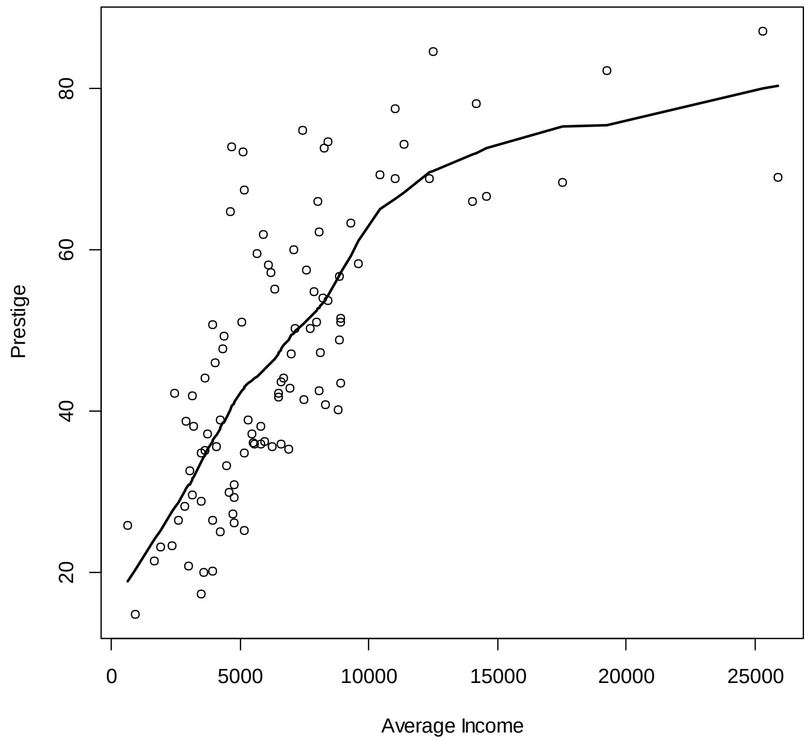
Local Polynomial Regression

- A pth-order weighted-least-squares polynomial regression

- $y = \beta_0 + \beta_1(x - x_1) \ + \beta_2 \ (x - x_1)^2$ + … $\beta_p \ (x - x_1)^2$ $+ \ \epsilon$

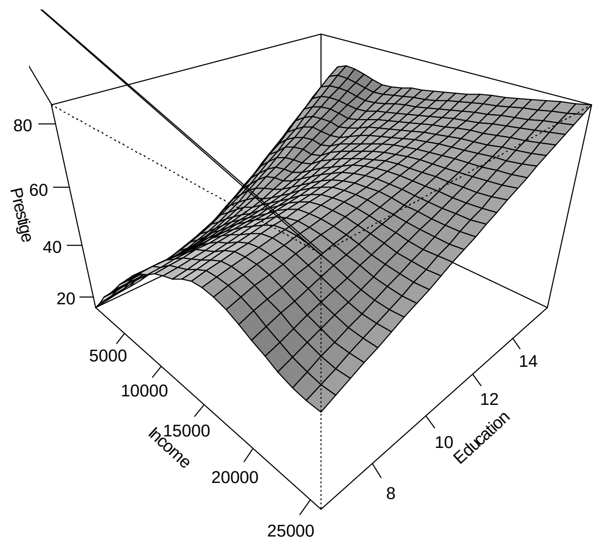
Canadian occupational-prestige data
Local Polynomial Regression - simple regression

- `library("carData") # for data sets`
- `plot(prestige ~ income, xlab="Average Income", ylab="Prestige", data=Prestige)`
- `with(Prestige, lines(lowess(income, prestige, f=0.5, iter=0), lwd=2))`



Multiple Regression

- `mod.lo <- loess(prestige ~ income + education, span=.5, degree=1, data=Prestige)`
- `summary(mod.lo)`
- `inc <- with(Prestige, seq(min(income), max(income), len=25))`
- `ed <- with(Prestige, seq(min(education), max(education), len=25))`
- `newdata <- expand.grid(income=inc, education=ed)`
- `fit.prestige <- matrix(predict(mod.lo, newdata), 25, 25)`
- `persp(inc, ed, fit.prestige, theta=45, phi=30, ticktype="detailed", xlab="Income", ylab="Education", zlab="Prestige", expand=2/3, shade=0.5)`



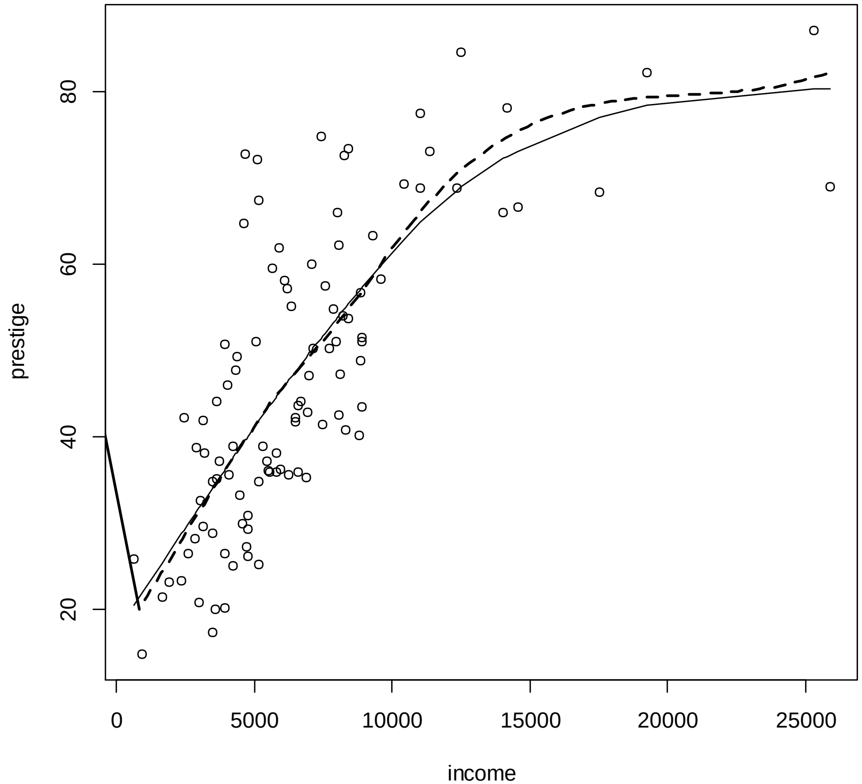
Smoothing Splines

- the penalized sum of squares
- = residual sum of squares + roughness penalty

- $SS(h) = \sum_{i=1}^{n}(y_i - m_i)^2 + h\int m''(x)^2\,dx$
  - h is a smoothing parameter
  - If h is large, m is selected, m'' = 0 everywhere, a globally linear least-squared fit to the data

Smoothing Splines

- `mod.lo.inc <- loess(prestige ~ income, span=.7, degree=1, data=Prestige) # omitting education`
- `plot(prestige ~ income, data=Prestige)`
- `inc.100 <- with(Prestige, seq(min(income), max(income), len=100)) # 100 x-values`
- `pres <- predict(mod.lo.inc, data.frame(income=inc.100)) # fitted values`
- `lines(inc.100, pres, lty=2, lwd=2) # loess curve`
- `lines(with(Prestige, smooth.spline(income, prestige, df=3.85),`
- `lwd=2)) # smoothing spline`

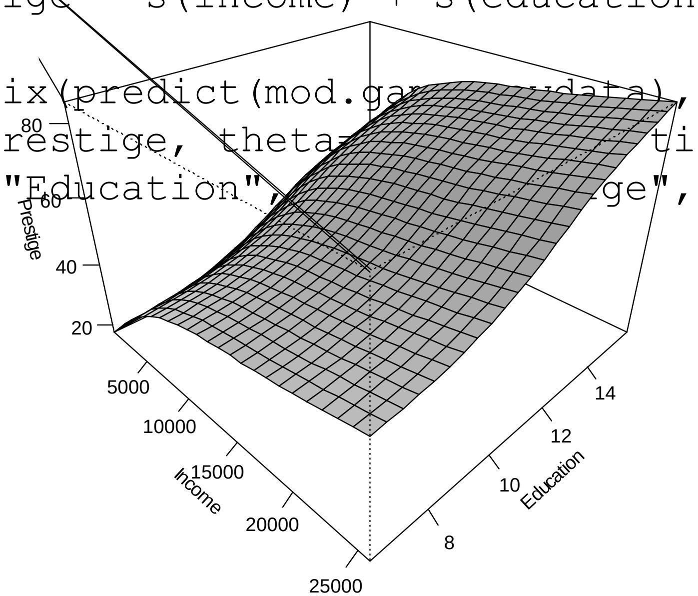


Additive Nonparametric Regression

- $y = m(x_1) + m(x_2) + \cdots m(x_p) + \epsilon$

- $\log\frac{p}{1-p} = m(x_1) + m(x_2) + \cdots m(x_p) + \epsilon$

```
• library("mgcv")
• mod.gam <- gam(prestige ~ s(income) + s(education), data=Prestige)
• summary(mod.gam)
• fit.prestige <- matrix(predict(mod.ga[illegible]data), 25, 25)
• persp(inc, ed, fit.prestige, theta=[illegible] ticktype="detailed",
  xlab="Income", ylab="Education", [illegible]ge", expand=2/3,
  shade=0.5)
```



Generalized Nonparametric Regression

```
• library("car")
• remove(list=objects()) # clean up everything
• Mroz$k5f <- factor(Mroz$k5)
• Mroz$k618f <- factor(Mroz$k618)
• Mroz$k5f <- recode(Mroz$k5f, "3 = 2")
• Mroz$k618f <- recode(Mroz$k618f, "6:8 = 5")
• mod.1 <- gam(lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc,
• family=binomial, data=Mroz)
• summary(mod.1)

• mod.2 <- gam(lfp ~ age + s(inc) + k5f + k618f + wc + hc,
• family=binomial, data=Mroz)
• anova(mod.2, mod.1, test="Chisq")

• mod.3 <- gam(lfp ~ s(age) + inc + k5f + k618f + wc + hc,
• family=binomial, data=Mroz)
• anova(mod.3, mod.1, test="Chisq")

• mod.4 <- update(mod.1, . ~ . - s(age))
• anova(mod.4, mod.1, test="Chisq")
```

A generalized linear model

- A transformation of the conditional expectation E[Y|X] is a linear function of X
- The logistic regression is
- $\log \frac{p}{1-p} = \mathrm{g}(\mathrm{E}[\mathrm{Y}|\mathrm{X}]) = \mathrm{x}\beta$

- y_i ~ some distribution with mean μ_i

$$g(\mu_i) = x_i\beta$$

- A GLM therefore consists of three components:
  - The systematic component, $x_i\beta$ (linear predictor)
  - The random component: the specified distribution for y_i
  - The link function g

exponential family

- A distribution falls into the exponential family if its distribution function can be written as

$$f(y|\theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\},$$

ϕ : Scale parameter

Link function

- y_i ~ some distribution with mean μ_i

$$g(\mu_i) = x_i\beta$$

  - the link component connects the random and systematic components: g

Example: Binomial distribution

- The pdf $\exp\left(\text{ylog}\left(\frac{p}{1-p}\right) + \log(1-p)\right)$ and it is the exponential family with

- $\theta = \log\frac{p}{1-p}$
- $\text{b}(\theta) = \log(1+\theta)$

- $\text{g(u)} = \log\frac{u}{1-u}$, [canonical link]
- $\theta = g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$

- then
- $\theta = h(\mu\quad) = h(h^{-1}(\eta)) = \eta = x_i\beta$

Example: Poisson regression

- In a disease epidemic, the rate at which new cases occur increases exponentially through time

- $\mu_i = \gamma e^{\delta t_i}$

- $\log(\mu_i) = \log(\gamma) + \delta t_i = \beta_0 + \beta_1 t_i$

- this model fits into the GLM framework with a Poisson outcome distribution, a log link, and a linear predictor of $\beta_0 + \beta_1 t_i$

- The pdf $\exp(\mathrm{y}\log(\mu) - \mu - log y!)$ and it is the exponential family with $\theta = \log(\mu), \mathrm{b}(\theta) = \exp(\theta)$

- for the exponential family
- $\mathrm{E(Y) = b'(\theta)}$
- $\mathrm{V(Y)} = \varphi \mathrm{b''(\theta)}$

  - the variance of Y depends on both the scale parameter and on a function of the mean (because θ is a function of μ), with b controlling the relationship between mean and variance